Part III: Causality

17

Instrumental Sets

Carlos Brito

1 Introduction

The research of Judea Pearl in the area of causality has been very much acclaimed. Here we highlight his contributions for the use of graphical languages to represent and reason about causal knowledge.[1]

The concept of causation seems to be fundamental to our understanding of the world. Philosophers like J. Carroll put it in these terms: "With regard to our total conceptual apparatus, causation is the center of the center" [Carroll 1994]. Perhaps more dramatically, David Hume states that causation together with resemblance and contiguity are "the only ties of our thoughts, ... for us the cement of the universe" [Hume 1978]. In view of these observations, the need for an adequate language to talk about causation becomes clear and evident.

The use of graphical languages was present in the early times of causal modelling. Already in 1934, Sewall Wright [Wright 1934] represented the causal relation among several variables with diagrams formed by points and arrows (i.e., a directed graph), and noted that the correlations observed between the variables could be associated with the various paths between them in the diagram. From this observation he obtained a method to estimate the strength of the causal connections known as The Method of Path Coefficients, or simply Path Analysis.

With the development of the research in the field, the graphical representation gave way to a mathematical language, in which causal relations are represented by equations of the form $Y = \alpha+\beta X+e$. This movement was probably motivated by an increasing interest in the quantitative aspects of the model, or by the rigorous and formal appearance offered by the mathematical language. However it may be, the consequence was a progressive departure from our basic causal intuitions. Today people ask whether such an equation represents a functional or a causal relation [Reiss 2005]. Sewall Wright and Judea Pearl would presumably answer: "Causal, of course!".

2 The Identification Problem

We explore the feasibility of inferring linear cause-effect relationships from various combinations of data and theoretical assumptions. The assumptions are represented

[1] This contribution is a simplified version of a joint paper with Judea Pearl in UAI 2002. A great deal of technicality was removed, and new discussion was added, in the hope that the reader will be able to easily follow and enjoy the argument.

Figure 1. (a) a bow-pattern; and (b) a bow-free model.

in the form of an acyclic causal diagram, which contains both arrows and bidirected arcs [Pearl 1995; Pearl 2000a]. The arrows represent the potential existence of direct causal relationships between the corresponding variables, and the bidirected arcs represent spurious correlations due to unmeasured common causes. All interactions among variables are assumed to be linear. Our task is to decide whether the assumptions represented in the diagram are sufficient for assessing the strength of causal effects from non-experimental data, and, if sufficiency is proven, to express the target causal effect in terms of estimable quantities.

This decision problem has been tackled in the past half century, primarily by econometricians and social scientists, under the rubric "The Identification Problem" [Fisher 1966] - it is still unsolved. Certain restricted classes of models are nevertheless known to be identifiable, and these are often assumed by social scientists as a matter of convenience or convention [Duncan 1975]. A hierarchy of three such classes is given in [McDonald 1997]: (1) no bidirected arcs, (2) bidirected arcs restricted to root variables, and (3) bidirected arcs restricted to variables that are not connected through directed paths.

In a further development [Brito and Pearl 2002], we have shown that the identification of the entire model is ensured if variables standing in direct causal relationship (i.e., variables connected by arrows in the diagram) do not have correlated errors; no restrictions need to be imposed on errors associated with indirect causes. This class of models was called "bow-free", since their associated causal diagrams are free of any "bow-pattern" [Pearl 2000a] (see Figure 1).

Most existing conditions for identification in general models are based on the concept of Instrumental Variables (IV) [Pearl 2000b; Bowden and Turkington 1984]. IV methods take advantage of conditional independence relations implied by the model to prove the identification of specific causal-effects. When the model is not rich in conditional independence relations, these methods are not informative. In [Brito and Pearl 2002] we proposed a new graphical criterion for identification which does not make direct use of conditional independence, and thus can be successfully applied to models in which the IV method would fail.

The result presented in this paper is a generalization of the graphical version

of the method of instrumental variables, offered by Judea Pearl [Pearl 2000a], to deal with several parameters of the model simultaneously. The traditional method of instrumental variables involves conditions on the independence of the relevant variables and on the rank of a certain matrix of correlations [McFadden]. The first of these is captured by the notion of d-separation. As for the second, since we know from [Wright 1934] that correlations correspond to paths in the causal diagram, we can investigate which structural properties of the model give rise to the proper conditions of the IV method. The results are graphical criteria that allow us to conclude the identification of some parameters from consideration of the qualitative information represented in the causal diagram.

3 Linear Models and Identification

A linear model for the random variables $Y_1, \ldots, Y_n$ is defined by a set of equations of the form:

$$(1) \quad Y_j = \sum_i c_{ji} Y_i + e_j, \quad j = 1, \ldots, n$$

An equation $Y = cX + e$ encodes two distinct assumptions: (1) the possible existence of (direct) causal influence of X on Y; and, (2) the absence of causal influence on Y of any variable that does not appear on the right-hand side of the equation. The parameter c quantifies the (direct) causal effect of X on Y. That is, the equation claims that a unit increase in X would result in c units increase of Y, assuming that everything else remains the same. The variable e is called an error or disturbance; it represents unobserved background factors that the modeler decides to keep unexplained; this variable is assumed to have a normal distribution with zero mean.

The specification of the equations and the pairs of error-terms (e_i, e_j) with non-zero correlation defines the structure of the model. This structure can be represented by a directed graph, called causal diagram, in which the set of nodes is defined by the variables $Y_1, \ldots, Y_n$, and there is a directed edge from Y_i to Y_j if Y_i appears on the right-hand side of the equation for Y_j. Additionally, if error-terms e_i and e_j are assumed to have non-zero correlation, we add a (dashed) bidirected edge between Y_i and Y_j. Figure 2 shows a model with the respective causal diagram.

In this work, we consider only recursive models, which are defined by the restriction that $c_{ji} = 0$, for all $i \geq j$. This simply means that the directed edges in the causal diagram do not form cycles.

The set of parameters of the model, denoted by Θ, is formed by the coefficients c_{ij} and the non-zero entries of the error covariance matrix Ψ, $[\Psi_{ij}] = cov(e_i, e_j)$.

Fixing the structure and assigning values to the parameters Θ, the model determines a unique covariance matrix Σ over the observed variables $Y_1, \ldots, Y_n$, given by (see [Bollen 1989], page 85):

$$(2) \quad \Sigma(\Theta) = (I - C)^{-1} \Psi [(I - C)^{-1}]'$$

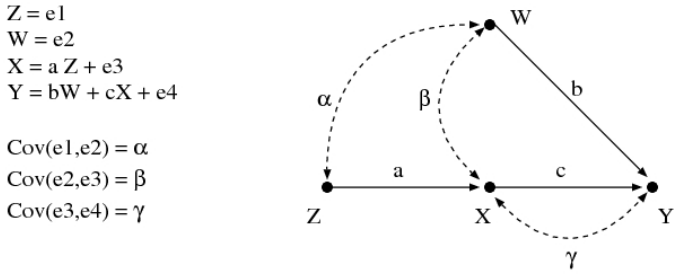


Figure 2. A simple linear model and its causal diagram.

where C is the matrix of coefficients c_{ji}.

Conversely, in the Identification Problem, after fixing the structure of the model, one attempts to solve for Θ in terms of the observed covariance Σ. This is not always possible. In some cases, no parametrization of the model is compatible with a given Σ. In other cases, the structure of the model may permit several distinct solutions for the parameters. In these cases, the model is called *non-identified*.

However, even if the model is non-identified, some parameters may still be uniquely determined by the given assumptions and data. Whenever this is the case, the specific parameters are said to be *identified*.

Finally, since the conditions we seek involve the structure of the model alone, and do not depend on the numerical values of the parameters Θ, we insist only on having identification almost everywhere, allowing few pathological exceptions. The concept of identification almost everywhere can be formalized as follows.

Let h denote the total number of parameters in the model. Then, each vector $\Theta \in \Re^h$ defines a parametrization of the model. For each parametrization Θ, the model G generates a unique covariance matrix $\Sigma(\Theta)$. Let $\Theta(\lambda_1, \ldots, \lambda_n)$ denotes the vector of values assigned by Θ to the parameters $\lambda_1, \ldots, \lambda_n$.

Parameters $\lambda_1, \ldots, \lambda_n$ are identified almost everywhere if

$$\Sigma(\Theta) = \Sigma(\Theta') \quad \text{implies} \quad \Theta(\lambda_1, \ldots, \lambda_n) = \Theta'(\lambda_1, \ldots, \lambda_n)$$

except when Θ resides on a subset of Lebesgue measure zero of $\Re^h$.

4 Graph Background

DEFINITION 1.

1. A *path* in a graph is a sequence of edges such that each pair of consecutive edges share a common node, and each node appears only once along the path.

2. A *directed path* is a path composed only by directed edges, all of them oriented

in the same direction. If there is a directed path going from X to Y we say that Y is a *descendant* of X.

3. A path is *closed* if it has a pair of consecutive edges pointing to their common node (e.g., $\ldots \rightarrow X \leftarrow \ldots$ or $\ldots \leftrightarrow X \leftarrow \ldots$). In this case, the common node is called a *collider*. A path is *open* if it is not closed.

DEFINITION 2. A path p is blocked by a set of nodes $\mathbf{Z}$ (possibly empty) if either

1. $\mathbf{Z}$ contains some non-collider node of p, or

2. at least one collider of p and all of its descendants are outside $\mathbf{Z}$.

The idea is simple. If the path is closed, then it is naturally blocked by its colliders. However, if a collider, or one of its descendants, belongs to $\mathbf{Z}$, then it ceases to be an obstruction. But if a non-collider of p belongs to $\mathbf{Z}$, then the path is definitely blocked.

DEFINITION 3. A set of nodes $\mathbf{Z}$ d-separates X and Y if $\mathbf{Z}$ simultaneously blocks all the paths between X and Y. If $\mathbf{Z}$ is empty, then we simply say that X and Y are d-separated.

The significance of this definition comes from a result showing that if X and Y are d-separated by $\mathbf{Z}$ in the causal diagram of a linear model, then the variables X and Y are conditionally independent given $\mathbf{Z}$ [Pearl 2000a]. It is this sort of result that makes the connection between the mathematical and graphical languages, and allows us to express our conditions for identification in graphical terms.

DEFINITION 4. Let $p_1, \ldots, p_n$ be unblocked paths connecting the variables $Z_1, \ldots, Z_n$ and the variables $X_1, \ldots, X_n$, respectively. We say that the set of paths $p_1, \ldots, p_n$ is incompatible if we cannot rearrange their edges to form a different set of unblocked paths $p'_1, \ldots, p'_n$ between the same variables.

A set of disjoint paths (i.e., paths with no common nodes) consists in a simple example of an incompatible set of paths.

5 Instrumental Variable Methods

5.1 Identification of a Single Parameter

The method of Instrumental Variables (IV) for the identification of causal effects is intended to address the situation where we cannot attribute the entire correlation between two variables, say X and Y, to their causal connection. That is, part of the correlation between X and Y is due to common causes and/or correlations between disturbances. Figure 3 shows examples of this situation.

In the simplest cases, like in Figure 3a, we can find a conditioning set $\mathbf{W}$ such that the partial correlation of X and Y given $\mathbf{W}$ can indeed be attributed to the causal relation. In this example, if we take $\mathbf{W} = \{W\}$ we eliminate the source

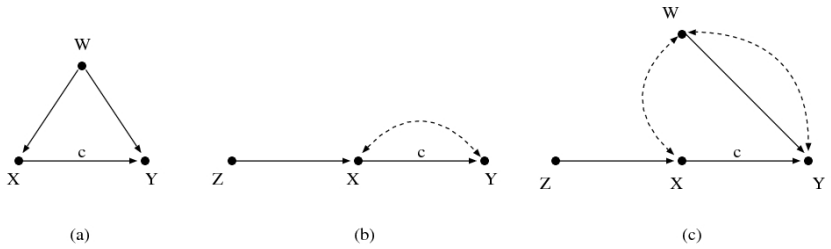


Figure 3. Models with spurious correlation between X and Y.

of spurious correlation. The causal effect of X on Y is identified and given by $c = \sigma_{XY.\mathbf{W}}$.

There are cases, however, where this idea does not work, either because the spurious correlation is originated by disturbances outside the model (Figure 3b), or else because the conditioning itself introduces spurious correlations (Figure 3c). In situations like these, the IV method asks us to look for a variable Z with the following properties [Bowden and Turkington 1984]:

IV-1. Z is not independent of X.

IV-2. Z is independent of all error terms that have an influence on Y that is not mediated by X.

The first condition simply states that there is a correlation between Z and X. The second condition says that the only source of correlation between Z and Y is due to a covariation bewteen Z and X that subsequently affects Y through the causal connection $X \xrightarrow{c} Y$.

If we can find a variable Z with these properties, then the causal effect of X on Y is identified and given by $c = \sigma_{ZY}/\sigma_{ZX}$.

Using the notion of d-separation we can express the conditions (1) and (2) of the IV method in graphical terms, thus obtaining a criterion for identification that can be applied directly to the causal diagram of the model. Let G be the graph representing the causal diagram of the model, and let G_c be the graph obtained after removing the edge $X \xrightarrow{c} Y$ from G (see Figure 4). Then, Z is an instrumental variable relative to $X \xrightarrow{c} Y$ if:

1. Z is not d-separated from X in G_c.

2. Z is d-separated from Y in G_c.

Using this criterion, it is easy to verify that Z is an instrumental variable relative to $X \xrightarrow{c} Y$ in the models of Figure 3b and c.

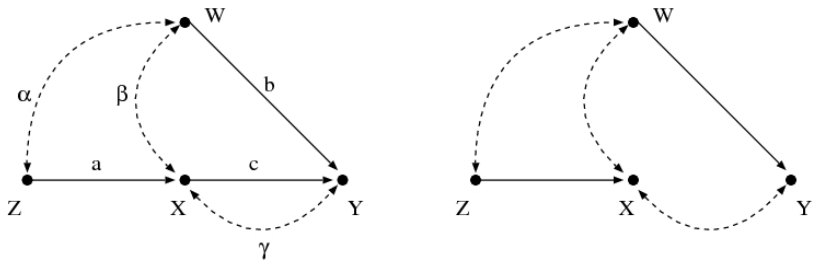


Figure 4. The causal diagram G of a linear model and the graph G_c.

5.2 Conditional Instrumental Variables

A generalization of the method of instrumental variables is offered through the use of conditioning. A conditional instrumental variable is a variable Z that may not have the properties (IV-1) and (IV-2) above, but after conditioning on a subset $\mathbf{W}$ these properties do hold. When such pair $(Z, \mathbf{W})$ is found, the causal effect of X on Y is identified and given by $c = \sigma_{ZY.\mathbf{W}}/\sigma_{ZX.\mathbf{W}}$.

Again, we obtain a graphical criterion for a conditional IV using the notion of d-separation. Variable Z is a conditional instrumental variable relative to $X \stackrel{c}{\rightarrow} Y$ given $\mathbf{W}$, if

1. $\mathbf{W}$ contains only non-descendants of Y.
2. $\mathbf{W}$ does not d-separate Z from X in G_c.
3. $\mathbf{W}$ d-separates Z from Y in G_c.

5.3 Identification of Multiple Parameters

So far we have been concerned with the identification of a single parameter of the model, but in its full version the method of instrumental variables allows to prove simultaneously the identification of several parameters in the same equation (i.e., the causal effects of several variables $X_1, \ldots, X_k$ on the same variable Y).

Following [McFadden], assume that we have the equation

$$Y = c_1 X_1 + \ldots + c_k X_k + e$$

in our linear model. The variables $Z_1, \ldots, Z_j$, with $j \geq k$, are called instruments if

1. The matrix of correlations between the variables $X_1, \ldots, X_k$ and the variables $Z_1, \ldots, Z_j$ is of maximum possible rank (i.e., rank k).
2. The variables $Z_1, \ldots, Z_j$ are uncorrelated with the error term e.

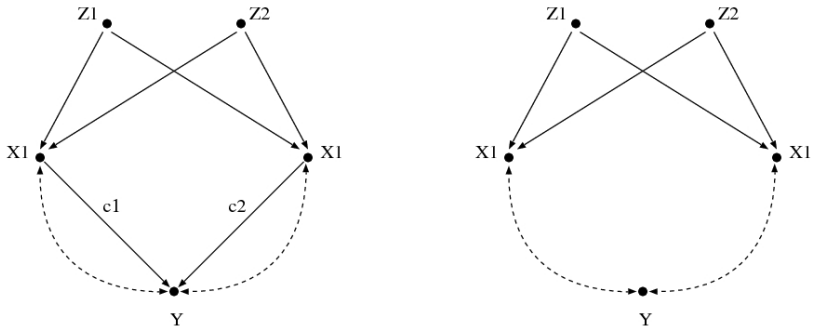


Figure 5. The causal diagram G of a linear model and the graph $\bar{G}$.

Next, we develop our graphical intuition and obtain a graphical criterion for identification that corresponds to the full version of the IV method.

Consider the model in Figure 5a. Here, the variables Z_1 and Z_2 do not qualify as instrumental variables (or even conditional IVs) with respect to either $X_1 \stackrel{c_1}{\rightarrow} Y$ or $X_2 \stackrel{c_2}{\rightarrow} Y$. But, following ideas similar to the ones developed in the previous sections, in Figure 5b we show the graph obtained by removing edges $X_1 \rightarrow Y$ and $X_2 \rightarrow Y$ from the causal diagram. Observe that now both d-separation conditions for an instrumental variable hold for Z_1 and Z_2. This leads to the idea that Z_1 and Z_2 could be used together as instruments to prove the identification of parameters c_1 and c_2. Indeed, next we give a graphical criterion that is sufficient to guarantee the identification of a subset of parameters of the model.

Fix a variable Y, and consider the edges $X_1 \stackrel{c_1}{\rightarrow} Y, \ldots, X_k \stackrel{c_k}{\rightarrow} Y$ in the causal diagram G of the model. Let $\bar{G}$ be the graph obtained after removing the edges $X_1 \rightarrow Y, \ldots, X_k \rightarrow Y$ from G. The variables $Z_1, \ldots, Z_k$ are instruments relative to $X_1 \stackrel{c_1}{\rightarrow} Y, \ldots, X_k \stackrel{c_k}{\rightarrow} Y$ if

1. There exists an incompatible set of unblocked paths $p_1, \ldots, p_k$ connecting the variables $Z_1, \ldots, Z_k$ to the variables $X_1, \ldots, X_k$.

2. The variables Z_i are d-separated from Y in $\bar{G}$.

3. Each variable Z_i is not d-separated from the corresponding variable X_i in $\bar{G}$. [2]

THEOREM 5. *If we can find variables $Z_1, \ldots, Z_k$ satisfying the conditions above, then the parameters $c_1, \ldots, c_k$ are identified almost everywhere, and can be computed by solving a system of linear equations.*

[2] Notice that this condition is redundant, since it follows from the first condition.

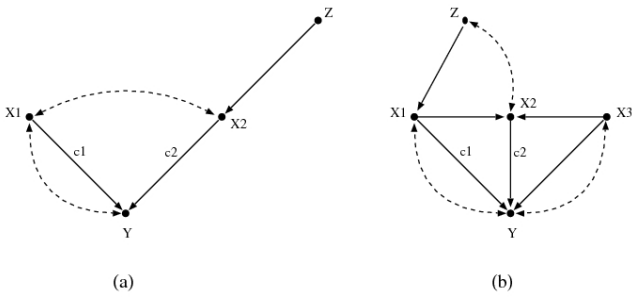


Figure 6. More examples of the new criterion.

Figure 6 shows more examples of application of the new graphical criterion. Model (a) illustrates an interesting case, in which variable X_2 is used as the instrumental variable for $X_1 \rightarrow Y$, while Z is the instrumental variable for $X_2 \rightarrow Y$. Finally, in model (b) we have an example in which the parameter of edge $X_3 \rightarrow Y$ is non-identified, and still the graphical criterion allows to show the identfication of c_1 and c_2.

6 Wright's Method of Path Coefficients

Here, we describe an important result introduced by Sewall Wright [Wright 1934], which is extensively explored in our proofs.

Given variables X and Y in a recursive linear model, the correlation coefficient of X and Y, denoted by ρ_{XY}, can be expressed as a polynomial on the parameters of the model. More precisely,

$$\sigma_{XY} = \sum_{p} T(p) \quad (3)$$

where the summation ranges over all unblocked paths p between X and Y, and each term $T(p)$ represents the contribution of the path p to the total correlation between X and Y. The term $T(p)$ is given by the product of the parameters of the edges along the path p. We refer to Equation 3 as Wright's equation for X and Y.

Wright's method of path coefficients for identification consists in forming Wright's equations for each pair of variables in the model, and then solving for the parameters in terms of the observed correlations. Whenever there is a unique solution for a parameter c, this parameter is identified.

7 Proof of Theorem 1

7.1 Notation

Fix a variable Y in the model. Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be the set of all non-descendants of Y which are connected to Y by an edge. Define the following set of edges incoming Y:

$$(4) \quad Inc(Y) = \{(X_i, Y) : X_i \in \mathbf{X}\}$$

Note that for some $X_i \in \mathbf{X}$ there may be more than one edge between X_i and Y (one directed and one bidirected). Thus, $|Inc(Y)| \geq |\mathbf{X}|$. Let $\lambda_1, \ldots, \lambda_m$, $m \geq k$, denote the parameters of the edges in $Inc(Y)$.

It follows that edges $X_1 \stackrel{c_1}{\rightarrow} Y, \ldots, X_k \stackrel{c_k}{\rightarrow} Y$ all belong to $Inc(Y)$, because $X_1, \ldots, X_k$ are clearly non-descendants of Y. We assume that $\lambda_i = c_i$, for $i = 1, \ldots, k$, while $\lambda_{k+1}, \ldots, \lambda_m$ are the parameters of the remaining edges of $Inc(Y)$.

Let Z be any non-descendant of Y. Wright's equation for the pair (Z, Y) is given by:

$$(5) \quad \sigma_{ZY} = \sum_p T(p)$$

where each term $T(p)$ corresponds to an unblocked path p between Z and Y. The next lemma proves a property of such paths.

LEMMA 6. *Any unblocked path between Y and one of its non-descendants Z must include exactly one edge from $Inc(Y)$.*

Lemma 6 allows us to write equation 4 as:

$$(6) \quad \sigma_{ZY} = \sum_{j=1}^{m} a_j \cdot \lambda_j$$

Thus, the correlation between Z and Y can be expressed as a linear function of the parameters $\lambda_1, \ldots, \lambda_m$, with no constant term. In addition, we can say something about the coefficients a_j. Each term in Equation 5 corresponds to an unblocked path that reaches Y through some egge, say $X_j \stackrel{\lambda_j}{\rightarrow} Y$. When we group the terms together according to the parameter λ_j and factor it out, we are, in a sense, removing the edge $X_j \rightarrow Y$ from those paths. Thus, each coefficient a_j in Equation 6 is a sum of terms associated with unblocked paths between Z and X_j.

7.2 Basic Linear Equations

We have just seen that the correlations between the instrumental variables Z_i and Y can be written as a linear function of the parameters $\lambda_1, \ldots, \lambda_m$:

$$(7) \quad \rho_{Z_iY} = \sum_{j=1}^{m} a_{ij} \cdot \lambda_j$$

Next, we prove an important result

LEMMA 7. *The coefficients $a_{i,k+1}, \ldots, a_{im}$ in Equation 7 are all identically zero.*

Proof. The fact that Z_i is d-separated from Y in $\bar{G}$ implies that $\rho_{Z_iY} = 0$ in any probability distribution compatible with $\bar{G}$. Hence, the expression for ρ_{Z_iY} must vanish when evaluated in the causal diagram $\bar{G}$. But this implies that each

coefficient a_{ij} in Equation 7 is identically zero, when the expression is evaluated in $\bar{G}$.

Next, we show that the only difference between the expression for ρ_{Z_iY} on the causal diagrams G and $\bar{G}$ are the coefficients of the parameters $\lambda_1, \ldots, \lambda_k$.

Recall from the previous section that each coefficient a_{ij} is a sum of terms associated with paths which can be extended by the edge $\stackrel{\lambda_j}{\rightarrow} Y$ to form an unblocked path between Z and Y.

Fixing $j > k$, we observe that the insertion of edges $x_1 \rightarrow Y, \ldots, X_k \rightarrow Y$ in $\bar{G}$ does not create any new such path (and clearly does not eliminate any existing one). Hence, for $j > k$, the coefficients a_{ij} in the expression for ρ_{Z_iY} in the causal diagrams G and $\bar{G}$ are exactly the same, namely, identically zero. □

The conclusion from Lemma 7 is that the expression for ρ_{Z_iY} is a linear function only of parameters $\lambda_1, \ldots, \lambda_k$:

$$\rho_{Z_iY} = \sum_{j=1}^{k} a_{ij} \cdot \lambda_j \tag{8}$$

7.3 System of Equations Φ

Writing Equation 8 for each instrumental variable Z_i, we obtain the following system of linear equations on the parameters $\lambda_1, \ldots, \lambda_k$:

$$\Phi = \begin{cases} \rho_{Z_1Y} = a_{11}\lambda_1 + \ldots, a_{1k}\lambda_k \\ \ldots \\ \rho_{Z_kY} = a_{k1}\lambda_1 + \ldots, a_{kk}\lambda_k \end{cases} \tag{9}$$

Our goal now is to show that Φ can be solved uniquely for the parameters λ_i, and so prove the identification of $\lambda_1, \ldots, \lambda_k$. Next lemma proves an important result in this direction.

Let A denote the matrix of coefficients of Φ.

LEMMA 8. *Det(A) is a non-trivial polynomial on the parameters of the model.*

Proof. The determinant of A is defined as the weighted sum, for all permutations π of $\langle 1, \ldots, k \rangle$, of the product of the entries selected by π. Entry a_{ij} is selected by a permutation π if the i^{th} element of π is j. The weights are either 1 or -1, depending on the parity of the permutation.

Now, observe that each diagonal entry a_{ii} is a sum of terms associated with unblocked paths between Z_i and X_i. Since p_i is one such path, we can write $a_{ii} = T(p_i) + \hat{a}_{ii}$. From this, it is easy to see that the term

$$T^* = \prod_{j=1}^{k} T(p_j) \tag{10}$$

appears in the product of permutation $\pi = \langle 1, \ldots, n \rangle$, which selects all the diagonal entries of A.

We prove that $det(A)$ does not vanish by showing that T^* is not cancelled out by any other term in the expression for det(A).

Let τ be any other term appearing in the summation that defines the determinant of A. This term appears in the product of some permutation π, and has as factors exactly one term from each entry a_{ij} selected by π. Thus, associated with such factor there is an unblocked path between Z_i and X_j. Let $p'_1, \ldots, p'_k$ be the unblocked paths associated with the factors of τ.

We conclude the proof observing that, since $p_1, \ldots, p_k$ is an incompatible set, its edges cannot be rearranged to form a different set of unblocked paths between the same variables, and so $\tau \neq T^*$. Hence, the term T^* is not cancelled out in the summation, and the expression for $det(A)$ does not vanish. □

7.4 Identification of $\lambda_1, \ldots, \lambda_k$

Lemma 8 gives that $det(Q)$ is a non-trivial polynomial on the parameters of the model. Thus, $det(Q)$ only vanishes on the roots of this polynomial. However, [Okamoto 1973] has shown that the set of roots of a polynomial has Lebesgue measure zero. Thus, the system Φ has unique solution almost everywhere.

It just remains to show that we can estimate the entries of the matrix of coefficients A from the data. But this is implied by the following observation.

Once again, coefficient a_{ij} is given by a sum of terms associated with unblocked paths between Z_i and X_j. But, in principle, not every unblocked path between Z_i and X_j would contribute with a term to the sum; just those which can be extended by the edge $X_j \rightarrow Y$ to form an unblocked path between Z_i and Y. However, since the edge $X_j \rightarrow Y$ does not point to X_j, every unblocked path between Z_i and X_j can be extended by the edge $X_j \rightarrow Y$ without creating a collider. Hence, the terms of all unblocked paths between Z_i and X_j appear in the expression for a_{ij}, and by the method of path coefficients, we have $a_{ij} = \rho_{Z_i X_j}$.

We conclude that each entry of matrix A can be estimated from data, and we can solve the system of linear equations Φ to obtain the parameters $\lambda_1, \ldots, \lambda_k$.

References

Bollen, K. (1989). *Structural Equations with Latent Variables.* John Wiley, New York.

Bowden, R. and D. Turkington (1984). *Instrumental Variables.* Cambridge Univ. Press.

Brito, C. and J. Pearl (2002). A graphical criterion for the identification of causal effects in linear models. *In Proc. of the AAAI Conference, Edmonton, Canada.*.

Carroll, J. (1994). *Laws of Nature.* Cambridge University Press.

Duncan, O. (1975). *Introduction to Structural Equation Models.* Academic Press.

Fisher, F. (1966). *The Identification Problem in Econometrics.* McGraw-Hill.

Hume, D. (1978). *A Treatise of Human Nature.* Oxford University Press.

McDonald, R. (1997). Haldane's lungs: A case study in path analysis. *Mult. Beh. Res.*, 1–38.

McFadden, D. *Lecture Notes for Econ 240b.* Dept of Economics, UC Berkeley.

Okamoto, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Annals of Statistics*, 763–765.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 669–710.

Pearl, J. (2000a). *Causality: Models, Reasoning and Inference.* Cambridge Press.

Pearl, J. (2000b). Parameter identification: A new perspective. *Technical Report R-276, UCLA.*

Reiss, J. (2005). Causal instrumental variables and interventions. *Philosophy of Science. 72*, 964–976.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Statistics.*, 161–215.

18

Seeing and Doing: The Pearlian Synthesis

PHILIP DAWID

1 Introduction

It is relatively recently that much attention has focused on what, for want of a better term, we might call "statistical causality", and the subject has developed in a somewhat haphazard way, without a very clear logical basis. There is in fact a variety of current conceptions and approaches [Campaner and Galavotti 2007; Hitchcock 2007; Galavotti 2008]—here we shall distinguish in particular *agency*, *graphical*, *probabilistic* and *modular* conceptions of causality—that tend to be mixed together in an informal and half-baked way, based on "definitions" that often do not withstand detailed scrutiny. In this article I try to unpick this tangle and expose the various different strands that contribute to it. Related points, with a somewhat different emphasis, are made in a companion paper [Dawid 2009].

The approach of Judea Pearl [2009] cuts through this Gordian knot like the sword of Alexander. Whereas other conceptions of causality may be philosophically questionable, definitionally unclear, pragmatically unhelpful, theoretically skimpy, or simply confused, Pearl's theory is none of these. It provides a valuable framework, founded on a rich and fruitful formal theory, by means of which causal assumptions about the world can be meaningfully represented, and their implications developed. Here we will examine both the relationships of Pearl's theory with the other conceptions considered, and its differences from them. We extract the essence of Pearl's approach as an assumption of "modularity", the transferability of certain probabilistic properties between observational and interventional regimes: so, in particular, forging a synthesis between the very different activities of "seeing" and "doing". And we describe a generalisation of this framework that releases it from any necessary connexion to graphical models.

The plan of the paper is as follows. In § 2, I describe the agency, graphical and probabilistic conceptions of causality, and their connexions and distinctions. Section 3 introduces Pearl's approach, showing its connexions with, and differences from, the other theories. Finally, in § 4, I present the generalisation of that approach, emphasising the modularity assumptions that underlie it, and the usefulness of the theory of "extended conditional independence" for describing and manipulating these.

Disclaimer I have argued elsewhere [Dawid 2000, 2007a, 2010] that it is important to distinguish arguments about "Effects of Causes" (EoC, otherwise termed "type", or "generic" causality"), from those about "Causes of Effects" (CoE, also termed "token", or "individual" causality); and that these demand different formal frameworks and analyses. My concern here will be entirely focused on problems of generic causality, EoC. A number of

the current frameworks for statistical causality, such as Rubin's "potential response models" [Rubin 1974, 1978], or Pearl's "probabilistic causal models" [Pearl 2009, Chapter 7], are more especially suited for handling CoE type problems, and will not be discussed further here. There are also numerous other conceptions of causality, such as *mechanistic causality* [Salmon 1984; Dowe 2000], that I shall not be considering here.

2 Some conceptions of causality

There is no generally agreed understanding of what "causality" is or how it should behave. There are two conceptions in particular that are especially relevant for "statistical causality": *Agency Causality* and *Probabilistic Causality*. The latter in turn is closely related to what we might term *Graphical Causality*.

2.1 Agency causality

The "agency" or "manipulability" interpretation of causality [Price 1991; Hausman 1998; Woodward 2003] depends on an assumed notion of external "manipulation" (or "intervention"), that might itself be taken as a primitive—at any rate we shall not try and explicate it further here. The basic idea is that causality is all about how an external manipulation that sets the value of some variable (or set of variables) X will affect some other (unmanipulated) "response variable" (or set of variables) Y. The emphasis is usually on comparison of the responses ensuing from different settings x for X: a version of the "contrastive" or "difference-making" understanding of causality. Much of Statistical Science—for example, the whole subfield of Experimental Design—aims to address exactly these kinds of questions about the comparative effects of interventions on a system, which are indeed a major object of all scientific enquiry.

We can define certain causal terms quite naturally within the agency theory [Woodward 2003]. Thus we could interpret the statement

"X *has no effect on* Y"[1]

as holding whenever, considering regimes that manipulate only X, the resulting value of Y (or some suitable codification of uncertainty about Y, such as its probability distribution) does not depend on the value x assigned to X. When this fails, X *has an effect on* Y; we might then go on to quantify this dependence in various ways.

We could likewise interpret

"X *has no (direct) effect on* Y*, after controlling for* W"

as the property that, considering regimes where we manipulate both W and X, when we manipulate W to some value w and X to some value x, the ensuing value (or uncertainty) for Y will depend only on w, and not further on x.

Now suppose that, explicitly or implicitly, we restrict consideration to some collection $\mathcal{V}$ of manipulable variables. Then we might interpret the statement

[1] Just as "zero" is fundamental to arithmetic and "independence" is fundamental to probability, so the concept of "no effect" is fundamental to causality.

"X *is a direct cause of* Y (relative to $\mathcal{V}$)"

(where $\mathcal{V}$ might be left unmentioned, but must be clearly understood) as the negation of "X has no direct effect on Y, after controlling for $\mathcal{V} \setminus \{X, Y\}$".[2]

It is important to bear in mind that all these assertions relate to properties of the real world under the various regimes considered: in particular, they can not be given purely mathematical definitions. And in real world problems there are typically various ways of manipulating variables, so we must be very clear as to exactly what is intended.

EXAMPLE 1. **Ideal gas law**

Consider the "ideal gas law":

(1) $\quad PV = kNT$

where P is the absolute pressure of the gas, V is its volume, N is the number of molecules of gas present, k is Boltzmann's constant, and T is the absolute temperature. For our current purposes this will be supposed to be universally valid , no matter how the values of the variables in (1) may have come to arise.

Taking a fixed quantity N of gas in an impermeable container, we might consider interventions on any of P, V and T. (Note however that, because of the constraint (1), we can not simultaneously and arbitrarily manipulate all three variables.)

An intervention that sets V to v and T to t will lead to the unique value $p = kNt/v$ for P. Because this depends on both v and t, we can say that there is a *direct effect* of each of V and T on P (relative to $\mathcal{V} = \{V, P, T\}$). Similarly, P has a direct effect on each of V and T.

What if we wish to quantify, say, "the causal effect of V on P"? Any attempt to do this must take account of the fact that the problem requires additional specification to be well-defined. Suppose the volume of the container can be altered by applying a force to a piston. Initially the gas has $V = v_0$, $P = p_0$, $T = t_0$. We wish to manipulate V to a new value v_1. If we do this *isothermally*, *i.e.* by sufficiently slow movement of the piston that, through flow of heat through the walls of the container, the temperature of the gas always remains the same as that of the surrounding heat bath, we will end up with $V = v_1$, $P = p_1 = v_0 p_0/v_1$, $T = t_1 = t_0$. But if we move the piston *adiabatically*, *i.e.* so fast that no heat can pass through the walls of the container, the relevant law is $PV^\gamma =$ constant, where $\gamma = 5/3$ for a monatomic gas. Then we get $V = v_1$, $P = p_1^* = p_0(v_0/v_1)^\gamma$, $T = t_1^* = p_1^* v_1/kN$.

2.2 Graphical causality

By *graphical causality* we shall refer to an interpretation of causality in terms of an underlying *directed acyclic graph* (DAG) (noting in passing that other graphical representations are also possible). As a basis for this, we suppose that there is a suitable "causal ambit"[3] $\mathcal{A}$ of variables (not all necessarily observable) that we regard as relevant, and a "causal DAG"

[2] Neapolitan [2003, p. 56] has a different and more complex interpretation of "direct cause".

[3] The importance of the causal ambit will become apparent later.

$\mathcal{D}$ over a collection $\mathcal{V} \subseteq \mathcal{A}$. These ingredients are "known to Nature", though not necessarily to us: $\mathcal{D}$ is "Nature's DAG". Given such a causal DAG $\mathcal{D}$, for $X, Y \in \mathcal{V}$ we interpret "X is a *direct cause* of Y" as synonymous with "X is a parent of Y in $\mathcal{D}$", and similarly equate "*cause*" with "ancestor in $\mathcal{D}$". One can also use the causal DAG to introduce further graphically defined causal terms, such as "causal chain", "intermediate variable", ...

The concepts of causal ambit and causal DAG might be regarded as primitive notions, or attempts might be made to define them in terms of pre-existing understandings of causal concepts. In either case, it would be good to have criteria to distinguish a putative causal ambit from a non-causal ambit, and a causal DAG from a non-causal DAG.

For example, we typically read [Hernán and Robins 2006]:

> "A *causal DAG* $\mathcal{D}$ is a DAG in which:
>
> (i). the lack of an arrow from V_j to V_m can be interpreted as the absence of a direct causal effect of V_j on V_m (relative to the other variables on the graph)
>
> (ii). all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph.[4]

If we start with a DAG $\mathcal{D}$ over $\mathcal{V}$ that we accept as being a causal DAG, and interpret "direct cause" *etc.* in terms of that, then conditions (i) and (ii) will be satisfied by definition. However, this begs the question of how we are to tell a causal from a non-causal DAG.

More constructively, suppose we start with a prior understanding of the term "direct cause" (relative to $\mathcal{V}$)—for example, though by no means necessarily,[5] based on the agency interpretation described in § 2.1 above. It appears that we could then use the above definition to check whether a proposed DAG $\mathcal{D}$ is indeed "causal". But while this is essentially straightforward so far as condition (i) is concerned (except that there is no obvious reason to require a DAG representation), interpretation and implementation of condition (ii) is more problematic. First, what is a "common cause"? Spirtes et al. [2000, p. 44] say that a variable X is a common cause of variables Y and Z if and only if X is both a direct cause of Y and a direct cause of Z — but in each case relative to the set $\{X, Y, Z\}$, so that this definition is not dependent on the causal ambit $\mathcal{V}$. Neapolitan [2003, p. 57] has a different interpretation, which apparently is relative to an essentially arbitrary set $\mathcal{V}$ — but then states that that problems can arise when at least one common cause is not in $\mathcal{V}$, a possibility that seems to be precluded by his definition.

As another attempt at clarification, Spirtes and Scheines [2004] require "that the set of variables in the causal graph be *causally sufficient*, *i.e.* if $\mathcal{V}$ is the set of variables in the causal graph, that there is no variable L not in $\mathcal{V}$ that is a direct cause (relative to $\mathcal{V} \cup \{L\}$) of two variables in $\mathcal{V}$". If "$L \notin \mathcal{V}$ is not a direct cause of $V \in \mathcal{V}$" is interpreted in agency terms, it would mean that V would not respond to manipulations of L, when holding fixed all the other variables in $\mathcal{V}$. But whatever the interpretation of direct cause, such a "definition" of causal sufficiency is ineffective when the range of possible choices

[4] The motivation for this requirement is not immediately obvious, but is related to the defensibility of the *causal Markov* property described in § 2.3 below.

[5] See § 2.2 below.

for the additional variable L is entirely unrestricted—for then how could we ever be sure that it holds, without conducting an infinite search over all unmentioned variables L? That is why we posit an appropriate clearly-defined "causal ambit" $\mathcal{A}$: we can then restrict the search to $L \in \mathcal{A}$.

It seems to me that we should, realistically, allow that "causality" can operate, in parallel, at several different levels of granularity. Thus while it may or may not be possible to describe the medical effects of aspirin treatment in terms of quantum theory, even if we could, it would be a category error to try and do so in the context of a clinical trial. So there may be various different causal descriptions of the world, all operating at different levels, each with its associated causal ambit $\mathcal{A}$ of variables and various causal DAGs $\mathcal{D}$ over sets $\mathcal{V} \subseteq \mathcal{A}$. The meaning of any causal terms used should then be understood in relation to the appropriate level of description.

The obvious questions to ask about graphical causality, which are however not at all easy to answer, are: "When can a collection $\mathcal{A}$ of variables be regarded as a causal ambit?", and "When can a DAG be regarded as a causal DAG?".

In summary, so long as we *start* with a DAG $\mathcal{D}$ over $\mathcal{V}$ that we are willing to accept as a *causal* DAG (taken as a primitive concept), we can take $\mathcal{V}$ itself as our causal ambit, and use the structure of $\mathcal{D}$ to *define* causal terms. Without having a prior primitive notion of what constitutes a "causal DAG", however, conditions such as (i) and (ii) are unsatisfactory as a definition. At the very least, they require that we have specified (but how?) an appropriate causal ambit $\mathcal{A}$, relevant to our desired level of description, and have a clear pre-existing understanding (*i.e.* not based on the structure of $\mathcal{D}$, since that would be logically circular) of the terms "direct causal effect", "common cause" (perhaps relative to a set $\mathcal{V}$).

Agency causality and graphical causality

It is tempting to use the agency theory as a basis for such prior causal understanding. However, graphical causality does not really sit well with agency causality. For, as seen clearly in Example 1, in the agency intepretation it is perfectly possible for two variables each to have a direct effect on the other—which could not hold under any DAG representation. Similarly [Halpern and Pearl 2005; Hall 2000] there is no obvious reason to expect agency causality to be a transitive relation, which would again be a requirement under the graphical conception. For better or worse, the agency theory does not currently seem to be endowed with a sufficiently rich axiomatic structure to guide manipulations of its causal properties; and however such a general axiomatic structure might look, it would seem unduly restrictive to relate it closely to DAG models.

2.3 Probabilistic causality

Probabilistic Causality [Reichenbach 1956; Suppes 1970; Spohn 2001] depends on the existence and properties of a probability distribution P over quantities of interest. At its (over-)simplest, it equates *causality* with *probability raising*: "A is a *cause* of B" (where A and B are events) if $P(B \mid A) > P(B)$. This is more usefully re-expressed in its null form, and referred to random variables X and Y: X is not a cause of Y if the distribution of Y given X is the same as the marginal distribution of Y; and this is equivalent to

probabilistic independence of Y from X: $Y \perp\!\!\!\perp X$. But this is clearly unsatisfactory as it stands, since we could have dependence between X and Y, $Y \not\perp\!\!\!\perp X$, with, at the same time, conditional independence given some other variable (or set of variables) Z: $Y \perp\!\!\!\perp X \mid Z$. If Z can be regarded as delimiting the context in which we are considering the relationship between X and Y, we might still regard X and Y as "causally unrelated". Thus probabilistic causality is based on *conditional* (in)dependence properties of probability distributions. However there remain obvious problems in simply equating the non-symmetrical relation of cause-and-effect with the symmetrical relation of probabilistic (in)dependence, and with clarifying what counts as an appropriate conditioning "context" variable Z, so that additional structure and assumptions (*e.g.* related to an assumed "causal order", possibly but not necessarily temporal) are required to complete the theory.

Most modern accounts locate probabilistic causality firmly within the graphical conception — so inheriting all the features and difficulties of that approach. It is *assumed* that there is a DAG $\mathcal{D}$, over a suitable collection $\mathcal{V}$ of variables, such that

(i). $\mathcal{D}$ can be interpreted as a *causal* DAG; and, in addition,

(ii). the joint probability distribution P of the variables in $\mathcal{V}$ is *Markov* over $\mathcal{D}$, *i.e.* its probabilistic conditional independence (CI) properties are represented by the same DAG $\mathcal{D}$, according to the "d-separation" semantics described by Pearl [1986], Verma and Pearl [1990], Lauritzen et al. [1990].

In particular, from (ii), for any $V \in \mathcal{V}$, V is independent of its non-descendants, $\mathrm{nd}(V)$, in $\mathcal{D}$, given its parents, $\mathrm{pa}(V)$, in $\mathcal{D}$. Given the further interpretation (i) of $\mathcal{D}$ as a causal DAG, this can be expressed as "V is independent of its non-effects, given its direct causes in $\mathcal{V}$"— the so-called *causal Markov* assumption. Also, (ii) implies that, for any sets of variables X and Y in $\mathcal{D}$, $X \perp\!\!\!\perp Y \mid \mathrm{an}(X) \cap \mathrm{an}(Y)$ (where $\mathrm{an}(X)$ denotes the set of ancestors of X in $\mathcal{D}$, including X itself): again with $\mathcal{D}$ interpreted as causal, this can be read as saying "X and Y are conditionally independent, given their common causes in $\mathcal{V}$". In particular, marginal independence (where $X \perp\!\!\!\perp Y$ is represented in $\mathcal{D}$) holds if and only if $\mathrm{an}(X) \cap \mathrm{an}(Y) = \emptyset$, *i.e.* (using (i)) "$X$ and Y have no common cause" (including each other) in $\mathcal{V}$; in the "if" direction, this has been termed the *weak causal Markov* assumption [Scheines and Spirtes 2008]. Many workers regard the causal and weak causal Markov assumptions as compelling—but this must depend on making the "right" choice for $\mathcal{V}$ (essentially, through appropriate delineation of the causal ambit.)

Note that this conception of causality involves, simultaneously, two very different ways of interpreting the DAG $\mathcal{D}$ (see Dawid [2009] for more on this). The d-separation semantics by means of which we relate $\mathcal{D}$ to conditional independence properties of the joint distribution P, while clearly defined, are somewhat subtle: in particular, the arrows in $\mathcal{D}$ are somewhat incidental "construction lines", that only play a small rôle in the semantics. But as soon as we also give $\mathcal{D}$ an interpretation as a "causal DAG" we are into a completely different way of interpreting it, where the arrows themselves are regarded as directly carrying causal meaning. Probabilistic causality can thus be thought of as the progeny of a shotgun wedding between two ill-matched parties.

Causal discovery

The enterprise of *Causal Discovery* [Spirtes et al. 2000; Glymour and Cooper 1999; Neapolitan 2003] is grounded in this probabilistic-cum-graphical conception of causality. There are many variations, but all share the same basic philosophy. Essentially, one analyses observational data in an attempt to identify conditional independencies (possibly involving unobserved variables) in the distribution from which they arise. Some of these might be discarded as "accidental" (perhaps because they are inconsistent with an *a priori* causal order); those that remain might be represented by a DAG. The hope is that this discovered conditional independence DAG can also be interpreted as a causal DAG. When, as is often the case, there are several Markov equivalent DAG representations of the discovered CI relationships, which, moreover, cannot be causally distinguished on *a priori* grounds (*e.g.* in terms of an assumed causal order), this hope can not be fully realised; but if we can assume that one of these, at least, is a causal DAG, then at least an arrow common to all of them can be interpreted causally.

2.4 A spot of bother

Spirtes et al. [2000] and Pearl [2009], among others, have stressed the fundamental importance of distinguishing between the activities of *Seeing* and *Doing*. *Seeing* involves passive observation of a system in its natural state. *Doing*, on the other hand, relates to the behaviour of the system in a disturbed state brought about by external intervention. As a simple point of pure logic, there is no reason for there to be any relationship between these two types of behaviour of a system.

The probabilistic interpretation of causality relates solely to the *seeing* regime, whereas the agency account focuses entirely on what happens in *doing* regimes. As such these two interpretations inhabit totally unrelated universes. There are non-trivial foundational difficulties with the probabilistic (or other graphical) interpretations of causality (what exactly is a causal DAG? how will we know when we have got one?); on the other hand agency causality, while less obviously problematic and perhaps more naturally appealing, does not currently appear to offer a rich enough theory to be very useful. Even at a purely technical level, agency and probabilistic causality have very little in common. Probabilistic causality, through its close ties with conditional independence, has at its disposal the well-developed theoretical machinery of that concept, while the associated graphical structure allows for ready interpretation of concepts such as "causal pathway". Such considerations are however of marginal relevance to agency causality, which need not involve any probabilistic or graphical connexions.

From the point of view of a statistician, this almost total disconnect between the causal theories relating to the regimes of seeing and doing is particularly worrying. For one of the major purposes of "causal inference" is to draw conclusions, from purely observational "seeing" data on a system, about "doing": how would the system behave were we to intervene in it in certain ways? But not only is there no necessary logical connexion between the behaviours in the different regimes, the very concepts and representations by which we try to understand causality in the different regimes are worlds apart.

3 The Pearlian Synthesis

Building on ideas introduced by Spirtes et al. [2000], Pearl's approach to causality, as laid out for example in his book [Pearl 2009],[6] attempts to square this circle: it combines the two apparently incommensurable approaches of agency causality and probabilistic causality[7] in a way that tries to bring together the best features of both, while avoiding many of their individual problems and pitfalls.

Pearl considers a type of stochastic model, described by a DAG $\mathcal{D}$ over a collection $\mathcal{V}$ of variables, that can be simultaneously interpreted in terms of both agency and probabilistic causality. We could, if we wished, think of $\mathcal{V}$ as a "causal ambit", and $\mathcal{D}$ as a "causal DAG", but little is gained (or lost) by doing so, since the interpretations of any causal terms we may employ are provided internally by the model, rather than built on any pre-existing causal conceptions.

In its probabilistic interpretation, such a DAG $\mathcal{D}$ represents the conditional independence properties of the undisturbed system, which is supposed Markov with respect to $\mathcal{D}$. In its agency interpretation, the same DAG $\mathcal{D}$ is used to describe precisely how the system responds, probabilistically, to external interventions that set the values of (an arbitrary collection of) its variables. Specifically, such a disturbed probability distribution is supposed still Markov with respect to $\mathcal{D}$, and the conditional distribution of any variable V in $\mathcal{V}$, given its parents in $\mathcal{D}$, is supposed the same in all regimes, seeing or doing (except of course those that directly set the value of V itself, say at v, for which that distribution is replaced by the 1-point distribution at v). The "parent-child" conditional distributions thus constitute invariant "modular components" that (with the noted exception) can be transferred unchanged from one regime to another.

We term such a causal DAG model "Pearlian". Whether or not a certain DAG $\mathcal{D}$ indeed supplies a Pearlian DAG model for a given system can never be a purely syntactical question about its graphical structure, but is, rather, a semantic question about its relationship with the real world: do the various regimes actually have the probabilistic properties and relationships asserted? This may be true or false, but at least it is a meaningful question, and it is clear in principle how it can be addressed in purely empirical fashion: by observing and comparing the behaviours of the system under the various regimes.[8] A Pearlian DAG

[6] We in fact shall deal only with Pearl's earlier, fully stochastic, theory. More recently (see the second-half of Pearl [2009], starting with Chapter 7), he has moved to an interpretation of DAG models based on deterministic functional relationships, with stochasticity deriving solely from unobserved exogenous variables. That interpretation does however imply all the properties of the stochastic theory, and can be regarded as a specialisation of it. We shall not here be considering any features (such as the possibility of counterfactual analysis) dependent on the additional structure of Pearl's deterministic approach, since these only become relevant when analysing "causes of effects"—see Dawid [2000, 2002] for more on this.

[7] We have already remarked that probabilistic causality is itself the issue of an uneasy alliance between two quite different ways of interpreting graphs. Further miscegenation with the agency conception of causality looks like a eugenically risky endeavour!

[8] For this to be effective, the variables in $\mathcal{V}$ should have clearly-defined meanings and be observable in the real-world. Some Pearlian models incorporate unobservable latent variables without clearly identified external referents, in which case only the implications of such a model for the behaviour of observables can be put to empirical test.

model thus has the great virtue, all too rare in treatments of causality, of being totally clear and explicit about what is being said—allowing one to accord it, in a principled way, acceptance or rejection, as deemed appropriate, in any given application. And when a system can indeed be described by a Pearlian DAG, it is straightforward to learn (not merely qualitatively, but quantitatively too), from purely observational data, about the (probabilistic) effects of any interventions on variables in the system.

3.1 Justification

The falsifiability of the property of being a Pearlian DAG (unlike, for example, the somewhat ill-defined property of being a "causal DAG") is at once a great strength of the theory (especially for those with a penchant for Karl Popper's "falsificationist" Philosophy of Science), and something of an Achilles' heel. For all too often it will be impossible, for a variety of pragamatic, ethical or financial reasons, to conduct the experiments that would be needed to falsify the Pearlian assumptions. A lazy reaction might then simply be to assume that a DAG found, perhaps by "causal discovery", to represent observational conditional independencies, but without any interventions having been applied, is indeed Pearlian—and so also describes what would happen under interventions. While this may well be an interesting working hypothesis to guide further experimental investigations, it would be an illogical and dangerous point at which to conclude our studies. In particular, further experimental investigations could well result in rejection of our assumed Pearlian model.

Nevertheless, if forced to make a tentative judgment on the Pearlian nature, or otherwise, of a putative DAG model[9] of a system, there are a number of more or less reasonable, more or less intuitive, arguments that can be brought to bear. As a very simple example, we would immediately reject any putative "Pearlian DAG" in which an arrow goes backwards in time,[10] or otherwise conflicts with an accepted causal order. As another, if an "observational" regime itself involves an imposed physical randomisation to generate the value of some variable X, in a way that might possibly take account of variables Z temporally prior to X, we might reasonably regard the conditional distribution of some later variable Y, given X and Z, as a modular component, that would be the same in a regime that intervenes to set the value of X as it is in the (observational) randomisation regime.[11] Such arguments can be further extended to "natural experiments", where it is Nature that imposed the external randomisation. This is the case for "Mendelian randomisation" [Didelez and Sheehan 2007], which capitalises on the random assortment of genes under Mendelian genetics. Other natural experiments rely on other causal assumptions about Nature: thus the "discontinuity design" [Trochim 1984] assumes that Nature supplies continuous dose-response cause-effect relationships. But all such justifications are, and must be, based on (what we think are) properties of the real world, and not solely on the internal structure of

[9] Assumed, for the sake of non-triviality, already to be a Markov model of its observational probabilistic properties.

[10] Assuming, as most would accept, that an intervention in a variable at some time can not affect any variable whose value is determined at an earlier time.

[11] See Dawid [2009] for an attempted argument for this, as well as caveats as to its general applicability.

the putative Pearlian DAG. In particular, they are founded on pre-existing ideas we have about causal and non-causal processes in the world, even though these ideas may remain unformalised and woolly: the important point is that we have enough, perhaps tacit, shared understanding of such processes to convince both ourselves and others that they can serve as external justification for a suggested Pearlian model. Unless we have sufficient justification of this kind, all the beautiful analysis (*e.g.* in Pearl [2009]) that develops the implications of a Pearlian model will be simply irrelevant. To echo Cartwright [1994, Chapter 2], "No causes in, no causes out".

4 Modularity, extended conditional independence and decision-theoretic causality

Although Pearlian causality as described above appears to be closely tied to graphical representation, this is really an irrelevance. We can strip it of its graphical clothing, laying bare its core ingredient: the property that certain conditional distributions[12] are the same across several different regimes. This *modular* conception provides us with yet another interpretation of causality. When, as here, the regimes considered encompass both observation (seeing) and intervention (doing), it has the great advantage over other theories of linking those disparate universes, thus supporting *causal inference*.

The modularity assumption can be conveniently expressed formally in the algebraic language of conditional independence, suitably interpreted [Dawid 1979, 2002, 2009], making no reference to graphs. Thus let F be a "regime indicator", a non-stochastic parameter variable, whose value indicates the regime whose probabilistic properties are under consideration. If X and Y are stochastic variables, the "extended conditional independence" (ECI) property

(2) $\quad Y \perp\!\!\!\perp F \mid X$

can be interpreted as asserting that the conditional distribution of Y, for specified regime $F = f$ and given observed value $X = x$, depends only on x and not further on the regime f that is operating: in terms of densities we could write $p(y \mid f, x) = p(y \mid x)$. If F had been a stochastic variable this would be entirely equivalent to stochastic conditional independence of Y and F given X; but it remains meaningful, with the above interpretation, even when F is a non-stochastic regime indicator: Indeed, it asserts exactly the modular nature of the conditional distribution $p(y \mid x)$, as being the same across all the regimes indicated by values of F. Such modularity properties, when expressed in terms of ECI, can be formally manipulated—and, in those special cases where this is possible and appropriate, represented and manipulated graphically—in essentially the same fashion as for regular probabilistic conditional independence.

For applications of ECI to causal inference, we would typically want one or more of the regimes indicated by F to represent the behaviour of the system when subjected to an intervention of a specified kind—thus linking up nicely with the agency interpretation; and one

[12]More generally, we could usefully identify features of the different regimes other than conditional distributions—for example, conditional expectations, or odds ratios—as modular components.

regime to describe the undisturbed system on which observations are made—thus allowing the possibility of "causal inference" and making links with probabilistic causality, but in a non-graphical setting. Modularity/ECI assumptions can now be introduced, as considered appropriate, and their implications extracted by algebraic or graphical manipulations, using the established theory of conditional independence. We emphasise that, although the notation and technical machinery of conditional independence is being used here, this is applied in a way that is very different from the approach of probabilistic causality: no assumptions need be made connecting causal relationships with ordinary probabilistic conditional independence.

Because it concerns the probabilistic behaviour of a system under interventions—a particular interpretation of agency causality—this general approach can be termed "decision-theoretic" causality. With the emphasis now on modularity, intuitive or graphically motivated causal terms such as "direct effect" or "causal pathway" are best dispensed with (and with them such assumptions as the causal Markov property). The decision-theoretic approach should not be regarded as providing a philosophical foundation for "causality", or even as a way of interpreting causal terms, but rather as very useful machinery for expressing and manipulating whatever modularity assertions one might regard as appropriate in a given problem.

4.1 Intervention DAGs

The assumptions that are implicit in a Pearlian model can be displayed very explicitly in the decision-theoretic framework, by associating a non-stochastic "intervention variable" F_X with each "domain variable" $X \in \mathcal{V}$. The assumed ECI properties are conveniently displayed by means of a DAG, $\mathcal{D}^*$, which extends the Pearlian DAG $\mathcal{D}$ by adding extra nodes for these regime indicators, and extra arrows, from F_X to X for each $X \in \mathcal{V}$ [Spohn 1976; Spirtes et al. 2000; Pearl 2009; Dawid 2002; Dawid 2009]. If $\mathcal{X}$ is the set of values for X, then that for F_X is $\mathcal{X} \cup \{\emptyset\}$: the intended interpretation is that $F_X = \emptyset$ (the "idle" regime) corresponds to the purely observational regime, while $F_X = x \in \mathcal{X}$ corresponds to "setting" X at x.

To be precise, we specify the distribution of $X \in \mathcal{V}$ given its parents $(\mathrm{pa}(X), F_X)$ in $\mathcal{D}^*$ (where $\mathrm{pa}(X)$ denotes the "domain" parents of X, in $\mathcal{D}$) as follows. When $F_X = \emptyset$, this is the same as the observational conditional distribution of X, given $\mathrm{pa}(X)$; and when $F_X = x$ it is just a 1-point distribution on x, irrespective of the values of $\mathrm{pa}(X)$. The extended DAG $\mathcal{D}^*$, supplied with these parent-child specifications, is the *intervention DAG* representation of the problem.

With this construction, for any settings of all the regime indicators, some to idle and some to fixed values, the implied joint distribution of all the domain variables in that regime is exactly as required for the Pearlian DAG interpretation. But a valuable added bonus of the intervention DAG representation is that the Pearlian assumptions are explicitly represented. For example, the standard d-separation semantics applied to $\mathcal{D}^*$ allows us to read off the ECI property $X \perp\!\!\!\perp \{F_Y : Y \neq X\} \mid (\mathrm{pa}(X), F_X)$, which asserts the modular property of the conditional distribution of X given $\mathrm{pa}(X)$: when $F_X = \emptyset$ (the only non-trivial case) the

conditional distribution of X given $\mathrm{pa}(X)$ is the same, no matter how the other variables are set (or left idle).

4.2 More general causal models

It is implicit in the Pearlian conception that every variable in $\mathcal{V}$ should be manipulable (the causal Markov property then follows). But there is no real reason to require this. We can instead introduce intervention variables for just those variables that we genuinely wish to consider as "settable". The advantage of this is that fewer assumptions need be made and justified, but useful conclusions can often still be drawn.

EXAMPLE 2. **(Instrumental variable)**

Suppose we are interested in the "causal effect" of a binary exposure variable X on some response Y. However we can not directly manipulate X. Moreover the observational relationship between X and Y may be distorted because of an unobserved "confounder" variable, U, associated with both X and Y. In an attempt to evade this difficulty, we also measure an "instrumental variable" Z.

To express our interest in the *causal* effect of X on Y, we introduce an intervention variable F_X associated with X, defined and interpreted exactly as in § 4.1 above. The aim of our causal inference is to make some kind of comparison between the distributions of the response Y in the interventional regimes, $F_X = 0$ and $F_X = 1$, corresponding to manipulating the value of X. The available data, however, are values of (X, Y, Z) generated under the observational regime, $F_X = \emptyset$. We must make some assumptions if we are to be able to use features of that observational joint distribution to address our causal question, and clearly these must involve some kind of transference of information across regimes.

A useful (when valid!) set of assumptions about the relationships between all the variables in the problem is embodied in the following set of ECI properties (the "core conditions"[13] for basing causal inferences on an instrumental variable):

$$(U, Z) \quad \perp\!\!\!\perp \quad F_X \tag{3}$$

$$U \quad \perp\!\!\!\perp \quad Z \mid F_X \tag{4}$$

$$Y \quad \perp\!\!\!\perp \quad F_X \mid (X, U) \tag{5}$$

$$Y \quad \perp\!\!\!\perp \quad Z \mid (X, U; F_X) \tag{6}$$

$$X \quad \not\perp\!\!\!\perp \quad Z \mid F_X = \emptyset \tag{7}$$

Property (3) is to be interpreted as saying that the joint distribution of (U, Z) is independent of the regime F_X: *i.e.*, it is the same in all three regimes. That is to say, it is entirely unaffected by whether, and if so how, we intervene to set the value of X. The identity of this joint distribution across the two interventional regimes, $F_X = 0$ and $F_X = 1$, can be interpreted as expressing a causal property: manipulating X has no (probabilistic) effect

[13] In addition to these core conditions, precise identification of a causal effect by means of an instrumental variable requires further modelling assumptions, such as linear regressions [Didelez and Sheehan 2007].

on the pair of variables (U, Z). Moreover, since this common joint distribution is also supposed the same in the idle regime, $F_X = \emptyset$, we could in principle use observational data to estimate it—thus opening up the possibility of causal inference.

Property (4) asserts that, in their (common) joint distribution in any regime, U and Z are independent (this however is a purely probabilistic, not a causal, property).

Property (5) says that the conditional distribution of Y given (X, U) is the same in both interventional regimes, as well as in the observational regime, and can thus be considered as a modular component, fully transferable between the three regimes—again, I regard this as expressing a causal property.

Property (6) asserts that this common conditional distribution is unaffected by further conditioning on Z (not in itself a causal property).

Finally, property (7) requires that Z be genuinely associated with X in the observational regime.

Of course, these ECI properties should not simply be assumed without some attempt at justification: for example, Mendelian randomisation attempts this in the case that Z is an inherited gene. But because we have no need to consider interventions at any node other than X, less by way of justification is required than if we were to do so.

Once expressed in terms of ECI, these core conditions can be manipulated algebraically using the general theory of conditional independence [Dawid 1979]. Depending on what further modelling assumptions are made, it may then be possible to identify, or to bound, the desired causal effect in terms of properties of the observational joint distribution of (X, Y, Z) [Dawid 2007b, Chapter 11].

In this particular case, although the required ECI conditions are expressed without reference to any graphical representation, it is possible (though not obligatory!) to give them one. This is shown in Figure 1. Properties (3)–(6) can be read off this DAG directly using the standard d-separation semantics. (Property (7) is only represented under a further assumption that the graphical representation is faithful.) We term such a DAG an *augmented DAG*: it differs from a Pearlian DAG in that some, but not necessarily all, variables have associated intervention indicators.

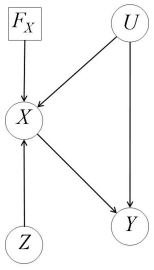


Figure 1. Instrumental variable: Augmented DAG representation

Just as for regular CI, it is possible for a collection of ECI properties, constituting a

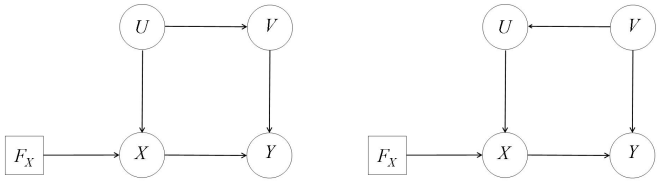


Figure 2. Two Markov-equivalent augmented DAGs

decision-theoretic causal model, to have no (augmented) DAG representation, or more than one. This latter is the case for Figure 2, where the direction of the arrow between U and V is not determined. This emphasises that, even when we do have an augmented DAG representation, we can not necessarily interpret the direction of an arrow in it as directly related to the direction of causality. Even in Figure 1 (and in spite of the natural connotation of the term "instrument"), the arrow pointing from Z to X is not be interpreted as necessarily causal, since the dependence between Z and X could be due to a "common cause" U^* without affecting the ECI properties (3)–(6) [Dawid 2009], and Figure 1 is merely a graphical representation of these properties, based on d-separation semantics. In particular, one should be cautious of using an augmented DAG, which is nothing but a way of representing certain ECI statements, to introduce graphically motivated concepts such as "causal pathway". The general decision-theoretic description of causality *via* modularity, expressed in terms of ECI properties, where there is no requirement that the assumptions be representable by means of an augmented DAG at all, allows us to evade some of the restrictions of graphical causality, while still retaining a useful "agency-cum-probabilistic" causal theory.

The concept of an "interventional regime" can be made much more general, and in particular we need not require that it have the properties assumed above for an intervention variable associated with a domain variable. We could, for example, incorporate "fat hand" interventions that do not totally succeed in their aim of setting a variable to a fixed value, or interventions (such as kicking the system) that simultaneously affect several domain variables [Duvenaud et al. 2009]. So long as we understand what such regimes refer to in the real world, and can make and justify assumptions of modularity of appropriate conditional distributions as we move across regimes, we can apply the decision-theoretic ECI machinery. And at this very general level we can even apply a variant of "causal discovery" algorithms—so long as we can make observations under all the regimes considered.[14] For example, if we can observe (X, Y) under the different regimes described by F, we can readily investigate the validity of the ECI property $X \perp\!\!\!\perp F \mid Y$ using standard tests (*e.g.*

[14]Or we might make parametric modelling assumptions about the relationships across regimes, to fill in for regimes we are not able to observe. This would be required for example when want to consider the effect of setting the value of a continuous "dose" variable. At this very general level we can even dispense entirely with the assumption of modular conditional distributions [Duvenaud et al. 2009].

the χ^2-test) for conditional independence. Such discovered ECI properties (whether or not they can be expressed graphically) can then be used to model the "causal structure" of the problem.

5 Conclusion

Over many years, Judea Pearl's original and insightful approach to understanding uncertainty and causality have had an enormous influence on these fields. They have certainly had a major influence on my own research directions: I have often—as evidenced by this paper—found myself following in his footsteps, picking up a few crumbs here and there for further digestion.

Pearl's ideas do not however exist in a vacuum, and I believe it is valuable both to relate them to their precursors and to assess the ways in which they may develop. In attempting this task I fully acknowledge the leadership of a peerless researcher, whom I feel honoured to count as a friend.

References

Campaner, R. and M. C. Galavotti (2007). Plurality in causality. In P. K. Machamer and G. Wolters (Eds.), *Thinking About Causes: From Greek Philosophy to Modern Physics*, pp. 178–199. Pittsburgh: University of Pittsburgh Press.

Cartwright, N. (1994). *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

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

Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association 95*, 407–448.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review 70*, 161–189. Corrigenda, *ibid*., 437.

Dawid, A. P. (2007a). Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. In F. Russo and J. Williamson (Eds.), *Causality and Probability in the Sciences*, Volume 5 of *Texts in Philosophy*, pp. 503–32. London: College Publications.

Dawid, A. P. (2007b). Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London. <http://www.ucl.ac.uk/Stats/research/reports/psfiles/rr279.pdf>

Dawid, A. P. (2010). Beware of the DAG! *Journal of Machine Learning Research*. To appear.

Dawid, A. P. (2010). The rôle of scientific and statistical evidence in assessing causality. In R. Goldberg, J. Paterson, and G. Gordon (Eds.), *Perspectives on Causation*, Oxford. Hart Publishing. To appear.

Didelez, V. and N. A. Sheehan (2007). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research 16*, 309–330.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Duvenaud, D., D. Eaton, K. Murphy, and M. Schmidt (2010). Causal learning without DAGs. *Journal of Machine Learning Research*. To appear.

Galavotti, M. C. (2008). Causal pluralism and context. In M. C. Galavotti, R. Scazzieri, and P. Suppes (Eds.), *Reasoning, Rationality and Probability*, Chapter 11, pp. 233–252. Chicago: The University of Chicago Press.

Glymour, C. and G. F. Cooper (Eds.) (1999). *Computation, Causation and Discovery*. Menlo Park, CA: AAAI Press.

Hall, N. (2000). Causation and the price of transitivity. *Journal of Philosophy XCVII*, 198–222.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science 56*, 843–887.

Hausman, D. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.

Hernán, M. A. and J. M. Robins (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology 17*, 360–372.

Hitchcock, C. (2007). How to be a causal pluralist. In P. K. Machamer and G. Wolters (Eds.), *Thinking About Causes: From Greek Philosophy to Modern Physics*, pp. 200–221. Pittsburgh: University of Pittsburgh Press.

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

Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Upper Saddle River, New Jersey: Prentice Hall.

Pearl, J. (1986). A constraint–propagation approach to probabilistic reasoning. In L. N. Kanal and J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence*, Amsterdam, pp. 357–370. North-Holland.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (Second ed.). Cambridge: Cambridge University Press.

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science 42*, 157–176.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of Los Angeles Press.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics 6*, 34–68.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Scheines, R. and P. Spirtes (2008). Causal structure search: Philosophical foundations and future problems. Paper presented at NIPS 2008 Workshop "Causality: Objectives and Assessment", Whistler, Canada.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction and Search* (Second ed.). New York: Springer-Verlag.

Spirtes, P. and R. Scheines (2004). Causal inference of ambiguous manipulations. *Philosophy of Science 71*, 833–845.

Spohn, W. (1976). *Grundlagen der Entscheidungstheorie*. Ph.D. thesis, University of Munich. (Published: Kronberg/Ts.: Scriptor, 1978).

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In M. C. Galavotti, P. Suppes, and D. Costantini (Eds.), *Stochastic Dependence and Causality*, Chapter 9, pp. 157–172. Chicago: University of Chicago Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North Holland.

Trochim, W. M. K. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. SAGE Publications.

Verma, T. and J. Pearl (1990). Causal networks: Semantics and expressiveness. In R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4*, Amsterdam, pp. 69–76. North-Holland.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

19

Effect Heterogeneity and Bias in Main-Effects-Only Regression Models

FELIX ELWERT AND CHRISTOPHER WINSHIP

1 Introduction

The overwhelming majority of OLS regression models estimated in the social sciences, and in sociology in particular, enter all independent variables as main effects. Few regression models contain many, if any, interaction terms. Most social scientists would probably agree that the assumption of constant effects that is embedded in main-effects-only regression models is theoretically implausible. Instead, they would maintain that regression effects are historically and contextually contingent; that effects vary across individuals, between groups, over time, and across space. In other words, social scientists doubt constant effects and believe in effect heterogeneity.

But why, if social scientists believe in effect heterogeneity, are they willing to substantively interpret main-effects-only regression models? The answer—not that it's been discussed explicitly—lies in the implicit assumption that the main-effects coefficients in linear regression represent straightforward averages of heterogeneous individual-level causal effects.

The belief in the averaging property of linear regression has previously been challenged. Angrist [1998] investigated OLS regression models that were correctly specified in all conventional respects except that effect heterogeneity in the main treatment of interest remained unmodeled. Angrist showed that the regression coefficient for this treatment variable gives a rather peculiar type of average—a conditional variance weighted average of the heterogeneous individual-level treatment effects in the sample. If the weights differ greatly across sample members, the coefficient on the treatment variable in an otherwise well-specified model may differ considerably from the arithmetic mean of the individual-level effects among sample members.

In this paper, we raise a new concern about main-effects-only regression models. Instead of considering models in which heterogeneity remains unmodeled in only one effect, we consider standard linear path models in which unmodeled heterogeneity is potentially pervasive.

Using simple examples, we show that unmodeled effect heterogeneity in more than one structural parameter may mask confounding and selection bias, and thus lead to biased estimates. In our simulations, this heterogeneity is indexed by latent (unobserved) group membership. We believe that this setup represents a fairly realistic scenario—one in which the analyst has no choice but to resort to a main-effects-only regression model because she cannot include the desired interaction terms since group-membership is un-

observed. Drawing on Judea Pearl's theory of directed acyclic graphs (DAG) [1995, 2009] and VanderWeele and Robins [2007], we then show that the specific biases we report can be predicted from an analysis of the appropriate DAG. This paper is intended as a serious warning to applied regression modelers to beware of unmodeled effect heterogeneity, as it may lead to gross misinterpretation of conventional path models.

We start with a brief discussion of conventional attitudes toward effect heterogeneity in the social sciences and in sociology in particular, formalize the notion of effect heterogeneity, and briefly review results of related work. In the core sections of the paper, we use simulations to demonstrate the failure of main-effects-only regression models to recover average causal effects in certain very basic three-variable path models where unmodeled effect heterogeneity is present in more than one structural parameter. Using DAGs, we explain which constellations of unmodeled effect heterogeneity will bias conventional regression estimates. We conclude with a summary of findings.

2 A Presumed Averaging Property of Main-Effects-Only Regression

2.1 Social Science Practice

The great majority of empirical work in the social sciences relies on the assumption of constant coefficients to estimate OLS regression models that contain nothing but main effect terms for all variables considered.[1] Of course, most researchers do not believe that real-life social processes follow the constant-coefficient ideal of conventional regression. For example, they aver that the effect of marital conflict on children's self-esteem is larger for boys than for girls [Amato and Booth 1997]; or that the death of a spouse increases mortality more for white widows than for African American widows [Elwert and Christakis 2006]. When pressed, social scientists would probably agree that the causal effect of almost any treatment on almost any outcome likely varies from group to group, and from person to person.

But if researchers are such firm believers in effect heterogeneity, why is the constant-coefficients regression model so firmly entrenched in empirical practice? The answer lies in the widespread belief that the coefficients of linear regression models estimate averages of heterogeneous parameters—average causal effects—representing the average of the individual-level causal effects across sample members. This (presumed) averaging property of standard regression models is important for empirical practice for at least three reasons. First, sample sizes in the social sciences are often too small to investigate effect heterogeneity by including interaction terms between the treatment and more than a few common effect modifiers (such as sex, race, education, income, or place of residence); second, the variables needed to explicitly model heterogeneity may well not have been measured; third, and most importantly, the complete list of effect modifiers along which the causal effect of treatment on the outcome varies is typically unknown (indeed, unknowable) to the analyst in any specific application. Analysts thus rely on faith that

[1] Whether a model requires an interaction depends on the functional form of the dependent and/or independent variables. For example, a model with no interactions in which the independent variables are entered in log form, would require a whole series of interactions in order to approximate this function if the independent variables where entered in nonlog form.

their failure to anticipate and incorporate all dimensions of effect heterogeneity into regression analysis simply shifts the interpretation of regression coefficients from individual-level causal effects to average causal effects, without imperiling the causal nature of the estimate.

2.2 Defining Effect Heterogeneity

We start by developing our analysis of the consequences of causal heterogeneity within the counterfactual (potential outcomes) model. For a continuous treatment $T \in (-\infty,\infty)$, let $T = t$ denote some specific treatment value and $T = 0$ the control condition. $Y(t)_i$ is the potential outcome of individual i for treatment $T = t$, and $Y(0)_i$ is the potential outcome of individual i for the control condition. For a particular individual, generally only one value of $Y(t)_i$ will be observed. The *individual-level causal effect* (ICE) of treatment level $T = t$ compared to $T = 0$ is then defined as: $\delta_{i,t} = Y(t)_i - Y(0)_i$ (or δ_i, for short, if T is binary).

Since $\delta_{i,t}$ is generally not directly estimable, researchers typically attempt estimating the *average causal effect* (ACE) for some sample or population:

$$\bar{\delta}_t = \sum_{i=1}^{N} \delta_{i,t} / N$$

We say that the effect of treatment T is *heterogeneous* if: $\delta_{i,t} \neq \bar{\delta}_t$ for at least one i.

In other words, effect heterogeneity exists if the causal effect of the treatment differs across individuals. The basic question of this paper is whether a regression estimate for the causal effect of the treatment can be interpreted as an average causal effect if effect heterogeneity is present.

2.3 Regression Estimates as Conditional Variance Weighted Average Causal Effects

The ability of regression to recover average causal effects under effect heterogeneity has previously been challenged by Angrist [1998].[2] Here, we briefly sketch the main result. For a binary treatment, T=0,1, Angrist assumed a model where treatment was ignorable given covariates X and the effect of treatment varied across strata defined by the values of X. He then analyzed the performance of an OLS regression model that properly controlled for confounding in X but was misspecified to include only a main effect term for T and no interactions between T and X. Angrist showed that the regression estimate for the main effect of treatment can be expressed as a weighted average of stratum-specific treatment effects, albeit one that is difficult to interpret. For each stratum defined by fixed values of X, the numerator of the OLS estimator has the form $\delta_x W_x P(X=x)$,[3] where δ_x is the stratum-specific causal effect and $P(X=x)$ is the relative size of the stratum in the sample. The weight, W_x, is a function of the propensity score, $P_x=P(T=1 \mid X)$, associated with the stratum, $W_x = P_x (1- P_x)$, which equals the stratum-specific variance of treatment. This variance, and hence the weight, is largest if $P_x=.5$ and smaller as P_x goes to 0 or 1.

[2]This presentation follows Angrist [1998] and Angrist and Pischke [2009].

[3]The denominator of the OLS estimator is just a normalizing constant that does not aid intuition.

If the treatment effect is constant across strata, these weights make good sense. OLS gives the minimum variance linear unbiased estimator of the model parameters under homoscedasticity assuming correct specification of the model. Thus in a model without interactions between treatment and covariates X the OLS estimator gives the most weight to strata with the smallest variance for the estimated within-stratum treatment effect, which, not considering the size of the strata, are those strata with the largest treatment variance, i.e. with the P_x that are closest to .5. However, if effects are heterogeneous across strata, this weighting scheme makes little substantive sense: in order to compute the average causal effect, $\bar{\delta}$, as defined above, we would want to give the same weight to every individual in the sample. As a variance-weighted estimator, however, regression estimates under conditions of unmodeled effect heterogeneity do not give the same weight to every individual in the sample and thus do not converge to the (unweighted) average treatment effect.

3 Path Models with Pervasive Effect Heterogeneity

Whereas Angrist analyzed a misspecified regression equation that incorrectly assumed no treatment-covariate interaction for a *single* treatment variable, we investigate the ability of a main-effects-only regression model to recover unbiased average causal effects in simple path models with unmodeled effect heterogeneity across *multiple* parameters.

Setup: To illustrate how misleading the belief in the averaging power of the constant-coefficient model can be in practice, we present simulations of basic linear path models, shown in summary in Figure 1 (where we have repressed the usual uncorrelated error terms).

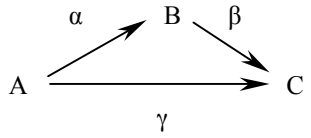


Figure 1. A simple linear path model

To introduce effect heterogeneity, let G = 0, 1 index membership in a latent group and permit the possibility that the three structural parameters α, β, and γ vary across (but not within) levels of G. The above path model can then be represented by two linear equations: $B = A\alpha_G + \varepsilon_B$ and $C = A\gamma_G + B\beta_G + \varepsilon_C$. In our simulations, we assume that A~N(0,1) and ε_B, and ε_C are iid N(0,1), and hence all variables are normally distributed. From these equations, we next simulate populations of N=100,000 observations, with P(G=1) = P(G=0) = 1/2. We start with a population in which all three parameters are constant across the two subgroups defined by G, and then systematically introduce effect heterogeneity by successively permitting the structural parameters to vary by group, yielding one population for each of the $2^3 = 8$ possible combinations of constant/varying parameters. To fix ideas, we choose the group-specific parameter values shown in Table

1. For simulations in which one or more parameters do not vary by group, we set the constant parameter(s) to the average of the group specific parameters, e.g. $\alpha = (\alpha_0 + \alpha_1)/2$.

Table 1: Group-specific structural parameters for simulations

	α_G	β_G	γ_G
Group:			
G=0	0.4	0.5	0.6
G=1	1.2	2.5	1.4
Average	0.8	1.5	1.0

Finally, we estimate a conventional linear regression model for the effects of A and B on C using the conventional default specification, in which all variables enter as main effects only, $C = A\gamma + B\beta + \varepsilon$. (Note that G is latent and therefore cannot be included in the model.) The parameter, γ refers to the direct effect of A on C holding B constant, and β refers to the total effect of B on C.[4] In much sociological and social science research, this main-effects regression model is intended to recover average structural (causal) effects, and is commonly believed to be well suited for the purpose.

Results: Table 2 shows the regression estimates for the main effect parameters across the eight scenarios of effect heterogeneity. We see that the main effects regression model correctly recovers the desired (average) parameters, γ=1 and β=1.5 if none of the parameters vary across groups (column 1), or if only one of the three parameters varies (columns 2-4).

Other constellations of effect heterogeneity, however, produce biased estimates. If α_G and β_G (column 5); or α_G and γ_G (column 6); or α_G, β_G, and γ_G (column 8) vary across groups, the main-effects-only regression model fails to recover the true (average) parameter values known to underlie the simulations. For our specific parameter values, the estimated (average) effect of B on C in these troubled scenarios is always too high, and the estimated average direct effect of A on C is either too high or too low. Indeed, if we set γ=0 but let α_G and β_G vary across groups, the estimate for γ in the main-effects-only regression model would suggest the presence of a direct effect of A on C even though it is known by design that no such direct effect exists (not shown).

Failure of the regression model to recover the known path parameters is not merely a function of the number of paths that vary. Although none of the scenarios in which fewer than two parameters vary yield incorrect estimates, and the scenario in which all three parameters vary is clearly biased, results differ for the three scenarios in which exactly two parameters vary. In two of these scenarios (columns 5 and 6), regression fails to recover the desired (average) parameters, while regression does recover the correct average parameters in the third scenario (column 7).

[4]The notion of direct and indirect effects is receiving deserved scrutiny in important recent work by Robins and Greenland [1992]; Pearl [2001]; Robins [2003]; Frangakis and Rubin [2002]; Sobel [2008]; and VanderWeele [2008].

Table 2: OLS regression estimates for the main effects of A and B on C across eight different combinations of effect heterogeneity in α, β,and/or γ

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Heterogeneity in: -		α	β	γ	α, β	α, γ	β, γ	α, β, γ
Group:	G0 G1	G0 G1	G0 G1	G0 G1	G0 G1	G0 G1	G0 G1	G0 G1
α	0.8	0.4 1.2	0.8	0.8	0.4 1.2	0.4 1.2	0.8	0.4 1.2
β	1.5	1.5	0.5 2.5	1.5	0.5 2.5	1.5	0.5 2.5	0.5 2.5
γ	1.0	1.0	1.0	0.6 1.4	1.0	0.6 1.4	0.6 1.4	0.6 1.4
Pooled OLS estimate:								
β	1.50	1.50	1.50	1.50	**1.77**	**1.64**	1.50	**1.91**
γ	1.00	1.00	1.00	1.00	**1.17**	**0.89**	1.00	**1.07**

Note: Bold estimates are biased for the true (average) parameters. Results from independent simulations of N=100,000 for each scenario using (group-specific) parameters listed above. See text for details.

In sum, the naïve main-effects-only linear regression model recovers the correct (average) parameter values only under certain conditions of limited effect heterogeneity, and it fails to recover the true average effects in certain other scenarios, including the scenario we consider most plausible in the majority of sociological applications, i.e., where all three parameters vary across groups. If group membership is latent—because group membership is unknown to or unmeasured by the analyst— and thus unmodeled, linear regression generally will fail to recover the true average effects.

4 DAGs to the Rescue

These results spell trouble for empirical practice in sociology. Judea Pearl's work on causality and directed acyclic graphs (DAGs) [1995, 2009] offers an elegant and powerful approach to understanding the problem. Focusing on the appropriate DAGs conveys the critical insight for the present discussion that effect heterogeneity, rather than being a nuisance that is easily averaged away, encodes structural information that analysts ignore at their peril.

Pearl's DAGs are nonparametric path models that encode causal dependence between variables: an arrow between two variables indicates that the second variable is causally dependent on the first (for detailed formal expositions of DAGs, see Pearl [1995, 2009]; for less technical introductions see Robins [2001]; Greenland, Pearl and Robins [1999] in epidemiology, and Morgan and Winship [2007] in sociology). For example, the DAG in Figure 2 indicates that Z is a function of X and Y, $Z= f(X,Y,\varepsilon_Z)$, where ε_Z is an unobserved error term independent of (X,Y).

In a non-parametric DAG—as opposed to a conventional social science path model—the term f( ) can be any function. Thus, the DAG in Figure 2 is consistent with a linear structural equation in which X only modifies (i.e. introduces heterogeneity into) the effect

of Y on Z, $Z=Y\xi + YX\psi + \varepsilon_Z$.[5] In the language of VanderWeele and Robins [2007], who provide the most extensive treatment of effect heterogeneity using DAGs to date, one may call X a "direct effect modifier" of the effect of Y on Z. The point is that a variable that modifies the effect of Y on Z is causally associated with Z, as represented by the arrow from X to Z.

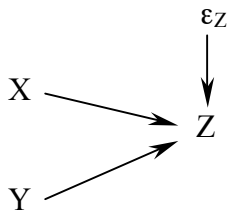


Figure 2. DAG illustrating direct effect modification of the effect of Y on Z in X

Returning to our simulation, one realizes that the social science path model of Figure 1, although a useful tool for informally illustrating the data generation process, does not, generally, provide a sufficiently rigorous description of the causal structure underlying the simulations. Figure 1, although truthfully representing the separate data generating mechanism for each group and each individual in the simulated population, is not the correct DAG for the pooled population containing groups G = 0 and G = 1 for all of the heterogeneity scenarios considered above. Specifically, in order to turn the informal social science path model of Figure 1 into a DAG, one would have to integrate the source of heterogeneity, G, into the picture. How this is to be done depends on the structure of heterogeneity. If only β_G (the effect of B on C) and/or γ_G (the direct effect of A on C holding B constant) varied with G, then one would add an arrow from G into C. If α_G (the effect of A on B) varied with G, then one would add an arrow from G into B. The DAG in Figure 3 thus represents those scenarios in which α_G as well as either β_G or γ_G, or both, vary with G (columns 5, 6, and 8). Interpreted in terms of a linear path model, this DAG is consistent with the following two structural equations: $B = A\alpha_0 + AG\alpha_1 + \varepsilon_{\mathbf{B}}$ and $C = A\gamma_0 + AG\gamma_1 + B\beta_0 + BG\beta_1 + \varepsilon_C$ (where the iid errors, ε, have been omitted from the DAG and are assumed to be uncorrelated).[6]

In our analysis, mimicking the reality of limited observational data with weak substantive theory, we have assumed that A, B, and C are observed, but that G is not observed. It is immediately apparent that the presence of G in Figure 3 means that, first, G is a confounder for the effect of B on C; and, second, that B is a "collider" [Pearl 2009] on

[5]It is also consistent with an equation that adds a main effect of X. For the purposes of this paper it does not matter whether the main effect is present.

[6]By construction of the example, we assume that A is randomized and thus marginally independent of G. Note, however, that even though G is mean independent of B and C (no main effect of G on either B or C), G is not marginally independent of B or C because $var(B|G=1)\neq var(B|G=0)$ and $var(C|G=1)\neq var(C|G=0)$, which explains the arrows from G into B and C. Adding main effects of G on B and C would not change the arguments presented here.

the path from A to C via B and G. Together, these two facts explain the failure of the main-effects-only regression model to recover the true parameters in panels 5, 6, and 8: First, in order to recover the effect of B on C, β, one would need to condition on the confounders A and G. But G is latent so it cannot be conditioned on. Second, conditioning on the collider B in the regression opens a "backdoor path" from A to C via B and G (when G is not conditioned on), i.e. it induces a non-causal association between A and C, creating selection bias in the estimate for the direct effect of A on C, γ [Pearl 1995, 2009; Hernán et al 2004]. Hence, both coefficients in the main-effects-only regression model will be biased for the true (average) parameters.

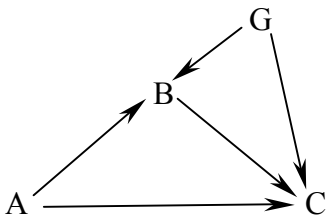


Figure 3. DAG consistent with effect modification of the effects of A on B, and B on C and/or A on C, in G

By contrast, if G modifies neither β nor γ, then the DAG would not contain an arrow from G into C; and if G does not modify α then the DAG would not contain an arrow from G into B. Either way, if either one (or both) of the arrows emanating from G are missing, then G is not a confounder for the effect of B on C, and conditioning on B will not induce selection bias by opening a backdoor path from A to C. Only then would the main effects regression model be unbiased and recover the true (average) parameters, as seen in panels 1-4 and 7.

In sum, Pearl's DAGs neatly display the structural information encoded in effect heterogeneity [VanderWeele and Robins 2007]. Consequently, Pearl's DAGs immediately draw attention to problems of confounding and selection bias that can occur when more than one effect in a causal system varies across sample members. Analyzing the appropriate DAG, the failure of main-effects-only regression models to recover average structural parameters in certain constellations of effect heterogeneity becomes predictable.

5 Conclusion

This paper considered a conventional structural model of a kind commonly used in the social sciences and explored its performance under various basic scenarios of effect heterogeneity. Simulations show that the standard social science strategy of dealing with effect heterogeneity—by ignoring it—is prone to failure. In certain situations, the main-effects-only regression model will recover the desired quantities, but in others it will not. We believe that effect heterogeneity in all arrows of a path model is plausible in many, if not most, substantive applications. Since the sources of heterogeneity are often not theorized, known, or measured, social scientists continue routinely to estimate main-effects-

only regression models in hopes of recovering average causal effects. Our examples demonstrate that the belief in the averaging powers of main-effects-only regression models may be misplaced if heterogeneity is pervasive, as estimates can be mildly or wildly off the mark. Judea Pearl's DAGs provide a straightforward explanation for these difficulties—DAGs remind analysts that effect heterogeneity may encode structural information about confounding and selection bias that requires consideration when designing statistical strategies for recovering the desired average causal effects.

Acknowledgments: We thank Jamie Robins for detailed comments on a draft version of this paper, and Michael Sobel, Stephen Morgan, Hyun Sik Kim, and Elizabeth Wrigley-Field for advice. Genevieve Butler provided editorial assistance.

References

Amato, Paul R., and Alan Booth. (1997). *A Generation at Risk: Growing Up in an Era of Family Upheaval*. Cambridge, MA: Harvard University Press.

Angrist, Joshua D. (1998). "Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Date on Military Applicants." *Econometrica 66*: 249-88.

Angrist, Joshua D. and Jörn-Steffen Pischke. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Elwert, Felix, and Nicholas A. Christakis. (2006). "Widowhood and Race." *American Sociological Review 71*: 16-41.

Frangakis, Constantine E., and Donald B. Rubin. (2002). "Principal Stratification in Causal Inference." *Biometrics 58*: 21–29.

Greenland, Sander, Judea Pearl, and James M. Robbins. (1999). "Causal Diagrams for Epidemiologic Research." *Epidemiology 10*: 37-48.

Hernán, Miguel A., Sonia Hernández-Diaz, and James M. Robins. (2004). "A Structural Approach to Section Bias." *Epidemiology 155* (2): 174-184.

Morgan, Stephen L. and Christopher Winship. (2007). *Counterfactuals and Causal Inference: Methods and Principles of Social Research.* Cambridge: Cambridge University Press.

Pearl, Judea. (1995). "Causal Diagrams for Empirical Research." *Biometrika 82* (4): 669-710.

Pearl, Judea. (2001). "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann, 411-420.

Pearl, Judea. (2009). *Causality: Models, Reasoning, and Inference*. Second Edition. Cambridge: Cambridge University Press.

Robins, James M. (2001). "Data, Design, and Background Knowledge in Etiologic Inference," *Epidemiology 11* (3): 313-320.

Robins, James M. (2003). "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." In: *Highly Structured Stochastic Systems*, P. Green, N. Hjort and S. Richardson, Eds. Oxford: Oxford University Press.

Robins, James M, and Sander Greenland. (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology 3*:143-155.

Sobel, Michael. (2008). "Identification of Causal Parameters in Randomized Studies with Mediating Variables," *Journal of Educational and Behavioral Statistics 33* (2): 230-251.

VanderWeele, Tyler J. (2008). "Simple Relations Between Principal Stratification and Direct and Indirect Effects." *Statistics and Probability Letters 78*: 2957-2962.

VanderWeele, Tyler J. and James M. Robins. (2007). "Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs." *Epidemiology 18* (5): 561-568.

20

Causal and Probabilistic Reasoning in P-log

Michael Gelfond and Nelson Rushton

1 Introduction

In this paper we give an overview of the knowledge representation (KR) language P-log [Baral, Gelfond, and Rushton 2009] whose design was greatly influenced by work of Judea Pearl. We introduce the syntax and semantics of P-log, give a number of examples of its use for knowledge representation, and discuss the role Pearl's ideas played in the design of the language. Most of the technical material presented in the paper is not new. There are however two novel technical contributions which could be of interest. First we expand P-log semantics to allow domains with infinite Herbrand bases. This allows us to represent infinite sequences of random variables and (indirectly) continuous random variables. Second we generalize the logical base of P-log which improves the degree of elaboration tolerance of the language.

The goal of the P-log designers was to create a KR-language allowing natural and elaboration tolerant representation of commonsense knowledge involving logic and probabilities. The logical framework of P-log is Answer Set Prolog (ASP) — a language for knowledge representation and reasoning based on the answer set semantics (*aka* stable model semantics) of logic programs [Gelfond and Lifschitz 1988; Gelfond and Lifschitz 1991]. ASP has roots in declarative programing, the syntax and semantics of standard Prolog, disjunctive databases, and non-monotonic logic. The semantics of ASP captures the notion of possible beliefs of a reasoner who adheres to the *rationality principle* which says that "One shall not believe anything one is not forced to believe". The entailment relation of ASP is non-monotonic[1], which facilitates a high degree of elaboration tolerance in ASP theories. ASP allows natural representation of defaults and their exceptions, causal relations (including effects of actions), agents' intentions and obligations, and other constructs of natural language. ASP has a number of efficient reasoning systems, a well developed mathematical theory, and a well tested methodology of representing and using knowledge for computational tasks (see, for instance, [Baral 2003]). This, together with the fact that some of the designers of P-log came from the ASP community made the choice of a logical foundation for P-log comparatively easy.

[1] Roughly speaking, a language L is *monotonic* if whenever Π_1 and Π_2 are collections of statements of L with $\Pi_1 \subset \Pi_2$, and W is a model of Π_2, then W is a model of Π_1. A language which is not monotonic is said to be *nonmonotonic*.

The choice of a probabilistic framework was more problematic and that is where Judea's ideas played a major role. Our first problem was to choose from among various conceptualizations of probability: classical, frequentist, subjective, etc. Understanding the intuitive readings of basic language constructs is crucial for a software/knowledge engineer — probably more so than for a mathematician who may be primarily interested in their mathematical properties. Judea Pearl in [Pearl 1988] introduced the authors to the subjective view of probability — i.e. understanding of probabilities as degrees of belief of a rational agent — and to the use of subjective probability in AI. This matched well with the ASP-based logic side of the language. The ASP part of a P-log program can be used for describing possible beliefs, while the probabilistic part would allow knowledge engineers to quantify the degrees of these beliefs.

After deciding on an intuitive reading of probabilities, the next question was *which sorts of probabilistic statements to allow.* Fortunately, the question of concise and transparent representation of probability distributions was already addressed by Judea in [Pearl 1988], where he showed how Bayesian nets can be successfully used for this purpose. The concept was extended in [Pearl 2000] where Pearl introduced the notion of Causal Bayesian Nets (CBN's). Pearl's definition of CBN's is pioneering in three respects. First, he gives a framework where nondeterministic causal relations are the primitive relations among random variables. Second, he shows how relationships of correlation and (classical) independence *emerge* from these causal relationships in a natural way; and third he shows how this emergence is faithful to our intuitions about the difference between causality and (mere) correlation.

As we mentioned above, one of the primary desired features in the design of P-log was elaboration tolerance — defined as the ability of a representation to incorporate new knowledge with minimal revision [McCarthy 1999]. P-log inherited from ASP the ability to naturally incorporate many forms of new logical knowledge. An extension of ASP, called CR-Prolog, further improved this ability [Balduccini and Gelfond 2003]. The term "elaboration tolerance" is less well known in the field of probabilistic reasoning, but one of the primary strengths of Bayes nets as a representation is the ability to systematically and smoothly incorporate new knowledge through conditioning, using Bayes Theorem as well as algorithms given by Pearl [Pearl 1988] and others. Causal Bayesian Nets carry this a step further, by allowing us to formalize interventions in addition to (and as distinct from) observations, and smoothly incorporate either kind of new knowledge in the form of updates. Thus from the standpoint of elaboration tolerance, CBN's were a natural choice as a probabilistic foundation for P-log.

Another reason for choosing CBN's is that we simply believe Pearl's distinction between observations and interventions to be central to commonsense probabilistic reasoning. It gives a precise mathematical basis for distinguishing between the following questions: (1) what can I expect to happen given that I observe $X = x$, and (2) what can I expect to happen if I *intervene in the normal operation of*

a probabilistic system by fixing value of variable X to x? These questions could in theory be answered using classical methods, but only by creating a separate probabilistic model for each question. In a CBN these two questions may be treated as conditional probabilities (one conditioned on an observation and the other on an action) of a single probabilistic model.

P-log carries things another step. There are many actions one could take to manipulate a system besides fixing the values of (otherwise random) variables — and the effects of such actions are well studied under headings associated with ASP. Moreover, besides actions, there are many sorts of information one might gain besides those which simply eliminate possible worlds: one may gain knowledge which introduces new possible worlds, alters the probabilities of possible worlds, introduces new logical rules, etc. ASP has been shown to be a good candidate for handling such updates in non-probabilistic settings, and our hypothesis was that it would serve as well when combined with a probabilistic representation. Thus some of the key advantages of Bayesian nets, which are amplified by CBN's, show plausible promise of being even further amplified by their combination with ASP. This is the methodology of P-log: to combine a well studied method for elaboration tolerant probabilistic representations (CBN's) with a well studied method for elaboration tolerant logical representations (ASP).

Finally let us say a few words about the current status of the language. It is comparatively new. The first publication on the subject appeared in [Baral, Gelfond, and Rushton 2004], and the full journal paper describing the language appeared only recently in [Baral, Gelfond, and Rushton 2009]. The use of P-log for knowledge representation was also explored in [Baral and Hunsaker 2007] and [Gelfond, Rushton, and Zhu 2006]. A prototype reasoning system based on ASP computation allowed the use of the language for a number of applications (see, for instance, [Baral, Gelfond, and Rushton 2009; Pereira and Ramli 2009]). We are currently working on the development and implementation of a more efficient system, and on expanding it to allow rules of CR-Prolog. Finding ways for effectively combining ASP-based computational methods of P-log with recent advanced algorithms for Bayesian nets is probably one of the most interesting open questions in this area.

The paper is organized as follows. Section 2 contains short introduction to ASP and CR-Prolog. Section 3 describes the syntax and informal semantics of P-log, illustrating both through a nontrivial example. Section 4 gives another example, similar in nature to Simpson's Paradox. Section 5 states a new theorem which extends the semantics of P-log from that given in [Baral, Gelfond, and Rushton 2009] to cover programs with infinitely many random variables. The basic idea of Section 5 is accessible to a general audience, but its technical details require an understanding of the material presented in [Baral, Gelfond, and Rushton 2009].

2 Preliminaries

This section contains a description of syntax and semantics of both ASP and CR-Prolog. In what follows we use a standard notion of a sorted signature from classical logic. Terms and atoms are defined as usual. An atom $p(\bar{t})$ and its negation $\neg p(\bar{t})$ are referred to as *literals*. Literals of the form $p(\bar{t})$ and $\neg p(\bar{t})$ are called *contrary*. ASP and CR-Prolog also contain connectives *not* and *or* which are called *default negation* and *epistemic disjunction* respectively. Literals possibly preceded by default negation are called *extended literals*.

An ASP program is a pair consisting of a signature σ and a collection of rules of the form

$$l_0 \ or \ \ldots \ or \ l_m \leftarrow l_{m+1}, \ldots, l_k, not \ l_{k+1}, \ldots, not \ l_n \tag{1}$$

where l's are literals. The right-hand side of of the rule is often referred to as the rule's *body*, the left-hand side as the rule's head.

The answer set semantics of a logic program Π assigns to Π a collection of *answer sets* – partial interpretations[2] corresponding to possible sets of beliefs which can be built by a rational reasoner on the basis of rules of Π. In the construction of such a set S, the reasoner is assumed to be guided by the following informal principles:

- S must satisfy the rules of Π;
- the reasoner should adhere to the *rationality principle*, which says that *one shall not believe anything one is not forced to believe*.

To understand the former let us consider a partial interpretation S viewed as a possible set of beliefs of our reasoner. A ground atom p is satisfied by S if $p \in S$, i.e., the reasoner believes p to be true. According to the semantics of our connectives $\neg p$ means that p is false. Consequently, $\neg p$ is satisfied by S iff $\neg p \in S$, i.e., the reasoner believes p to be false. Unlike $\neg p$, *not* p has an epistemic character and is read as *there is no reason to believe that* p *is true*. Accordingly, S satisfies *not* l if $l \notin S$. (Note that it is possible for the reasoner to believe neither p nor $\neg p$). An epistemic disjunction l_1 *or* l_2 is satisfied by S if $l_1 \in S$ or $l_2 \in S$, i.e., the reasoner believes at least one of the disjuncts to be true. Finally, S satisfies the body (resp., head) of rule (1) if S satisfies all of the extended literals occurring in its body (resp., head); and S satisfies rule (1) if S satisfies its head or does not satisfy its body.

What is left is to capture the intuition behind the rationality principle. This will be done in two steps.

DEFINITION 1 (Answer Sets, Part I). Let program Π consist of rules of the form:

$$l_0 \ or \ \ldots \ or \ l_i \leftarrow l_{i+1}, \ldots, l_m.$$

An answer set of Π is a consistent set S of ground literals such that:

[2] By partial interpretation we mean a consistent set of ground literals of $\sigma(\Pi)$.

- S satisfies the rules of Π.
- S is minimal; i.e., no proper subset of S satisfies the rules of Π.

The rationality principle here is captured by the minimality condition. For example, it is easy to see that { } is the only answer set of program consisting of the single rule $p \leftarrow p$, and hence the reasoner associated with it knows nothing about the truth or falsity of p. The program consisting of rules

$p(a)$.
$q(a)$ *or* $q(b) \leftarrow p(a)$.

has two answer sets: $\{p(a), q(a)\}$ and $\{p(a), q(b)\}$. Note that no rule requires the reasoner to believe in both $q(a)$ and $q(b)$. Hence he believes that the two formulas $p(a)$ and ($q(a)$ *or* $q(b)$) are true, and that $\neg p(a)$ is false. He remains undecided, however, about, say, the two formulas $p(b)$ and ($\neg q(a)$ *or* $\neg q(b)$). Now let us consider an arbitrary program:

DEFINITION 2 (Answer Sets, Part II). Let Π be an arbitrary collection of rules (1) and S a set of literals. By Π^S we denote the program obtained from Π by

1. removing all rules containing *not* l such that $l \in S$;
2. removing all other premises containing *not* .

S is an answer set of Π iff S is an answer set of Π^S.

To illustrate the definition let us consider a program

$p(a)$.
$p(b)$.
$\neg p(X) \leftarrow$ *not* $p(X)$.

where p is a unary predicate whose domain is the set $\{a, b, c\}$. The last rule, which says that if X is not believed to satisfy p then $p(X)$ is false, is the ASP formalization of a Closed World Assumption for a relation p [Reiter 1978]. It is easy to see that $\{p(a), p(b), \neg p(c)\}$ is the only answer set of this program. If we later learn that c satisfies p, this information can be simply added to the program as $p(c)$. The default for c will be defeated and the only answer set of the new program will be $\{p(a), p(b), p(c)\}$.

The next example illustrates the ASP formalization of a more general default. Consider a statement: "Normally, computer science courses are taught only by computer science professors. The logic course is an exception to this rule. It may be taught by faculty from the math department." This is a typical *default* with a *weak exception*[3] which can be represented in ASP by the rules:

[3]An exception to a default is called *weak* if it stops application of the default without defeating its conclusion. Otherwise it is called *strong*.

$$\begin{array}{lcl} \neg may\_teach(P,C) & \leftarrow & \neg member(P,cs), \\ & & course(C,cs), \\ & & not\ ab(d_1(P,C)), \\ & & not\ may\_teach(P,C). \\ ab(d_1(P,logic)) & \leftarrow & not\ \neg member(P,math). \end{array}$$

Here $d_1(P,C)$ is the name of the default rule and $ab(d_1(P,C))$ says that default $d_1(P,C)$ is not applicable to the pair $\langle P,C\rangle$. The second rule above stops the application of the default in cases where the class is *logic* and P *may be* a math professor. Used in conjunction with rules:

$member(john,cs)$.
$member(mary,math)$.
$member(bob,ee)$.
$\neg member(P,D) \leftarrow not\ member(P,D)$.
$course(logic,cs)$.
$course(data\_structures,cs)$.

the program will entail that Mary does not teach data structures while she may teach logic; Bob teaches neither logic nor data structures, and John may teach both classes.

The previous examples illustrate the representation of defaults and their strong and weak exceptions. There is another type of possible exception to defaults, sometimes referred to as an **indirect exception**. Intuitively, these are rare exceptions that come into play only as a last resort, to restore the consistency of the agent's world view when all else fails. The representation of indirect exceptions seems to be beyond the power of ASP. This observation led to the development of a simple but powerful extension of ASP called **CR-Prolog** (or ASP with consistency-restoring rules). To illustrate the problem let us consider the following example.

Consider an ASP representation of the default "elements of class c normally have property p":

$$\begin{array}{lcl} p(X) & \leftarrow & c(X), \\ & & not\ ab(d(X)), \\ & & not\ \neg p(X). \end{array}$$

together with the rule

$$q(X) \leftarrow p(X).$$

and the facts $c(a)$ and $\neg q(a)$. Let us denote this program by E, where E stands for "exception".

It is not difficult to check that E is inconsistent. No rules allow the reasoner to prove that the default is not applicable to a (i.e. to prove $ab(d(a))$) or that a does not have property p. Hence the default must conclude $p(a)$. The second rule implies $q(a)$ which contradicts one of the facts. However, there seems to exists a

commonsense argument which may allow a reasoner to avoid inconsistency, and to conclude that a is an indirect exception to the default. The argument is based on the **Contingency Axiom** for default $d(X)$ which says that *any element of class c can be an exception to the default $d(X)$ above, but such a possibility is very rare, and, whenever possible, should be ignored*. One may informally argue that since the application of the default to a leads to a contradiction, the possibility of a being an exception to $d(a)$ cannot be ignored and hence a must satisfy this rare property.

In what follows we give a brief description of CR-Prolog — an extension of ASP capable of encoding and reasoning about such rare events.

A program of CR-Prolog is a four-tuple consisting of

1. A (possibly sorted) signature.
2. A collection of regular rules of ASP.
3. A collection of rules of the form

$$l_0 \stackrel{+}{\leftarrow} l_1, \ldots, l_k, not\ l_{k+1}, \ldots, not\ l_n \tag{2}$$

   where l's are literals. Rules of this type are called *consistency restoring* rules (CR-rules).

4. A partial order, $\leq$, defined on sets of CR-rules. This partial order is often referred to as a **preference relation**.

Intuitively, rule (2) says that if the reasoner associated with the program believes the body of the rule, then he "may possibly" believe its head. However, this possibility may be used only if there is no way to obtain a consistent set of beliefs by using only regular rules of the program. The partial order over sets of CR-rules will be used to select preferred possible resolutions of the conflict. Currently the inference engine of CR-Prolog [Balduccini 2007] supports two such relations, denoted $\leq_1$ and $\leq_2$. One is based on the set-theoretic inclusion ($R_1 \leq_1 R_2$ holds iff $R_1 \subseteq R_2$). The other is defined by the cardinality of the corresponding sets ($R_1 \leq_2 R_2$ holds iff $|R_1| \leq |R_2|$). To give the precise semantics we will need some terminology and notation.

The set of regular rules of a CR-Prolog program Π will be denoted by Π^r, and the set of CR-rules of Π will be denoted by Π^{cr}. By $\alpha(r)$ we denote a regular rule obtained from a consistency restoring rule r by replacing $\stackrel{+}{\leftarrow}$ by $\leftarrow$. If R is a set of CR-rules then $\alpha(R) = \{\alpha(r) \ : \ r \in R\}$. As in the case of ASP, the semantics of CR-Prolog will be given for ground programs. A rule with variables will be viewed as a shorthand for a set of ground rules.

DEFINITION 3. (Abductive Support)
A minimal (with respect to the preference relation of the program) collection R of

CR-rules of Π such that $\Pi^r \cup \alpha(R)$ is consistent (i.e. has an answer set) is called an **abductive support** of Π.

DEFINITION 4. (Answer Sets of CR-Prolog)
A set A is called an *answer set* of Π if it is an answer set of a regular program $\Pi^r \cup \alpha(R)$ for some abductive support R of Π.

Now let us show how CR-Prolog can be used to represent defaults and their indirect exceptions. The CR-Prolog representation of the default $d(X)$, which we attempted to represent in ASP program E, may look as follows

$$\begin{array}{rcl} p(X) & \leftarrow & c(X), \\ & & not\ ab(d(X)), \\ & & not\ \neg p(X). \\ \neg p(X) & \stackrel{+}{\leftarrow} & c(X). \end{array}$$

The first rule is the standard ASP representation of the default, while the second rule expresses the Contingency Axiom for the default $d(X)$[4]. Consider now a program obtained by combining these two rules with an atom $c(a)$.

Assuming that a is the only constant in the signature of this program, the program's unique answer set will be $\{c(a), p(a)\}$. Of course this is also the answer set of the regular part of our program. (Since the regular part is consistent, the Contingency Axiom is ignored.) Let us now expand this program by the rules

$q(X) \leftarrow p(X)$.
$\neg q(a)$.

The regular part of the new program is inconsistent. To save the day we need to use the Contingency Axiom for $d(a)$ to form the abductive support of the program. As a result the new program has the answer set $\{\neg q(a), c(a), \neg p(a))\}$. The new information does not produce inconsistency, as it did in ASP program E. Instead the program withdraws its previous conclusion and recognizes a as a (strong) exception to default $d(a)$.

3 The Language

A P-log program consists of its *declarations*, *logical rules*, *random selection rules*, *probability atoms*, *observations*, and *actions*. We will begin this section with a brief description of the syntax and informal readings of these components of the programs, and then proceed to an illustrative example.

The declarations of a P-log program give the types of objects and functions in the program. Logical rules are "ordinary" rules of the underlying logical language

[4] In this form of Contingency Axiom, we treat X as a strong exception to the default. Sometimes it may be useful to also allow weak indirect exceptions; this can be achieved by adding the rule $ab(d(X)) \stackrel{+}{\leftarrow} c(X)$.

written using light syntactic sugar. For purposes of this paper, the underlying logical language is CR-Prolog.

P-log uses *random selection rules* to declare random attributes (essentially random variables) of the form $a(\bar{t})$, where a is the name of the attribute and $\bar{t}$ is a vector of zero or more parameters. In this paper we consider random selection rules of the form

$$[\ r\ ]\ random(a(\bar{t})) \leftarrow B. \tag{3}$$

where r is a term used to name the random causal process associated with the rule and B is a conjunction of zero or more extended literals. The name $[\ r\ ]$ is optional and can be omitted if the program contains exactly one random selection rule for $a(\bar{t})$. Statement (3) says that *if B were to hold, the value of $a(\bar{t})$ would be selected at random from its range by process r, unless this value is fixed by a deliberate action.* More general forms of random selection rules, where the values may be selected from a range which depends on context, are discussed in [Baral, Gelfond, and Rushton 2009].

Knowledge of the numeric probabilities of possible values of random attributes is expressed through *causal probability atoms*, or *pr*-atoms. A *pr*-atom takes the form

$$pr_r(a(\bar{t}) = y|_c\ B) = v$$

where $a(\bar{t})$ is a random attribute, B a conjunction of literals, r is a causal process, $v \in [0, 1]$, and y is a possible value of $a(\bar{t})$. The statement says that *if the value of $a(\bar{t})$ is fixed by process r, and B holds, then the probability that r causes $a(\bar{t}) = y$ is v*. If r is uniquely determined by the program then it can be omitted. The "causal stroke" '$|_c$' and the "rule body" B may also be omitted in case B is empty.

Observations and actions of a P-log program are written, respectively, as

$$obs(l). \qquad\qquad do(a(\bar{t}) = y)).$$

where l is a literal, $a(\bar{t})$ a random attribute, and y a possible value of $a(\bar{t})$. $obs(l)$ is read *l is observed to be true*. The action $do(a(\bar{t}) = y)$ is read *the value of $a(\bar{t})$, instead of being random, is set to y by a deliberate action.*

This completes a general introductory description of P-log. Next we give an example to illustrate this description. The example shows how certain forms of knowledge may be represented, including deterministic causal knowledge, probabilistic causal knowledge, and strict and defeasible logical rules (a rule is *defeasible* if it states an overridable presumption; otherwise it is *strict*). We will use this example to illustrate the syntax of P-log, and, afterward, to provide an indication of the formal semantics. Complete syntax and semantics are given in [Baral, Gelfond, and Rushton 2009], and the reader is invited to refer there for more details.

EXAMPLE 5. [Circuit]
A circuit has a motor, a breaker, and a switch. The switch may be open or closed. The breaker may be tripped or not; and the motor may be turning or not. The operator may toggle the switch or reset the breaker. If the switch is closed and the system is functioning normally, the motor turns. The motor never turns when the switch is open, the breaker is tripped, or the motor is burned out. The system may break and if so the break could consist of a tripped breaker, a burned out motor, or both, with respective probabilities .9, .09, and .01. Breaking, however, is rare, and should be considered only in the absence of other explanations.

Let us show how to represent this knowledge in P-log. First we give declarations of sorts and functions relevant to the domain. As typical for representation of dynamic domains we will have sorts for actions, fluents (properties of the domain which can be changed by actions), and time steps. Fluents will be partitioned into inertial fluents and defined fluents. The former are subject to the law of inertia [Hayes and McCarthy 1969] (which says that things stay the same by default), while the latter are specified by explicit definitions in terms of already defined fluents. We will also have a sort for possible types of breaks which may occur in the system. In addition to declared sorts P-log contains a number of predefined sorts, e.g. a sort *boolean*. Here are the sorts of the domain for the circuit example:

$action = \{toggle, reset, break\}$.

$inertial\_fluent = \{closed, tripped, burned\}$.

$defined\_fluent = \{turning, faulty\}$.

$fluent = inertial\_fluent \cup defined\_fluent$.

$step = \{0, 1\}$.

$breaks = \{trip, burn, both\}$.

In addition to sorts we need to declare functions (referred in P-log as *attributes*) relevant to our domain.

$holds \ : \ fluent \times step \rightarrow boolean$.

$occurs \ : \ action \times step \rightarrow boolean$.

Here $holds(f, T)$ says that fluent f is true at time step T and $occurs(a, T)$ indicates that action a was executed at T.

The last function we need to declare is a random attribute $type\_of\_break(T)$ which denotes the type of an occurrence of action *break* at step T.

$type\_of\_break \ : \ step \rightarrow breaks$.

The first two logical rules of the program define the direct effects of action *toggle*.

$$\begin{array}{lcl} holds(closed, T+1) & \leftarrow & occurs(toggle, T), \\ & & \neg holds(closed, T). \\ \neg holds(closed, T+1) & \leftarrow & occurs(toggle, T), \\ & & holds(closed, T). \end{array}$$

They simply say that toggling opens and closes the switch. The next rule says that resetting the breaker untrips it.

$$\neg holds(tripped, T+1) \quad \leftarrow \quad occurs(reset, T).$$

The effects of action *break* are described by the rules

$$\begin{array}{lcl} holds(tripped, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = trip. \\ holds(burned, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = burn. \\ holds(tripped, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = both. \\ holds(burned, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = both. \end{array}$$

The next two rules express the inertia axiom which says that *by default, things stay as they are.* They use default negation *not* — the main nonmonotonic connective of ASP —, and can be viewed as typical representations of defaults in ASP and its extensions.

$$\begin{array}{lcl} holds(F, T+1) & \leftarrow & inertial\_fluent(F), \\ & & holds(F, T), \\ & & not\ \neg holds(F, T+1). \\ \neg holds(F, T+1) & \leftarrow & inertial\_fluent(F), \\ & & \neg holds(F, T), \\ & & not\ holds(F, T+1). \end{array}$$

Next we explicitly define fluents *faulty* and *turning*.

$$\begin{array}{lcl} holds(faulty, T) & \leftarrow & holds(tripped, T). \\ holds(faulty, T) & \leftarrow & holds(burned, T). \\ \neg holds(faulty, T) & \leftarrow & not\ holds(faulty, T). \end{array}$$

The rules above say that the system is functioning abnormally if and only if the breaker is tripped or the motor is burned out. Similarly the next definition says that the motor turns if and only if the switch is closed and the system is functioning normally.

$$\begin{array}{lcl} holds(turning, T) & \leftarrow & holds(closed, T), \\ & & \neg holds(faulty, T). \\ \neg holds(turning, T) & \leftarrow & not\ holds(turning, T). \end{array}$$

The above rules are sufficient to define causal effects of actions. For instance if we assume that at Step 0 the motor is turning and the breaker is tripped, i.e.

action *break* of the type *trip* occurred at 0, then in the resulting state we will have $holds(tripped, 1)$ as the direct effect of this action; while $\neg holds(turning, 1)$ will be its indirect effect[5].

We will next have a default saying that for each action A and time step T, in the absence of a reason to believe otherwise we assume A does not occur at T.

$\neg occurs(A, T) \leftarrow action(A), not\ occurs(A, T).$

We next state a CR-rule representing possible exceptions to this default. The rule says that a break to the system may be considered if necessary (that is, necessary in order to reach a consistent set of beliefs).

$occurs(break, 0) \stackrel{+}{\leftarrow}$.

The next collection of facts describes the initial situation of our story.

$\neg holds(closed, 0).\quad \neg holds(burned, 0).\quad \neg holds(tripped, 0).\quad occurs(toggle, 0).$

Next, we state a random selection rule which captures the non-determinism in the description of our circuit.

$random(type\_of\_break(T)) \leftarrow occurs(break, T).$

The rule says that if action *break* occurs at step T then the type of break will be selected at random from the range of possible types of breaks, unless this type is fixed by a deliberate action. Intuitively, *break* can be viewed as a non-deterministic action, with non-determinism coming from the lack of knowledge about the precise type of *break*.

Let π_0 be the circuit program given so far. Next we will give a sketch of the formal semantics of P-log, using π_0 as an illustrative example.

The *logical part* of a P-log program Π consists of its declarations, logical rules, random selection rules, observations, and actions; while its *probabilistic part* consists of its *pr*-atoms (though the above program does not have any). The semantics of P-log describes a translation of the logical part of Π into an "ordinary" CR-Prolog program $\tau(\Pi)$. The semantics of Π is then given by

[5]It is worth noticing that, though short, our formalization of the circuit is non-trivial. It is obtained using the general methodology of representing dynamic systems modeled by transition diagrams whose nodes correspond to physically possible states of the system and whose arcs are labeled by actions. A transition $\langle \sigma_0, a, \sigma_1 \rangle$ indicates that state σ_1 may be a result of execution of a in σ_0. The problem of finding concise and mathematically accurate description of such diagrams has been a subject of research for over 30 years. Its solution requires a good understanding of the nature of causal effects of actions in the presence of complex interrelations between fluents. An additional level of complexity is added by the need to specify what is not changed by actions. As noticed by John McCarthy, the latter, known as the Frame Problem, can be reduced to finding a representation of the Inertia Axiom which requires the ability to represent defaults and to do non-monotonic reasoning. The representation of this axiom as well as that of the interrelations between fluents we used in this example is a simple special case of general theory of action and change based on logic programming under the answer set semantics.

1. a collection of answer sets of $\tau(\Pi)$ viewed as the set of possible worlds of a rational agent associated with Π, along with

2. a probability measure over these possible worlds, determined by the collection of the probability atoms of Π.

To obtain $\tau(\pi_0)$ we represent sorts as collections of facts. For instance, sort *step* would be represented in CR-Prolog as

$step(0). \quad step(1).$

For a non-boolean function *type_of_break* the occurrences of atoms of the form $type\_of\_break(T) = trip$ in π_0 are replaced by $type\_of\_break(T, trip)$. Similarly for *burn* and *both*. The translation also contains the axiom

$$\begin{array}{rcl} \neg type\_of\_break(T, V_1) & \leftarrow & breaks(V_1), breaks(V_2), V_1 \neq V_2, \\ & & type\_of\_break(T, V_2). \end{array}$$

to guarantee that *type_of_break* is a function. In general, the same transformation is performed for all non-boolean functions.

Logical rules of π_0 are simply inserted into $\tau(\pi_0)$. Finally, the random selection rule is transformed into

$$\begin{array}{l} type\_of\_break(T, trip) \text{ or } type\_of\_break(T, burn) \text{ or } type\_of\_break(T, both) \leftarrow \\ \qquad\qquad occurs(break, T), \\ \qquad\qquad not\ intervene(type\_of\_break(T)). \end{array}$$

It is worth pointing out here that while CBN's represent the notion of intervention in terms of transformations on graphs, P-log axiomatizes the semantics of intervention by including *not intervene*(...) in the body of the translation of each random selection rule. This amounts to a *default presumption* of randomness, overridable by intervention. We will see next how actions using *do* can defeat this presumption.

Observations and actions are translated as follows. For each literal l in π_0, $\tau(\pi_0)$ contains the rule

$\leftarrow obs(l), not\ l.$

For each atom $a(\bar{t}) = y$, $\tau(\pi)$ contains the rules

$a(\bar{t}, y) \leftarrow do(a(\bar{t}, y)).$

and

$intervene(a(\bar{t})) \leftarrow do(a(\bar{t}, Y)).$

The first rule eliminates possible worlds of the program failing to satisfy l. The second rule makes sure that interventions affect their intervened-upon variables in the expected way. The third rule defines the relation *intervene* which, for each action, cancels the randomness of the corresponding attribute.

It is not difficult to check that under the semantics of CR-Prolog, $\tau(\pi_0)$ has a unique possible world W containing $holds(closed, 1)$ and $holds(turning, 1)$, the direct and indirect effects, respectively, of the action *close*. Note that the collection of regular ASP rules of $\tau(\pi_0)$ is consistent, i.e., has an answer set. This means that CR-rule $occurs(break, 0) \stackrel{+}{\leftarrow}$ is not activated, break does not occur, and the program contains no randomness.

Now we will discuss how probabilities are computed in P-log. Let Π be a P-log program containing the random selection rule $[r]\ random(a(\bar{t})) \leftarrow B_1$ and the *pr*-atom $pr_r(a(\bar{t}) = y \mid_c B_2) = v$. Then if W is a possible world of Π satisfying B_1 and B_2, the *assigned probability* of $a(\bar{t}) = y$ in W is defined [6] to be v. In case W satisfies B_1 and $a(\bar{t}) = y$, but there is no *pr*-atom $pr_r(a(\bar{t} = y \mid_c B_2) = v$ of Π such that W satisfies B_2, then the *default probability* of $a(\bar{t}) = y$ in W is computed using the "indifference principle", which says that two possible values of a random selection are equally likely if we have no reason to prefer one to the other (see [Baral, Gelfond, and Rushton 2009] for details). The *probability* of each random atom $a(\bar{t}) = y$ occurring in each possible world W of program Π, written $P_\Pi(W, a(\bar{t}) = y)$, is now defined to be the assigned probability or the default probability, as appropriate.

Let W be a possible world of Π. The *unnormalized probability*, $\hat{\mu}_\Pi(W)$, of a possible world W *induced by* Π is

$$\hat{\mu}_\Pi(W) =_{def} \prod_{a(\bar{t},y) \in\ W} P_\Pi(W, a(\bar{t}) = y)$$

where the product is taken only over atoms for which $P(W, a(\bar{t}) = y)$ is defined.

Suppose Π is a P-log program having at least one possible world with nonzero unnormalized probability, and let Ω be the set of possible worlds of Π. The *measure*, $\mu_\Pi(W)$, of a possible world W *induced by* Π is the unnormalized probability of W divided by the sum of the unnormalized probabilities of all possible worlds of Π, i.e.,

$$\mu_\Pi(W) =_{def} \frac{\hat{\mu}_\Pi(W)}{\sum_{W_i \in \Omega} \hat{\mu}_\Pi(W_i)}$$

When the program Π is clear from context we may simply write $\hat{\mu}$ and μ instead of $\hat{\mu}_\Pi$ and μ_Π respectively.

This completes the discussion of how probabilities of possible worlds are defined in P-log. Now let us return to the circuit example. Let program π_1 be the union of π_0 with the single observation

$obs(\neg holds(turning, 1))$

The observation contradicts our previous conclusion $holds(turning, 1)$ reached by using the effect axiom for *toggle*, the definitions of $faulty$ and $turning$, and the

[6] For the sake of well definiteness, we consider only programs in which at most one v satisfies this definition.

inertia axiom for *tripped* and *burned*. The program $\tau(\pi_1)$ will resolve this contradiction by using the CR-rule $occurs(break, 0) \stackrel{+}{\leftarrow}$ to conclude that the action *break* occurred at Step 0. Now *type_of_break* randomly takes one of its possible values. Accordingly, $\tau(\pi_1)$ has three answer sets: W_1, W_2, and W_3. All of them contain $occurs(break, 0)$, $holds(faulty, 1)$, $\neg holds(turning, 1)$. One, say W_1 will contain

$type\_of\_break(1, trip)$, $holds(tripped, 1)$, $\neg holds(burned, 1)$

W_2 and W_3 will respectively contain

$type\_of\_break(1, burn)$, $\neg holds(tripped, 1)$, $holds(burned, 1)$

and

$type\_of\_break(1, both)$, $holds(tripped, 1)$, $holds(burned, 1)$

In accordance with our general definition, π_1 will have three possible worlds, W_1, W_2, and W_3. The probabilities of each of these three possible worlds can be computed as 1/3, using the indifference principle.

Now let us add some quantitative probabilities to our program. If π_2 is the union of π_1 with the following three *pr*-atoms

$pr(type\_of\_break(T) = trip \mid_c break(T)) = 0.9$
$pr(type\_of\_break(T) = burned \mid_c break(T)) = 0.09$
$pr(type\_of\_break(T) = both \mid_c break(T)) = 0.01$

then program π_2 has the same possible worlds as Π_1. Not surprisingly, $P_{\pi_2}(W_1) = 0.9$. Similarly $P_{\pi_2}(W_2) = 0.09$ and $P_{\pi_2}(W_3) = 0.01$. This demonstrates how a P-log program may be written in stages, with quantitative probabilities added as they are needed or become available.

Typically we are interested not just in the probabilities of individual possible worlds, but in the probabilities of certain interesting sets of possible worlds described, e.g., those described by formulae. For current purposes a rather simple definition suffices. Viz., recalling that possible worlds are sets of literals, for an arbitrary set C of literals we define

$$P_\pi(C) =_{def} P_\pi(\{W : C \subseteq W\}).$$

For example, $P_{\pi_1}(holds(turning, 1)) = 0$, $P_{\pi_1}(holds(tripped, 1)) = 1/3$,
and $P_{\pi_2}(holds(tripped, 1)) = 0.91$.

Our example is in some respects rather simple. For instance, every possible world of our program contains at most one atom of the form $a(\bar{t}) = y$ where $a(\bar{t})$ is a random attribute. We hope, however, that this example gives a reader some insight in the syntax and semantics of P-log. It is worth noticing that the example shows the ability of P-log to mix logical and probabilistic reasoning, including reasoning about causal effects of actions and explanations of observations. In addition it

demonstrates the non-monotonic character of P-log, i.e. its ability to react to new knowledge by changing probabilistic models of the domain and creating new possible worlds.

The ability to introduce new possible worlds as a result of conditioning is of interest from two standpoints. First, it reflects the common sense semantics of utterances such as "the motor might be burned out." Such a sentence does not eliminate existing possible beliefs, and so there is no classical (i.e., monotonic) semantics in which the statement would be informative. If it is informative, as common sense suggests, then its content seems to introduce new possibilities into the listener's thought process.

Second, nonmonotonicity can improve performance. Possible worlds tend to proliferate exponentially with the size of a program, quickly making computations intractable. The ability to consider only those random selections which may explain our abnormal observations may make computations tractable for larger programs. Even though our current solver is in its early stages of development, it is based on well researched answer set solvers which efficiently eliminate impossible worlds from consideration based on logical reasoning. Thus even our early prototype has shown promising performance on problems where logic may be used to exclude possible worlds from consideration in the computation of probabilities [Gelfond, Rushton, and Zhu 2006].

4 Spider Example

In this section, we consider a variant of Simpson's paradox, to illustrate the formalization of interventions in P-log. The story we would like to formalize is as follows:

In Stan's home town there are two kinds of poisonous spider, the creeper and the spinner. Bites from the two are equally common in Stan's area — though spinner bites are more common on a worldwide basis. An experimental anti-venom has been developed to treat bites from either kind of spider, but its effectiveness is questionable.

One morning Stan wakes to find he has a bite on his ankle, and drives to the emergency room. A doctor examines the bite, and concludes it is a bite from either a creeper or a spinner. In deciding whether to administer the anti-venom, the doctor examines the data he has on bites from the two kinds of spiders: out of 416 people bitten by the creeper worldwide, 312 received the anti-venom and 104 did not. Among those who received the anti-venom, 187 survived; while 73 survived who did not receive anti-venom. The spinner is more deadly and tends to inhabit areas where the treatment is less available. Of 924 people bitten by the spinner, 168 received the anti-venom, 34 of whom survived. Of the 756 spinner bite victims who did not receive the experimental treatment, only 227 survived.

For a random individual bitten by a creeper or spinner, let s, a, and c denote the

events of *survival, administering anti-venom,* and *creeper bite.* Based on the fact that the two sorts of bites are equally common in Stan's region, the doctor assigns a 0.5 probability to either kind of bite. He also computes a probability of survival, with and without treatment, from each kind of bite, based on the sampling distribution of the available data. He similarly computes the probabilities that victims of each kind of bite received the anti-venom. We may now imagine the doctor uses Bayes' Theorem to compute $P(s \mid a) = 0.522$ and $P(s \mid \neg a) = 0.394$.

Thus we see that if we choose a historical victim, in such a way that he has a 50/50 chance of either kind of bite, those who received anti-venom would have a substantially higher chance of survival. Stan is in the situation of having a 50/50 chance of either sort of bite; however, he is *not* a historical victim. Since we are intervening in the decision of whether he receives anti-venom, the computation above is not germane (as readers of [Pearl 2000] already know) — though we can easily imagine the doctor making such a mistake. A correct solution is as follows. Formalizing the relevant parts of the story in a P-log program Π gives

$survive, antivenom \ : \ boolean.$
$spider \ : \ \{creeper, spinner\}.$

$random(spider).$
$random(survive).$
$random(antivenom).$

$pr(spider = creeper) = 0.5.$

$pr(survive \mid_c spider = creeper, antivenom) = 0.6.$
$pr(survive \mid_c spider = creeper, \neg antivenom) = 0.7.$
$pr(survive \mid_c spider = spinner, antivenom) = 0.2.$
$pr(survive \mid_c spider = spinner, \neg antivenom) = 0.3.$

and so, according to our semantics,

$P_{\Pi \cup \{do(antivenom\}}(survive) = 0.4$
$P_{\Pi \cup \{do(\neg antivenom\}}(survive) = 0.5$

Thus, the correct decision, assuming we want to intervene to maximize Stan's chance of survival, is to not administer antivenom.

In order to reach this conclusion by classical probability, we would need to consider separate probability measures P_1 and P_2, on the sets of patients who received or did not receive antivenom, respectively. If this is done correctly, we obtain $P_1(s) = 0.4$ and $P_2(s) = 0.5$, as in the P-log program.

Thus we can get a correct classical solution using separate probability measures. Note however, that we could also get an *incorrect* classical solution using separate measures, since there exist probability measures $\hat{P}_1$ and $\hat{P}_2$ on the sets of historical bite victims which capture classical conditional probabilities given a and $\neg a$ respectively. We may define

$\hat{P}_1(E) =_{def} \frac{P(E \cap a)}{0.3582}$

$\hat{P}_2(E) =_{def} \frac{P(E \cap \neg a)}{0.6418}$

It is well known that each of these is a probability measure. They are seldom seen only because classical conditional probability gives us simple notations for them *in terms of a single measure capturing common background knowledge*. This allows us to refer to probabilities conditioned on observations without defining a new measure for each such observation. What we do not have, classically, is a similar mechanism for probabilities conditioned on intervention — which is sometimes of interest as the example shows. The ability to condition on interventions in this way has been a fundamental contribution of Pearl; and the inclusion in P-log of such conditioning-on-intervention is a direct result of the authors' reading of his book.

5 Infinite Programs

The definitions given so far for P-log apply only to programs with finite numbers of random selection rules. In this section we state a theorem which allows us to extend these semantics to programs which may contain infinitely many random selection rules. No changes are required from the syntax given in [Baral, Gelfond, and Rushton 2009], and the probability measure described here agrees with the one in [Baral, Gelfond, and Rushton 2009] whenever the former is defined.

We begin by defining the class of programs for which the new semantics are applicable. The reader is referred to [Baral, Gelfond, and Rushton 2009] for the definitions of *causally ordered*, *unitary*, and *strict probabilistic levelling*.

DEFINITION 6. [Admissible Program]
A P-log program is *admissible* if it is causally ordered and unitary, and if there exists a strict probabilistic levelling $||$ on Π such that no ground literal occurs in the heads of rules in infinitely many Π_i with respect to $||$.

The condition of admissibility, and the definitions it relies on, are all rather involved to state precisely, but the intuition is as follows. Basically, a program is unitary if the probabilities assigned to the possible outcomes of each selection rule are either all assigned and sum to 1, or are not all assigned and their sum does not exceed 1. The program is causally ordered if its causal dependencies are acyclic and if the only nondeterminism in it is a result of random selection rules. A strict probabilistic levelling is a well ordering of the selection rules of a program which witnesses the fact that it is causally ordered. Finally, a program which meets these conditions is admissible if every ground literal in the program logically depends on only finitely many random experiments. For example, the following program is not unitary:

$random(a) : boolean.$
$pr(a) = 1/2.$
$pr(\neg a) = 2/3.$

The following program is not causally ordered:

$random(a) : boolean.$
$random(b) : boolean.$
$pr_r(a|_c\ b) = 1/3.$
$pr_r(a|_c\ \neg b) = 2/3.$
$pr_r(b|_c\ a) = 1/5.$

and neither is the following:

$p\ \leftarrow not\ q.$
$q\ \leftarrow not\ p.$

since it has two answer sets which arise from circularity of defaults, rather than random selections. The following program is both unitary and causally ordered, but not admissible, because $atLeastOneTail$ depends on infinitely many coin tosses.

$coin\_toss : positive\_integer \rightarrow \{head, tail\}.$
$atLeastOneTail : boolean.$
$random(coin\_toss(N)).$
$atLeastOneTail \leftarrow coin\_toss(N) = tail.$

We need one more definition before stating the main theorem:

DEFINITION 7. [Cylinder algebra of Π]
Let Π be a countably infinite P-log program with random attributes $a_i(t)$, $i > 0$, and let C be the collection of sets of the form $\{\omega\ : a_i(t) = y \in \omega\}$ for arbitrary t, i, and y. The sigma algebra generated by C will be called the *cylinder algebra* of program Π.

Intuitively, the cylinder algebra of a program Π is the collection of sets which can be formed by performing countably many set operations (union, intersection, and complement) upon sets whose probabilities are defined by finite subprograms. We are now ready to state the main proposition of this section.

PROPOSITION 8. *[Admissible programs]*
Let Π *be an admissible P-log program with at most countably infinitely many ground rules, and let A be the cylinder algebra of* Π*. Then there exists a unique probability measure* P_Π *defined on A such that whenever* $[r]\ random(a(\bar{t})) \leftarrow B_1$ *and* $pr_r(a(\bar{t}) = y \mid B_2) = v$ *occur in* Π*, and* $P_\Pi(B_1 \wedge B_2) > 0$*, we have* $P_\Pi(a(\bar{t}) = y \mid B_1 \wedge B_2) = v$*.*

Recall that the semantic value of a P-log program Π consists of (1) a set of possible worlds of Π and (2) a probability measure on those possible worlds. The proposition now puts us in position to give semantics for programs with infinitely many random

selection rules. The possible worlds of the program are the answer sets of the associated (infinite) CR-Prolog program, as determined by the usual definition — while the probability measure is P_Π, as defined in Proposition 8.

We next give an example which exercises the proposition, in a form of a novel paradox. Imagine a casino which offers an infinite sequence of games, of which our agent may decide to play as many or as few as he wishes. For the n^{th} game, a fair coin is tossed n times. If the agent chooses to play the n^{th} game, then the agent wins $2^{n+1}+1$ dollars if all tosses made in the n^{th} game are heads and otherwise loses one dollar.

We can formalize this game as an infinite P-log program Π. First, we declare a countable sequence of games and an integer valued variable, representing the player's net winnings after each game.

$game : positive\_integer.$
$winnings : game \rightarrow integer.$
$play : game \rightarrow boolean.$
$coin : \{\langle M, N\rangle \mid 1 \leq M \leq N\} \rightarrow \{head, tail\}.$

Note that the declaration for *coin* is not written in the current syntax of P-log; but to save space we use set-builder notation here as a shorthand for the more lengthy formal declaration. Similarly, the notation $\langle M, N\rangle$ is also a shorthand. From this point on we will write $coin(M, N)$ instead of $coin(\langle M, N\rangle)$.

Π also contains a declaration to say that the throws are random and the coin is known to be fair:

$random(coin(M, N)).$
$pr(coin(M, N) = head) = 1/2.$

The conditions of winning the N^{th} game are described as follows:

$lose(N) \leftarrow play(N),\ coin(N, M) = tail.$
$win(N) \leftarrow play(N),\ not\ lose(N).$

The amount the agent wins or loses on each game is given by

$winnings(0) = 0.$
$winnings(N+1) = winnings(N) + 1 + 2^{N+1} \leftarrow win(N).$
$winnings(N+1) = winnings(N) - 1 \leftarrow lose(N).$
$winnings(N+1) = winnings(N) \leftarrow \neg play(N).$

Finally the program contains rules which describe the agent's strategy in choosing which games to play. Note that the agent's expected winnings in the N^{th} game are given by $(1/2^N)(1+2^{N+1}) - (1-1/2^N) = 1$, so each game has positive expectation for the player. Thus a reasonable strategy might be to play every game, represented as

$play(N)$.

This completes program Π. It can be shown to be admissible, and hence there is a unique probability measure P_Π satisfying the conclusion of Proposition 1. Thus, for example, $P_\Pi(coin(3,2) = head) = 1/2$, and $P_\Pi(win(10)) = 1/2^{10}$. Each of these probabilities can be computed from finite sub-programs. As more interesting example, let S be the set of possible worlds in which the agent wins infinitely many games. The probability of this event cannot be computed from any finite sub-program of Π. However, S is a countable intersection of countable unions of sets whose probabilities are defined by finite subprograms. In particular,

$$S = \bigcap_{N=1}^{\infty} \bigcup_{J=N}^{\infty} \{W \mid win(J) \in W\}$$

and therefore, S is in the cylinder algebra of Π and so its probability is given by the measure defined in Proposition 1.

So where is the Paradox? To see this, let us compute the probability of S. Since P_Π is a probability measure, it is monotonic in the sense that no set has greater probability than any of its subsets. P_Π must also be *countably subadditive*, meaning that the probability of a countable union of sets cannot exceed the sum of their probabilities. Thus, from the above we get for every N,

$$\begin{aligned} P_\Pi(S) &< P_\Pi(\bigcup_{J=N}^{\infty} \{W \mid win(J) \in W\} \\ &\leq \sum_{J=N}^{\infty} P_\Pi(\{W \mid win(J) \in W\}) \\ &= \sum_{J=N}^{\infty} 1/2^J \\ &= 1/2^N \end{aligned}$$

Now since right hand side can be made arbitrarily small by choosing a sufficiently large N, it follows that $P_\Pi(S) = 0$. Consequently, with probability 1, our agent will *lose* all but finitely many of the games he plays. Since he loses one dollar per play indefinitely after his final win, his winnings converge to $-\infty$ with probability 1, even though each of his wagers has positive expectation!

Acknowledgement
The first author was partially supported in this research by iARPA.

References

Balduccini, M. (2007). CR-MODELS: An inference engine for CR-Prolog. In C. Baral, G. Brewka, and J. Schlipf (Eds.), *Proceedings of the 9th Inter-*

national Conference on Logic Programming and Non-Monotonic Reasoning (LPNMR'07), Volume 3662 of *Lecture Notes in Artificial Intelligence*, pp. 18–30. Springer.

Balduccini, M. and M. Gelfond (2003, Mar). Logic Programs with Consistency-Restoring Rules. In P. Doherty, J. McCarthy, and M.-A. Williams (Eds.), *International Symposium on Logical Formalization of Commonsense Reasoning*, AAAI 2003 Spring Symposium Series, pp. 9–18.

Baral, C. (2003). *Knowledge representation, reasoning and declarative problem solving with answer sets*. Cambridge University Press.

Baral, C., M. Gelfond, and N. Rushton (2004, Jan). Probabilistic Reasoning with Answer Sets. In *Proceedings of LPNMR-7*.

Baral, C., M. Gelfond, and N. Rushton (2009). Probabilistic reasoning with answer sets. *Journal of Theory and Practice of Logic Programming (TPLP) 9*(1), 57–144.

Baral, C. and M. Hunsaker (2007). Using the probabilistic logic programming language p-log for causal and counterfactual reasoning and non-naive conditioning. In *Proceedings of IJCAI-2007*, pp. 243–249.

Gelfond, M. and V. Lifschitz (1988). The stable model semantics for logic programming. In *Proceedings of ICLP-88*, pp. 1070–1080.

Gelfond, M. and V. Lifschitz (1991). Classical negation in logic programs and disjunctive databases. *New Generation Computing 9*(3/4), 365–386.

Gelfond, M., N. Rushton, and W. Zhu (2006). Combining logical and probabilistic reasoning. AAAI 2006 Spring Symposium Series, pp. 50–55.

Hayes, P. J. and J. McCarthy (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*, pp. 463–502. Edinburgh University Press.

McCarthy, J. (1999). Elaboration tolerance. In progress.

Pearl, J. (1988). *Probabistic reasoning in intelligent systems: networks of plausable inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Pereira, L. M. and C. Ramli (2009). Modelling decision making with probabilistic causation. *Intelligent Decision Technologies (IDT)*. to appear.

Reiter, R. (1978). *On Closed World Data Bases*, pp. 119–140. Logic and Data Bases. Plenum Press.

21

On Computers Diagnosing Computers

Moises Goldszmidt

1 Introduction

I came to UCLA in the fall of 1987 and immediately enrolled in the course titled "Probabilistic Reasoning in Intelligent Systems" where we, as a class, went over the draft of Judea's book of the same title [Pearl 1988]. The class meetings were fun and intense. Everybody came prepared, having read the draft of the appropriate chapter and having struggled through the list of homework exercises that were due that day. There was a high degree of discussion and participation, and I was very impressed by Judea's attentiveness and interest in our suggestions. He was fully engaged in these discussions and was ready to incorporate our comments and change the text accordingly. The following year, I was a teaching assistant (TA) for that class. The tasks involved with being a TA gave me a chance to rethink and really digest the contents of the course. It dawned on me then what a terrific insight Judea had to focus on formalizing the notion of conditional independence: All the "juice" he got in terms of making "reasoning under uncertainty" computationally effective came from that formalization. Shortly thereafter, I had a chance to chat with Judea about these and related thoughts. I was in need of formalizing a notion of "relevance" for my own research and thought that I could adapt some ideas from the graphoid models [Pearl 1988]. In that opportunity Judea shared another of his great insights with me. After hearing me out, Judea said one word: "causality". I don't remember the exact words he used to elaborate, but the gist of what he said to me was: "we as humans perform extraordinarily complex reasoning tasks, being able to select the relevant variables, circumscribe the appropriate context, and reduce the number of factors that we should manipulate. I believe that our intuitive notions of causality enable us to do so. Causality is the holly grail [for Artificial Intelligence]".

In this short note, I would like to pay tribute to Judea's scientific work by speculating on the very realistic possibility of computers using his formalization of causality for automatically performing a nontrivial reasoning task commonly reserved for humans. Namely designing, generating, and executing experiments in order to conduct a proper diagnosis and identify the causes of performance problems on code being executed in large clusters of computers. What follows in the next two sections is not a philosophical exposition on the meaning of "causality" or on the reasoning powers of automatons. It is rather a brief description of the current state of the art

in programming large clusters of computers and then, a brief account argumenting that the conditions are ripe for embarking on this research path.

2 Programming large clusters of computers made easy

There has been a recent research surge in systems directed at providing programmers with the ability to write efficient parallel and distributed applications [Hadoop 2008; Dean and Ghemawat 2004; Isard et al. 2007]. Programs written in these environments are automatically parallelized and executed on large clusters of commodity machines. The tasks of enabling programmers to effectively write and deploy parallel and distributed application has of course been a long-standing problem. Yet, the relatively recent emergence of large-scale internet services, which depend on clusters of hundreds of thousands of general purpose servers, have given the area a forceful push. Indeed, this is not merely an academic exercise; code written in these environments has been deployed and is very much in everyday use at companies such as Google, Microsoft, and Yahoo (and many others). These programs process web pages in order to feed the appropriate data to the search and news summarization engines; render maps for route planning services; and update usage and other statistics from these services. Year old figures estimate that Dryad, the specific such environment created at Microsoft [Isard et al. 2007], is used to crunch on the order of a petabyte a day at Microsoft. In addition, in our lab at Microsoft Research, a cluster of 256 machines controlled by Dryad runs daily at a 100% utilization. This cluster mostly runs tests and experiments on research algorithms in machine learning, privacy, and security that process very large amounts of data.

The intended model in Dryad is for the programmer to build code as if she were programming one computer. The system then takes care of a) distributing the code to the actual cluster and b) managing the execution of the code in the cluster. All aspects of execution, including data partition, communications, and fault tolerance, are the responsibility of Dryad.

With these new capabilities comes the need for new tools for debugging code, profiling execution performance, and diagnosing system faults. By the mere fact that clusters of large numbers of computers are being employed, rare bugs will manifest themselves more often, and devices will fail in more runs (due to both software and hardware problems). In addition, as the code will be executed in a networked environment and the data will be partitioned (usually according to some hash function), communication bandwidth, data location, contention for shared disks, and data skewness will impact the performance of the programs. Most of the times the impact of these factors will be hard to reproduce in a single machine, making it an imperative that the diagnosis, profiling, and debugging be performed in the same environment and conditions as those in which the code is running.

3 Computers diagnosing computers

The good news is that the same infrastructure that enables the programming and control of these clusters can be used for debugging and diagnosis. Normally the computation proceeds in stages where the different nodes in the cluster perform the same computation in parallel on different portions of the data. For purposes of fault tolerance, there are mechanisms in Dryad to monitor the execution time of each node at any computation stage. It is therefore possible to gather robust statistics about the expected execution time of any particular node at a given stage and identify especially slow nodes. Currently, this information is used to restart those nodes or to migrate the computation to other nodes.

We can take this further and collect the copious amount of data that is generated by the various built-in monitors looking at things such as cpu utilization, memory utilization, garbage collection, disk utilization, and statistics on I/O.[1] The statistical analysis of these signals may provide clues pointing at the probable causes of poor performance and even of failures. Indeed we have built a system called Artemis [Creţu-Ciocârlie et al. 2008], that takes advantage of the Dryad infrastructure to collect and preprocess the data from these signals in a distributed and opportunistic fashion. Once the data is gathered, Artemis will run a set of statistical and machine learning algorithms ranging from summarizations to regression and pattern classification. In this paper we propose one more step. We can imagine a system that guided with the information from these analyses, performs active experiments on the execution of the code. The objective will be to causally diagnose problems, and properly profile dependencies between the various factors affecting the performance of the computations.

Let us ground this idea in a realistic example. Suppose that through the analysis of the execution logs of some large task we identify that, on a computationally intensive stage, a small number of machines performed significantly worse that the average/median (in terms of overall processing speed). Through further analysis, for example logistic regression with L1 regularization, we are able to identify the factors that differentiate the slower machines. Thus, we narrow down the possibilities and determine that the main difference between these machines and the machines that performed well is the speed at which the data is read by the slower machines.[2] Further factors influencing this speed are whether the data resides on a local disk and whether there are other computational nodes that share that disk (and introduce contention), and on the speed of the network. Figure 1 shows a (simplified) causal model of this scenario depicting two processing nodes. The dark nodes represent factors/variables that can be controlled or where intervention is possible. Conducting controlled experiments guided by this graph would enable the

[1] The number of counters and other signals that these monitors yield can easily reach on the order of hundreds per machine.

[2] This particular case was encountered by the author while running a benchmark based on Terasort on a cluster with hundreds of machines [Creţu-Ciocârlie et al. 2008].

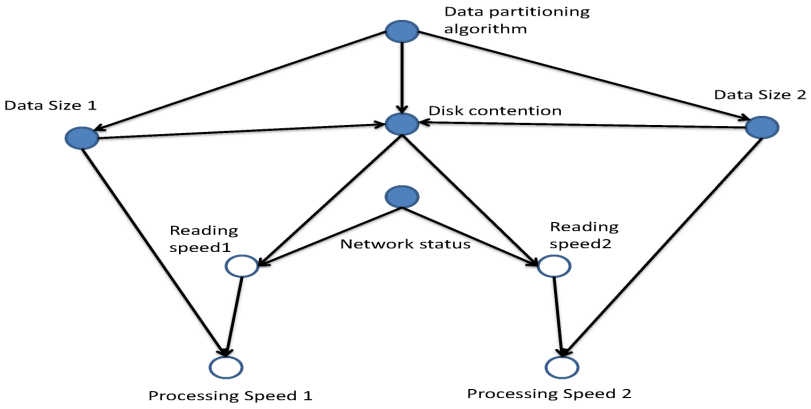


Figure 1. Simplified causal network depicting the processing speed scenario. Dark nodes represent the factor/variables than can be controlled or where intervention is possible.

precise characterization of the relationship between processing speed, data skewness, and disk contention, so that we can figure out how to partition and locate the data more efficiently and avoid having slow processing nodes. As the causal graph clearly exposes, controlling the data sizes for the computing nodes is not enough: if they reside on the same disk, contention may still cause slowdowns. This is obvious from the representation and algebra proposed by Judea in [Pearl 2000], as applied to this graph. This model also makes clear that intervening directly on the level of contention in the disk will indeed eliminate the dependency between the reading speed and the size of the data.

The idea of using graphical models for diagnosing computer systems goes back at least to [Breese and Heckerman 1996; Blake and Breese 1995]. It took close to 10 years after those papers for the first publication reporting the use of Bayesian networks for diagnosis in a nontrivial system in production to appear in a top tier systems conference [Cohen et al. 2004]. The methods in [Cohen et al. 2004] involve passive observation, and the authors make very clear that inferences concern correlation and not necessarily causation. However, hinting at root cause through correlation may not be enough in the very near future. Complexity and scale in current networked distributed systems keeps on increasing at a rapid pace. Because of service availability and reliability requirements, root cause analysis pointing at effective repair actions and accurate empirical characterization of dependencies between the different factors affecting computation are rapidly becoming a must.

Systems such as Dryad[Isard et al. 2007] enable the effective programming of large cluster of computers. In addition, they provide effective mechanisms for con-

trolling the "variables" of interest and setting up experiments in these clusters. Systems such as Artemis [Creţu-Ciocârlie et al. 2008] enable efficient collection and processing of extensive monitoring data, including the recording of the system state for recreating particular troublesome scenarios. The final ingredient for having machines automatically set up and conduct experiments is a language to describe these experiments and an algebra to reason about them in order to guarantee that the right variables are being controlled, and that we are intervening in the right spots in order to get to the correct conclusions. Through his seminal work in [Pearl 2000] and follow up papers, Judea Pearl has already given us that ingredient.

Acknowledgments: The author wishes to thank Mihai Budiu for numerous technical discussions on the topics of this paper, Joe Halpern for his help with the presentation, and very especially Judea Pearl for his continuous inspiration in the relentless and honest search for scientific truth.

References

Blake, R. and J. Breese (1995). Automatic bottleneck detection. In *UAI'95: Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Breese, J. and D. Heckerman (1996). Decision theoretic troubleshooting. In *UAI'96: Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Cohen, I., M. Goldszmidt, T. Kelly, J. Symons, and J. Chase (2004). Correlating instrumentation data to systems states: A building block for automated diagnosis and control. In *OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*. USENIX Association.

Creţu-Ciocârlie, G. F., M. Budiu, and M. Goldszmidt (2008). Hunting for problems with Artemis. In *USENIX Workshop on the Analysis of System Logs (WASL)*.

Dean, J. and S. Ghemawat (2004). Mapreduce: simplified data processing on large clusters. In *OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*. USENIX Association.

Hadoop (2008). The hadoop project. http://hadoop.apache.org.

Isard, M., M. Budiu, Y. Yu, A. Birrell, and D. Fetterly (2007). Dryad: distributed data-parallel programs from sequential building blocks. In *EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*. ACM.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press.

22

Overthrowing the Tyranny of Null Hypotheses Hidden in Causal Diagrams

SANDER GREENLAND

1 Introduction

Graphical models have a long history before and outside of causal modeling. Mathematical graph theory extends back to the 1700s and was used for circuit analysis in the 19th century. Its application in probability and computer science dates back at least to the 1960s (Biggs et al., 1986), and by the 1980s graphical models had become fully developed tools for these fields (e.g., Pearl, 1988; Hajek et al., 1992; Lauritzen, 1996).

As *Bayesian networks*, graphical models are carriers of direct conditional independence judgments, and thus represent a collection of assumptions that confine prior support to a lower dimensional manifold of the space of prior distributions over the nodes. Such dimensionality reduction was recognized as essential in formulating explicit and computable algorithms for digital-machine inference, an essential task of artificial-intelligence (AI) research. By the 1990s, these models had been merged with causal path diagrams long used in observational health and social science (OHSS) (Wright, 1934; Duncan, 1975), resulting in a formal theory of causal diagrams (Spirtes et al., 1993; Pearl, 1995, 2000).

It should be no surprise that some of the most valuable and profound contributions to these *developments* were from Judea Pearl, a renowned AI theorist. He motivated causal diagrams as *causal* Bayesian networks (Pearl, 2000), in which the basis for the dimensionality reduction is grounded in judgments of causal independence (and especially, autonomy) rather than mere probabilistic independence. Beyond his extensive technical and philosophical contributions, Pearl fought steadfastly to roll back prejudice against causal modeling and causal graphs in statistics. Today, only a few statisticians still regard causality as a metaphysical notion to be banned from formal modeling (Lad, 1999). While a larger minority still reject some aspects of causal-diagram or potential-outcome theory (e.g., Dawid, 2000, 2008; Shafer, 2002), the spreading wake of applications display the practical value of these theories, and formal causal diagrams have advanced into applied journals and books (e.g., Greenland et al., 1999; Cole and Hernán, 2002; Hernán et al., 2002; Jewell, 2004; Morgan and Winship, 2007; Glymour and Greenland, 2008) – although their rapid acceptance in OHSS may well have been facilitated by the longstanding informal use of path diagrams to represent qualities of causal systems (e.g., Susser, 1973; Duncan, 1975).

Graphs are unsurpassed tools for illustrating certain mathematical results that hold in functional systems (whether stochastic or not, or causal or not). Nonetheless, it is essential to recognize that many if not most causal judgments in OHSS are based on

observational (purely associational) data, with little or nothing in the way of manipulative (or "surgical") experiment to test these judgments. Time order is usually known, which insures that the chosen arrow directions are correct; but rarely is there a sound basis for deleting an arrow, leaving autonomy in question. When all empirical constraints encoded by the causal network come from passive frequency observations rather than experiments, the primacy of causal independence judgments has to be questioned. In these situations (which characterize observational research), we should not neglect associational models (including graphs) that encode frequency-based judgments, for these models may be all that are identified by available data. Indeed, a deep philosophical commitment to statistically identified quantities seems to drive the arguments of certain critics of potential outcomes and causal diagrams (Dawid, 2000, 2008). Even if we reject this philosophy, however, we should retain the distinction between levels of identification provided by our data, for even experimental data will not identify everything we would like to know.

I will argue that, in some ways, the distinction of nonidentification from identification is as fundamental to modeling and statistical inference about causal effects as is the distinction of causation from association (Gustafson, 2005; Greenland, 2005a, 2009a, 2009b). Indeed, I believe that some of the controversy and confusion over causation versus association stems from the inability of statistical observations to point identify (consistently estimate) many of the causal parameters that astute scientists legitimately ask about. Furthermore, if we consider strategies that force identification from available data (such as node or arrow deletions from graphical models) we will find that identification may arise only by declaring some types of joint frequencies as justifying the corresponding conditional independence assumptions. This leads directly into the complex topic of pruning algorithms, including the choice of target or loss function.

I will outline these problems in their most basic forms, for I think that in the rush to adopt causal diagrams some realism has been lost by neglecting problems of nonidentification and pruning. My exposition will take the form of a series of vignettes that illustrate some basic points of concern. I will not address equally important concerns that many of the nodes offered as "treatments" may be ill-defined or nonmanipulable, or may correspond poorly to the treatments they ostensibly represent (Greenland, 2005b; Hernán, 2005; Cole and Frangakis, 2009; VanderWeele, 2009).

2 Nonidentification from Unfaithfulness in a Randomized Trial

Nonidentification can be seen and has caused controversy in the simplest causal-inference settings. Consider an experiment that randomizes a node R. Inferences on causal effects of R from subsequent associations of R with later events would then be justified, since R would be an exogenous node. R would also be an instrumental variable for certain descendants under further conditional-independence assumptions.

A key problem is how one could justify removing arrows along the line of descent from R to another node Y, even if R is exogenous. The overwhelmingly dominant approach licenses such removal if the observed R-Y association fails to meet some criterion for departure from pure randomness. This schematic for a causal-graph pruning

algorithm was employed by Spirtes et al. (1993), unfortunately with a very naïve Neyman-Pearsonian criterion (basically, allowing removal of arrows when a *P*-value exceeds an α level). These and related graphical algorithms (Pearl and Verma, 1991) produce what appear to be results in conflict with practical intuitions, namely causal "discovery" algorithms for single observational data sets, with no need for experimental evidence. These algorithms have been criticized philosophically on grounds related to the identification problem (Freedman and Humphreys, 1999; Robins and Wasserman, 1999ab), and there are also objections based on statistical theory (Robins et al., 2003).

One controversial assumption in these algorithms is *faithfulness* (or stability) that all connected nodes are associated. Although arguments have been put forward in its favor (e.g., Spirtes et al., 1993; Pearl, 2000, p. 63), this assumption coheres poorly with prior beliefs of some experienced researchers. Without faithfulness, two nodes may be independent even if there is an arrow linking them directly, if that arrow represents the presence of causal effects among units in a target population. A classic example of such unfaithfulness appeared in the debates between Fisher and Neyman in the 1930s, in which they disagreed on how to formulate the causal null hypothesis (Senn, 2004). The framework of their debate would be recognized today as the *potential-outcome* or counterfactual model, although in that era the model (when named) was called the randomization model. This model illustrates the benefit of randomization as a means of detecting a signal by injecting white noise into a system to drown out uncontrolled influences.

To describe the model, suppose we are to study the effect of a treatment X on an outcome Y_{obs} observable on units in a specific target population. Suppose further we can fully randomize X, so X will equal the randomized node R. In the potential-outcome formulation, the outcome becomes a vector **Y** indexed by X. Specifically, X determines which component Y_x of **Y** is observable conditional on X=x: $Y_{obs} = Y_x$ given X=x. To say X can causally affect a unit makes no reference to observation, however; it merely means that some components of **Y** are unequal. With a binary treatment and outcome, there are four types of units in the target population about a binary treatment X which indexes a binary potential-outcome vector **Y** (Copas, 1973):

1) Noncausal units with outcomes **Y**=(1,1) under X=1,0 ("doomed' to Y_{obs}=1);
2) Causal units with outcomes **Y**=(1,0) under X=1,0 (X=1 causes Y_{obs}=1);
3) Causal units with outcomes **Y**=(0,1) under X=1,0 (X=1 prevents Y_{obs}=1); and
4) Noncausal units with outcomes **Y**=(0,0) under X=1,0 ("immune" to Y_{obs}=1).

Suppose the proportion of type *i* in the trial population is p_i. There are now two null hypotheses:

H_s: There are no causal units: $p_2=p_3=0$ (sharp or strong null),

H_w: There is no net effect of treatment on the distribution of Y_{obs}: $p_2=p_3$ (weak null).

Under the randomization distribution we have

$E(Y_{obs}|X=1) = Pr(Y_{obs}=1|do[X=1]) = Pr(Y_1=1) = p_1+p_2$ and

$E(Y_{obs}|X=0) = Pr(Y_{obs}=1|do[X=0]) = Pr(Y_0=1) = p_1+p_3$;

hence H_w: $p_2=p_3$ is equivalent to the hypothesis that the expected outcome is the same for both treatment groups, and that the proportions with Y_{obs}=1 under the extreme population

intervention do[X=1] to every unit and do[X=0] to every unit are equal. Note however that only H_s entails that the proportion with $Y_{obs}=1$ would be the same under *every* possible allocation of treatment X among the units; this property implies that the Y margin is fixed under H_s, and thus provides a direct causal rationale for Fisher's exact test of H_s (Greenland, 1991).

H_s also entails H_w (or, in terms of parameter subspaces, $H_s \subset H_w$). The converse is false; but, under any of the "optimal" statistical tests that can be formulated from data on X and Y_{obs} only, power is identical to the test size on all alternatives to the sharp null with $p_2=p_3$, i.e., H_s is not identifiable within H_w, so within H_w the power of any valid test of H_s will not exceed its nominal alpha level. Thus, following Neyman, it is only relevant to think in terms of H_w, because H_w could be rejected whenever H_s could be rejected. Furthermore, some later authors would disallow $H_w - H_s$: $p_2 = p_3 \neq 0$ because it violates faithfulness (Spirtes et al., 2001) or because it represents an extreme treatment-by-unit interaction with no main effect (Senn, 2004).

There is also a Bayesian argument for focusing exclusively on H_w. H_w is of Lebesgue measure zero, so under the randomization model, distinctions within H_w can be ignored by inferences based on an absolutely continuous prior on $\mathbf{p} = (p_1,p_2,p_3)$ (Spirtes et al., 1993). More generally, any distinction that remains *a posteriori* can be traced to the prior. A more radical stance would dismiss both H_s and the model defined by 1-4 above as "metaphysical," because it invokes constraints on the joint distribution of the components Y_1 and Y_0, and that joint distribution is not identified by randomization of X if only X and Y_{obs} are observed (Dawid, 2000).

On the other hand, following Fisher one can argue that the null of key scientific and practical interest is H_s, and that $H_w - H_s$: $p_2 = p_3 \neq 0$ is a scientifically important and distinct hypothesis. For instance, $p_2>0$, $p_3>0$ entails the existence of units who should be treated quite differently, and provides an imperative to seek covariates that discriminate between the two causal types, even if $p_2=p_3$. Furthermore, rejection of the stronger H_s is a *weaker* inference than rejection of the weaker H_w, and thus rejecting only H_s would be a conservative interpretation of a "significant" test statistic. Thus, focusing on H_s is compatible with a strictly falsificationist view of testing in which acceptance of the null is disallowed. Finally, there are real examples in which X=1 causes Y=1 in some units and causes Y=0 in others; in some of these cases there may be near-perfect balance of causation and prevention, as predicted by certain physical explanations for the observations (e.g., as in Neutra et al., 1980).

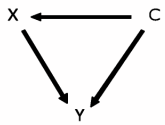
To summarize, identification problems arose in the earliest days of formal causal modeling, even when considering only the simplest of trials. Those problems pivoted not on whether one should attempt formal modeling of causation as distinct from association, but rather on what could be identified by standard experimental designs. In the face of limited (and limiting) design strategies, these problems initiated a long history of attempts to banish identification problems based on idealized inference systems and absolute philosophical assertions. But a counter-tradition of arguments, both practical and philosophical, has regarded identification problems as carriers of valuable scientific information: They are signs of study limitations which need to be recognized and can

only be dealt with effectively by innovative data collection (e.g., measuring more covariates or deploying new study designs), instead of by increasing sample sizes and defining the problems away so that "identical replications" are sufficient to narrow inferences.

3 Causal Diagrams Encode Numerous Uncertain Null Hypotheses

To move to the observational setting that is my main concern, consider figure 1, a typical causal diagram used to illustrate assumptions used by methods for estimating "the effect of X on Y" from observational data.

Figure 1: Naïve causal diagram



The first point to note is that this diagram is woefully incomplete relative to the epidemiologic reality, because it ignores

a) unmodeled confounders (variables not in the graph that affect more than one node in the graph),
b) selection effects (effects of factors in the graph on selection), and
c) measurement errors (which require addition of measurement nodes for each imperfectly measured node).

Put another way, typical causal DAGs like that in figure 1 are full of hidden, assumed null hypotheses, in the form of assumptions that imply problems a, b, and c are absent. For example, a causal DAG assumes that for **every** node pair (A,B) in the DAG,

1) there is **no** shared ancestor not in graph (not A↔B),
2) there is **no** unmarked conditioning event that has opened a path between A and B (not A─B),
3) if A and B are nonadjacent (neither A→B nor A←B), there is **no** mechanism that leads directly from one node to another (thus bypassing other nodes in the graph).

Not every study will seriously violate all of these assumptions. But in most studies in OHSS, none of the nulls 1-3 will have convincing support, and any purported test of a causal effect will really be a test of these 3 nulls as well as the specified causal null. This fact is just a special case of longstanding observations that statistical tests are really tests of all assumptions used in the test, not just the particular null of interest (Fisher, 1943; Box, 1980). In this regard, note that absence of arrows between nodes (3) encodes particularly strong nulls that are routinely presumed but rarely have supporting data. More often in OHSS, we observe only a conditional temporal sequence such as "A precedes B," which may be due to A→B, A↔B, A─B or some combination.

While sensitivity analysis is often recommended to examine the impact of deviations from assumptions, it becomes unintelligible if not infeasible as the number of assumptions (or corresponding parameters) increase. Then too, some causal inferences will display unlimited sensitivity to certain assumptions, requiring the introduction of priors on the corresponding parameters in order to salvage any inference (Greenland, 1998, 2005a; Gustafson, 2005). This problem arises in the model given below.

4 Eliminating Unsupported Nulls (graphical realism)

Let conditioning be denoted with square brackets around the conditioned event node. Then, in contrast to Figure 1, realistic causal graphs for OHSS will have

1) numerous unobserved (latent) nodes, often more of them than observed nodes,
2) few node pairs without an arc between them,
3) no **observed** set of variables sufficient for bias control, and
4) a selection node S that is bracketed and potentially affected by most other nodes.

In particular, when all variables are subject to measurement error, a realistic causal model for a single exposure-disease analysis will have at least:

X = Exposure, X*: measured X

Y = Outcome, Y*: measured Y

C = Known antecedents, C*: measured C

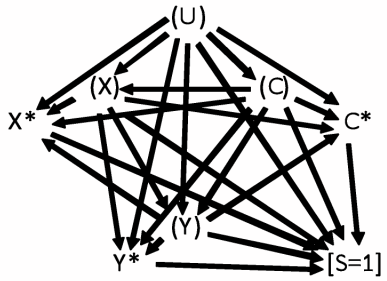
U = Other antecedents (unmeasured and possibly unknown)

S = Selection into the analysis (from selection into the study plus exclusions).

Because analysis is always conditioned on S=1, we should always show this conditioning event on the graph with a circle or brackets around it, e.g., as [S=1].

As an example, fig. 2 shows what I'd consider a **minimal** realistic causal graph for a typical case-control study of a life history and a degenerative disease outcome (e.g, nicotine intake X and Alzheimer's disease Y), which has 25 of 28 possible adjacencies.

Figure 2: Realistic causal diagram



What can fig. 2 provide if further assumptions cannot be justified? The only observed distribution is p(c*,x*,y*|S=1), which is not a factor in the causal Markov decomposition entailed by the graph,

$p(u,c,x,y,c^*,x^*,y^*,s) =$

$p(u)p(c|u)p(x|u,c)p(y|u,c,x)p(c^*|u,c,x,y)p(x^*|u,c,x,y)p(y^*|u,c,x,y)p(s|u,c,x,y,c^*,x^*,y^*)$,

which involves both S=0 events (not selected) and S=1 events (selected), i.e., the lowercase "s" is used when S can be either 0 or 1.

The marginal (total-population) potential-outcome distribution for Y after intervention on X, $p(y_x)$, equals $p(y|do[X=x])$, which under fig. 2 equals the standardized (mixing) distribution of Y given X standardized to (weighted by or mixed over) $p(c,u) = p(c|u)p(u)$:

$p(y_x) = p(y|do[x]) = \sum_{u,c} p(y|u,c,x)p(c|u)p(u)$.

This estimand involves only three factors in the decomposition, but none of them are identified if U is unobserved and no further assumptions are made. Analysis of the causal estimand $p(y_x)$ must somehow relate it to the observed distribution $p(c^*,x^*,y^*|S=1)$ using known or estimable quantities, or else remain purely speculative (i.e., a sensitivity analysis).

It is a long, hard road from $p(c^*,x^*,y^*|S=1)$ to $p(y_x)$, much longer than the current "causal inference" literature often makes it look. To appreciate the distance, rewrite the summand of the standardization formula for $p(y_x)$ as an inverse-probability-weighted (IPW) term derived from an observation $(c^*,x^*,y^*|S=1)$: From fig. 2,

$p(y|u,c,x)p(c|u)p(u) =$

$p(c^*,x^*,y^*|S=1)p(S=1)p(u,c,x,y|c^*,x^*,y^*,S=1)/$

$p(x|u,c)p(c^*|u,c,x,y)p(x^*|u,c,x,y)p(y^*|u,c,x,y)p(S=1|u,c,x,y,c^*,x^*,y^*)$.

The latter expression includes

1) the exposure dependence on its parents, $p(x|u,c)$;
2) the measurement distributions $p(c^*|u,c,x,y)$, $p(x^*|u,c,x,y)$, $p(y^*|u,c,x,y)$; and
3) the fully conditioned selection probability $p(S=1|u,c,x,y,c^*,x^*,y^*)$.

The absence of effects corresponding to 1−3 from graphs offered as "causal" suggests that "causal inference" from observational data using formal causal models remains a theoretical and largely speculative exercise (albeit often presented without explicit acknowledgement of that fact).

When adjustments for these effects are attempted, we are usually forced to use crude empirical counterparts of terms like those in 1−3, with each substitution demanding nonidentified assumptions. Consider that, for valid inference under figure 2,

1) Propensity scoring and IPW for treatment need $p(x|u,c)$, but all we get from data is $p(x^*|c^*)$. Absence of u and c is usually glossed over by assuming "no unmeasured confounders" or "no residual confounding." These are not credible assumptions in OHSS.
2) IPW for selection and censoring needs $p(S=1|u,c,x,y,c^*,x^*,y^*)$, but usually the most we get from a cohort study or nested study is $p(S=1|c^*,x^*)$. We do not even get that much in a case-control study.
3) Measurement-error correction needs conditional distributions from $p(c^*,x^*,y^*,u,c,x,y|S=1)$, but even when a "validation" study is done, we obtain only alternative measurements $c^\dagger,x^\dagger,y^\dagger$ (which are rarely error-free) on a tiny and

biased subset. So we end up with observations from $p(c^{\dagger},x^{\dagger},y^{\dagger},c^*,x^*,y^*|S=1,V=1)$ where V is the validation indicator.

4) Consistency between the observed X and the intervention variable, in the sense t hat $P(Y|X=x) = P(Y|do[X=x],X=x)$. This can be hard to believe for common variables such as smoking, body-mass index, and blood pressure, even if do[X=x] is well-defined (which is not usually the case).

In the face of these realities, standard practice seems to be: Present wildly hypothetical analyses that pretend the observed distribution $p(c^*,x^*,y^*|S=1)$, perhaps along with $p(c^{\dagger},x^{\dagger},y^{\dagger},c^*,x^*,y^*|S=1,V=1)$ or $p(S=1|c^*,x^*)$, is sufficient for causal inference. The massive gaps are filled in with models or assumptions, which are priors that reduce dimensionality of the problem to something within computing bounds. For example, use of IPW with $p(S=1|c^*,x^*)$ to adjust for selection bias (as when 1−S is a censoring indicator) depends crucially on a nonidentified ignorability assumption that $S \perp\!\!\!\perp (U,C,X,Y)|(C^*,X^*)$, i.e., that selection S is independent of the latent variables U,C,X,Y given the observed variables C*,X*. We should expect this condition to be violated whenever a latent variable affects selection directly or shares unobserved causes with selection. If such effects are exist but are missing from the analysis graph, then by some definitions the graph (and hence the resulting analysis) isn't causal, no matter how much propensity scoring (PS), marginal structural modeling (MSM), inverse-probability weighting (IPW), or other causal-modeling procedures we apply to the observations $(c^*,x^*,y^*|S=1)$.

Of course, the overwhelming dimensionality of typical OHSS problems virtually guarantees that arbitrary constraints will enter at some point, and forces even the best scientists to rely on a tiny subset of all the models or explanations consistent with available facts. Personal bias in determining this subset may be unavoidable due to strong cultural influences (such as adherence to received theories, as well as moral strictures and financial incentives), which can also lead to biased censoring of observations (Greenland, 2009c). One means of coping with such bias is by being aware of it, then trying to test it against the facts one can muster (which are often few).

The remaining sections sketch some alternatives to pretending we can identify unbiased or assuredly valid estimators of causal effects in observational data, as opposed to within hypothetical models for data generation (Greenland, 1990; Robins, 2001). In these approaches, both frequentist and Bayesian analyses are viewed as hypotheticals conditioned on a data-generation model of unknown validity. Frequentist analysis provides only inferences of the form "if the data-generation process behaves like this, here is how the proposed decision rule would perform," while Bayesian analysis provides only inferences of the form "if I knew that its data-generation process behaves like this, here is how this study would alter my bets."[1] If we aren't sure how the data-generation

[1]This statement describes Bayes factors (Good, 1983) conditioned on the model. That model may include an unknown parameter that indexes a finite number of submodels scattered over some high-dimensional subspace, in which case the Bayesian analysis is called "model averaging," usually with an implicit uniform prior over the models. Model averaging may also operate over continuous parameters via priors on those parameters.

process behaves, no statistical analysis can provide more, no matter how much causal modeling is done.

5 Predictive Analysis

If current models for observed-data generators (whether logistic, structural, or propensity models) can't be taken seriously as "causal", what can we make of their outputs? It is hard to believe the usual excuses offered for regression outputs (e.g., that they are "descriptive") when the fitted model is asserted to be causal or "structural." Are we to consider the outputs of (say) and IPW-fitted MSM to be some sort of data summary? Or will it function as some kind of optimal predictor of outcomes in a purely predictive context? No serious case has been made for causal models in either role, and it seems that some important technical improvements are needed before causal modeling methods become credible predictive tools.

Nonetheless, graphical models remain useful (and might be less misleading) even when they are not "causal," serving instead as mere carriers of conditional independence assumptions within a time-ordered framework. In this usage, one may still employ presumed causal independencies as prior judgments for specification. In particular, for predictive purposes, some or all of the arrows in the graph may retain informal causal interpretations; but they may be causally wrong, and yet the graph can still be correct for predictive purposes.

In this regard, most of the graphical modeling literature in statistics imposes little in the way of causal burden on the graph, as when graphs are used as influence diagrams, belief and information networks, and so on without formal causal interpretation (that is, without representing a formal causal model, e.g., Pearl, 1988; Hajek et al., 1992; Cox and Wermuth, 1996; Lauritzen, 1996). DAG rules remain valid for prediction if the absence of an open path from X to Y is interpreted as entailing $X \perp\!\!\!\perp Y$, or equivalently if the absence of a directed path from X to Y (in causal terms, X is not a cause of Y; equivalently, Y is not affected by X) is interpreted as entailing $X \perp\!\!\!\perp Y|\mathbf{pa}_X$, the noncausal Markov condition (where $\mathbf{pa}_X$ is the set of parents of X). In that case, $X \rightarrow Y$ can be used in the graph even if X has no effect on Y, or vice-versa.

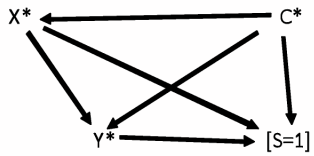
As an example, suppose X and Y are never observed without them affecting selection S, as when X is affects miscarriage S and Y is congenital malformation. If the target population is births, X predicts malformations Y among births (which have S=1). As another example, suppose X and Y are never observed without an uncontrolled, ungraphed confounder U, as when X is diet and Y is health status. If one wishes to target those at high risk for screening or actuarial purposes it does not matter if $X \rightarrow Y$ represents a causally confounded relation. Lack of a directed path from X to Y now corresponds to lack of additional predictive value for Y from X given $\mathbf{pa}_X$. Arrow directions in temporal (time-ordered) predictive graphs correspond to point priors about time order, just as they do in causal graphs.

Of course, if misinterpreted as causal, predictive inferences from graphs (or any predictive modeling) may be potentially disastrous for judging interventions on X. But, in OHSS, the causality represented by a directed path in a so-called causal diagram rarely

corresponds to more than a hypothesis, plausible perhaps but only one among a myriad of others. If most arrows shown in a graph encode no real data other than an observed conditional temporal sequencing, then labeling the graph as a "causal diagram" sets the stage for the disaster.

Figure 3 is the temporal predictive diagram for the observables in the earlier example, assuming those events occur in the order C*, X*, Y*, [S=1].

Figure 3: Temporally predictive diagram



Comparison to the causal diagram in figure 2 illustrates how a temporal predictive diagram for an observable frequency distribution may be derived from an underlying causal diagram for a nonidentified theory. Figure 3 is saturated in the sense that all nodes are connected by an edge, but this need not be so for a predictive diagram derived from a causal one. If there is temporal ambiguity among the observables, there may be multiple predictive diagrams compatible with the causal diagram (which will form a subset of the multiple probability graphs compatible with the causal diagram).

If we treat causal models as carriers of prior information about conditional independencies, they appear as legitimate candidates to consider as predictive models. For example, MSMs can be evaluated as devices for prediction from fixed sequences and structural nested models can be evaluated as devices for prediction from stochastic processes. I would thus offer this challenge to the current "longitudinal causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify predictive links among observed quantities, are there predictive advantages of structural modeling (modeling potential outcomes as well as observed outcomes)? If not, what precisely is the advantage of fitting such models (compared to noncausal models) when effects are not identified?

I believe there *are* advantages of causal models, precisely as described by Pearl (2000): They provide an encoding for qualitative (structural) prior information expressed in terms of "cause" and "effect." But in current practice, fitting methods for complex causal models are quite primitive, and need to incorporate properly smoothness and other information that can be freely assumed in purely predictive-modeling approaches. This is a general problem of semi-parametric theory: It necessarily focuses sharp constraints in some dimensions and none in "most" dimensions (represented by the infinite-dimensional time component in standard Cox models). When relevant dimensions for constraint (those

where much background information is available) are not well represented by the dimensions constrained by the model, considerably efficiency can be lost for estimating parameters of interest. A striking example given by Whittemore and Keller (1986) displayed the poor small-sample performance for estimating a survival curve when using an unsmoothed nonparametric hazard estimator (Kaplan-Meier or Nelson-Altschuler estimation), relative to spline smoothing of the hazard.

6 Pruning the Identified Portion of the Model

Over recent decades, great strides have been made in creating predictive algorithms; the question remains however, what role should these algorithms play in causal inference? It would seem that these algorithms can be beneficially applied to fitting the marginal distribution identified by the observations. Nonetheless, the targets of causal inference in observational studies lie beyond the identified margin, and thus beyond the reach of these algorithms. At best, then, the algorithms can provide the identified foundation for building into unobserved dimensions of the phenomena under study.

Even if we focus only on the identified margin, however, there may be far more nodes and edges than seem practical to allow in the final model. A prominent feature of modern predictive algorithms is that they start with an impractically large number of terms and then aggressively prune the model, and may re-grow and re-prune repeatedly (Hastie et al., 2009). These strategies coincide with the intuition that omitting a term is justified when its contribution is too small to stand out against bias and background noise; e.g., we do not include variables like patient identification number because we know that are usually pure noise.

Nonetheless, automated algorithms often delete variables or connections that prior information instead suggests are relevant or related; thus shields from pruning are often warranted. Furthermore, a deleted node or arrow may indeed be important from a contextual perspective even if does not meet algorithmic retention criteria. Thus, model simplification strategies such as pruning may be justified by a need for dimensionality reduction, but should be recognized as part of algorithmic compression or computational prediction, not as a mode of inference about structural models.

Apart from these vague cautions, it has long been recognized that if our goal is to evaluate causal effects, different loss functions are needed from those in the pruning algorithms commonly applied by researchers. Specifically, the loss or benefit entailed by pruning needs to be evaluated in reference to the target effect under study, and not simply successful prediction of identified quantities. Operationalizing this imperative requires building out into the nonidentified (latent) realm of the target effects, which is the focus of *bias modeling*.

7 Modeling Latent Causal Structures (Bias Modeling)

The target effects in causal inference are functions of unobserved dimensions of the data-generating process, which consist primarily of bias sources (Greenland, 2005a). Once we recognize the nonidentification this structure entails, the major analysis task shifts away

from mathematical statistics to prior specification, because with nonidentification only proper priors on nonidentified parameters can lead to proper posteriors.

Even the simplest point-exposure case can involve complexities that transform simple and precise-looking conventional results into complex and utterly ambiguous posteriors (Greenland, 2009a, 2009b). In a model complex enough to reflect Figure 2, there are far too many elements of specification to contextually justify them all in detail. For example, one could only rarely justify fewer than two free structural parameters per arrow, and the distributional form for each parameter prior would call for at least two hyperparameters per parameter (e.g., a mean and a variance), leading to at least 50 parameters and 100 hyperparameters in a graph with 25 arrows. Allowing but one prior association parameter (e.g., a correlation) per parameter pair adds over 1,000 (50 choose 2) more hyperparameters.

As a consequence of the exponential complexity of realistic models, prior specification is difficult, ugly, ad hoc, highly subjective, and tentative in the extreme. In addition, the hard-won model will lack generalizability and elegance, making it distasteful to both the applied scientist and the theoretical statistician. Nor will it please the applied statistician concerned with "data analysis," for the analysis will instead revolve around numerous contextual judgments that enlist diverse external sources of information. In contrast to the experimental setting (in which the data-generation model may be dictated entirely by the design), the usually sharp distinction between prior and data information will be blurred by the dependence of the data-generation model on external information.

These facts raise another challenge to the current "causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify predictive links among observed quantities, how can we incorporate simultaneously all error sources (systematic as well as random) known to be important into a complex longitudinal framework involving mismeasurement of entire sequences of exposures and confounders? Some progress on this front has been made, but primarily in contexts with validation data available (Cole et al., 2010), which is not the usual case.

8 The Descriptive Alternative

In the face of the extraordinary complexity of realistic models for OHSS, it should be an option of each study to focus on describing the study and its data thoroughly, sparing us attempts at inference about nonidentified quantities such as "causal effects." This option will likely never be popular, but should be allowed and even encouraged (Greenland et al., 2004). After all, why should I care about your causal inferences, especially if they are based on or grossly over-weighted by the one or few studies that you happened to be involved in? If I am interested in forming my own inferences, I do want to see your data and get an accurate narrative of the physical processes that produced them. In this regard, statistics may supply data summaries. Nonetheless, it must be made clear exactly how the statistics offered reflect the data as opposed to some hypothesis about the population from which they came; *P*-values do not satisfy this requirement (Greenland, 1993; Poole, 2001).

Here then is a final challenge to the "causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify associations among observed quantities, how can we interpret the outputs of "structural modeling" (such as confidence limits for ostensibly causal estimands which are not in fact identified) as data summaries? We should want to see answers that are sensible when the targets are effects in a context at least as complex as in fig. 2.

9 What is a Causal Diagram?

The above considerations call into question some epidemiological accounts of causal diagrams. Pearl (2000) describes a causal model *M* as a formal functional system giving relations among a set of variables. *M* defines a joint probability distribution p() and an intervention operator do[] on the variables. A causal diagram is then a directed graph *G* that implies the usual Markov decomposition for p() and displays additional properties relating p() and do[]. In particular, each child-parent family {X, $\mathbf{pa}_X$} in *G* satisfies

1) p(x|do[$\mathbf{pa}_X$=a]) = p(x|$\mathbf{pa}_X$=a), and
2) if Z is not in {X, $\mathbf{pa}_X$}, p(x|do[Z=z],$\mathbf{pa}_X$=a) = p(x|$\mathbf{pa}_X$=a).

(e.g., see Pearl, 2000, p. 24). These properties stake out *G* as an illustration (mapping) of structure within *M*.

Condition 1 is often described as stating that the association of each node X with its parent vector $\mathbf{pa}_X$ is unconfounded given *M*. Condition 2 says that, given *M*, the only variables in *G* that affect a node X are its parents, and is often called the causal Markov condition (CMC). Nonetheless, as seems to happen often as time passes and methods become widely adopted, details have gotten lost. In the more applied literature, causal diagrams have come to be described as "unconfounded graphs" without reference to an underlying causal model (e.g., Hernán et al., 2004; VanderWeele and Robins, 2007; Glymour and Greenland, 2008). This description not only misses the CMC (2) but, taken literally, means that all shared causes are in the graph.

Condition 1 is a property relating two mathematical objects, *G* and *M*. To claim a diagram is unconfounded is to instead make a claim about the relation of *G* the real world, thus inviting confusion between a *model* for causal processes and the actual processes. For many experts in OHSS, the claim of unconfoundedness has zero probability of being correct because of its highly restrictive empirical content (e.g., see Robins and Wasserman, 1999ab). At best, we can only hope that the diagram provides a useful computing aid for predicting the outcomes of intervention strategies.

As with regression models, causal models in OHSS are always false. Because we can never know we have a correct model (and in fact in OHSS we can't even know if we are very close), to say *G* is causal if unconfounded is a scientifically vacuous definition: It is saying the graph is causal if the causal model it represents is correct. This is akin to saying a monotone increasing function from the range of X to [0,1] is not a probability distribution if it is not in fact how X is distributed; thus a normal(μ,σ^2) cumulative function wouldn't be a probability distribution unless it is *the* actual probability distribution for X (whether that distribution is an objective event generator or a subjective betting schedule).

So, to repeat: To describe a causal diagram as an "unconfounded graph" blurs the distinction between models and reality. Model-based deductions are logical conditionals of the form "model *M* deductively yields these conclusions," and have complete certainty *given* the model *M*. But the model, and hence reality, is never known with certainty, and in OHSS cannot be claimed as known except in the most crude fashion. The point is brought home above by appreciating just how unrealistic all causal models and diagrams in OHSS must be. Thus I would encourage the description of causal diagrams as graphical causal models (or more precisely, graphical representations of certain equivalence classes of causal models), rather than as "unconfounded graphs" (or similar phrases). This usage might even be acceptable to some critics of the current causal-modeling literature (Dawid, 2008).

10 Summary and Conclusions

I would be among the last to deny the utility of causal diagrams; but I argue that their practical utility in OHSS is limited to (i) compact and visually immediate representation of assumptions, and (ii) illustration of sources of nonidentification and bias given realistic assumptions. Converse claims about their utility for identification seem only the latest in a long line of promises to "solve" the problem of causal inference. These promises are akin to claims of preventing and curing all cancers; while progress is possible, the enormous complexity of real systems should leave us skeptical about claims of "solutions" to the real problem.

Many authors have recognized that the problem of effect identification is unsolvable in principle. Although this logical impossibility led some to deny the scientific merit of causal thinking, it has not prevented development of useful tools that have causal-modeling components. Nonetheless, the most precision we can realistically hope for estimating effects in OHSS is about one-digit accuracy, and in many problems even that seems too optimistic. Thus some practical sense is needed to determine what is and isn't important to include as model components. Yet, despite the crudeness of OHSS, good sense seems to lead almost inevitably to including more components than can be identified by available data.

My main point is that effect identification (in the frequentist sense of identification by the observed data) should be abandoned as a primary goal in causal modeling in OHSS. My reasons are practical: Identification will often demand dropping too much of importance from the model, thus imposing null hypotheses that have no justification in either past frequency observations or in priors about mechanisms generating the observations, thus leading to overconfident and biased inferences. In particular, defining a graph as "causal" if it is unconfounded assumes a possibly large set of causal null hypotheses (at least two for every pair of nodes in the graph: no shared causes or conditioned descendants not in the graph). In OHSS, the only graphs that satisfy such a definition will need many latent nodes to be "causal" in this sense, and as a consequence will reveal the nonidentified nature of target effects. Inference may then proceed by imposing contextually defensible priors or penalties (Greenland, 2005a, 2009a, 2009b, 2010).

Despite my view and similar ones (e.g., Gustafson, 2005), I suspect the bulk of causal-inference statistics will trundle on relying exclusively on artificially identified models. It will thus be particularly important to remember that just because a method is labeled a "causal modeling" method does not mean it gives us estimates and tests of actual causal effects. For those who find identification too hard to abandon in formal analysis, the only honest recourse is to separate identified and nonidentified components of the model, focus technique on the identified portion, and leave the latent residual as a topic for sensitivity analysis, speculative modeling, and further study. In this task, graphs can be used without the burden of causality if we allow them a role as pure prediction tools, and they can also be used as causal diagrams of the largely latent structure that generates the data.

Acknowledgments: I am most grateful to Tyler VanderWeele, Jay Kaufman, and Onyebuchi Arah for their extensive and useful comments on this chapter.

References

Biggs, N., Lloyd, E. and Wilson, R. (1986). *Graph Theory, 1736-1936*. Oxford University Press.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society*, *Series A* **143**, 383–430.

Cole S.R. and M.A. Hernán (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31**, 163–165.

Cole, S.R. and C.E. Frangakis (2009). The consistency assumption in causal inference: a definition or an assumption? *Epidemiology* **20**, 3–5.

Cole, S.R., L.P. Jacobson, P.C. Tien, L. Kingsley, J.S. Chmiel and K. Anastos (2010). Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *American Journal of Epidemiology* **171**, 113-122.

Copas, J.G. (1973). Randomization models for matched and unmatched 2x2 tables. *Biometrika* **60**, 267-276.

Cox, D.R. and N. Wermuth. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. Boca Raton, FL: CRC/Chapman and Hall.

Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association* **95**, 407-448.

Dawid, A.P. (2008). Beware of the DAG! In: *NIPS 2008 Workshop Causality: Objectives and Assessment*. JMLR Workshop and Conference Proceedings.

Duncan, O.D. (1975). *Introduction to Structural Equation Models*. New York: Academic Press.

Fisher, R.A. (1943; reprinted 2003). Note on Dr. Berkson's criticism of tests of significance. *Journal of the American Statistical Association* **38**, 103–104. Reprinted in the *International Journal of Epidemiology* **32**, 692.

Freedman, D.A. and Humphreys, P. (1999). Are there algorithms that discover causal structure? *Synthese* **121**, 29–54.

Glymour, M.M. and S. Greenland (2008). Causal diagrams. Ch. 12 in: Rothman, K.J., S. Greenland and T.L. Lash, eds. *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott.

Good, I.J. (1983). *Good thinking*. Minneapolis: U. Minnesota Press.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology* **1**, 421-429.

Greenland, S. (1991). On the logical justification of conditional tests for two-by-two-contingency tables. *The American Statistician* **45**, 248-251.

Greenland, S. (1993). Summarization, smoothing, and inference. *Scandinavian Journal of Social Medicine* **21**, 227-232.

Greenland, S. (1998). The sensitivity of a sensitivity analysis. In: 1997 Proceedings of the Biometrics Section. Alexandria, VA: American Statistical Association, 19-21.

Greenland, S. (2005a). Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). *Emerging Themes in Epidemiology* (online journal) 2:1–4. (Originally published as "Causality theory for policy uses of epidemiologic measures," Chapter 6.2 in: Murray, C.J.L., J.A. Salomon, C.D. Mathers and A.D. Lopez, eds. (2002) *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/WHO, 291-302.)

Greenland, S. (2005b). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A* **168**, 267–308.

Greenland, S. (2009a). Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. *International Journal of Epidemiology* **38**, 1662–1673.

Greenland, S. (2009b). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Science* **24**, 195-210.

Greenland, S. (2009c). Dealing with uncertainty about investigator bias: disclosure is informative. *Journal of Epidemiology and Community Health* **63**, 593-598.

Greenland, S. (2010). The need for syncretism in applied statistics (comment on "The future of indirect evidence" by Bradley Efron). *Statistical Science* **25**, in press.

Greenland, S., J. Pearl, and J.M. Robins (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37-48.

Greenland, S., M. Gago-Dominguez, and J.E. Castellao (2004). The value of risk-factor ("black-box") epidemiology (with discussion). *Epidemiology* **15**, 519-535.

Gustafson, P. (2005). On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion). *Statistical Science* **20**, 111-140.

Hajek, P., T. Havranek and R. Jirousek (1992). *Uncertain Information Processing in Expert Systems*. Boca Raton, FL: CRC Press.

Hastie, T., R. Tibshirani and J. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. New York: Springer.

Hernán, M.A. (2005). Hypothetical interventions to define causal effects—afterthought or prerequisite? *American Journal of Epidemiology* **162**, 618–620.

Hernán M.A., S. Hernandez-Diaz, M.M. Werler and A.A. Mitchell. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology* **155**, 176–184.

Hernán M.A., S. Hernandez-Diaz and J.M. Robins (2004). A structural approach to selection bias. *Epidemiology* **15**, 615-625.

Jewell, N. (2004). *Statistics for Epidemiology*. Boca Raton, FL: Chapman and Hall/CRC.

Lad, F. (1999). Assessing the foundations of Bayesian networks: A challenge to the principles and the practice. *Soft Computing* **3**, 174-180.

Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon Press.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Morgan, S.L. and C. Winship. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Neutra, R.R., S. Greenland, and E.A. Friedman (1980). The effect of fetal monitoring on cesarean section rates. *Obstetrics and Gynecology* **55**, 175-180.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika* **82**, 669-710.

Pearl, J. (2000; 2nd ed. 2009). *Causality*. New York: Cambridge University Press.

Pearl, J. and P. Verma (1991). A theory of inferred causation. In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Ed. J.A. Allen, R. Filkes and E. Sandewall. San Francisco: Morgan Kaufmann, 441-452.

Poole, C. (2001). Poole C. Low P-values or narrow confidence intervals: Which are more durable? *Epidemiology* **12**, 291–294.

Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12**, 313–320.

Robins, J.M. and L. Wasserman (1999a). On the impossibility of inferring causation from association without background knowledge. In: *Computation, Causation, and Discovery*. Glymour, C. and Cooper, G., eds. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, pp. 305-321.

Robins, J.M. and L. Wasserman (1999b). Rejoinder to Glymour and Spirtes. In: *Computation, Causation, and Discovery*. Glymour, C. and Cooper, G., eds. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, pp. 333-342.

Robins, J.M., R. Scheines, P. Spirtes and L. Wasserman (2003). Uniform consistency in causal inference. *Biometrika* **90**, 491-515.

Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine* **23**, 3729–3753.

Shafer, G. (2002). Comment on "Estimating causal effects," by George Maldonado and Sander Greenland. *International Journal of Epidemiology* **31**, 434-435.

Spirtes, P., C. Glymour and R. Scheines (1993; 2nd ed. 2001). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Susser, M. (1973). *Causal Thinking in the Health Sciences*. New York: Oxford University Press.

VanderWeele, T.J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology* **20**, 880-883.

VanderWeele, T.J. and J.M. Robins (2007). Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *American Journal of Epidemiology* **166**, 1096-1104.

Whittemore, A.S. and J.B. Keller (1986). Survival estimation using splines. *Biometrics* **42**, 495-506.

Wright, S., (1934). The method of path coefficients. *Annals of Mathematical Statistics* **5**,161-215.

23

Actual Causation and the Art of Modeling

Joseph Y. Halpern and Christopher Hitchcock

1 Introduction

In *The Graduate*, Benjamin Braddock (Dustin Hoffman) is told that the future can be summed up in one word: "Plastics". One of us (Halpern) recalls that in roughly 1990, Judea Pearl told him that the future was in causality. Pearl's own research was largely focused on causality in the years after that; his seminal contributions are widely known. We were among the many influenced by his work. We discuss one aspect of it, *actual causation*, in this article, although a number of our comments apply to causal modeling more generally.

Pearl introduced a novel account of actual causation in Chapter 10 of *Causality*, which was later revised in collaboration with one of us [Halpern and Pearl 2005]. In some ways, Pearl's approach to actual causation can be seen as a contribution to the philosophical project of trying to analyze actual causation in terms of counterfactuals, a project associated most strongly with David Lewis [1973a]. But Pearl's account was novel in at least two important ways. The first was his use of *structural equations* as a tool for modeling causality. In the philosophical literature, causal structures were often represented using so-called *neuron diagrams*, but these are not (and were never intended to be) all-purpose representational tools. (See [Hitchcock 2007b] for a detailed discussion of the limitations of neuron diagrams.) We believe that the lack of a more adequate representational tool had been a serious obstacle to progress. Second, while the philosophical literature on causality has focused almost exclusively on actual causality, for Pearl, actual causation was a rather specialized topic within the study of causation, peripheral to many issues involving causal reasoning and inference. Thus, Pearl's work placed the study of actual causation within a much broader context.

The use of structural equations as a model for causal relationships was well known long before Pearl came on the scene; it seems to go back to the work of Sewall Wright in the 1920s (see [Goldberger 1972] for a discussion). However, the details of the framework that have proved so influential are due to Pearl. Besides the Halpern-Pearl approach mentioned above, there have been a number of other closely-related approaches for using structural equations to model actual causation; see, for example, [Glymour and Wimberly 2007; Hall 2007; Hitchcock 2001; Hitchcock 2007a; Woodward 2003]. The goal of this paper is to look more carefully at the modeling of causality using structural equations. For definiteness, we use the

Halpern-Pearl (HP) version [Halpern and Pearl 2005] here, but our comments apply equally well to the other variants.

It is clear that the structural equations can have a major impact on the conclusions we draw about causality—it is the equations that allow us to conclude that lower air pressure is the cause of the lower barometer reading, and not the other way around; increasing the barometer reading will not result in higher air pressure. The structural equations express the effects of *interventions*: what happens to the bottle if it is hit with a hammer; what happens to a patient if she is treated with a high dose of the drug, and so on. These effects are, in principle, objective; the structural equations can be viewed as describing objective features of the world. However, as pointed out by Halpern and Pearl [2005] and reiterated by others [Hall 2007; Hitchcock 2001; Hitchcock 2007a], the choice of variables and their values can also have a significant impact on causality. Moreover, these choices are, to some extent, subjective. This, in turn, means that judgments of actual causation are subjective.

Our view of actual causation being at least partly subjective stands in contrast to the prevailing view in the philosophy literature, where the assumption is that the job of the philosopher is to analyze the (objective) notion of causation, rather like that of a chemist analyzing the structure of a molecule. This may stem, at least in part, from failing to appreciate one of Pearl's lessons: actual causality is only part of the bigger picture of causality. There can be an element of subjectivity in ascriptions of actual causality without causation itself being completely subjective. In any case, the experimental evidence certainly suggests that people's views of causality are subjective, even when there is no disagreement about the relevant structural equations. For example, a number of experiments show that broadly normative considerations, including the subject's own moral beliefs, affect causal judgment. (See, for example, [Alicke 1992; Cushman 2009; Cushman, Knobe, and Sinnott-Armstrong 2008; Hitchcock and Knobe 2009; Knobe and Fraser 2008].) Even in relatively non-controversial cases, people may want to focus on different aspects of a problem, and thus give different answers to questions about causality. For example, suppose that we ask for the cause of a serious traffic accident. A traffic engineer might say that the bad road design was the cause; an educator might focus on poor driver's education; a sociologist might point to the pub near the highway where the driver got drunk; a psychologist might say that the cause is the driver's recent breakup with his girlfriend.[1] Each of these answers is reasonable. By appropriately choosing the variables, the structural equations framework can accommodate them all.

Note that we said above "by appropriately choosing the variables". An obvious question is "What counts as an appropriate choice?". More generally, what makes a model an appropriate model? While we do want to allow for subjectivity, we need

[1] This is a variant of an example originally due to Hanson [1958].

to be able to justify the modeling choices made. A lawyer in court trying to argue that faulty brakes were the cause of the accident needs to be able to justify his model; similarly, his opponent will need to understand what counts as a legitimate attack on the model. In this paper we discuss what we believe are reasonable bases for such justifications. Issues such as model stability and interactions between the events corresponding to variables turn out to be important.

Another focus of the paper is the use of defaults in causal reasoning. As we hinted above, the basic structural equations model does not seem to suffice to completely capture all aspects of causal reasoning. To explain why, we need to briefly outline how actual causality is defined in the structural equations framework. Like many other definitions of causality (see, for example, [Hume 1739; Lewis 1973b]), the HP definition is based on counterfactual dependence. Roughly speaking, A is a cause of B if, had A not happened (this is the counterfactual condition, since A did in fact happen) then B would not have happened. As is well known, this naive definition does not capture all the subtleties involved with causality. Consider the following example (due to Hall [2004]): Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had Suzy not thrown. Thus, according to the naive counterfactual definition, Suzy's throw is not a cause of the bottle shattering. This certainly seems counterintuitive.

The HP definition deals with this problem by taking A to be a cause of B if B counterfactually depends on A *under some contingency*. For example, Suzy's throw is the cause of the bottle shattering because the bottle shattering counterfactually depends on Suzy's throw, under the contingency that Billy doesn't throw. (As we will see below, there are further subtleties in the definition that guarantee that, if things are modeled appropriately, Billy's throw is not also a cause.)

While the definition of actual causation in terms of structural equations has been successful at dealing with many of the problems of causality, examples of Hall [2007], Hiddleston [2005], and Hitchcock [2007a] show that it gives inappropriate answers in cases that have structural equations isomorphic to ones where it arguably gives the appropriate answer. This means that, no matter how we define actual causality in the structural-equations framework, the definition must involve more than just the structural equations. Recently, Hall [2007], Halpern [2008], and Hitchcock [2007a] have suggested that using defaults might be a way of dealing with the problem. As the psychologists Kahneman and Miller [1986, p. 143] observe, "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it". This intuition is also present in the legal literature. Hart and Honoré [1985] observe that the statement "It was the presence of oxygen that caused the fire" makes sense only if there were reasons to view the presence of oxygen as abnormal.

As shown by Halpern [2008], we can model this intuition formally by combining a well-known approach to modeling defaults and normality, due to Kraus, Lehmann,

and Magidor [1990] with the structural-equation model. Moreover, doing this leads to a straightforward solution to the problem above. The idea is that, when showing that if A hadn't happened then B would not have happened, we consider only contingencies that are at least as normal as the actual world. For example, if someone typically leaves work at 5:30 PM and arrives home at 6, but, due to unusually bad traffic, arrives home at 6:10, the bad traffic is typically viewed as the cause of his being late, not the fact that he left at 5:30 (rather than 5:20).

But once we add defaults to the model, the problem of justifying the model becomes even more acute. We not only have to justify the structural equations and the choice of variables, but also the default theory. The problem is exacerbated by the fact that default and "normality" have a number of interpretations. Among other things, they can represent moral obligations, societal conventions, prototypicality information, and statistical information. All of these interpretations are relevant to understanding causality; this makes justifying default choices somewhat subtle.

The rest of this paper is organized as follows. In Sections 2 and 3, we review the notion of causal model and the HP definition of actual cause; most of this material is taken from [Halpern and Pearl 2005]. In Section 4, we discuss some issues involved in the choice of variables in a model. In Section 5, we review the approach of [Halpern 2008] for adding considerations of normality to the HP framework, and discuss some modeling issues that arise when we do so. We conclude in Section 6.

2 Causal Models

In this section, we briefly review the HP definition of causality. The description of causal models given here is taken from [Halpern 2008], which in turn is based on that of [Halpern and Pearl 2005].

The HP approach assumes that the world is described in terms of random variables and their values. For example, if we are trying to determine whether a forest fire was caused by lightning or an arsonist, we can take the world to be described by three random variables:

- F for forest fire, where $F = 1$ if there is a forest fire and $F = 0$ otherwise;
- L for lightning, where $L = 1$ if lightning occurred and $L = 0$ otherwise;
- ML for match (dropped by arsonist), where $ML = 1$ if the arsonist drops a lit match, and $ML = 0$ otherwise.

Some random variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. For example, to model the fact that if either a match is lit or lightning strikes, then a fire starts, we could use the random variables ML, F, and L as above, with the equation $F = \max(L, ML)$. (Alternately, if a fire requires both causes to be present, the equation for F becomes $F = \min(L, ML)$.) The equality sign in this equation should be thought of more like an assignment statement in programming languages; once we set the values of F

and L, then the value of F is set to their maximum. However, despite the equality, if a forest fire starts some other way, that does not force the value of either *ML* or L to be 1.

It is conceptually useful to split the random variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. For example, in the forest-fire example, the variables *ML*, L, and F are endogenous. However, we want to take as given that there is enough oxygen for the fire and that the wood is sufficiently dry to burn. In addition, we do not want to concern ourselves with the factors that make the arsonist drop the match or the factors that cause lightning. These factors are all determined by the exogenous variables.

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *modifiable structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (that is, the set of values over which Y *ranges*). $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u, y, z) = u + y$, which is usually written as $X \leftarrow U + Y$.[2] Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.

In the running forest-fire example, suppose that we have an exogenous random variable U that determines the values of L and *ML*. Thus, U has four possible values of the form (i, j), where both of i and j are either 0 or 1. The i value determines the value of L and the j value determines the value of *ML*. Although F_L gets as arguments the vale of U, *ML*, and F, in fact, it depends only on the (first component of) the value of U; that is, $F_L((i, j), m, f) = i$. Similarly, $F_{ML}((i, j), l, f) = j$. The value of F depends only on the value of L and *ML*. *How* it depends on them depends on whether either cause by itself is sufficient for the forest fire or whether both are necessary. If either one suffices, then $F_F((i, j), l, m) = \max(l, m)$, or, perhaps more comprehensibly, $F = \max(L, ML)$; if both are needed, then $F = \min(L, ML)$. For future reference, call the former model the *disjunctive* model, and the latter the *conjunctive* model.

The key role of the structural equations is to define what happens in the presence of external interventions. For example, we can explain what happens if the arsonist

[2]The fact that X is assigned $U + Y$ (i.e., the value of X is the sum of the values of U and Y) does not imply that Y is assigned $X - U$; that is, $F_Y(U, X, Z) = X - U$ does not necessarily hold.

does *not* drop the match. In the disjunctive model, there is a forest fire exactly if there is lightning; in the conjunctive model, there is definitely no fire. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model denoted $M_{X \leftarrow x}$. In the new causal model, the equation for X is very simple: X is just set to x; the remaining equations are unchanged. More formally, $M_{X \leftarrow x} = (\mathcal{S}, \mathcal{F}^{X \leftarrow x})$, where $\mathcal{F}^{X \leftarrow x}$ is the result of replacing the equation for X in $\mathcal{F}$ by $X = x$.

The structural equations describe *objective* information about the results of interventions, that can, in principle, be checked. Once the modeler has selected a set of variables to include in the model, *the world* determines which equations among those variables correctly represent the effects of interventions.[3] By contrast, the *choice* of variables is subjective; in general, there need be no objectively "right" set of exogenous and endogenous variables to use in modeling a problem. We return to this issue in Section 4.

It may seem somewhat circular to use causal models, which clearly already encode causal information, to define actual causation. Nevertheless, as we shall see, there is no circularity. The equations of a causal model do not represent relations of *actual causation*, the very concept that we are using them to define. Rather, the equations characterize the results of *all possible* interventions (or at any rate, all of the interventions that can be represented in the model) without regard to what actually happened. Specifically, the equations do not depend upon the actual values realized by the variables. For example, the equation $F = \max(L, ML)$, by itself, does not say anything about whether the forest fire was actually caused by lightning or by an arsonist, or, for that matter, whether a fire even occurred. By contrast, relations of actual causation depend crucially on how things actually play out.

A sequence of endogenous $X_1, \ldots, X_n$ of is a *directed path* from X_1 to X_n if the value of X_{i+1} (as given by $F_{X_{i+1}}$) depends on the value of X_i, for $1 = 1, \ldots, n-1$. In this paper, following HP, we restrict our discussion to *acyclic* causal models, where causal influence can be represented by an acyclic Bayesian network. That is, there is no cycle $X_1, \ldots, X_n, X_1$ of endogenous variables that forms a directed path from X_1 to itself. If M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, there is a unique solution for all the equations.

[3] In general, there may be uncertainty about the causal model, as well as about the true setting of the exogenous variables in a causal model. Thus, we may be uncertain about whether smoking causes cancer (this represents uncertainty about the causal model) and uncertain about whether a particular patient actually smoked (this is uncertainty about the value of the exogenous variable that determines whether the patient smokes). This uncertainty can be described by putting a probability on causal models and on the values of the exogenous variables. We can then talk about the probability that A is a cause of B.

3 The HP Definition of Actual Cause

3.1 A language for describing causes

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$, where ϕ is a Boolean combination of primitive events, $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\phi$. The special case where $k = 0$ is abbreviated as ϕ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$ says that ϕ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A causal formula ψ is true or false in a causal model, given a context. As usual, we write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in causal model M given context $\vec{u}$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (that is, the unique vector of values for the endogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\phi$ if $(M_{\vec{Y} \leftarrow \vec{y}}, \vec{u}) \models \phi$. We write $M \models \phi$ if $(M, \vec{u}) \models \phi$ for all contexts $\vec{u}$.

For example, if M is the disjunctive causal model for the forest fire, and u is the context where there is lightning and the arsonist drops the lit match, then $(M, u) \models [ML \leftarrow 0](F = 1)$, since even if the arsonist is somehow prevented from dropping the match, the forest burns (thanks to the lightning); similarly, $(M, u) \models [L \leftarrow 0](F = 1)$. However, $(M, u) \models [L \leftarrow 0; ML \leftarrow 0](F = 0)$: if the arsonist does not drop the lit match and the lightning does not strike, then the forest does not burn.

3.2 A preliminary definition of causality

The HP definition of causality, like many others, is based on counterfactuals. The idea is that if A and B both occur, then A is a cause of B if, if A hadn't occurred, then B would not have occurred. This idea goes back to at least Hume [1748, Section VIII], who said:

> We may define a cause to be an object followed by another, ..., where, if the first object had not been, the second never had existed.

This is essentially the *but-for* test, perhaps the most widely used test of actual causation in tort adjudication. The but-for test states that an act is a cause of injury if and only if, but for the act (i.e., had the the act not occurred), the injury would not have occurred.

There are two well-known problems with this definition. The first can be seen by considering the disjunctive causal model for the forest fire again. Suppose that the arsonist drops a match and lightning strikes. Which is the cause? According to a naive interpretation of the counterfactual definition, neither is. If the match hadn't dropped, then the lightning would still have struck, so there would have been

a forest fire anyway. Similarly, if the lightning had not occurred, there still would have been a forest fire. As we shall see, the HP definition declares both lightning and the arsonist causes of the fire. (In general, there may be more than one actual cause of an outcome.)

A more subtle problem is what philosophers have called *preemption*, which is illustrated by the rock-throwing example from the introduction. As we observed, according to a naive counterfactual definition of causality, Suzy's throw would not be a cause.

The HP definition deals with the first problem by defining causality as counterfactual dependency *under certain contingencies*. In the forest-fire example, the forest fire does counterfactually depend on the lightning under the contingency that the arsonist does not drop the match; similarly, the forest fire depends counterfactually on the dropping of the match under the contingency that the lightning does not strike.

Unfortunately, we cannot use this simple solution to treat the case of preemption. We do not want to make Billy's throw the cause of the bottle shattering by considering the contingency that Suzy does not throw. So if our account is to yield the correct verdict in this case, it will be necessary to limit the contingencies that can be considered. The reason that we consider Suzy's throw to be the cause and Billy's throw not to be the cause is that Suzy's rock hit the bottle, while Billy's did not. Somehow the definition of actual cause must capture this obvious intuition.

With this background, we now give the preliminary version of the HP definition of causality. Although the definition is labeled "preliminary", it is quite close to the final definition, which is given in Section 5. The definition is relative to a causal model (and a context); A may be a cause of B in one causal model but not in another. The definition consists of three clauses. The first and third are quite simple; all the work is going on in the second clause.

The types of events that the HP definition allows as actual causes are ones of the form $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$—that is, conjunctions of primitive events; this is often abbreviated as $\vec{X} = \vec{x}$. The events that can be caused are arbitrary Boolean combinations of primitive events. The definition does not allow statements of the form "A or A' is a cause of B", although this could be treated as being equivalent to "either A is a cause of B or A' is a cause of B". On the other hand, statements such as "A is a cause of B or B'" are allowed; this is not equivalent to "either A is a cause of B or A is a cause of B'".

DEFINITION 1. (Actual cause; preliminary version) [Halpern and Pearl 2005] $\vec{X} = \vec{x}$ is an *actual cause of* ϕ *in* $(M, \vec{u})$ if the following three conditions hold:

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \phi$.

AC2. There is a partition of $\mathcal{V}$ (the set of endogenous variables) into two subsets $\vec{Z}$ and $\vec{W}$ with $\vec{X} \subseteq \vec{Z}$, and a setting $\vec{x}'$ and $\vec{w}$ of the variables in $\vec{X}$ and $\vec{W}$,

respectively, such that if $(M, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\phi$.

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\phi$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$, where we abuse notation and write $\vec{W}' \leftarrow \vec{w}$ to denote the assignment where the variables in $\vec{W}'$ get the same values as they would in the assignment $\vec{W} \leftarrow \vec{w}$.

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2.

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of ϕ unless both $\vec{X} = \vec{x}$ and ϕ actually happen. AC3 is a minimality condition, which ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing ϕ in AC2(a) are considered part of a cause; inessential elements are pruned. Without AC3, if dropping a lit match qualified as a cause of the forest fire, then dropping a match and sneezing would also pass the tests of AC1 and AC2. AC3 serves here to strip "sneezing" and other irrelevant, over-specific details from the cause. Clearly, all the "action" in the definition occurs in AC2. We can think of the variables in $\vec{Z}$ as making up the "causal path" from $\vec{X}$ to ϕ, consisting of one or more directed paths from variables in $\vec{X}$ to variables in ϕ. Intuitively, changing the value(s) of some variable(s) in $\vec{X}$ results in changing the value(s) of some variable(s) in $\vec{Z}$, which results in the value(s) of some other variable(s) in $\vec{Z}$ being changed, which finally results in the truth value of ϕ changing. The remaining endogenous variables, the ones in $\vec{W}$, are off to the side, so to speak, but may still have an indirect effect on what happens. AC2(a) is essentially the standard counterfactual definition of causality, but with a twist. If we want to show that $\vec{X} = \vec{x}$ is a cause of ϕ, we must show (in part) that if $\vec{X}$ had a different value, then ϕ would have been false. However, this effect of the value of $\vec{X}$ on the truth value of ϕ may not hold in the actual context; the value of $\vec{W}$ may have to be different to allow this effect to manifest itself. For example, consider the context where both the lightning strikes and the arsonist drops a match in the disjunctive model of the forest fire. Stopping the arsonist from dropping the match will not prevent the forest fire. The counterfactual effect of the arsonist on the forest fire manifests itself only in a situation where the lightning does not strike (i.e., where L is set to 0). AC2(a) is what allows us to call both the lightning and the arsonist causes of the forest fire. Essentially, it ensures that $\vec{X}$ alone suffices to bring about the change from ϕ to $\neg\phi$; setting $\vec{W}$ to $\vec{w}$ merely eliminates possibly spurious side effects that may mask the effect of changing the value of $\vec{X}$. Moreover, when $\vec{X} = \vec{x}$, although the values of variables on the causal path (i.e., the variables $\vec{Z}$) may be perturbed by the change to $\vec{W}$, this perturbation has no impact on the value of ϕ. If $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then z^* is the value of the variable Z in the context $\vec{u}$. We capture the fact that the perturbation has no impact on the value of ϕ by saying that if some variables Z on

the causal path were set to their original values in the context $\vec{u}$, ϕ would still be true, as long as $\vec{X} = \vec{x}$.

EXAMPLE 2. For the forest-fire example, let M be the disjunctive model for the forest fire sketched earlier, with endogenous variables L, ML, and F. We want to show that $L = 1$ is an actual cause of $F = 1$. Clearly $(M, (1, 1)) \models F = 1$ and $(M, (1, 1)) \models L = 1$; in the context (1,1), the lightning strikes and the forest burns down. Thus, AC1 is satisfied. AC3 is trivially satisfied, since $\vec{X}$ consists of only one element, L, so must be minimal. For AC2, take $\vec{Z} = \{L, F\}$ and take $\vec{W} = \{ML\}$, let $x' = 0$, and let $w = 0$. Clearly, $(M, (1, 1)) \models [L \leftarrow 0, ML \leftarrow 0](F \neq 1)$; if the lightning does not strike and the match is not dropped, the forest does not burn down, so AC2(a) is satisfied. To see the effect of the lightning, we must consider the contingency where the match is not dropped; the definition allows us to do that by setting ML to 0. (Note that here setting L and ML to 0 overrides the effects of U; this is critical.) Moreover, $(M, (1, 1)) \models [L \leftarrow 1, ML \leftarrow 0](F = 1)$; if the lightning strikes, then the forest burns down even if the lit match is not dropped, so AC2(b) is satisfied. (Note that since $\vec{Z} = \{L, F\}$, the only subsets of $\vec{Z} - \vec{X}$ are the empty set and the singleton set consisting of just F.)

It is also straightforward to show that the lightning and the dropped match are also causes of the forest fire in the context where $U = (1, 1)$ in the conjunctive model. Again, AC1 and AC3 are trivially satisfied and, again, to show that AC2 holds in the case of lightning we can take $\vec{Z} = \{L, F\}$, $\vec{W} = \{ML\}$, and $x' = 0$, but now we let $w = 1$. In the conjunctive scenario, if there is no lightning, there is no forest fire, while if there is lightning (and the match is dropped) there is a forest fire, so AC2(a) and AC2(b) are satisfied; similarly for the dropped match.

EXAMPLE 3. Now consider the Suzy-Billy example.[4] We get the desired result—that Suzy's throw is a cause, but Billy's is not—but only if we model the story appropriately. Consider first a coarse causal model, with three endogenous variables:

- ST for "Suzy throws", with values 0 (Suzy does not throw) and 1 (she does);
- BT for "Billy throws", with values 0 (he doesn't) and 1 (he does);
- BS for "bottle shatters", with values 0 (it doesn't shatter) and 1 (it does).

(We omit the exogenous variable here; it determines whether Billy and Suzy throw.) Take the formula for BS to be such that the bottle shatters if either Billy or Suzy throw; that is $BS = \max(BT, ST)$. (We assume that Suzy and Billy will not miss if they throw.) BT and ST play symmetric roles in this model; there is nothing to distinguish them. Not surprisingly, both Billy's throw and Suzy's throw are classified as causes of the bottle shattering in this model. The argument is essentially identical to that in the disjunctive model of the forest-fire example in

[4]The discussion of this example is taken almost verbatim from HP.

the context $U = (1, 1)$, where both the lightning and the dropped match are causes of the fire.

The trouble with this model is that it cannot distinguish the case where both rocks hit the bottle simultaneously (in which case it would be reasonable to say that both $ST = 1$ and $BT = 1$ are causes of $BS = 1$) from the case where Suzy's rock hits first. To allow the model to express this distinction, we add two new variables to the model:

- BH for "Billy's rock hits the (intact) bottle", with values 0 (it doesn't) and 1 (it does); and
- SH for "Suzy's rock hits the bottle", again with values 0 and 1.

Now our equations will include:

- $SH = ST$;
- $BH = \min(BT, 1 - SH)$; and
- $BS = \max(SH, BH)$.

Now it is the case that, in the context where both Billy and Suzy throw, $ST = 1$ is a cause of $BS = 1$, but $BT = 1$ is not. To see that $ST = 1$ is a cause, note that, as usual, it is immediate that AC1 and AC3 hold. For AC2, choose $\vec{Z} = \{ST, SH, BH, BS\}$, $\vec{W} = \{BT\}$, and $w = 0$. When BT is set to 0, BS tracks ST: if Suzy throws, the bottle shatters and if she doesn't throw, the bottle does not shatter. To see that $BT = 1$ is *not* a cause of $BS = 1$, we must check that there is no partition $\vec{Z} \cup \vec{W}$ of the endogenous variables that satisfies AC2. Attempting the symmetric choice with $\vec{Z} = \{BT, BH, SH, BS\}$, $\vec{W} = \{ST\}$, and $w = 0$ violates AC2(b). To see this, take $\vec{Z}' = \{BH\}$. In the context where Suzy and Billy both throw, $BH = 0$. If BH is set to 0, the bottle does not shatter if Billy throws and Suzy does not. It is precisely because, in this context, Suzy's throw hits the bottle and Billy's does not that we declare Suzy's throw to be the cause of the bottle shattering. AC2(b) captures that intuition by allowing us to consider the contingency where $BH = 0$, despite the fact that Billy throws. We leave it to the reader to check that no other partition of the endogenous variables satisfies AC2 either.

This example emphasizes an important moral. If we want to argue in a case of preemption that $X = x$ is the cause of ϕ rather than $Y = y$, then there must be a random variable (BH in this case) that takes on different values depending on whether $X = x$ or $Y = y$ is the actual cause. If the model does not contain such a variable, then it will not be possible to determine which one is in fact the cause. This is certainly consistent with intuition and the way we present evidence. If we want to argue (say, in a court of law) that it was A's shot that killed C rather than B's, then we present evidence such as the bullet entering C from the left side (rather

than the right side, which is how it would have entered had B's shot been the lethal one). The side from which the shot entered is the relevant random variable in this case. Note that the random variable may involve temporal evidence (if Y's shot had been the lethal one, the death would have occurred a few seconds later), but it certainly does not have to.

4 The Choice of Variables

A modeler has considerable leeway in choosing which variables to include in a model. Nature does not provide a uniquely correct set of variables. Nonetheless, there are a number of considerations that guide variable selection. While these will not usually suffice to single out one choice of variables, they can provide a framework for the rational evaluation of models, including resources for motivating and defending certain choices of variables, and criticizing others.

The problem of choosing a set of variables for inclusion in a model has many dimensions. One set of issues concerns the question of how many variables to include in a model. If the modeler begins with a set of variables, how can she know whether she should add additional variables to the model? Given that it is always possible to add additional variables, is there a point at which the model contains "enough" variables? Is it ever possible for a model to have "too many" variables? Can the addition of further variables ever do positive harm to a model?

Another set of issues concerns the values of variables. Say that variable X' is a *refinement* of X if, for each value x in the range of X, there is some subset S of the range of X' such that $X = x$ just in case X' is in S. When is it appropriate or desirable to replace a variable with a refinement? Can it ever lead to problems if a variable is too fine-grained? Similarly, are there considerations that would lead us to prefer a model that replaced X with a new variable X'', whose range is a proper subset or superset of the range of X?

Finally, are there constraints on the set of variables in a model over and above those we might impose on individual variables? For instance, can the choice to include a particular variable X within a model require us to include another variable Y, or to exclude a particular variable Z?

While we cannot provide complete answers to all of these questions, we believe a good deal can be said to reduce the arbitrariness of the choice of variables. The most plausible way to motivate guidelines for the selection of variables is to show how inappropriate choices give rise to systems of equations that are inaccurate, misleading, or incomplete in their predictions of observations and interventions. In the next three subsections, we present several examples to show how such considerations can be brought to bear on the problem of variable choice.

4.1 The Number of Variables

We already saw in Example 3 that it is important to choose the variables correctly. Adding more variables can clearly affect whether A is a cause of B. When is it

appropriate or necessary to add further variables to a model?[5] Suppose that we have an infinite sequence of models $M^1, M^2, \ldots$ such that the variables in M^i are $X_0, \ldots, X_{i+1}, Y$, and $M^{i+1}_{X_{i+1} \leftarrow 1} = M_i$ (so that M^{i+1} can be viewed as an extension of M^i). Is it possible that whether $X_0 = 1$ is a cause of $Y = 1$ can alternate as we go through this sequence? This would indicate a certain "instability" in the causality. In this circumstance, a lawyer should certainly be able to argue against using, say, M^7 as a model to show that $X_0 = 1$ is cause of $Y = 1$. On the other hand, if the sequence stabilizes, that is, if there is some k such that for all $i \geq k$, M^i delivers the same verdict on some causal claim of interest, that would provide a strong reason to accept M^k as sufficient.

Compare Example 2 with Example 3. In Example 2, we were able to adequately model the scenario using only three endogenous variables: L, ML, and F. By contrast, in Example 3, the model containing only three endogenous variables, BT, ST, and BS, was inadequate. What is the difference between the two scenarios? One difference we have already mentioned is that there seems to be an important feature of the second scenario that cannot be captured in the three-variable model: Suzy's rock hit the bottle before Billy's did. There is also a significant "topological" difference between the two scenarios. In the forest-fire example, there are two directed paths into the variable F. We could interpolate additional variables along these two paths. We could, for instance, interpolate a variable representing the occurrence of a small brush fire. But doing so would not fundamentally change the causal structure: there would still be just two directed paths into F. In the case of preemption, however, adding the additional variables SH and BH created an additional directed path that was not there before. The three-variable model contained just two directed paths: one from ST to BS, and one from BT to BS. However, once the variables SH and BH were added, there were three directed paths: $\{ST, SH, BS\}$, $\{BT, BH, BS\}$, and $\{ST, SH, BH, BS\}$. The intuition, then, is that adding additional variables to a model will not affect the relations of actual causation that hold in the model unless the addition of those variables changes the "topology" of the model. A more complete mathematical characterization of the conditions under which the verdicts of actual causality remain stable under the addition of further variables strikes us as a worthwhile research project that has not yet been undertaken.

4.2 The Ranges of Variables

Not surprisingly, the set of possible values of a variable must also be chosen appropriately. Consider, for example, a case of "trumping", introduced by Schaffer [2000]. Suppose that a group of soldiers is very well trained, so that they will obey any order given by a superior officer; in the case of conflicting orders, they obey the

[5] Although his model of causality is quite different from ours, Spohn [2003] also considers the effect of adding or removing variables, and discusses how a model with fewer variables should be related to one with more variables.

highest-ranking officer. Both a sergeant and a major issue the order to march, and the soldiers march. Let us put aside the morals that Schaffer attempts to draw from this example (with which we disagree; see [Halpern and Pearl 2005] and [Hitchcock 2010]), and consider only the modeling problem. We will presumably want variables S, M, and A, corresponding to the sergeant's order, the major's order, and the soldiers' action. We might let $S = 1$ represent the sergeant's giving the order to march and $S = 0$ represent the sergeant's giving no order; likewise for M and A. But this would not be adequate. If the only possible order is the order to march, then there is no way to capture the principle that in the case of conflicting orders, the soldiers obey the major. One way to do this is to replace the variables M, S, and A by variables M', S' and A' that take on three possible values. Like M, $M' = 0$ if the major gives no order and $M' = 1$ if the major gives the order to march. But now we allow $M' = 2$, which corresponds to the major giving some other order. S' and A' are defined similarly. We can now write an equation to capture the fact that if $M' = 1$ and $S' = 2$, then the soldiers march, while if $M' = 2$ and $S' = 1$, then the soldiers do not march.

The appropriate set of values of a variable will depend on the other variables in the picture, and the relationship between them. Suppose, for example, that a hapless homeowner comes home from a trip to find that his front door is stuck. If he pushes on it with a normal force then the door will not open. However, if he leans his shoulder against it and gives a solid push, then the door will open. To model this, it suffices to have a variable O with values either 0 or 1, depending on whether the door opens, and a variable P, with values 0 or 1 depending on whether or not the homeowner gives a solid push.

On the other hand, suppose that the homeowner also forgot to disarm the security system, and that the system is very sensitive, so that it will be tripped by any push on the door, regardless of whether the door opens. Let $A = 1$ if the alarm goes off, $A = 0$ otherwise. Now if we try to model the situation with the same variable P, we will not be able to express the dependence of the alarm on the homeowner's push. To deal with both O and A, we need to extend P to a 3-valued variable P', with values 0 if the homeowner does not push the door, 1 if he pushes it with normal force, and 2 if he gives it a solid push.

These considerations parallel issues that arise in philosophical discussions about the metaphysics of "events".[6] Suppose that our homeowner pushed on the door with enough force to open it. Is there just one event, the push, that can be described at various levels of detail, such as a "push" or a "hard push"? This is the view of Davidson [1967]. Or are there rather many different events corresponding to these different descriptions, as argued by Kim [1973] and Lewis [1986b]? And if we take the latter view, which of the many events that occur should be counted as causes of the door's opening? These strike us as pseudoproblems. We believe that questions

[6]This philosophical usage of the word "event" is different from the typical usage of the word in computer science and probability, where an event is just a subset of the state space.

about causality are best addressed by dealing with the methodological problem of constructing a model that correctly describes the effects of interventions in a way that is not misleading or ambiguous.

A slightly different way in which one variable may constrain the values that another may take is by its implicit presuppositions. For example, a counterfactual theory of causation seems to have the somewhat counterintuitive consequence that one's birth is a cause of one's death. This sounds a little odd. If Jones dies suddenly one night, shortly before his 80th birthday, the coroner's inquest is unlikely to list "birth" as among the causes of his death. Typically, when we investigate the causes of death, we are interested in what makes the difference between a person's dying and his surviving. So our model might include a variable D such $D = 1$ holds if Jones dies shortly before his 80th birthday, and $D = 0$ holds if he continues to live. If our model also includes a variable B, taking the value 1 if Jones is born, 0 otherwise, then there simply is no value that D would take if $B = 0$. Both $D = 0$ and $D = 1$ implicitly presuppose that Jones was born (i.e., $B = 1$). Our conclusion is that if we have chosen to include a variable such as D in our model, then we cannot conclude that Jones' birth is a cause of his death!

4.3 Dependence and Independence

Lewis [1986a] added a constraint to his counterfactual theory of causation. In order for event c to be a cause of event e, the two events cannot be logically related. Suppose for instance, that Martha says "hello" loudly. If she had not said "hello", then she certainly could not have said "hello" loudly. But her saying "hello" is not a cause of her saying "hello" loudly. The counterfactual dependence results from a logical, rather than a causal, relationship between the two events.

We must impose a similar constraint upon causal models. Values of different variables should not correspond to events that are logically related. But now, rather than being an *ad hoc* restriction, it has a clear rationale. For suppose that we had a model with variable H_1 and H_2, where H_1 represents "Martha says 'hello'" (i.e., $H_1 = 1$ if Martha says "hello" and $H_1 = 0$ otherwise), and H_2 represents "Martha says 'hello' loudly". The intervention $H_1 = 0 \wedge H_2 = 1$ is meaningless; it is logically impossible for Martha not to say "hello" and to say ''hello" loudly.

We doubt that any careful modeler would choose variables that have logically related values. However, the converse of this principle, that the different values of any particular variable *should* be logically related (in fact, mutually exclusive), is less obvious and equally important. Consider Example 3. While, in the actual context, Billy's rock will hit the bottle just in case Suzy's doesn't, this is not a necessary relationship. Suppose that, instead of using two variables SH and BH, we try to model the scenario with a variable H that takes the value 1 if Suzy's rock hits, and and 0 if Billy's rock hits. The reader can verify that, in this model, there is no contingency such that the bottle's shattering depends upon Suzy's throw. The problem, as we said, is that $H = 0$ and $H = 1$ are *not* mutually exclusive; there are

possible situations in which both rocks hit or neither rock hits the bottle. In particular, this representation does not allow us to consider independent interventions on the rocks hitting the bottle. As the discussion in Example 3 shows, it is precisely such an intervention that is needed to establish that Suzy's throw (and not Billy's) is the actual cause of the bottle shattering.

While these rules are simple in principle, their application is not always transparent.

EXAMPLE 4. Consider cases of "switching", which have been much discussed in the philosophical literature. A train is heading toward the station. An engineer throws a switch, directing the train down the left track, rather than the right track. The tracks re-converge before the station, and the train arrives as scheduled. Was throwing the switch a cause of the train's arrival? HP consider two causal models of this scenario. In the first, there is a random variable S which is 1 if the switch is thrown (so the train goes down the left track) and 0 otherwise. In the second, in addition to S, there are variables LT and RT, indicating whether or not the train goes down the left track and right track, respectively. Note that with the first representation, there is no way to model the train not making it to the arrival point. With the second representation, we have the problem that $LT = 1$ and $RT = 1$ are arguably not independent; the train cannot be on both tracks at once. If we want to model the possibility of one track or another being blocked, we should use, instead of LT and RT, variables LB and RB, which indicate whether the left track or right track, respectively, are blocked. This allows us to represent all the relevant possibilities without running into independence problems. Note that if we have only S as a random variable, then $S = 1$ cannot be a cause of the train arriving; it would have arrived no matter what. With RB in the picture, the preliminary HP definition of actual cause rules that $S = 1$ can be an actual cause of the train's arrival; for example, under the contingency that $RB = 1$, the train does not arrive if $S = 0$. (However, once we extend the definition to include defaults, as we will in the next section, it becomes possible once again to block this conclusion.)

These rules will have particular consequences for how we should represent events that might occur at different times. Consider the following simplification of an example introduced by Bennett [1987], and also considered in HP.

EXAMPLE 5. Suppose that the Careless Camper (CC for short) has plans to go camping on the first weekend in June. He will go camping unless there is a fire in the forest in May. If he goes camping, he will leave a campfire unattended, and there will be a forest fire. Let the variable C take the value 1 if CC goes camping, and 0 otherwise. How should we represent the state of the forest?

There appear to be at least three alternatives. The simplest proposal would be to use a variable F that takes the value 1 if there is a forest fire at some time, and 0 otherwise.[7] But now how are we to represent the dependency relations between F

[7] This is, in effect, how effects have been represented using "neuron diagrams" in late preemption

and C? Since CC will go camping only if there is no fire (in May), we would want to have an equation such as $C = 1 - F$. On the other hand, since there will be a fire (in June) just in case CC goes camping, we will also need $F = C$. This representation is clearly not rich enough, since it does not let us make the clearly relevant distinction between whether the forest fire occurs in May or June. The problem is manifested in the fact that the equations are cyclic, and have no consistent solution.[8]

A second alternative, adopted by Halpern and Pearl [2005, p. 860], would be to use a variable F' that takes the value 0 if there is no fire, 1 if there is a fire in May, and 2 if there is a fire in June. Now how should we write our equations? Since CC will go camping unless there is a fire in May, the equation for C should say that $C = 0$ iff $F' = 1$. And since there will be a fire in June if CC goes camping, the equation for F' should say that $F' = 2$ if $C = 1$ and $F' = 0$ otherwise. These equations are cyclic. Moreover, while they do have a consistent solution, they are highly misleading in what they predict about the effects of interventions. For example, the first equation tells us that intervening to create a forest fire in June would cause CC to go camping in the beginning of June. But this seems to get the causal order backwards!

The third way to model the scenario is to use two separate variables, F_1 and F_2, to represent the state of the forest at separate times. $F_1 = 1$ will represent a fire in May, and $F_1 = 0$ represents no fire in May; $F_2 = 1$ represents a fire in June and $F_2 = 0$ represents no fire in June. Now we can write our equations as $C = 1 - F_1$ and $F_2 = C \times (1 - F_1)$. This representation is free from the defects that plague the other two representations. We have no cycles, and hence there will be a consistent solution for any value of the exogenous variables. Moreover, this model correctly tells us that only an intervention on the state of the forest in May will affect CC's camping plans.

Once again, our discussion of the methodology of modeling parallels certain metaphysical discussions in the philosophy literature. If heavy rains delay the onset of a fire, is it the same fire that would have occurred without the rains, or a different fire? It is hard to see how to gain traction on such an issue by direct metaphysical speculation. By contrast, when we recast the issue as one about what kinds of variables to include in causal models, it is possible to say exactly how the models will mislead you if you make the wrong choice.

cases. See Hitchcock [2007b, pp. 85–88] for discussion.

[8] Careful readers will note the the preemption case of Example 3 is modeled in this way. In that model, BH is a cause of BS, even though it is the earlier shattering of the bottle that prevents Billy's rock from hitting. Halpern and Pearl [2005] note this problem and offer a dynamic model akin to the one recommended below. As it turns out, this does not affect the analysis of the example offered above.

5 Dealing with normality and typicality

While the definition of causality given in Definition 1 works well in many cases, it does not always deliver answers that agree with (most people's) intuition. Consider the following example, taken from Hitchcock [2007a], based on an example due to Hiddleston [2005].

EXAMPLE 6. Assassin is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Victim's coffee. Bodyguard puts antidote in the coffee, which would have neutralized the poison had there been any. Victim drinks the coffee and survives. Is Bodyguard's putting in the antidote a cause of Victim surviving? Most people would say no, but according to the preliminary HP definition, it is. For in the contingency where Assassin puts in the poison, Victim survives iff Bodyguard puts in the antidote.

Example 6 illustrates an even deeper problem with Definition 1. The structural equations for Example 6 are *isomorphic* to those in the forest-fire example, provided that we interpret the variables appropriately. Specifically, take the endogenous variables in Example 6 to be A (for "assassin does not put in poison"), B (for "bodyguard puts in antidote"), and VS (for "victim survives"). Then A, B, and VS satisfy exactly the same equations as L, ML, and F, respectively. In the context where there is lightning and the arsonists drops a lit match, both the lightning and the match are causes of the forest fire, which seems reasonable. But here it does not seem reasonable that Bodyguard's putting in the antidote is a cause. Nevertheless, any definition that just depends on the structural equations is bound to give the same answers in these two examples. (An example illustrating the same phenomenon is given by Hall [2007].) This suggests that there must be more to causality than just the structural equations. And, indeed, the final HP definition of causality allows certain contingencies to be labeled as "unreasonable" or "too farfetched"; these contingencies are then not considered in AC2(a) or AC2(b). As discussed by Halpern [2008], there are problems with the HP account; we present here the approach used in [Halpern 2008] for dealing with these problems, which involves assuming that an agent has, in addition to a theory of causality (as modeled by the structural equations), a theory of "normality" or "typicality". (The need to consider normality was also stressed by Hitchcock [2007a] and Hall [2007], and further explored by Hitchcock and Knobe [2009].) This theory would include statements like "typically, people do not put poison in coffee" and "typically doctors do not treat patients to whom they are not assigned". There are many ways of giving semantics to such typicality statements (e.g., [Adams 1975; Kraus, Lehmann, and Magidor 1990; Spohn 2009]). For definiteness, we use *ranking functions* [Spohn 2009] here.

Take a *world* to be a complete description of the values of all the random variables. we assume that each world has associated with it a *rank*, which is just a natural number or ∞. Intuitively, the higher the rank, the less "normal" or "typical" the

world. A world with a rank of 0 is reasonably normal, one with a rank of 1 is somewhat normal, one with a rank of 2 is quite abnormal, and so on. Given a ranking on worlds, the statement "if p then typically q" is true if in all the worlds of least rank where p is true, q is also true. Thus, in one model where people do not typically put either poison or antidote in coffee, the worlds where neither poison nor antidote is put in the coffee have rank 0, worlds where either poison or antidote is put in the coffee have rank 1, and worlds where both poison and antidote are put in the coffee have rank 2.

Take an *extended causal model* to be a tuple $M = (\mathcal{S}, \mathcal{F}, \kappa)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and κ is a *ranking function* that associates with each world a rank. In an acyclic extended causal model, a context $\vec{u}$ determines a world, denoted $s_{\vec{u}}$. $\vec{X} = \vec{x}$ is a *cause of ϕ in an extended model M and context $\vec{u}$* if $\vec{X} = \vec{x}$ is a cause of ϕ according to Definition 1, except that in AC2(a), there must be a world s such that $\kappa(s) \leq \kappa(s_{\vec{u}})$ and $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}$ is true at s. This can be viewed as a formalization of Kahneman and Miller's [1986] observation that "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it".

This definition deals well with all the problematic examples in the literature. Consider Example 6. Using the ranking described above, Bodyguard is not a cause of Victim's survival because the world that would need to be considered in AC2(a), where Assassin poisons the coffee, is less normal than the actual world, where he does not. We consider just one other example here (see [Halpern 2008] for further discussion).

EXAMPLE 7. Consider the following story, taken from (an early version of) [Hall 2004]: Suppose that Billy is hospitalized with a mild illness on Monday; he is treated and recovers. In the obvious causal model, the doctor's treatment is a cause of Billy's recovery. Moreover, if the doctor does *not* treat Billy on Monday, then the doctor's omission to treat Billy is a cause of Billy's being sick on Tuesday. But now suppose that there are 100 doctors in the hospital. Although only doctor 1 is assigned to Billy (and he forgot to give medication), in principle, any of the other 99 doctors could have given Billy his medication. Is the nontreatment by doctors 2–100 also a cause of Billy's being sick on Tuesday?

Suppose that in fact the hospital has 100 doctors and there are variables $A_1, \ldots, A_{100}$ and $T_1, \ldots, T_{100}$ in the causal model, where $A_i = 1$ if doctor i is assigned to treat Billy and $A_i = 0$ if he is not, and $T_i = 1$ if doctor i actually treats Billy on Monday, and $T_i = 0$ if he does not. Doctor 1 is assigned to treat Billy; the others are not. However, in fact, no doctor treats Billy. Further assume that, typically, no doctor is assigned to a given patient; if doctor i is not assigned to treat Billy, then typically doctor i does not treat Billy; and if doctor i *is* assigned to Billy, then typically doctor i treats Billy. We can capture this in an extended causal model where the world where no doctor is assigned to Billy and no doctor

treats him has rank 0; the 100 worlds where exactly one doctor is assigned to Billy, and that doctor treats him, have rank 1; the 100 worlds where exactly one doctor is assigned to Billy and no one treats him have rank 2; and the 100×99 worlds where exactly one doctor is assigned to Billy but some other doctor treats him have rank 3. (The ranking given to other worlds is irrelevant.) In this extended model, in the context where doctor i is assigned to Billy but no one treats him, i is the cause of Billy's sickness (the world where i treats Billy has lower rank than the world where i is assigned to Billy but no one treats him), but no other doctor is a cause of Billy's sickness. Moreover, in the context where i is assigned to Billy and treats him, then i is the cause of Billy's recovery (for AC2(a), consider the world where no doctor is assigned to Billy and none treat him).

Adding a normality theory to the model gives the HP account of actual causation greater flexibility to deal with these kinds of cases. This raises the worry, however, that this gives the modeler too much flexibility. After all, the modeler can now render any claim that A is an actual cause of B false, simply by choosing a normality order that assigns the actual world $s_{\vec{u}}$ a lower rank than any world s needed to satisfy AC2. Thus, the introduction of normality exacerbates the problem of motivating and defending a particular choice of model. Fortunately, the literature on the psychology of counterfactual reasoning and causal judgment goes some way toward enumerating the sorts of factors that constitute normality. (See, for example, [Alicke 1992; Cushman 2009; Cushman, Knobe, and Sinnott-Armstrong 2008; Hitchcock and Knobe 2009; Kahneman and Miller 1986; Knobe and Fraser 2008; Kahneman and Tversky 1982; Mandel, Hilton, and Catellani 1985; Roese 1997].) These factors include the following:

- Statistical norms concern what happens most often, or with the greatest frequency. Kahneman and Tversky [1982] gave subjects a story in which Mr. Jones usually leaves work at 5:30, but occasionally leaves early to run errands. Thus, a 5:30 departure is (statistically) "normal", and an earlier departure "abnormal". This difference affected which alternate possibilities subjects were willing to consider when reflecting on the causes of an accident in which Mr. Jones was involved.

- Norms can involve moral judgments. Cushman, Knobe, and Sinnott-Armstrong [2008] showed that people with different views about the morality of abortion have different views about the abnormality of insufficient care for a fetus, and this can lead them to make different judgments about the cause of a miscarriage.

- Policies adopted by social institutions can also be norms. For instance, Knobe and Fraser [2008] presented subjects with a hypothetical situation in which a department had implemented a policy allowing administrative assistants to take pens from the department office, but prohibiting faculty from doing

so. Subjects were more likely to attribute causality to a professor's taking a pen than to an assistant's taking one, even when the situation was otherwise similar.

- There can also be norms of "proper functioning" governing the operations of biological organs or mechanical parts: there are certain ways that hearts and spark plugs are "supposed" to operate. Hitchcock and Knobe [2009] show that these kinds of norms can also affect causal judgments.

The law suggests a variety of principles for determining the norms that are used in the evaluation of actual causation. In criminal law, norms are determined by direct legislation. For example, if there are legal standards for the strength of seat belts in an automobile, a seat belt that did not meet this standard could be judged a cause of a traffic fatality. By contrast, if a seat belt complied with the legal standard, but nonetheless broke because of the extreme forces it was subjected to during a particular accident, the fatality would be blamed on the circumstances of the accident, rather than the seat belt. In such a case, the manufacturers of the seat belt would not be guilty of criminal negligence. In contract law, compliance with the terms of a contract has the force of a norm. In tort law, actions are often judged against the standard of "the reasonable person". For instance, if a bystander was harmed when a pedestrian who was legally crossing the street suddenly jumped out of the way of an oncoming car, the pedestrian would not be held liable for damages to the bystander, since he acted as the hypothetical "reasonable person" would have done in similar circumstances. (See, for example, [Hart and Honoré 1985, pp. 142ff.] for discussion.) There are also a number of circumstances in which deliberate malicious acts of third parties are considered to be "abnormal" interventions, and affect the assessment of causation. (See, for example, [Hart and Honoré 1985, pp. 68ff.].)

As with the choice of variables, we do not expect that these considerations will always suffice to pick out a uniquely correct theory of normality for a causal model. They do, however, provide resources for a rational critique of models.

6 Conclusion

As HP stress, causality is relative to a model. That makes it particularly important to justify whatever model is chosen, and to enunciate principles for what makes a reasonable causal model. We have taken some preliminary steps in investigating this issue with regard to the choice of variables and the choice of defaults. However, we hope that we have convinced the reader that far more needs to be done if causal models are actually going to be used in applications.

Acknowledgments: We thank Wolfgang Spohn for useful comments. Joseph Halpern was supported in part by NSF grants IIS-0534064 and IIS-0812045, and by AFOSR grants FA9550-08-1-0438 and FA9550-05-1-0055.

References

Adams, E. (1975). *The Logic of Conditionals*. Dordrecht, Netherlands: Reidel.

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology 63*, 368–378.

Bennett, J. (1987). Event causation: the counterfactual analysis. In *Philosophical Perspectives, Vol. 1, Metaphysics*, pp. 367–386. Atascadero, CA: Ridgeview Publishing Company.

Cushman, F. (2009). The role of moral judgment in causal and intentional attribution: What we say or how we think?". Unpublished manuscript.

Cushman, F., J. Knobe, and W. Sinnott-Armstrong (2008). Moral appraisals affect doing/allowing judgments. *Cognition 108*(1), 281–289.

Davidson, D. (1967). Causal relations. *Journal of Philosophy LXIV*(21), 691–703.

Glymour, C. and F. Wimberly (2007). Actual causes and thought experiments. In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*, pp. 43–67. Cambridge, MA: MIT Press.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica 40*(6), 979–1001.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies 132*, 109–136.

Halpern, J. Y. (2008). Defaults and normality in causal structures. In *Principles of Knowledge Representation and Reasoning: Proc. Eleventh International Conference (KR '08)*, pp. 198–208.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science 56*(4), 843–887.

Hansson, R. N. (1958). *Patterns of Discovery*. Cambridge, U.K.: Cambridge University Press.

Hart, H. L. A. and T. Honoré (1985). *Causation in the Law* (second ed.). Oxford, U.K.: Oxford University Press.

Hiddleston, E. (2005). Causal powers. *British Journal for Philosophy of Science 56*, 27–59.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy XCVIII*(6), 273–299.

Hitchcock, C. (2007a). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review 116*, 495–532.

Hitchcock, C. (2007b). What's wrong with neuron diagrams? In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*, pp. 69–92. Cambridge, MA: MIT Press.

Hitchcock, C. (2010). Trumping and contrastive causation. *Synthese*. To appear.

Hitchcock, C. and J. Knobe (2009). Cause and norm. *Journal of Philosophy*. To appear.

Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Reprinted by Open Court Press, LaSalle, IL, 1958.

Kahneman, D. and D. T. Miller (1986). Norm theory: comparing reality to its alternatives. *Psychological Review 94*(2), 136–153.

Kahneman, D. and A. Tversky (1982). The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment Under Incertainty: Heuristics and Biases*, pp. 201–210. Cambridge/New York: Cambridge University Press.

Kim, J. (1973). Causes, nomic subsumption, and the concept of event. *Journal of Philosophy LXX*, 217–236.

Knobe, J. and B. Fraser (2008). Causal judgment and moral judgment: two experiments. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447. Cambridge, MA: MIT Press.

Kraus, S., D. Lehmann, and M. Magidor (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence 44*, 167–207.

Lewis, D. (1973a). Causation. *Journal of Philosophy 70*, 113–126. Reprinted with added "Postscripts" in D. Lewis, *Philosophical Papers*, Volume II, Oxford University Press, 1986, pp. 159–213.

Lewis, D. (1986a). Causation. In *Philosophical Papers*, Volume II, pp. 159–213. New York: Oxford University Press. The original version of this paper, without numerous postscripts, appeared in the *Journal of Philosophy* **70**, 1973, pp. 113–126.

Lewis, D. (1986b). Events. In *Philosophical Papers*, Volume II, pp. 241–270. New York: Oxford University Press.

Lewis, D. K. (1973b). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Mandel, D. R., D. J. Hilton, and P. Catellani (Eds.) (1985). *The Psychology of Counterfactual Thinking*. New York: Routledge.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin CXXI*, 133–148.

Schaffer, J. (2000). Trumping preemption. *Journal of Philosophy XCVII*(4), 165–181. Reprinted in J. Collins and N. Hall and L. A. Paul (eds.), *Causation and Counterfactuals*, MIT Press, 2002.

Spohn, W. (2003). Dependency equilibria and the causal structure of decision and game situations. In *Homo Oeconomicus XX*, pp. 195–255.

Spohn, W. (2009). A survey of ranking theory. In F. Huber and C. Schmidt-Petri (Eds.), *Degrees of Belief. An Anthology*, pp. 185–228. Dordrecht, Netherlands: Springer.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford, U.K.: Oxford University Press.

24

From C-Believed Propositions to the Causal Calculator

Vladimir Lifschitz

1 Introduction

Default rules, unlike inference rules of classical logic, allow us to derive a new conclusion only when it does not conflict with the other available information. The best known example is the so-called commonsense law of inertia: in the absence of information to the contrary, properties of the world can be presumed to be the same as they were in the past. Making the idea of commonsense inertia precise is known as the frame problem [Shanahan 1997]. Default reasoning is nonmonotonic, in the sense that we may be forced to retract a conclusion derived using a default when additional information becomes available.

The idea of a default first attracted the attention of AI researchers in the 1970s. Developing a formal semantics of defaults turned out to be a difficult task. For instance, the attempt to describe commonsense inertia in terms of circumscription outlined in [McCarthy 1986] was unsatisfactory, as we learned from the Yale Shooting example [Hanks and McDermott 1987].

In this note, we trace the line of work on the semantics of defaults that started with Judea Pearl's 1988 paper on the difference between "E-believed" and "C-believed" propositions. That paper has led other researchers first to the invention of several theories of nonmonotonic causal reasoning, then to designing action languages $\mathcal{C}$ and $\mathcal{C}+$, and then to the creation of the Causal Calculator—a software system for automated reasoning about action and change.

2 Starting Point: Labels E and C

The paper *Embracing Causality in Default Reasoning* [Pearl 1988] begins with the observation that

> almost every default rule falls into one of two categories: expectation-evoking or explanation-evoking. The former describes association among events in the outside world (e.g., fire is typically accompanied by smoke); the latter describes how we reason about the world (e.g., smoke normally suggests fire).

Thus the rule `fire` $\Rightarrow$ `smoke` is an expectation-evoking, or "causal" default; the rule `smoke` $\Rightarrow$ `fire` is explanation-evoking, or "evidential." To take another example,

(1) $\texttt{rained} \Rightarrow \texttt{grass\_wet}$

is a causal default;

(2) $\texttt{grass\_wet} \Rightarrow \texttt{sprinkler\_on}$

is an evidential default.

To discuss the distinction between properties of causal and evidential defaults, Pearl labels believed propositions by distinguishing symbols C and E. A proposition P is E-believed, written $E(P)$, if it is a direct consequence of some evidential rule. Otherwise, if P can be established as a direct consequence of only causal rules, it is said to be C-believed, written $C(P)$. The labels are used to prevent certain types of inference chains; in particular, C-believed propositions are prevented in Pearl's paper from triggering evidential defaults. For example, both causal rule (1) and evidential rule (2) are reasonable, but using them to infer `sprinkler_on` from `rained` is not.

We will see that the idea of using the distinguishing symbols C and E had a significant effect on the study of commonsense reasoning over the next twenty years.

3 "Explained" as a Modal Operator

The story continues with Hector Geffner's proposal to turn the label C into a modal operator and to treat Pearl's causal rules as formulas of modal logic. A formula F is considered "explained" if the formula $\mathsf{C}F$ holds.

> A rule such as "rain causes the grass to be wet" may thus be expressed as a sentence
>
> $$\texttt{rain} \rightarrow \mathsf{C}\,\texttt{grass\_wet},$$
>
> which can then be read as saying that if `rain` is true, `grass_wet` is explained [Geffner 1990].

The paper defined, for a set of axioms of this kind, which propositions are "causally entailed" by it.

Geffner showed how this modal language can be used to describe effects of actions. We can express that $e(x)$ is an effect of an action $a(x)$ with precondition $p(x)$ by the axiom

(3) $p(x)_t \wedge a(x)_t \rightarrow \mathsf{C}e(x)_{t+1},$

where $p(x)_t$ expresses that fluent $p(x)$ holds at time t, and $e(x)_{t+1}$ is understood in a similar way; $a(x)_t$ expresses that action $a(x)$ is executed between times t and $t+1$.

Such axioms explain the value of a fluent at some point in time ($t+1$ in the consequent of the implication) in terms of the past (t in the antecedent). Geffner gives also an example of explaining the value of a fluent in terms of the values of other fluents at the same point in time: if all ducts are blocked at time t, that causes

the room to be stuffy at time t. Such "static" causal dependencies are instrumental when actions with indirect effects are involved. For instance, blocking a duct can indirectly cause the room to become stuffy. We will see another example of this kind in the next section.

4 Predicate "Caused"

Fangzhen Lin showed a few years later that the intuitions explored by Pearl and Geffner can be made precise without introducing a new nonmonotonic semantics. Circumscription [McCarthy 1986] will do if we employ, instead of the modal operator C, a new predicate.

> Technically, we introduce a new ternary predicate *Caused* into the situation calculus: $Caused(p, v, s)$ if the proposition p is caused (by something unspecified) to have the truth value v in the state s [Lin 1995].

The counterpart of formula (3) in this language is

(4) $p(x, s) \rightarrow Caused(e(x), true, do(a(x), s))$.

Lin acknowledges his intellectual debt to [Pearl 1988] by noting that his approach echoes the theme of Pearl's paper—the need for a primitive notion of causality in default reasoning.

The proposal to circumscribe *Caused* was a major event in the history of research on the use of circumscription for solving the frame problem. As we mentioned before, the original method [McCarthy 1986] turned out to be unsatisfactory; the improvement described in [Haugh 1987; Lifschitz 1987] is only applicable when actions have no indirect effects. The method of [Lin 1995] is free of this limitation. The main example of that paper is a suitcase with two locks and a spring loaded mechanism that opens the suitcase instantaneously when both locks are in the up position; opening the suitcase may thus become an indirect effect of toggling a switch. The static causal relationship between the fluents $up(l)$ and *open* is expressed in Lin's language by the axiom

(5) $up(L1, s) \wedge up(L2, s) \rightarrow Caused(open, true, s)$.

5 Principle of Universal Causation

Yet another important modification of Geffner's theory was proposed in [McCain and Turner 1997]. That approach was originally limited to formulas of the form

$$F \rightarrow \mathsf{C}G,$$

where F and G do not contain C. (Such formulas are particularly useful; for instance, (3) has this form.) The authors wrote such a formula as

(6) $F \Rightarrow G$,

so that the thick arrow $\Rightarrow$ represented in their paper a combination of material implication $\rightarrow$ with the modal operator C. In [Turner 1999], that method was extended to the full language of [Geffner 1990].

The key idea of this theory of causal knowledge is described in [McCain and Turner 1997] as follows:

> Intuitively, in a causally possible world history every fact that is caused obtains. We assume in addition the *principle of universal causation*, according to which—in a causally possible world history—every fact that obtains is caused. In sum, we say that a world history is causally possible if exactly the facts that obtain in it are caused in it.

The authors note that the principle of universal causation represents a strong philosophical commitment that is rewarded by the mathematical simplicity of the nonmonotonic semantics that it leads to. The definition of their semantics is indeed surprisingly simple, or at least short. They note also that in applications this strong commitment can be easily relaxed.

The extension of [McCain and Turner 1997] described in [Giunchiglia, Lee, Lifschitz, McCain, and Turner 2004] allows F and G in (6) to be slightly more general than propositional formulas, which is convenient when non-Boolean fluents are involved. In the language of that paper we can write, for instance,

(7) $\quad a_t \Rightarrow f_{t+1} = v$

to express that executing action a causes fluent f to take value v.

6 Action Descriptions

An action description is a formal expression representing a transition system—a directed graph such that its vertices can be interpreted as states of the world, with edges corresponding to the transitions caused by the execution of actions. In [Giunchiglia and Lifschitz 1998], the nonmonotonic causal logic from [McCain and Turner 1997] was used to define an action description language, called $\mathcal{C}$. The language $\mathcal{C}+$ [Giunchiglia, Lee, Lifschitz, McCain, and Turner 2004] is an extension of $\mathcal{C}$ that accomodates non-Boolean fluents and is also more expressive in some other ways.

The distinguishing syntactic feature of action description languages is that they do not involve symbols for time instants. For example, the counterpart of (7) in $\mathcal{C}+$ is

$$a \textbf{ causes } f = v.$$

The $\mathcal{C}+$ keyword **causes** implicitly indicates a shift from the time instant t when the execution of action a begins to the next time instant $t+1$ when fluent f is evaluated. This keyword represents a combination of three elements: material implication, the Pearl-Geffner causal operator, and time shift.

7 The Causal Calculator

Literal completion, defined in [McCain and Turner 1997], is a modification of the completion process familiar from logic programming [Clark 1978]. It is applicable to any finite set T of causal laws (6) whose heads G are literals, and produces a set of propositional formulas such that its models in the sense of propositional logic are identical to the models of T in the sense of the McCain-Turner causal logic. Literal completion can be used to reduce some computational problems involving $\mathcal{C}$ action descriptions to the propositional satisfiability problem.

This idea is used in the design of the Causal Calculator (CCALC)—a software system that reasons about actions in domains described in a subset of $\mathcal{C}$ [McCain 1997]. CCALC performs search by invoking a SAT solver in the spirit of the "planning as satisfiability" method of [Kautz and Selman 1992]. Version 2 of CCALC [Lee 2005] extends it to $\mathcal{C}+$ action descriptions.

The Causal Calculator has been succesfully applied to several challenge problems in the theory of commonsense reasoning [Lifschitz, McCain, Remolina, and Tacchella 2000], [Lifschitz 2000], [Akman, Erdoğan, Lee, Lifschitz, and Turner 2004]. More recently, it was used for the executable specification of norm-governed computational societies [Artikis, Sergot, and Pitt 2009] and for the automatic analysis of business processes under authorization constraints [Armando, Giunchiglia, and Ponta 2009].

8 Conclusion

As we have seen, Judea Pearl's idea of labeling the propositions that are derived using causal rules has suggested to Geffner, Lin and others that the condition

G is caused (by something unspecified) if F holds

can be sometimes used as an approximation to

G is caused by F.

Eliminating the binary "is caused by" in favor of the unary "is caused" turned out to be a remarkably useful technical device.

9 Acknowledgements

Thanks to Selim Erdoğan, Hector Geffner, and Joohyung Lee for comments on a draft of this note. This work was partially supported by the National Science Foundation under Grant IIS-0712113.

References

Akman, V., S. Erdoğan, J. Lee, V. Lifschitz, and H. Turner (2004). Representing the Zoo World and the Traffic World in the language of the Causal Calculator. *Artificial Intelligence 153(1–2)*, 105–140.

Armando, A., E. Giunchiglia, and S. E. Ponta (2009). Formal specification and automatic analysis of business processes under authorization constraints: an action-based approach. In *Proceedings of the 6th International Conference on Trust, Privacy and Security in Digital Business (TrustBus'09)*.

Artikis, A., M. Sergot, and J. Pitt (2009). Specifying norm-governed computational societies. *ACM Transactions on Computational Logic 9*(1).

Clark, K. (1978). Negation as failure. In H. Gallaire and J. Minker (Eds.), *Logic and Data Bases*, pp. 293–322. New York: Plenum Press.

Geffner, H. (1990). Causal theories for nonmonotonic reasoning. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, pp. 524–530. AAAI Press.

Giunchiglia, E., J. Lee, V. Lifschitz, N. McCain, and H. Turner (2004). Nonmonotonic causal theories. *Artificial Intelligence 153(1–2)*, 49–104.

Giunchiglia, E. and V. Lifschitz (1998). An action language based on causal explanation: Preliminary report. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, pp. 623–630. AAAI Press.

Hanks, S. and D. McDermott (1987). Nonmonotonic logic and temporal projection. *Artificial Intelligence 33*(3), 379–412.

Haugh, B. (1987). Simple causal minimizations for temporal persistence and projection. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, pp. 218–223.

Kautz, H. and B. Selman (1992). Planning as satisfiability. In *Proceedings of European Conference on Artificial Intelligence (ECAI)*, pp. 359–363.

Lee, J. (2005). *Automated Reasoning about Actions.*[1] Ph.D. thesis, University of Texas at Austin.

Lifschitz, V. (1987). Formal theories of action (preliminary report). In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 966–972.

Lifschitz, V. (2000). Missionaries and cannibals in the Causal Calculator. In *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pp. 85–96.

Lifschitz, V., N. McCain, E. Remolina, and A. Tacchella (2000). Getting to the airport: The oldest planning problem in AI. In J. Minker (Ed.), *Logic-Based Artificial Intelligence*, pp. 147–165. Kluwer.

Lin, F. (1995). Embracing causality in specifying the indirect effects of actions. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1985–1991.

[1] `http://peace.eas.asu.edu/joolee/papers/dissertation.pdf` .

McCain, N. (1997). *Causality in Commonsense Reasoning about Actions.*[2] Ph.D. thesis, University of Texas at Austin.

McCain, N. and H. Turner (1997). Causal theories of action and change. In *Proceedings of National Conference on Artificial Intelligence (AAAI)*, pp. 460–465.

McCarthy, J. (1986). Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence 26*(3), 89–116.

Pearl, J. (1988). Embracing causality in default reasoning (research note). *Artificial Intelligence 35(2)*, 259–271.

Shanahan, M. (1997). *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. MIT Press.

Turner, H. (1999). A logic of universal causation. *Artificial Intelligence 113*, 87–123.

[2] `ftp://ftp.cs.utexas.edu/pub/techreports/tr97-25.ps.gz` .

25

Analysis of the Binary Instrumental Variable Model

Thomas S. Richardson and James M. Robins

1 Introduction

Pearl's seminal work on instrumental variables [Chickering and Pearl 1996; Balke and Pearl 1997] for discrete data represented a leap forwards in terms of understanding: Pearl showed that, contrary to what many had supposed based on linear models, in the discrete case the assumption that a variable was an instrument could be subjected to empirical test. In addition, Pearl improved on earlier bounds [Robins 1989] for the average causal effect (ACE) in the absence of any monotonicity assumptions. Pearl's approach was also innovative insofar as he employed a computer algebra system to derive analytic expressions for the upper and lower bounds.

In this paper we build on and extend Pearl's work in two ways. First we show the geometry underlying Pearl's bounds. As a consequence we are able to derive bounds on the average causal effect for all four compliance types. Our analysis also makes it possible to perform a sensitivity analysis using the distribution over compliance types. Second our analysis provides a clear geometric picture of the instrumental inequalities, and allows us to isolate the counterfactual assumptions necessary for deriving these tests. This may be seen as analogous to the geometric study of models for two-way tables [Fienberg and Gilbert 1970; Erosheva 2005]. Among other things this allows us to clarify which are the alternative hypotheses against which Pearl's test has power. We also relate these tests to recent work of Pearl's on bounding direct effects [Cai, Kuroki, Pearl, and Tian 2008].

2 Background

We consider three binary variables, X, Y and Z. Where:

Z is the instrument, presumed to be randomized e.g. the assigned treatment;

X is the treatment received;

Y is the response.

For X and Z, we will use 0 to indicate placebo, and 1 to indicate drug. For Y we take 1 to indicate a desirable outcome, such as survival. X_z is the treatment a

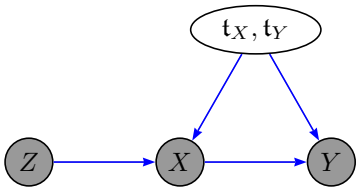


Figure 1. Graphical representation of the IV model given by assumptions (1) and (2).The shaded nodes are observed.

patient would receive if assigned to $Z = z$. We follow convention by referring to the four *compliance* types:

$X_{z=0}$	$X_{z=1}$	Compliance Type	
0	0	Never Taker	NT
0	1	Complier	CO
1	0	Defier	DE
1	1	Always Taker	AT

Since we suppose the counterfactuals are well-defined, if $Z = z$ then $X = X_z$. Similarly we consider counterfactuals Y_{xz} for Y. Except where explicitly noted we will make the exclusion restrictions:

$$Y_{x=0,z=0} = Y_{x=0,z=1} \qquad Y_{x=1,z=0} = Y_{x=1,z=1} \tag{1}$$

for each patient, so that a patient's outcome only depends on treatment assigned via the treatment received. One consequence of the analysis below is that these equations may be tested separately. We may thus similarly enumerate four types of patient in terms of their *response* to received treatment:

$Y_{x=0}$	$Y_{x=1}$	Response Type	
0	0	Never Recover	*NR*
0	1	Helped	*HE*
1	0	Hurt	*HU*
1	1	Always Recover	*AR*

As before, it is implicit in our notation that if $X = x$, then $Y_x = Y$; this is referred to as the 'consistency assumption' (or axiom) by Pearl among others. In what follows we will use $\mathfrak{t}_X$ to denote a generic compliance type in the set $\mathbb{D}_X$, and $\mathfrak{t}_Y$ to denote a generic response type in the set $\mathbb{D}_Y$. We thus have 16 patient types:

$$\langle \mathfrak{t}_X, \mathfrak{t}_Y \rangle \in \{\text{NT, CO, DE, AT}\} \times \{\textit{NR, HE, HU, AR}\} \equiv \mathbb{D}_X \times \mathbb{D}_Y \equiv \mathbb{D}.$$

(Here and elsewhere we use angle brackets $\langle \mathfrak{t}_X, \mathfrak{t}_Y \rangle$ to indicate an ordered pair.) Let $\pi_{\mathfrak{t}_X} \equiv p(\mathfrak{t}_X)$ denote the marginal probability of a given compliance type $\mathfrak{t}_X \in \mathbb{D}_X$,

and let

$$\pi_X \equiv \{\pi_{\mathfrak{t}_X} \mid \mathfrak{t}_X \in \mathbb{D}_X\}$$

denote a marginal distribution on $\mathbb{D}_X$. Similarly we use $\pi_{\mathfrak{t}_Y|\mathfrak{t}_X} \equiv p(\mathfrak{t}_Y \mid \mathfrak{t}_X)$ to denote the probability of a given response type within the sub-population of individuals of compliance type $\mathfrak{t}_X$, and $\pi_{Y|X}$ to indicate a specification of all these conditional probabilities:

$$\pi_{Y|X} \equiv \{\pi_{\mathfrak{t}_Y|\mathfrak{t}_X} \mid \mathfrak{t}_X \in \mathbb{D}_X, \mathfrak{t}_Y \in \mathbb{D}_Y\}.$$

We will use π to indicate a joint distribution $p(\mathfrak{t}_X, \mathfrak{t}_Y)$ on $\mathbb{D}$.

Except where explicitly noted we will make the randomization assumption that the distribution of types $\langle \mathfrak{t}_X, \mathfrak{t}_Y \rangle$ is the same in both arms:

$$Z \perp\!\!\!\perp \{X_{z=0}, X_{z=1}, Y_{x=0}, Y_{x=1}\}. \tag{2}$$

A graph corresponding to the model given by (1) and (2) is shown in Figure 1.

Notation

In places we will make use of the following compact notation for probability distributions:

$$\begin{aligned} p_{y_k|x_j z_i} &\equiv p(Y = k \mid X = j, Z = i), \\ p_{x_j|z_i} &\equiv p(X = j \mid Z = i), \\ p_{y_k x_j|z_i} &\equiv p(Y = k, X = j \mid Z = i). \end{aligned}$$

There are several simple geometric constructions that we will use repeatedly. In consequence we introduce these in a generic setting.

2.1 Joints compatible with fixed margins

Consider a bivariate random variable $U = \langle U_1, U_2 \rangle \in \{0,1\} \times \{0,1\}$. Now for fixed $c_1, c_2 \in [0,1]$ consider the set

$$\mathcal{P}_{c_1,c_2} = \left\{ p \;\middle|\; \sum_{u_2} p(1,u_2) = c_1 \;;\; \sum_{u_1} p(u_1,1) = c_2 \right\}$$

in other words, $\mathcal{P}_{c_1,c_2}$ is the set of joint distributions on U compatible with fixed margins $p(U_i = 1) = c_i$, $i = 1, 2$.

It is not hard to see that $\mathcal{P}_{c_1,c_2}$ is a one-dimensional subset (line segment) of the 3-dimensional simplex of distributions for U. We may describe it explicitly as follows:

$$\left\{ \begin{array}{lcll} p(1,1) &=& t & \\ p(1,0) &=& c_1 - t & \\ p(0,1) &=& c_2 - t & t \in \left[\max\{0, (c_1+c_2)-1\}, \min\{c_1, c_2\}\right] \\ p(0,0) &=& 1 - c_1 - c_2 + t & \end{array} \right\}. \tag{3}$$

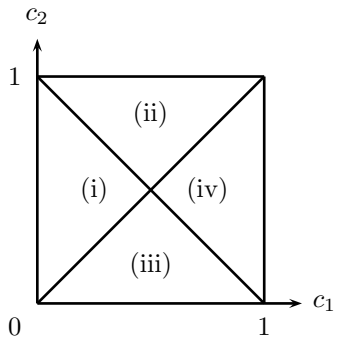


Figure 2. The four regions corresponding to different supports for t in (3); see Table 1.

See also [Pearl 2000] Theorem 9.2.10. The range of t, or equivalently the support for $p(1,1)$, is one of four intervals, as shown in Table 1. These cases correspond to

	$c_1 \leq 1 - c_2$		$c_1 \geq 1 - c_2$	
$c_1 \leq c_2$	(i)	$t \in [0, c_1]$	(ii)	$t \in [c_1 + c_2 - 1, c_1]$
$c_1 \geq c_2$	(iii)	$t \in [0, c_2]$	(iv)	$t \in [c_1 + c_2 - 1, c_2]$

Table 1. The support for t in (3) in each of the four cases relating c_1 and c_2.

the four regions show in Figure 2.

Finally, we note that since for $c_1, c_2 \in [0,1]$, $\max\{0, (c_1 + c_2) - 1\} \leq \min\{c_1, c_2\}$, it follows that $\{\langle c_1, c_2\rangle \mid \mathcal{P}_{c_1,c_2} \neq \emptyset\} = [0,1]^2$. Thus for every pair of values $\langle c_1, c_2\rangle$ there exists a joint distribution $p(U_1, U_2)$ for which $p(U_i = 1) = c_i$, $i = 1, 2$.

2.2 Two quantities with a specified average

We now consider the set:

$$\mathcal{Q}_{c,\alpha} = \{\langle u, v\rangle \mid \alpha u + (1-\alpha)v = c,\ u, v \in [0,1]\}$$

where $c, \alpha \in [0,1]$. In words, $\mathcal{Q}_{c,\alpha}$ is the set of pairs of values $\langle u, v\rangle$ in $[0,1]$ which are such that the weighted average $\alpha u + (1-\alpha)v$ is c.

It is simple to see that this describes a line segment in the unit square. Further consideration shows that for any value of $\alpha \in [0,1]$, the segment will pass through the point $\langle c, c\rangle$ and will be contained within the union of two rectangles:

$$([c,1] \times [0,c]) \cup ([0,c] \times [1,c]).$$

The slope of the line is negative for $\alpha \in (0,1)$. For $\alpha \in (0,1)$ the line segment may

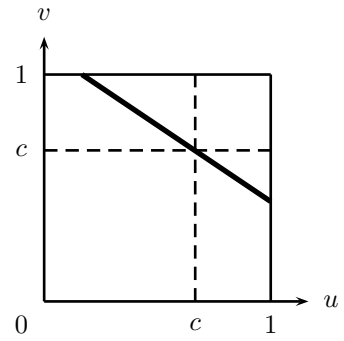


Figure 3. Illustration of $\mathcal{Q}_{c,\alpha}$.

be parametrized as follows:

$$\left\{ \begin{array}{lcl} u & = & (c - t(1-\alpha))/\alpha, \\ v & = & t, \end{array} \quad t \in \left[\max\left(0, \frac{c-\alpha}{1-\alpha}\right), \min\left(\frac{c}{1-\alpha}, 1\right) \right] \right\}.$$

The left and right endpoints of the line segment are:

$$\langle u, v \rangle = \left\langle \max\big(0, 1 + (c-1)/\alpha\big)\,,\ \min\big(c/(1-\alpha), 1\big) \right\rangle$$

and

$$\langle u, v \rangle = \left\langle \min\big(c/\alpha, 1\big)\,,\ \max\big(0, (c-\alpha)/(1-\alpha)\big) \right\rangle$$

respectively. See Figure 3.

2.3 Three quantities with two averages specified

We now extend the discussion in the previous section to consider the set:

$$\begin{aligned} \mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} = \{ \langle u, v, w \rangle \mid\ & \alpha_1 u + (1-\alpha_1) w = c_1, \\ & \alpha_2 v + (1-\alpha_2) w = c_2,\ u, v, w \in [0,1] \}. \end{aligned}$$

In words, this consists of the set of triples $\langle u, v, w \rangle \in [0,1]^3$ for which pre-specified averages of u and w (via α_1), and v and w (via α_2) are equal to c_1 and c_2 respectively.

If this set is not empty, it is a line segment in $[0,1]^3$ obtained by the intersection of two rectangles:

$$\Big(\{ \langle u, w \rangle \in \mathcal{Q}_{c_1,\alpha_1} \} \times \{ v \in [0,1] \} \Big) \cap \Big(\{ \langle v, w \rangle \in \mathcal{Q}_{c_2,\alpha_2} \} \times \{ u \in [0,1] \} \Big); \qquad (4)$$

see Figures 4 and 5. For $\alpha_1, \alpha_2 \in (0,1)$ we may parametrize the line segment (4) as follows:

$$\left\{ \begin{array}{lcll} u & = & (c_1 - t(1-\alpha_1))/\alpha_1, & \\ v & = & (c_2 - t(1-\alpha_2))/\alpha_2, & t \in [t_l, t_u] \\ w & = & t, & \end{array} \right\},$$

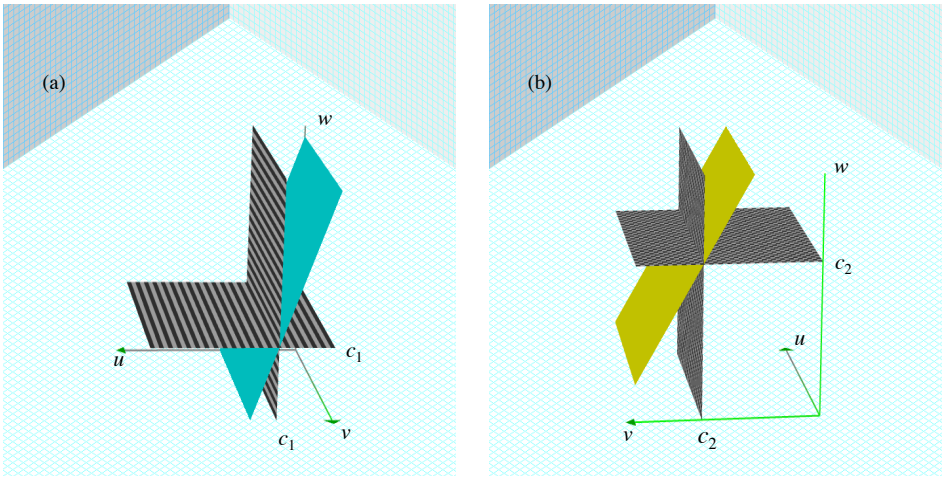


Figure 4. (a) The plane without stripes is $\alpha_1 u + (1-\alpha_1)w = c_1$. (b) The plane without checks is $\alpha_2 v + (1-\alpha_2)w = c_2$.

where

$$t_l \equiv \max\left\{0, \frac{c_1-\alpha_1}{1-\alpha_1}, \frac{c_2-\alpha_2}{1-\alpha_2}\right\}, \qquad t_u \equiv \min\left\{1, \frac{c_1}{1-\alpha_1}, \frac{c_2}{1-\alpha_2}\right\}.$$

Thus $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} \neq \emptyset$ if and only if $t_l \leq t_u$. It follows directly that for fixed c_1, c_2 the set of pairs $\langle \alpha_1, \alpha_2 \rangle \in [0,1]^2$ for which $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$ is not empty may be characterized thus:

$$\begin{aligned} \mathcal{R}_{c_1,c_2} &\equiv \left\{ \langle \alpha_1, \alpha_2 \rangle \,\middle|\, \mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} \neq \emptyset \right\} \\ &= [0,1]^2 \cap \bigcap_{\substack{i\in\{1,2\} \\ i^*=3-i}} \{\langle \alpha_1, \alpha_2 \rangle \mid (\alpha_i - c_i)(\alpha_{i^*} - (1-c_{i^*})) \leq c_i^*(1-c_i)\}. \end{aligned} \tag{5}$$

In fact, as shown in Figure 6 at most one constraint is active, so simplification is possible: let $k = \arg\max_j c_j$, and $k^* = 3-k$, then

$$\mathcal{R}_{c_1,c_2} = [0,1]^2 \cap \{\langle \alpha_1, \alpha_2 \rangle \mid (\alpha_k - c_k)(\alpha_{k^*} - (1-c_{k^*})) \leq c_k^*(1-c_k)\}.$$

(If $c_1 = c_2$ then $\mathcal{R}_{c_1,c_2} = [0,1]^2$.)

In the two dimensional analysis in §2.2 we observed that for fixed c, as α varied, the line segment would always remain inside two rectangles, as shown in Figure 3. In the three dimensional situation, the line segment (4) will stay within three boxes:

(i) If $c_1 < c_2$ then the line segment (4) is within:

$$([0,c_1] \times [0,c_2] \times [c_2,1]) \cup ([0,c_1] \times [c_2,1] \times [c_1,c_2]) \cup ([c_1,1] \times [c_2,1] \times [0,c_1]).$$

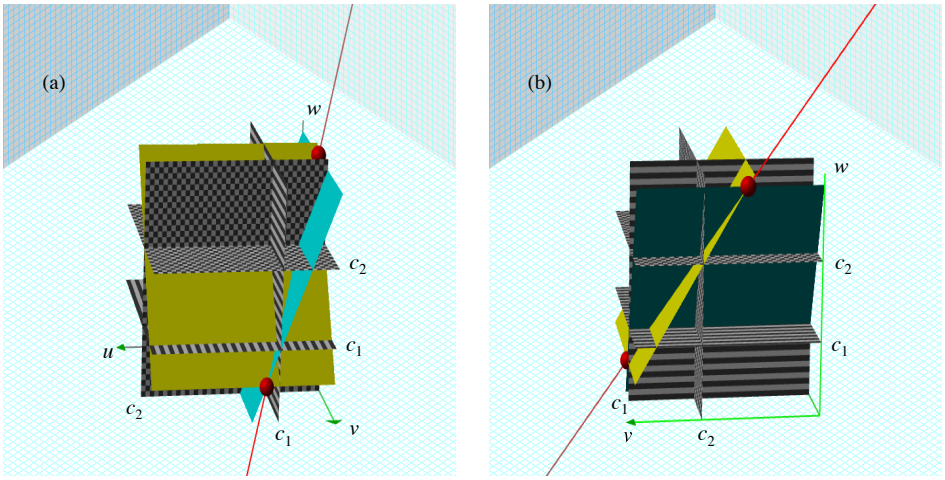


Figure 5. $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$ corresponds to the section of the line between the two marked points; (a) view towards u-w plane; (b) view from v-w plane. (Here $c_1 < c_2$.)

This may be seen as a 'staircase' with a 'corner' consisting of three blocks, descending clockwise from $\langle 0, 0, 1\rangle$ to $\langle 1, 1, 0\rangle$; see Figure 7(a). The first and second boxes intersect in the line segment joining the points $\langle 0, c_2, c_2\rangle$ and $\langle c_1, c_2, c_2\rangle$; the second and third intersect in the line segment joining $\langle c_1, c_2, c_1\rangle$ and $\langle c_1, 1, c_1\rangle$.

(ii) If $c_1 > c_2$ then the line segment is within:

$$([0, c_1] \times [0, c_2] \times [c_1, 1]) \cup ([c_1, 1] \times [0, c_2] \times [c_2, c_1]) \cup ([c_1, 1] \times [c_2, 1] \times [0, c_2]).$$

This is a 'staircase' of three blocks, descending counter-clockwise from $\langle 0, 0, 1\rangle$ to $\langle 1, 1, 0\rangle$; see Figure 7(b). The first and second boxes intersect in the line segment joining the points $\langle c_1, 0, c_1\rangle$ and $\langle c_1, c_2, c_1\rangle$; the second and third intersect in the line segment joining $\langle c_1, c_2, c_2\rangle$ and $\langle 1, c_2, c_2\rangle$.

(iii) If $c_1 = c_2 = c$ then the 'middle' box disappears and we are left with

$$([0, c] \times [0, c] \times [c, 1]) \cup ([c, 1] \times [c, 1] \times [0, c]).$$

In this case the two boxes touch at the point $\langle c, c, c\rangle$.

Note however, that the number of 'boxes' within which the line segment (4) lies may be 1, 2 or 3 (or 0 if $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} = \emptyset$). This is in contrast to the simpler case considered in §2.2 where the line segment $\mathcal{Q}_{c,\alpha}$ always intersected exactly two rectangles; see Figure 3.

3 Characterization of compatible distributions of type

Returning to the Instrumental Variable model introduced in §2, for a given patient the values taken by Y and X are deterministic functions of Z, t_X and t_Y. Conse-

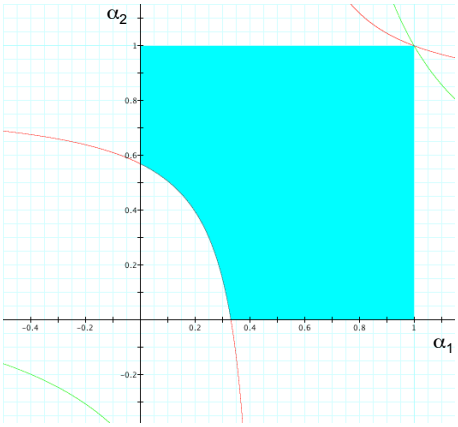


Figure 6. $\mathcal{R}_{c_1,c_2}$ corresponds to the shaded region. The hyperbola of which one arm forms a boundary of this region corresponds to the active constraint; the other hyperbola to the inactive constraint.

quently, under randomization (2), a distribution over $\mathbb{D}$ determines the conditional distributions $p(x, y \mid z)$ for $z \in \{0, 1\}$. However, since distributions on $\mathbb{D}$ form a 15 dimensional simplex, while $p(x, y \mid z)$ is of dimension 6, it is clear that the reverse does not hold; thus many different distributions over $\mathbb{D}$ give rise to the same distributions $p(x, y \mid z)$. In what follows we precisely characterize the set of distributions over $\mathbb{D}$ corresponding to a given distribution $p(x, y \mid z)$.

We will accomplish this in the following steps:

1. We first characterize the set of distributions π_X on $\mathbb{D}_X$ compatible with a given distribution $p(x \mid z)$.

2. Next we use the technique used for Step 1 to reduce the problem of characterizing distributions $\pi_{Y|X}$ compatible with $p(x, y \mid z)$ to that of characterizing the values of $p(y_x = 1 \mid \mathbf{t}_X)$ compatible with $p(x, y \mid z)$.

3. For a fixed marginal distribution π_X on $\mathbb{D}_X$ we then describe the set of values for $p(y_x = 1 \mid x, \mathbf{t}_X)$ compatible with the observed distribution $p(y \mid x, z)$.

4. In general, some distributions π_X on $\mathbb{D}_X$ and observed distributions $p(y \mid x, z)$ may be incompatible in that there are no compatible values for $p(y_x = 1 \mid \mathbf{t}_X)$. We use this to find the set of distributions π_X on $\mathbb{D}_X$ compatible with $p(y, x \mid z)$ (by restricting the set of distributions found at step 1).

5. Finally we describe the values for $p(y_x = 1 \mid \mathbf{t}_X)$ compatible with the distributions π over $\mathbb{D}_X$ found at the previous step.

We now proceed with the analysis.

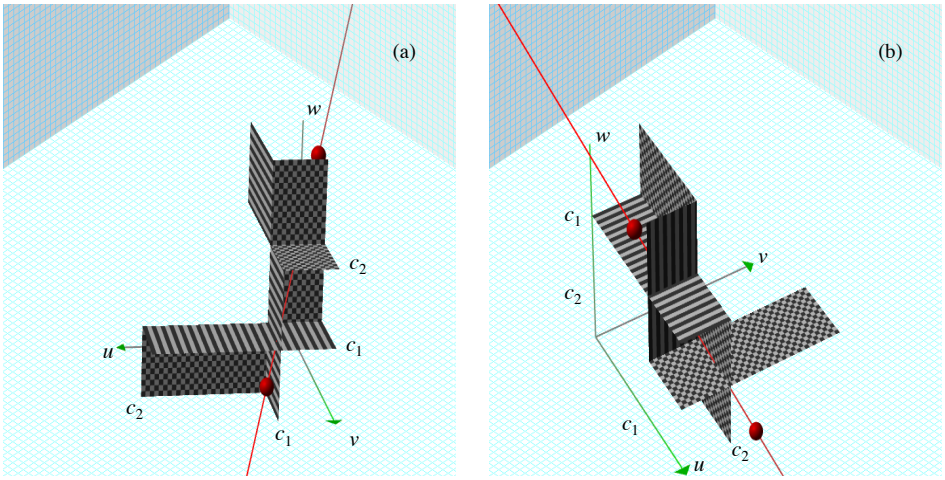


Figure 7. 'Staircases' of three boxes illustrating the possible support for $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$; (a) $c_1 < c_2$; (b) $c_2 < c_1$. Sides of the boxes that are formed by (subsets of) faces of the unit cube are not shown. The line segments shown are illustrative; in general they may not intersect all 3 boxes.

3.1 Distributions π_X on $\mathbb{D}_X$ compatible with $p(x \mid z)$

Under random assignment we have

$$\begin{aligned} p(x=1 \mid z=0) &= p(X_{z=0}=1, X_{z=1}=0) + p(X_{z=0}=1, X_{z=1}=1) \\ &= p(\mathrm{DE}) + p(\mathrm{AT}), \end{aligned}$$

$$\begin{aligned} p(x=1 \mid z=1) &= p(X_{z=0}=0, X_{z=1}=1) + p(X_{z=0}=1, X_{z=1}=1) \\ &= p(\mathrm{CO}) + p(\mathrm{AT}). \end{aligned}$$

Letting $U_{i+1} = X_{z=i}$, $i = 0, 1$ and $c_{j+1} = p(x = 1 \mid z = j)$, $j = 0, 1$, it follows directly from the analysis in §2.1 that the set of distributions π_X on $\mathbb{D}_X$ that are compatible with $p(x \mid z)$ are thus given by

$$\mathcal{P}_{c_1,c_2} = \left\{ \begin{array}{lcll} \pi_{\mathrm{AT}} &=& t, & \\ \pi_{\mathrm{DE}} &=& c_1 - t, & t \in \left[\max\left\{0, (c_1+c_2)-1\right\}, \min\left\{c_1, c_2\right\}\right] \\ \pi_{\mathrm{CO}} &=& c_2 - t, & \\ \pi_{\mathrm{NT}} &=& 1 - c_1 - c_2 + t, & \end{array} \right\}. \tag{6}$$

3.2 Reduction step in characterizing distributions $\pi_{Y|X}$ compatible with $p(x, y \mid z)$

Suppose that we were able to ascertain the set of possible values for the eight quantities:

$$\gamma^i_{\mathbf{t}_X} \equiv p(y_{x=i} = 1 \mid \mathbf{t}_X), \text{ for } i \in \{0,1\} \text{ and } \mathbf{t}_X \in \mathbb{D}_X,$$

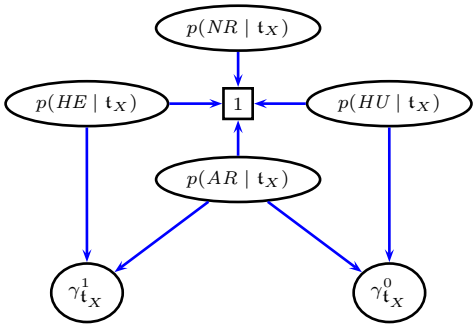


Figure 8. A graph representing the functional dependencies used in the reduction step in §3.2. The rectangular node indicates that the probabilities are required to sum to 1.

that are compatible with $p(x, y \mid z)$. Note that $p(y_{x=i} = 1 \mid \mathfrak{t}_X)$ is written as $p(y = 1 \mid \mathrm{do}(x = i), \mathfrak{t}_X)$ using Pearl's do$(\cdot)$ notation. It is then clear that the set of possible distributions $\pi_{Y|X}$ that are compatible with $p(x, y \mid z)$ simply follows from the analysis in §2.1, since

$$\begin{aligned} \gamma^0_{\mathfrak{t}_X} &= p(y_{x=0} = 1 \mid \mathfrak{t}_X) \\ &= p(HU \mid \mathfrak{t}_X) + p(AR \mid \mathfrak{t}_X), \end{aligned}$$

$$\begin{aligned} \gamma^1_{\mathfrak{t}_X} &= p(y_{x=1} = 1 \mid \mathfrak{t}_X) \\ &= p(HE \mid \mathfrak{t}_X) + p(AR \mid \mathfrak{t}_X). \end{aligned}$$

These relationships are also displayed graphically in Figure 8: in this particular graph all children are simple sums of their parents; the boxed 1 represents the 'sum to 1' constraint.

Thus, by §2.1, for given values of $\gamma^i_{\mathfrak{t}_X}$ the set of distributions $\pi_{Y|X}$ is given by:

$$\left\{ \begin{array}{lcl} p(AR \mid \mathfrak{t}_X) & \in & \left[\max\left\{0, (\gamma^0_{\mathfrak{t}_X} + \gamma^1_{\mathfrak{t}_X}) - 1\right\}, \min\left\{\gamma^0_{\mathfrak{t}_X}, \gamma^1_{\mathfrak{t}_X}\right\}\right], \\ p(NR \mid \mathfrak{t}_X) & = & 1 - \gamma^0_{\mathfrak{t}_X} - \gamma^1_{\mathfrak{t}_X} + p(AR \mid \mathfrak{t}_X), \\ p(HE \mid \mathfrak{t}_X) & = & \gamma^1_{\mathfrak{t}_X} - p(AR \mid \mathfrak{t}_X), \\ p(HU \mid \mathfrak{t}_X) & = & \gamma^0_{\mathfrak{t}_X} - p(AR \mid \mathfrak{t}_X) \end{array} \right\}. \quad (7)$$

It follows from the discussion at the end of §2.1 that the values of $\gamma^0_{\mathfrak{t}_X}$ and $\gamma^1_{\mathfrak{t}_X}$ are not restricted by the requirement that there exists a distribution $p(\cdot \mid \mathfrak{t}_X)$ on $\mathbb{D}_Y$. Consequently we may proceed in two steps: first we derive the set of values for the eight parameters $\{\gamma^i_{\mathfrak{t}_X}\}$ and the distribution on π_X (jointly) without consideration of the parameters for $\pi_{Y|X}$; second we then derive the parameters $\pi_{Y|X}$, as described above.

Finally we note that many causal quantities of interest, such as the average causal effect (ACE), and relative risk (RR) of X on Y, for a given response type $\mathfrak{t}_X$, may

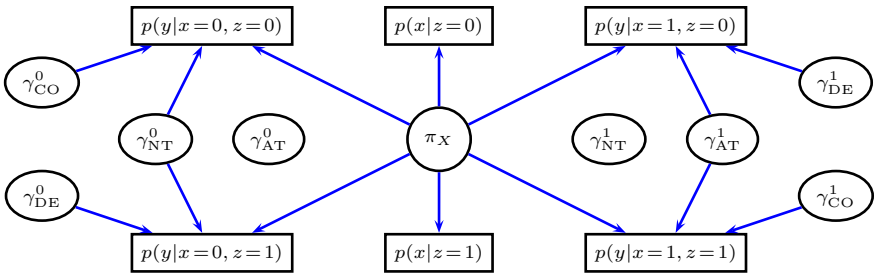


Figure 9. A graph representing the functional dependencies in the analysis of the binary IV model. Rectangular nodes are observed; oval nodes are unknown parameters. See text for further explanation.

be expressed in terms of the $\gamma^i_{\mathfrak{t}_X}$ parameters:

$$\mathrm{ACE}(\mathfrak{t}_X) = \gamma^1_{\mathfrak{t}_X} - \gamma^0_{\mathfrak{t}_X}, \qquad \mathrm{RR}(\mathfrak{t}_X) = \gamma^1_{\mathfrak{t}_X}/\gamma^0_{\mathfrak{t}_X}.$$

Consequently, for many purposes it may be unnecessary to consider the parameters $\pi_{Y|X}$ at all.

3.3 Values for $\{\gamma^i_{\mathfrak{t}_X}\}$ compatible with π_X and $p(y \mid x, z)$

We will call a specification of values for π_X, *feasible for the observed distribution* if (a) π_X lies within the set described in §3.1 of distributions compatible with $p(x \mid z)$ and (b) there exists a set of values for $\gamma^i_{\mathfrak{t}_X}$ which results in the distribution $p(y \mid x, z)$.

In the next section we give an explicit characterization of the set of feasible distributions π_X; in this section we characterize the set of values of $\gamma^i_{\mathfrak{t}_X}$ compatible with a fixed feasible distribution π_X and $p(y \mid x, z)$.

PROPOSITION 1. *The following equations relate* π_X, γ^0_{CO}, γ^0_{DE}, γ^0_{NT} *to* $p(y \mid x{=}0, z)$*:*

$$p(y{=}1 \mid x{=}0, z{=}0) = (\gamma^0_{\mathrm{CO}}\pi_{\mathrm{CO}} + \gamma^0_{\mathrm{NT}}\pi_{\mathrm{NT}})/(\pi_{\mathrm{CO}} + \pi_{\mathrm{NT}}), \tag{8}$$

$$p(y{=}1 \mid x{=}0, z{=}1) = (\gamma^0_{\mathrm{DE}}\pi_{\mathrm{DE}} + \gamma^0_{\mathrm{NT}}\pi_{\mathrm{NT}})/(\pi_{\mathrm{DE}} + \pi_{\mathrm{NT}}), \tag{9}$$

Similarly, the following relate π_X, γ^1_{CO}, γ^1_{DE}, γ^1_{AT} *to* $p(y \mid x{=}1, z)$*:*

$$p(y{=}1 \mid x{=}1, z{=}0) = (\gamma^1_{\mathrm{DE}}\pi_{\mathrm{DE}} + \gamma^1_{\mathrm{AT}}\pi_{\mathrm{AT}})/(\pi_{\mathrm{DE}} + \pi_{\mathrm{AT}}), \tag{10}$$

$$p(y{=}1 \mid x{=}1, z{=}1) = (\gamma^1_{\mathrm{CO}}\pi_{\mathrm{CO}} + \gamma^1_{\mathrm{AT}}\pi_{\mathrm{AT}})/(\pi_{\mathrm{CO}} + \pi_{\mathrm{AT}}). \tag{11}$$

Equations (8)–(11) are represented in Figure 9. Note that the parameters γ^0_{AT} and γ^1_{NT} are completely unconstrained by the observed distribution since they describe, respectively, the effect of non-exposure ($X = 0$) on Always Takers, and exposure ($X = 1$) on Never Takers, neither of which ever occur. Consequently, the set

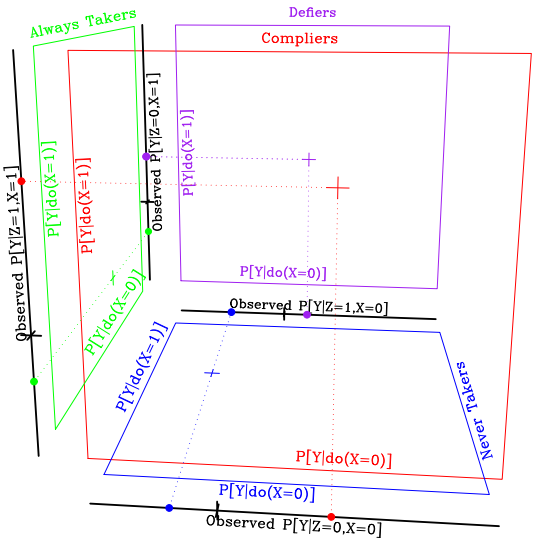


Figure 10. Geometric picture illustrating the relation between the $\gamma^i_{\mathbf{t}_X}$ parameters and $p(y \mid x, z)$. See also Figure 9.

of possible values for each of these parameters is always $[0, 1]$. Graphically this corresponds to the disconnection of γ^0_{AT} and γ^1_{NT} from the remainder of the graph.

As shown in Proposition 1 the remaining six parameters may be divided into two groups, $\{\gamma^0_{\mathrm{NT}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{CO}}\}$ and $\{\gamma^1_{\mathrm{AT}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{CO}}\}$, depending on whether they relate to unexposed subjects, or exposed subjects. Furthermore, as the graph indicates, for a fixed feasible value of π_X, compatible with the observed distribution $p(x, y \mid z)$ (assuming such exists), these two sets are variation independent. Thus, for a fixed feasible value of π_X we may analyze each of these sets separately.

A geometric picture of equations (8)–(11) is given in Figure 10: there is one square for each compliance type, with axes corresponding to $\gamma^0_{\mathbf{t}_X}$ and $\gamma^1_{\mathbf{t}_X}$; the specific value of $\langle \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X} \rangle$ is given by a cross in the square. There are four lines corresponding to the four observed quantities $p(y = 1 \mid x, z)$. Each of these observed quantities, which is denoted by a cross on the respective line, is a weighted average of two $\gamma^i_{\mathbf{t}_X}$ parameters, with weights given by π_X (the weights are not depicted explicitly).

Proof of Proposition 1: We prove (8); the other proofs are similar. Subjects for

whom $X = 0$ and $Z = 0$ are either Never Takers or Compliers. Hence

$$\begin{aligned} p(y{=}1 \mid x{=}0, z{=}0) &= p(y{=}1 \mid x{=}0, z{=}0, \mathfrak{t}_X{=}\mathrm{NT})p(\mathfrak{t}_X{=}\mathrm{NT} \mid x{=}0, z{=}0) \\ &\quad + p(y{=}1 \mid x{=}0, z{=}0, \mathfrak{t}_X{=}\mathrm{CO})p(\mathfrak{t}_X{=}\mathrm{CO} \mid x{=}0, z{=}0) \\ &= p(y_{x=0}{=}1 \mid x{=}0, z{=}0, \mathfrak{t}_X{=}\mathrm{NT})p(\mathfrak{t}_X{=}\mathrm{NT} \mid \mathfrak{t}_X \in \{\mathrm{CO}, \mathrm{NT}\}) \\ &\quad + p(y_{x=0}{=}1 \mid x{=}0, z{=}0, \mathfrak{t}_X{=}\mathrm{CO})p(\mathfrak{t}_X{=}\mathrm{CO} \mid \mathfrak{t}_X \in \{\mathrm{CO}, \mathrm{NT}\}) \\ &= p(y_{x=0}{=}1 \mid z{=}0, \mathfrak{t}_X{=}\mathrm{NT}) \times \pi_{\mathrm{NT}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\ &\quad + p(y_{x=0}{=}1 \mid z{=}0, \mathfrak{t}_X{=}\mathrm{CO}) \times \pi_{\mathrm{CO}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\ &= p(y_{x=0}{=}1 \mid \mathfrak{t}_X{=}\mathrm{NT}) \times \pi_{\mathrm{NT}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\ &\quad + p(y_{x=0}{=}1 \mid \mathfrak{t}_X{=}\mathrm{CO}) \times \pi_{\mathrm{CO}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\ &= (\gamma^0_{\mathrm{CO}}\pi_{\mathrm{CO}} + \gamma^0_{\mathrm{NT}}\pi_{\mathrm{NT}})/(\pi_{\mathrm{CO}} + \pi_{\mathrm{NT}}). \end{aligned}$$

Here the first equality is by the chain rule of probability; the second follows by consistency; the third follows since Compliers and Never Takers have $X = 0$ when $Z = 0$; the fourth follows by randomization (2). □

Values for $\gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}}$ compatible with a feasible π_X

Since (8) and (9) correspond to three quantities with two averages specified, we may apply the analysis in §2.3, taking $\alpha_1 = \pi_{\mathrm{CO}}/(\pi_{\mathrm{CO}} + \pi_{\mathrm{NT}})$, $\alpha_2 = \pi_{\mathrm{DE}}/(\pi_{\mathrm{DE}} + \pi_{\mathrm{NT}})$, $c_i = p(y = 1 \mid x = 0, z = i - 1)$ for $i = 1, 2$, $u = \gamma^0_{\mathrm{CO}}$, $v = \gamma^0_{\mathrm{DE}}$ and $w = \gamma^0_{\mathrm{NT}}$. Under this substitution, the set of possible values for $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ is then given by $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$.

Values for $\gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}}$ compatible with a feasible π_X

Likewise since (10) and (11) contain three quantities with two averages specified we again apply the analysis from §2.3, taking $\alpha_1 = \pi_{\mathrm{CO}}/(\pi_{\mathrm{CO}}+\pi_{\mathrm{AT}})$, $\alpha_2 = \pi_{\mathrm{DE}}/(\pi_{\mathrm{DE}}+\pi_{\mathrm{AT}})$, $c_i = p(y{=}1 \mid x{=}1, z{=}2 - i)$ for $i = 1, 2$, $u = \gamma^1_{\mathrm{CO}}$, $v = \gamma^1_{\mathrm{DE}}$ and $w = \gamma^1_{\mathrm{AT}}$. The set of possible values for $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$ is then given by $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$.

3.4 Values of π_X compatible with $p(x, y \mid z)$

In §3.1 we characterized the distributions π_X compatible with $p(x \mid z)$ as a one dimensional subspace of the three dimensional simplex, parameterized in terms of $t \equiv \pi_{\mathrm{AT}}$; see (6). We now incorporate the additional constraints on π_X that arise from $p(y \mid x, z)$. These occur because some distributions π_X, though compatible with $p(x \mid z)$, lead to an empty set of values for $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$ or $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ and thus are infeasible.

Constraints on π_X arising from $p(y \mid x = 0, z)$

Building on the analysis in §3.3 the set of values for

$$\begin{aligned} \langle \alpha_1, \alpha_2 \rangle &= \langle \pi_{\mathrm{CO}}/(\pi_{\mathrm{CO}} + \pi_{\mathrm{NT}}), \pi_{\mathrm{DE}}/(\pi_{\mathrm{DE}} + \pi_{\mathrm{NT}}) \rangle \\ &= \langle \pi_{\mathrm{CO}}/p_{x_0|z_0}, \pi_{\mathrm{DE}}/p_{x_0|z_0} \rangle \end{aligned} \tag{12}$$

compatible with $p(y \mid x=0,z)$ (i.e. for which the corresponding set of values for $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT}\rangle$ is non-empty) is given by $\mathcal{R}_{c_1^*,c_2^*}$, where $c_i^* = p(y=1 \mid x=0, z=i-1)$, $i=1,2$ (see §2.3). The inequalities defining $\mathcal{R}_{c_1^*,c_2^*}$ may be translated into upper bounds on $t \equiv \pi_{\rm AT}$ in (6), as follows:

$$t \leq \min\left\{ 1 - \sum_{j\in\{0,1\}} p(y=j, x=0 \mid z=j),\ 1 - \sum_{k\in\{0,1\}} p(y=k, x=0 \mid z=1-k) \right\}. \quad (13)$$

Proof: The analysis in §3.3 implied that for $\mathcal{R}_{c_1^*,c_2^*} \neq \emptyset$ we require

$$\frac{c_1^* - \alpha_1}{1-\alpha_1} \leq \frac{c_2^*}{1-\alpha_2} \quad \text{and} \quad \frac{c_2^* - \alpha_2}{1-\alpha_2} \leq \frac{c_1^*}{1-\alpha_1}. \quad (14)$$

Taking the first of these and plugging in the definitions of c_1^*, c_2^*, α_1 and α_2 from (12) gives:

$$\begin{aligned}
& \frac{p_{y_1|x_0,z_0} - (\pi_{\rm CO}/p_{x_0|z_0})}{1-(\pi_{\rm CO}/p_{x_0|z_0})} \leq \frac{p_{y_1|x_0,z_1}}{1-(\pi_{\rm DE}/p_{x_0|z_1})} \\
(\Leftrightarrow)\quad & (p_{y_1|x_0,z_0} - (\pi_{\rm CO}/p_{x_0|z_0}))(1-(\pi_{\rm DE}/p_{x_0|z_1})) \leq p_{y_1|x_0,z_1}(1-(\pi_{\rm CO}/p_{x_0|z_0})) \\
(\Leftrightarrow)\quad & (p_{y_1,x_0|z_0} - \pi_{\rm CO})(p_{x_0|z_1} - \pi_{\rm DE}) \leq p_{y_1,x_0|z_1}(p_{x_0|z_0} - \pi_{\rm CO}).
\end{aligned}$$

But $p_{x_0|z_1} - \pi_{\rm DE} = p_{x_0|z_0} - \pi_{\rm CO} = \pi_{\rm NT}$, hence these terms may be cancelled to give:

$$\begin{aligned}
(p_{y_1,x_0|z_0} - \pi_{\rm CO}) &\leq p_{y_1,x_0|z_1} \\
(\Leftrightarrow)\quad \pi_{\rm AT} - p_{x_1|z_1} &\leq p_{y_1,x_0|z_1} - p_{y_1,x_0|z_0} \\
(\Leftrightarrow)\quad \pi_{\rm AT} &\leq 1 - p_{y_0,x_0|z_1} - p_{y_1,x_0|z_0}.
\end{aligned}$$

A similar argument applied to the second constraint in (14) to derive that

$$\pi_{\rm AT} \leq 1 - p_{y_0,x_0|z_0} - p_{y_1,x_0|z_1},$$

as required. □

Constraints on π_X arising from $p(y \mid x=1,z)$

Similarly using the analysis in §3.3 the set of values for

$$\langle \alpha_1, \alpha_2 \rangle = \langle \pi_{\rm CO}/(\pi_{\rm CO}+\pi_{\rm AT}), \pi_{\rm DE}/(\pi_{\rm DE}+\pi_{\rm AT}) \rangle$$

compatible with $p(y \mid x=1,z)$ (i.e. that the corresponding set of values for $\langle \gamma^1_{\rm CO}, \gamma^1_{\rm DE}, \gamma^1_{\rm AT}\rangle$ is non-empty) is given by $\mathcal{R}_{c_1^{**},c_2^{**}}$, where $c_i^{**} = p(y=1 \mid x=1, z=2-i)$, $i=1,2$ (see §2.3). Again, we translate the inequalities which define $\mathcal{R}_{c_1^{**},c_2^{**}}$ into further upper bounds on $t = \pi_{\rm AT}$ in (6):

$$t \leq \min\left\{ \sum_{j\in\{0,1\}} p(y=j, x=1 \mid z=j),\ \sum_{k\in\{0,1\}} p(y=k, x=1 \mid z=1-k) \right\}. \quad (15)$$

The proof that these inequalities are implied, is very similar to the derivation of the upper bounds on π_{AT} arising from $p(y \mid x = 0, z)$ considered above.

The distributions π_X compatible with the observed distribution

It follows that the set of distributions on $\mathbb{D}_X$ that are compatible with the observed distribution, which we denote $\mathcal{P}_X$, may be given thus:

$$\mathcal{P}_X = \left\{ \begin{array}{rcl} \pi_{\mathrm{AT}} & \in & [l\pi_{\mathrm{AT}}, u\pi_{\mathrm{AT}}], \\ \pi_{\mathrm{NT}}(\pi_{\mathrm{AT}}) & = & 1 - p(x = 1 \mid z = 0) - p(x = 1 \mid z = 1) + \pi_{\mathrm{AT}}, \\ \pi_{\mathrm{CO}}(\pi_{\mathrm{AT}}) & = & p(x = 1 \mid z = 1) - \pi_{\mathrm{AT}}, \\ \pi_{\mathrm{DE}}(\pi_{\mathrm{AT}}) & = & p(x = 1 \mid z = 0) - \pi_{\mathrm{AT}} \end{array} \right\}, \quad (16)$$

where

$$l\pi_{\mathrm{AT}} = \max\{0, p(x = 1 \mid z = 0) + p(x = 1 \mid z = 1) - 1\};$$

$$u\pi_{\mathrm{AT}} = \min \left\{ \begin{array}{ll} p(x = 1 \mid z = 0), & p(x = 1 \mid z = 1), \\ 1 - \sum_j p(y{=}j, x{=}0 \mid z{=}j), & 1 - \sum_k p(y{=}k, x{=}0 \mid z{=}1{-}k), \\ \sum_j p(y{=}j, x{=}1 \mid z{=}j), & \sum_k p(y{=}k, x{=}1 \mid z{=}1{-}k) \end{array} \right\}.$$

Observe that unlike the upper bound, the lower bound on π_{AT} (and π_{NT}) obtained from $p(x, y \mid z)$ is the same as the lower bound derived from $p(x \mid z)$ alone.
We define $\pi_X(\pi_{\mathrm{AT}}) \equiv \langle \pi_{\mathrm{NT}}(\pi_{\mathrm{AT}}), \pi_{\mathrm{CO}}(\pi_{\mathrm{AT}}), \pi_{\mathrm{DE}}(\pi_{\mathrm{AT}}), \pi_{\mathrm{AT}} \rangle$, for use below. Note the following:

PROPOSITION 2. *When π_{AT} (equivalently π_{NT}) is minimized then either $\pi_{\mathrm{NT}} = 0$ or $\pi_{\mathrm{AT}} = 0$.*

Proof: This follows because, by the expression for $l\pi_{\mathrm{AT}}$, either $l\pi_{\mathrm{AT}} = 0$, or $l\pi_{\mathrm{AT}} = p(x = 1 \mid z = 0) + p(x = 1 \mid z = 1) - 1$, in which case $l\pi_{\mathrm{NT}} = 0$ by (16). □

4 Projections

The analysis in §3 provides a complete description of the set of distributions over $\mathbb{D}$ compatible with a given observed distribution. In particular, equation (16) describes the one dimensional set of compatible distributions over $\mathbb{D}_X$; in §3.3 we first gave a description of the one dimensional set of values over $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ compatible with the observed distribution and a specific feasible distribution π_X over $\mathbb{D}_X$; we then described the one dimensional set of values for $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$. Varying π_X over the set $\mathcal{P}_X$ of feasible distributions over $\mathbb{D}_X$, describes a set of lines, forming two two-dimensional manifolds which represent the space of possible values for $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ and likewise for $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$. As noted previously, the parameters γ^0_{AT} and γ^1_{NT} are unconstrained by the observed data. Finally, if there is interest in distributions over response types, there is a one-dimensional set

of such distributions associated with each possible pair of values from $\gamma^0_{\mathfrak{t}_X}$ and $\gamma^1_{\mathfrak{t}_X}$.

For the purposes of visualization it is useful to look at projections. There are many such projections that could be considered, here we focus on projections that display the relation between the possible values for π_X and $\gamma^x_{\mathfrak{t}_X}$. See Figure 11.

We make the following definition:

$$\alpha^{ij}_{\mathfrak{t}_X}(\pi_X) \equiv p(\mathfrak{t}_X \mid X_{z=i} = j),$$

where $\pi_X = \langle \pi_{\rm NT}, \pi_{\rm CO}, \pi_{\rm DE}, \pi_{\rm AT} \rangle \in \mathcal{P}_X$, as before. For example, $\alpha^{00}_{\rm NT}(\pi_X) = \pi_{\rm NT}/(\pi_{\rm NT} + \pi_{\rm CO})$, $\alpha^{10}_{\rm NT}(\pi_X) = \pi_{\rm NT}/(\pi_{\rm NT} + \pi_{\rm DE})$.

4.1 Upper and Lower bounds on $\gamma^x_{\mathfrak{t}_X}$ as a function of π_X

We use the following notation to refer to the upper and lower bounds on $\gamma^0_{\rm NT}$ and $\gamma^1_{\rm AT}$ that were derived earlier. If π_X is such that $\pi_{\rm NT} > 0$, so $\alpha^{00}_{\rm NT}, \alpha^{10}_{\rm NT} > 0$ then we define:

$$\begin{aligned} l\gamma^0_{\rm NT}(\pi_X) &\equiv \max\left\{0, \frac{p_{y_1|x_0z_0} - \alpha^{00}_{\rm CO}(\pi_X)}{\alpha^{00}_{\rm NT}(\pi_X)}, \frac{p_{y_1|x_0z_1} - \alpha^{10}_{\rm DE}(\pi_X)}{\alpha^{10}_{\rm NT}(\pi_X)}\right\}, \\ u\gamma^0_{\rm NT}(\pi_X) &\equiv \min\left\{\frac{p_{y_1|x_0z_0}}{\alpha^{00}_{\rm NT}(\pi_X)}, \frac{p_{y_1|x_0z_1}}{\alpha^{10}_{\rm NT}(\pi_X)}, 1\right\}, \end{aligned}$$

while if $\pi_{\rm NT} = 0$ then we define $l\gamma^0_{\rm NT}(\pi_X) \equiv 0$ and $u\gamma^0_{\rm NT}(\pi_X) \equiv 1$. Similarly, if π_X is such that $\pi_{\rm AT} > 0$ then we define:

$$\begin{aligned} l\gamma^1_{\rm AT}(\pi_X) &\equiv \max\left\{0, \frac{p_{y_1|x_1z_1} - \alpha^{11}_{\rm CO}(\pi_X)}{\alpha^{11}_{\rm AT}(\pi_X)}, \frac{p_{y_1|x_1z_0} - \alpha^{01}_{\rm DE}(\pi_X)}{\alpha^{01}_{\rm AT}(\pi_X)}\right\}, \\ u\gamma^1_{\rm AT}(\pi_X) &\equiv \min\left\{\frac{p_{y_1|x_1z_1}}{\alpha^{11}_{\rm AT}(\pi_X)}, \frac{p_{y_1|x_1z_0}}{\alpha^{01}_{\rm AT}(\pi_X)}, 1\right\}, \end{aligned}$$

while if $\pi_{\rm AT} = 0$ then let $l\gamma^1_{\rm AT}(\pi_X) \equiv 0$ and $u\gamma^1_{\rm AT}(\pi_X) \equiv 1$.

We note that Table 2 summarizes the upper and lower bounds, as a function of $\pi_X \in \mathcal{P}_X$, on each of the eight parameters $\gamma^x_{\mathfrak{t}_X}$ that were derived earlier in §3.3. These are shown by the thicker lines on each of the plots forming the upper and lower boundaries in Figure 11 ($\gamma^0_{\rm AT}$ and $\gamma^1_{\rm NT}$ are not shown in the Figure).

The upper and lower bounds on $\gamma^0_{\rm NT}$ and $\gamma^1_{\rm AT}$ are relatively simple:

PROPOSITION 3. *$l\gamma^0_{\rm NT}(\pi_X)$ and $l\gamma^1_{\rm AT}(\pi_X)$ are non-decreasing in $\pi_{\rm AT}$ and $\pi_{\rm NT}$. Likewise $u\gamma^0_{\rm NT}(\pi_X)$ and $u\gamma^1_{\rm AT}(\pi_X)$ are non-increasing in $\pi_{\rm AT}$ and $\pi_{\rm NT}$.*

Proof: We first consider $l\gamma^0_{\rm NT}$. By (16), $\pi_{\rm NT} = 1 - p(x = 1 \mid z = 0) - p(x = 1 \mid z = 1) + \pi_{\rm AT}$, hence a function is non-increasing [non-decreasing] in $\pi_{\rm AT}$ iff it is non-increasing [non-decreasing] in $\pi_{\rm NT}$. Observe that for $\pi_{\rm NT} > 0$,

$$\begin{aligned} (p_{y_1|x_0z_0} - \alpha^{00}_{\rm CO}(\pi_X))/\alpha^{00}_{\rm NT}(\pi_X) &= \left(p_{y_1|x_0z_0}(\pi_{\rm NT} + \pi_{\rm CO}) - \pi_{\rm CO}\right)/\pi_{\rm NT} \\ &= p_{y_1|x_0z_0} - p_{y_0|x_0z_0}(\pi_{\rm CO}/\pi_{\rm NT}) \\ &= p_{y_1|x_0z_0} + p_{y_0|x_0z_0}(1 - (p_{x_0|z_0}/\pi_{\rm NT})) \end{aligned}$$

	Lower Bound	Upper Bound
γ^0_{NT}	$l\gamma^0_{\mathrm{NT}}(\pi_X)$	$u\gamma^0_{\mathrm{NT}}(\pi_X)$
γ^0_{CO}	$(p_{y_1\|x_0z_0} - u\gamma^0_{\mathrm{NT}}(\pi_X)\cdot\alpha^{00}_{\mathrm{NT}})/\alpha^{00}_{\mathrm{CO}}$	$(p_{y_1\|x_0z_0} - l\gamma^0_{\mathrm{NT}}(\pi_X)\cdot\alpha^{00}_{\mathrm{NT}})/\alpha^{00}_{\mathrm{CO}}$
γ^0_{DE}	$(p_{y_1\|x_0z_1} - u\gamma^0_{\mathrm{NT}}(\pi_X)\cdot\alpha^{10}_{\mathrm{NT}})/\alpha^{10}_{\mathrm{DE}}$	$(p_{y_1\|x_0z_1} - l\gamma^0_{\mathrm{NT}}(\pi_X)\cdot\alpha^{10}_{\mathrm{NT}})/\alpha^{10}_{\mathrm{DE}}$
γ^0_{AT}	0	1
γ^1_{NT}	0	1
γ^1_{CO}	$(p_{y_1\|x_1z_1} - u\gamma^1_{\mathrm{AT}}(\pi_X)\cdot\alpha^{11}_{\mathrm{AT}})/\alpha^{11}_{\mathrm{CO}}$	$(p_{y_1\|x_1z_1} - l\gamma^1_{\mathrm{AT}}(\pi_X)\cdot\alpha^{11}_{\mathrm{AT}})/\alpha^{11}_{\mathrm{CO}}$
γ^1_{DE}	$(p_{y_1\|x_1z_0} - u\gamma^1_{\mathrm{AT}}(\pi_X)\cdot\alpha^{01}_{\mathrm{AT}})/\alpha^{01}_{\mathrm{DE}}$	$(p_{y_1\|x_1z_0} - l\gamma^1_{\mathrm{AT}}(\pi_X)\cdot\alpha^{01}_{\mathrm{AT}})/\alpha^{01}_{\mathrm{DE}}$
γ^1_{AT}	$l\gamma^1_{\mathrm{AT}}(\pi_X)$	$u\gamma^1_{\mathrm{AT}}(\pi_X)$

Table 2. Upper and Lower bounds on $\gamma^x_{\mathbf{t}_X}$, as a function of $\pi_X \in \mathcal{P}_X$. If for some π_X an expression giving a lower bound for a quantity is undefined then the lower bound is 0; conversely if an expression for an upper bound is undefined then the upper bound is 1.

which is non-decreasing in π_{NT}. Similarly,

$$(p_{y_1|x_0z_1} - \alpha^{10}_{\mathrm{DE}}(\pi_X))/\alpha^{10}_{\mathrm{NT}}(\pi_X) = p_{y_1|x_0z_1} + p_{y_0|x_0z_1}(1 - (p_{x_0|z_1}/\pi_{\mathrm{NT}})).$$

The conclusion follows since the maximum of a set of non-decreasing functions is non-decreasing.

The other arguments are similar. □

We note that the bounds on γ^x_{CO} and γ^x_{DE} need not be monotonic in π_{AT}.

PROPOSITION 4. *Let $\pi^{\min}_X$ be the distribution in $\mathcal{P}_X$ for which π_{AT} and π_{NT} are minimized then either:*

(1) $\pi^{\min}_{\mathrm{NT}} = 0$, *hence* $l\gamma^0_{\mathrm{NT}}(\pi^{\min}_X) = 0$ *and* $u\gamma^0_{\mathrm{NT}}(\pi^{\min}_X) = 1$*; or*

(2) $\pi^{\min}_{\mathrm{AT}} = 0$, *hence* $l\gamma^1_{\mathrm{AT}}(\pi^{\min}_X) = 0$ *and* $u\gamma^1_{\mathrm{AT}}(\pi^{\min}_X) = 1$.

Proof: This follows from Proposition 2, and the fact that if $\pi_{\mathbf{t}_X} = 0$ then $\gamma^i_{\mathbf{t}_X}$ is not identified (for any i). □

4.2 Upper and Lower bounds on $p(\mathrm{AT})$ as a function of γ^0_{NT}

The expressions given in Table 2 allow the range of values for each $\gamma^i_{\mathbf{t}_X}$ to be determined as a function of π_X, giving the upper and lower bounding curves in Figure 11. However it follows directly from (8) and (9) that there is a bijection between the three shapes shown for γ^0_{CO}, γ^0_{DE} and γ^0_{NT} (top row of Figure 11).

In this section we describe this bijection by deriving curves corresponding to fixed values of $\gamma^0_{\rm NT}$ that are displayed in the plots for $\gamma^0_{\rm CO}$ and $\gamma^0_{\rm DE}$. Similarly it follows from (10) and (11) that there is a bijection between the three shapes shown for $\gamma^1_{\rm CO}$, $\gamma^1_{\rm DE}$, $\gamma^1_{\rm AT}$ (bottom row of Figure 11). Correspondingly we add curves to the plots for $\gamma^1_{\rm CO}$ and $\gamma^1_{\rm DE}$ corresponding to fixed values of $\gamma^1_{\rm AT}$. (The expressions in this section are used solely to add these curves and are not used elsewhere.)

As described earlier, for a given distribution $\pi_X \in \mathcal{P}_X$ the set of values for $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT}\rangle$ forms a one dimensional subspace. For a given π_X if $\pi_{\rm CO} > 0$ then $\gamma^0_{\rm CO}$ is a deterministic function of $\gamma^0_{\rm NT}$, likewise for $\gamma^0_{\rm DE}$.

It follows from Proposition 3 that the range of values for $\gamma^0_{\rm NT}$ when $\pi_X = \pi^{\min}_X$ contains the range of possible values for $\gamma^0_{\rm NT}$ for any other $\pi_X \in \mathcal{P}_X$. The same holds for $\gamma^1_{\rm AT}$. Thus for any given possible value of $\gamma^0_{\rm NT}$, the minimum compatible value of $\pi_{\rm AT} = l\pi_{\rm AT} \equiv \max\left\{0, p_{x_1|z_0} + p_{x_1|z_1} - 1\right\}$. This is reflected in the plots in Figure 11 for $\gamma^0_{\rm NT}$ and $\gamma^1_{\rm AT}$ in that the left hand endpoints of the thinner lines (lying between the upper and lower bounds) all lie on the same vertical line for which $\pi_{\rm AT}$ is minimized.

In contrast the upper bounds on $\pi_{\rm AT}$ vary as a function of $\gamma^0_{\rm NT}$ (also $\gamma^1_{\rm AT}$). The upper bound for $\pi_{\rm AT}$ as a function of $\gamma^0_{\rm NT}$ occurs when one of the thinner horizontal lines in the plot for $\gamma^0_{\rm NT}$ in Figure 11 intersects either $u\gamma^0_{\rm NT}(\pi_X)$, $l\gamma^0_{\rm NT}(\pi_X)$, or the vertical line given by the global upper bound, $u\pi_{\rm AT}$, on $\pi_{\rm AT}$:

$$
\begin{aligned}
u\pi_{\rm AT}(\gamma^0_{\rm NT}) &\equiv \max\left\{\pi_{\rm AT} \mid \gamma^0_{\rm NT} \in [l\gamma^0_{\rm NT}(\pi_X), u\gamma^0_{\rm NT}(\pi_X)]\right\} \\
&= \min\left\{ p_{x_1|z_1} - p_{x_0|z_0}\left(1 - \frac{p_{y_1|x_0z_0}}{\gamma^0_{\rm NT}}\right),\ p_{x_1|z_0} - p_{x_0|z_1}\left(1 - \frac{p_{y_1|x_0z_1}}{\gamma^0_{\rm NT}}\right),\right. \\
&\qquad \left. p_{x_1|z_1} - p_{x_0|z_0}\left(1 - \frac{p_{y_0|x_0z_0}}{1-\gamma^0_{\rm NT}}\right),\ p_{x_1|z_0} - p_{x_0|z_1}\left(1 - \frac{p_{y_0|x_0z_1}}{1-\gamma^0_{\rm NT}}\right), u\pi_{\rm AT}\right\};
\end{aligned}
$$

similarly we have

$$
\begin{aligned}
u\pi_{\rm AT}(\gamma^1_{\rm AT}) &\equiv \max\left\{\pi_{\rm AT} \mid \gamma^1_{\rm AT} \in [l\gamma^0_{\rm AT}(\pi_X), u\gamma^1_{\rm AT}(\pi_X)]\right\} \\
&= \min\left\{u\pi_{\rm AT}, \frac{p_{x_1|z_1}p_{y_1|x_1z_1}}{\gamma^1_{\rm AT}}, \frac{p_{x_1|z_0}p_{y_1|x_1z_0}}{\gamma^1_{\rm AT}}, \frac{p_{x_1|z_1}p_{y_0|x_1z_1}}{1-\gamma^1_{\rm AT}}, \frac{p_{x_1|z_0}p_{y_0|x_1z_0}}{1-\gamma^1_{\rm AT}}\right\}.
\end{aligned}
$$

The curves added to the unexposed plots for Compliers and Defiers in Figure 11 are as follows:

$$
\begin{aligned}
\gamma^0_{\rm CO}(\pi_X, \gamma^0_{\rm NT}) &\equiv (p_{y_1|x_0z_0} - \gamma^0_{\rm NT}\cdot\alpha^{00}_{\rm NT})/\alpha^{00}_{\rm CO}, \\
c\gamma^0_{\rm CO}(\pi_{\rm AT}, \gamma^0_{\rm NT}) &\equiv \{\langle \pi_{\rm AT}, \gamma^0_{\rm CO}(\pi_X(\pi_{\rm AT}), \gamma^0_{\rm NT})\rangle\}; \qquad (17)
\end{aligned}
$$

$$
\begin{aligned}
\gamma^0_{\rm DE}(\pi_X, \gamma^0_{\rm NT}) &\equiv (p_{y_1|x_0z_1} - \gamma^0_{\rm NT}\cdot\alpha^{10}_{\rm NT})/\alpha^{10}_{\rm DE}, \\
c\gamma^0_{\rm DE}(\pi_{\rm AT}, \gamma^0_{\rm NT}) &\equiv \{\langle \pi_{\rm AT}, \gamma^0_{\rm DE}(\pi_X(\pi_{\rm AT}), \gamma^0_{\rm NT})\rangle\}; \qquad (18)
\end{aligned}
$$

for $\gamma^0_{\rm NT} \in [l\gamma^0_{\rm NT}(\pi^{\min}_X), u\gamma^0_{\rm NT}(\pi^{\min}_X)]$; $\pi_{\rm AT} \in [l\pi_{\rm AT}, u\pi_{\rm AT}(\gamma^0_{\rm NT})]$. The curves added

Table 3. Flu Vaccine Data from [McDonald, Hiu, and Tierney 1992].

Z	X	Y	count
0	0	0	99
0	0	1	1027
0	1	0	30
0	1	1	233
1	0	0	84
1	0	1	935
1	1	0	31
1	1	1	422
			2,861

to the exposed plots for Compliers and Defiers in Figure 11 are given by:

$$\begin{aligned} \gamma^1_{\mathrm{CO}}(\pi_X, \gamma^1_{\mathrm{AT}}) &\equiv (p_{y_1|x_1z_1} - \gamma^1_{\mathrm{AT}} \cdot \alpha^{11}_{\mathrm{AT}})/\alpha^{11}_{\mathrm{CO}}, \\ c\gamma^1_{\mathrm{DE}}(\pi_{\mathrm{AT}}, \gamma^1_{\mathrm{AT}}) &\equiv \{\langle \pi_{\mathrm{AT}}, \gamma^1_{\mathrm{CO}}(\pi_X(\pi_{\mathrm{AT}}), \gamma^1_{\mathrm{AT}})\rangle\}; \end{aligned} \tag{19}$$

$$\begin{aligned} \gamma^1_{\mathrm{DE}}(\pi_X, \gamma^1_{\mathrm{AT}}) &\equiv (p_{y_1|x_1z_0} - \gamma^1_{\mathrm{AT}} \cdot \alpha^{01}_{\mathrm{AT}})/\alpha^{01}_{\mathrm{DE}}, \\ c\gamma^1_{\mathrm{DE}}(\pi_{\mathrm{AT}}, \gamma^1_{\mathrm{AT}}) &\equiv \{\langle \pi_{\mathrm{AT}}, \gamma^1_{\mathrm{DE}}(\pi_X(\pi_{\mathrm{AT}}), \gamma^1_{\mathrm{AT}})\rangle\}; \end{aligned} \tag{20}$$

for $\gamma^1_{\mathrm{AT}} \in [l\gamma^1_{\mathrm{AT}}(\pi^{\min}_X), u\gamma^1_{\mathrm{AT}}(\pi^{\min}_X)]$; $\pi_{\mathrm{AT}} \in [l\pi_{\mathrm{AT}}, u\pi_{\mathrm{AT}}(\gamma^1_{\mathrm{AT}})]$.

4.3 Example: Flu Data

To illustrate some of the constructions described we consider the influenza vaccine dataset [McDonald, Hiu, and Tierney 1992] previously analyzed by [Hirano, Imbens, Rubin, and Zhou 2000]; see Table 3. Here the instrument Z was whether a patient's physician was sent a card asking him to remind patients to obtain flu shots, or not; X is whether or not the patient did in fact get a flu shot. Finally $Y = 1$ indicates that a patient was *not* hospitalized. Unlike the analysis of [Hirano, Imbens, Rubin, and Zhou 2000] we ignore baseline covariates, and restrict attention to displaying the set of parameters of the IV model that are compatible with the empirical distribution.

The set of values for π_X vs. $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ (upper row), and π_X vs. $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$ corresponding to the empirical distribution for $p(x, y \mid z)$ are shown in Figure 11. The empirical distribution is not consistent with there being no Defiers (though the scales in Figure 11 show 0 as one endpoint for the proportion π_{DE} this is merely a consequence of the significant digits displayed; in fact the true lower bound on this proportion is 0.0005).

We emphasize that this analysis merely derives the logical consequences of the empirical distribution under the IV model and ignores sampling variability.

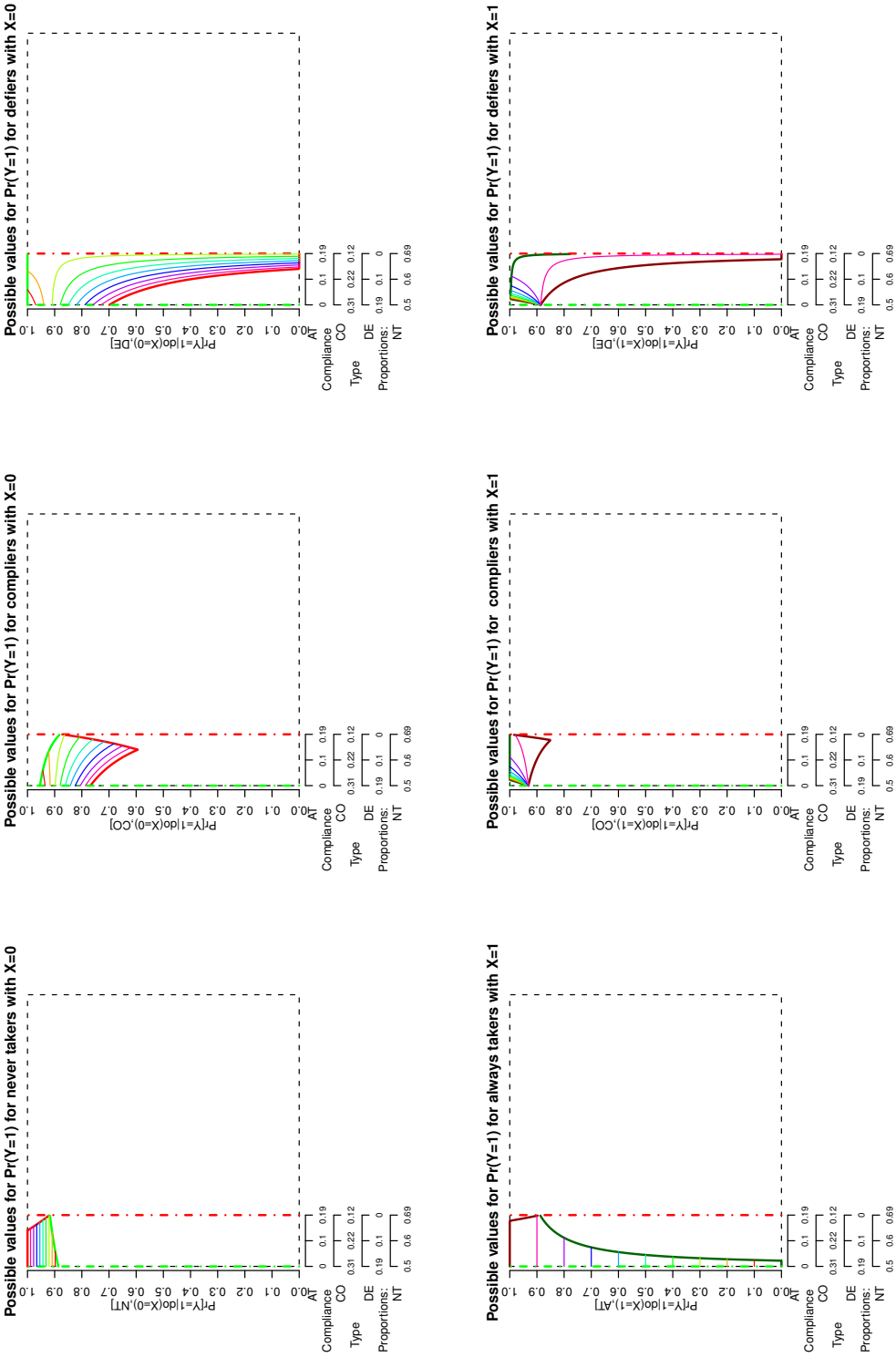


Figure 11. Depiction of the set of values for π_X vs. $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ (upper row), and π_X vs. $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$ for the flu data.

5 Bounding Average Causal Effects

We may use the results above to obtain bounds on average causal effects, for different complier strata:

$$\begin{aligned}
\mathrm{ACE}_{\mathbf{t}_X}(\pi_X, \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}) &\equiv \gamma^1_{\mathbf{t}_X}(\pi_X) - \gamma^0_{\mathbf{t}_X}(\pi_X), \\
l\mathrm{ACE}_{\mathbf{t}_X}(\pi_X) &\equiv \min_{\gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}} \mathrm{ACE}_{\mathbf{t}_X}(\pi_X, \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}), \\
u\mathrm{ACE}_{\mathbf{t}_X}(\pi_X) &\equiv \max_{\gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}} \mathrm{ACE}_{\mathbf{t}_X}(\pi_X, \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}),
\end{aligned}$$

as a function of a feasible distribution π_X; see Table 5. As shown in the table, the values of γ^0_{NT} and γ^1_{AT} which maximize (minimize) $\mathrm{ACE}_{\mathrm{CO}}$ and $\mathrm{ACE}_{\mathrm{DE}}$ are those which minimize (maximize) $\mathrm{ACE}_{\mathrm{NT}}$ and $\mathrm{ACE}_{\mathrm{AT}}$; this is an immediate consequence of the negative coefficients for γ^0_{NT} and γ^1_{AT} in the bounds for γ^x_{CO} and γ^x_{DE} in Table 2.

ACE bounds for the four compliance types are shown for the flu data in Figure 12. The ACE bounds for Compliers indicate that, under the observed distribution, the possibility of a zero ACE for Compliers is consistent with all feasible distributions over compliance types, except those for which the proportion of Defiers in the population is small.

Following [Pearl 2000; Robins 1989; Manski 1990; Robins and Rotnitzky 2004] we also consider the average causal effect on the entire population:

$$\mathrm{ACE}_{\mathrm{global}}(\pi_X, \{\gamma^x_{\mathbf{t}_X}\}) \equiv \sum_{\mathbf{t}_X \in \mathbb{D}_X} (\gamma^1_{\mathbf{t}_X}(\pi_X) - \gamma^0_{\mathbf{t}_X}(\pi_X))\pi_{\mathbf{t}_X};$$

upper and lower bounds taken over $\{\gamma^x_{\mathbf{t}_X}\}$ are defined similarly. The bounds given for $\mathrm{ACE}_{\mathbf{t}_X}$ in Table 5 are an immediate consequence of equations (8)–(11) which relate $p(y \mid x, z)$ to π_X and $\{\gamma^x_{\mathbf{t}_X}\}$. Before deriving the ACE bounds we need the following observation:

LEMMA 5. *For a given feasible π_X and $p(y, x \mid z)$,*

$$\begin{aligned}
&\mathrm{ACE}_{\mathrm{global}}(\pi_X, \{\gamma^x_{\mathbf{t}_X}\}) \\
&= p_{y_1,x_1|z_1} - p_{y_1,x_0|z_0} + \pi_{\mathrm{DE}}(\gamma^1_{\mathrm{DE}} - \gamma^0_{\mathrm{DE}}) + \pi_{\mathrm{NT}}\gamma^1_{\mathrm{NT}} - \pi_{\mathrm{AT}}\gamma^0_{\mathrm{AT}} \qquad (21) \\
&= p_{y_1,x_1|z_0} - p_{y_1,x_0|z_1} + \pi_{\mathrm{CO}}(\gamma^1_{\mathrm{CO}} - \gamma^0_{\mathrm{CO}}) + \pi_{\mathrm{NT}}\gamma^1_{\mathrm{NT}} - \pi_{\mathrm{AT}}\gamma^0_{\mathrm{AT}}. \qquad (22)
\end{aligned}$$

Proof: (21) follows from the definition of $\mathrm{ACE}_{\mathrm{global}}$ and the observation that $p_{y_1,x_1|z_1} = \pi_{\mathrm{CO}}\gamma^1_{\mathrm{CO}} + \pi_{\mathrm{AT}}\gamma^1_{\mathrm{AT}}$ and $p_{y_1,x_0|z_0} = \pi_{\mathrm{CO}}\gamma^0_{\mathrm{CO}} + \pi_{\mathrm{NT}}\gamma^0_{\mathrm{NT}}$. The proof of (22) is similar. □

PROPOSITION 6. *For a given feasible π_X and $p(y, x \mid z)$, the compatible distribution which minimizes* [*maximizes*] $\mathrm{ACE}_{\mathrm{global}}$ *has*

Group	ACE Lower Bound	ACE Upper Bound
NT	$0 - u\gamma^0_{\rm NT}(\pi_X)$	$1 - l\gamma^0_{\rm NT}(\pi_X)$
CO	$l\gamma^1_{\rm CO}(\pi_X) - u\gamma^0_{\rm CO}(\pi_X)$ $= \gamma^1_{\rm CO}(\pi_X, u\gamma^1_{\rm AT}(\pi_X))$ $\quad -\gamma^0_{\rm CO}(\pi_X, l\gamma^0_{\rm NT}(\pi_X))$	$u\gamma^1_{\rm CO}(\pi_X) - l\gamma^0_{\rm CO}(\pi_X)$ $= \gamma^1_{\rm CO}(\pi_X, l\gamma^1_{\rm AT}(\pi_X))$ $\quad -\gamma^0_{\rm CO}(\pi_X, u\gamma^0_{\rm NT}(\pi_X))$
DE	$l\gamma^1_{\rm DE}(\pi_X) - u\gamma^0_{\rm DE}(\pi_X)$ $= \gamma^1_{\rm DE}(\pi_X, u\gamma^1_{\rm AT}(\pi_X))$ $\quad -\gamma^0_{\rm DE}(\pi_X, l\gamma^0_{\rm NT}(\pi_X))$	$u\gamma^1_{\rm DE}(\pi_X) - l\gamma^0_{\rm DE}(\pi_X)$ $= \gamma^1_{\rm DE}(\pi_X, l\gamma^1_{\rm AT}(\pi_X))$ $\quad -\gamma^0_{\rm DE}(\pi_X, u\gamma^0_{\rm NT}(\pi_X))$
AT	$l\gamma^1_{\rm AT}(\pi_X) - 1$	$u\gamma^1_{\rm AT}(\pi_X) - 0$
global	$p_{y_1,x_1\|z_1} - p_{y_1,x_0\|z_0}$ $\quad + \pi_{\rm DE} \cdot l{\rm ACE}_{\rm DE}(\pi_X) - \pi_{\rm AT}$ $= p_{y_1,x_1\|z_0} - p_{y_1,x_0\|z_1}$ $\quad + \pi_{\rm CO} \cdot l{\rm ACE}_{\rm CO}(\pi_X) - \pi_{\rm AT}$	$p_{y_1,x_1\|z_1} - p_{y_1,x_0\|z_0}$ $\quad + \pi_{\rm DE} \cdot u{\rm ACE}_{\rm DE}(\pi_X) + \pi_{\rm NT}$ $= p_{y_1,x_1\|z_0} - p_{y_1,x_0\|z_1}$ $\quad + \pi_{\rm CO} \cdot u{\rm ACE}_{\rm CO}(\pi_X) + \pi_{\rm NT}$

Table 4. Upper and Lower bounds on average causal effects for different groups, as a function of a feasible π_X. Here $\pi^c_{\rm NT} \equiv 1 - \pi_{\rm NT}$

$$\begin{aligned} \langle \gamma^0_{\rm NT}, \gamma^1_{\rm AT} \rangle &= \langle l\gamma^0_{\rm NT}, u\gamma^1_{\rm AT} \rangle \quad [\langle u\gamma^0_{\rm NT}, l\gamma^1_{\rm AT} \rangle] \\ \langle \gamma^1_{\rm NT}, \gamma^0_{\rm AT} \rangle &= \langle 0, 1 \rangle \quad [\langle 1, 0 \rangle] \end{aligned}$$

thus also minimizes [*maximizes*] $\text{ACE}_{\rm CO}$ *and* $\text{ACE}_{\rm DE}$*, and conversely maximizes* [*minimizes*] $\text{ACE}_{\rm AT}$ *and* $\text{ACE}_{\rm NT}$.

Proof: The claims follow from equations (21) and (22), together with the fact that $\gamma^0_{\rm AT}$ and $\gamma^1_{\rm NT}$ are unconstrained, so $\text{ACE}_{\rm global}$ is minimized by taking $\gamma^0_{\rm AT} = 1$ and $\gamma^1_{\rm NT} = 0$, and maximized by taking $\gamma^0_{\rm AT} = 0$ and $\gamma^1_{\rm NT} = 1$. □

It is of interest here that although the definition of $\text{ACE}_{\rm global}$ treats the four compliance types symmetrically, the compatible distribution which minimizes [maximizes] this quantity (for a given π_X) does not: it always corresponds to the scenario in which the treatment has the smallest [greatest] effect on Compliers and Defiers.

The bounds on the global ACE for the flu vaccine data of [Hirano, Imbens, Rubin, and Zhou 2000] are shown are shown in Figure 13.

Finally we note that it would be simple to develop similar bounds for other measures such as the Causal Relative Risk and Causal Odds Ratio.

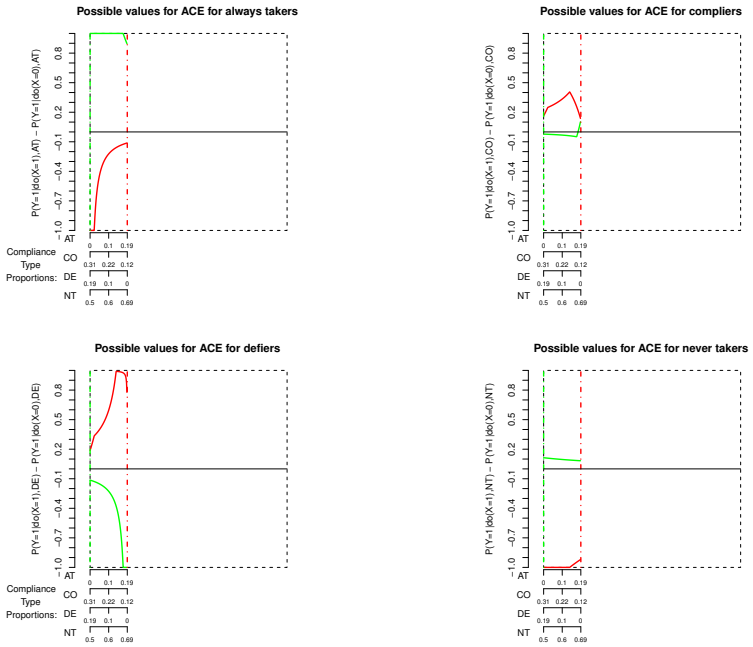


Figure 12. Depiction of the set of values for π_X vs. $\mathrm{ACE}_{\mathbf{t}_X}(\pi_X)$ for $\mathbf{t}_X \in \mathbb{D}_X$ for the flu data.

6 Instrumental inequalities

The expressions involved in the upper bound on π_{AT} in (16) appear similar to those which occur in Pearl's instrumental inequalities. Here we show that the requirement that $\mathcal{P}_X \neq \emptyset$, or equivalently, $l\pi_{\mathrm{AT}} \leq u\pi_{\mathrm{AT}}$ is in fact equivalent to the instrumental inequality. This also provides an interpretation as to what may be inferred from the violation of a specific inequality.

THEOREM 7. *The following conditions place equivalent restrictions on* $p(x \mid z)$ *and* $p(y \mid x{=}0, z)$*:*

(a1) $\max\{0,\ p(x=1 \mid z=0) + p(x=1 \mid z=1) - 1\} \leq$
$\min\left\{1 - \sum_j p(y{=}j, x{=}0 \mid z{=}j),\ 1 - \sum_k p(y{=}k, x{=}0 \mid z{=}1{-}k)\right\}$;

(a2) $\max\left\{\sum_j p(y{=}j, x{=}0 \mid z{=}j),\ \sum_k p(y{=}k, x{=}0 \mid z{=}1-k)\right\} \leq 1$.

Similarly, the following place equivalent restrictions on $p(x \mid z)$ *and* $p(y \mid x{=}1, z)$*:*

(b1) $\max\{0, p(x=1 \mid z=0) + p(x=1 \mid z=1) - 1\} \leq$
$\min\left\{\sum_j p(y{=}j, x{=}1 \mid z{=}j),\ \sum_k p(y{=}k, x{=}1 \mid z{=}1{-}k)\right\}$;

(b2) $\max\left\{\sum_j p(y{=}j, x{=}1 \mid z{=}j),\ \sum_k p(y{=}k, x{=}1 \mid z{=}1-k)\right\} \leq 1$.

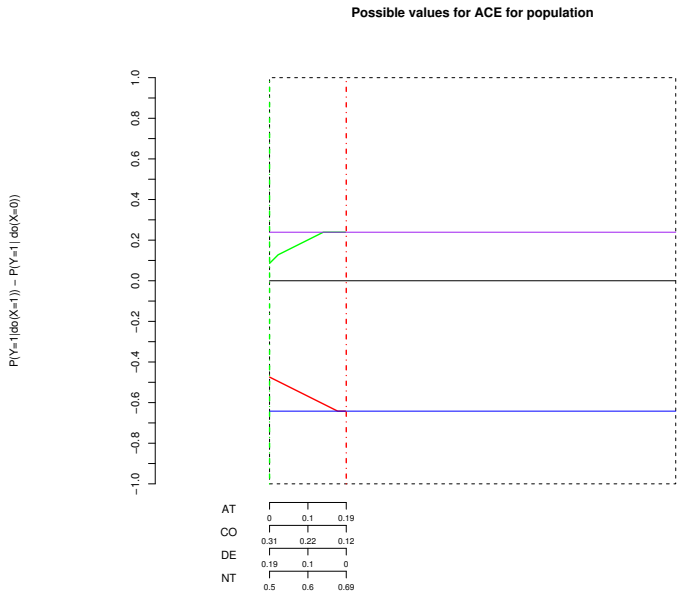


Figure 13. Depiction of the set of values for π_X vs. the global ACE for the flu data. The horizontal lines represent the overall bounds on the global ACE due to Pearl.

Thus the instrumental inequality (a2) corresponds to the requirement that the upper bounds on $p(\mathrm{AT})$ resulting from $p(x \mid z)$ and $p(y=1 \mid x=0,z)$ be greater than the lower bound on $p(\mathrm{AT})$ (derived solely from $p(x \mid z)$). Similarly for (b2) and the upper bounds on $p(\mathrm{AT})$ resulting from $p(y=1 \mid x=1,z)$.

Proof: [(a1) $\Leftrightarrow$ (a2)] We first note that:

$$\begin{aligned} & 1-\textstyle\sum_j p(y=j,x=0 \mid z=j) \geq \left(\sum_j p(x=1 \mid z=j)\right) - 1 \\ \Leftrightarrow \quad & \textstyle\sum_j \left(1-p(y=j,x=0 \mid z=j)\right) \geq \sum_j p(x=1 \mid z=j) \\ \Leftrightarrow \quad & \textstyle\sum_j \left(p(y=1-j,x=0 \mid z=j) + p(x=1 \mid z=j)\right) \geq \sum_j p(x=1 \mid z=j) \\ \Leftrightarrow \quad & \textstyle\sum_j p(y=j,x=0 \mid z=j) \geq 0. \end{aligned}$$

which always holds. By a symmetric argument we can show that it always holds that:

$$1-\textstyle\sum_j p(y=j,x=0 \mid z=1-j) \geq \left(\sum_j p(x=1 \mid z=j)\right) - 1.$$

Thus if (a1) does not hold then $\max\{0,\ p(x=1 \mid z=0) + p(x=1 \mid z=1) - 1\} = 0$. It is then simple to see that (a1) does not hold iff (a2) does not hold.

[(b1) $\Leftrightarrow$ (b2)] It is clear that neither of the sums on the RHS of (b1) are negative, hence if (b1) does not hold then $\max\{0, p(x=1 \mid z=0) + p(x=1 \mid z=1) - 1\} =$

$\left(\sum_j p(x=1 \mid z=j)\right) - 1$. Now

$$\begin{aligned} &\textstyle\sum_j p(y=j, x=1 \mid z=j) < \left(\sum_j p(x=1 \mid z=j)\right) - 1 \\ \Leftrightarrow \quad &1 < \textstyle\sum_j p(y=j, x=1 \mid z=1-j). \end{aligned}$$

Likewise

$$\begin{aligned} &\textstyle\sum_j p(y=j, x=1 \mid z=1-j) < \left(\sum_j p(x=1 \mid z=j)\right) - 1 \\ \Leftrightarrow \quad &1 < \textstyle\sum_j p(y=j, x=1 \mid z=j). \end{aligned}$$

Thus (b1) fails if and only if (b2) fails. □

This equivalence should not be seen as surprising since [Bonet 2001] states that the instrument inequalities (a2) and (b2) are sufficient for a distribution to be compatible with the binary IV model. This is not the case if, for example, X takes more than 2 states.

6.1 Which alternatives does a test of the instrument inequalities have power against?

[Pearl 2000] proposed testing the instrument inequalities (a2) and (b2) as a means of testing the IV model; [Ramsahai 2008] develops tests and analyzes their properties. It is then natural to ask what should be inferred from the failure of a specific instrumental inequality. It is, of course, always possible that randomization has failed. If randomization is not in doubt, then the exclusion restriction (1) must have failed in some way. The next result implies that tests of the inequalities (a2) and (b2) have power, respectively, against failures of the exclusion restriction for Never Takers (with $X = 0$) and Always Takers (with $X = 1$):

THEOREM 8. *The conditions* (RX), ($\mathrm{RY}_{X=0}$) *and* ($\mathrm{E}_{X=0}$) *described below imply* (a2)*; similarly* (RX), ($\mathrm{RY}_{X=1}$) *and* ($\mathrm{E}_{X=1}$) *imply* (b2).

(RX) $\quad Z \perp\!\!\!\perp \mathfrak{t}_X$ *equivalently* $Z \perp\!\!\!\perp X_{z=0}, X_{z=1}$:

($\mathrm{RY}_{X=0}$) $\quad Z \perp\!\!\!\perp Y_{x=0,z=0} \mid \mathfrak{t}_X = \mathrm{NT}; \quad Z \perp\!\!\!\perp Y_{x=0,z=1} \mid \mathfrak{t}_X = \mathrm{NT};$

($\mathrm{RY}_{X=1}$) $\quad Z \perp\!\!\!\perp Y_{x=1,z=1} \mid \mathfrak{t}_X = \mathrm{AT}; \quad Z \perp\!\!\!\perp Y_{x=1,z=1} \mid \mathfrak{t}_X = \mathrm{AT};$

($\mathrm{E}_{X=0}$) $\quad p(Y_{x=0,z=0} = Y_{x=0,z=1} \mid \mathfrak{t}_X = \mathrm{NT}) = 1;$

($\mathrm{E}_{X=1}$) $\quad p(Y_{x=1,z=0} = Y_{x=1,z=1} \mid \mathfrak{t}_X = \mathrm{AT}) = 1.$

Conditions (RX) and ($\mathrm{RY}_{X=x}$) correspond to the assumption of randomization with respect to compliance type and response type. For the purposes of technical clarity we have stated condition ($\mathrm{RY}_{X=x}$) in the weakest form possible. However, we know of no subject matter knowledge which would lead one to believe that (RX) and ($\mathrm{RY}_{X=x}$) held, without also implying the stronger assumption (2). In contrast, the exclusion restrictions ($\mathrm{E}_{X=x}$) are significantly weaker than (1), e.g. one could conceive of situations where assignment had an effect on the outcome for Always

Takers, but not for Compliers. It should be noted that tests of the instrument inequalities have no power to detect failures of the exclusion restriction for Compliers or defier.

We first prove the following Lemma, which also provides another characterization of the instrument inequalities:

LEMMA 9. *Suppose* (RX) *holds and* $Y \perp\!\!\!\perp Z \mid \mathfrak{t}_X = \text{NT}$ *then* (a2) *holds. Similarly, if* (RX) *holds and* $Y \perp\!\!\!\perp Z \mid \mathfrak{t}_X = \text{AT}$ *then* (b2) *holds.*

Note that the conditions in the antecedent make no assumption regarding the existence of counterfactuals for Y.

Proof: We prove the result for Never Takers; the other proof is similar. By hypothesis we have:

$$p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \text{NT}) = p(Y = 1 \mid Z = 1, \mathfrak{t}_X = \text{NT}) \equiv \gamma^0_{\text{NT}}. \tag{23}$$

In addition,

$$\begin{aligned}
&p(Y = 1 \mid Z = 0, X = 0) \\
&\quad = \quad p(Y = 1 \mid Z = 0, X = 0, X_{z=0} = 0) \\
&\quad = \quad p(Y = 1 \mid Z = 0, X_{z=0} = 0) \\
&\quad = \quad p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \text{CO})\, p(\mathfrak{t}_X = \text{CO} \mid Z = 0, X_{z=0} = 0) \\
&\qquad\quad + p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \text{NT})\, p(\mathfrak{t}_X = \text{NT} \mid Z = 0, X_{z=0} = 0) \\
&\quad = \quad p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \text{CO})\, p(\mathfrak{t}_X = \text{CO} \mid X_{z=0} = 0) \\
&\qquad\quad + \gamma^0_{\text{NT}}\, p(\mathfrak{t}_X = \text{NT} \mid X_{z=0} = 0).
\end{aligned} \tag{24}$$

The first three equalities here follow from consistency, the definition of the compliance types and the law of total probability. The final equality uses (RX). Similarly, it may be shown that

$$\begin{aligned}
&p(Y = 1 \mid Z = 1, X = 0) \\
&\quad = \quad p(Y = 1 \mid Z = 1, \mathfrak{t}_X = \text{DE}) p(\mathfrak{t}_X = \text{DE} \mid X_{z=1} = 0) \\
&\qquad\quad + \gamma^0_{\text{NT}}\, p(\mathfrak{t}_X = \text{NT} \mid X_{z=1} = 0).
\end{aligned} \tag{25}$$

Equations (24) and (25) specify two averages of three quantities, thus taking $u = p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \text{CO})$, $v = p(Y = 1 \mid Z = 1, \mathfrak{t}_X = \text{DE})$ and $w = \gamma^0_{\text{NT}}$, we may apply the analysis of §2.3. This then leads to the upper bound on π_{AT} given by equation (15). (Note that the lower bounds on π_{AT} are derived from $p(x \mid z)$ and hence are unaffected by dropping the exclusion restriction.) The requirement that there exist some feasible distribution π_X then implies equation (a2) which is shown in Theorem 7 to be equivalent to (b2) as required. □

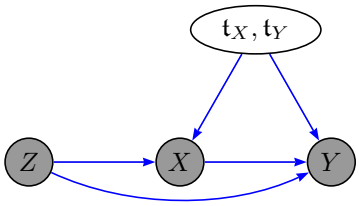


Figure 14. Graphical representation of the model given by the randomization assumption (2) alone. It is no longer assumed that Z does not have a direct effect on Y.

Proof of Theorem 8: We establish that (RX), ($\text{RY}_{X=0}$), ($\text{E}_{X=0}$) $\Rightarrow$ (a2). The proof of the other implication is similar. By Lemma 9 it is sufficient to establish that $Y \perp\!\!\!\perp Z \mid \mathfrak{t}_X = \text{NT}$.

$$
\begin{aligned}
& p(Y=1 \mid Z=0, \mathfrak{t}_X = \text{NT}) \\
&\quad = \quad p(Y=1 \mid Z=0, X=0, \mathfrak{t}_X = \text{NT}) && \text{definition of NT;} \\
&\quad = \quad p(Y_{x=0,z=0}=1 \mid Z=0, X=0, \mathfrak{t}_X = \text{NT}) && \text{consistency;} \\
&\quad = \quad p(Y_{x=0,z=0}=1 \mid Z=0, \mathfrak{t}_X = \text{NT}) && \text{definition of NT;} \\
&\quad = \quad p(Y_{x=0,z=0}=1 \mid \mathfrak{t}_X = \text{NT}) && \text{by } (\text{RY}_{X=0}); \\
&\quad = \quad p(Y_{x=0,z=1}=1 \mid \mathfrak{t}_X = \text{NT}) && \text{by } (\text{E}_{X=0}); \\
&\quad = \quad p(Y_{x=0,z=1}=1 \mid Z=1, \mathfrak{t}_X = \text{NT}) && \text{by } (\text{RY}_{X=0}); \\
&\quad = \quad p(Y=1 \mid Z=1, \mathfrak{t}_X = \text{NT}) && \text{consistency, NT.}
\end{aligned}
$$

□

A similar result is given in [Cai, Kuroki, Pearl, and Tian 2008], who consider the *Average Controlled Direct Effect*, given by:

$$\text{ACDE}(x) \equiv p(Y_{x,z=1}=1) \; - \; p(Y_{x,z=0}=1),$$

under the model given solely by the equation (2), which corresponds to the graph in Figure 14. Cai *et al.* prove that under this model the following bounds obtain:

$$\text{ACDE}(x) \quad \geq \quad p(y=0, x \mid z=0) + p(y=1, x \mid z=1) - 1, \tag{26}$$

$$\text{ACDE}(x) \quad \leq \quad 1 - p(y=0, x \mid z=1) - p(y=1, x \mid z=0). \tag{27}$$

It is simple to see that ACDE(x) will be bounded away from 0 for some x iff one of the instrumental inequalities is violated. This is as we would expect: the IV model of Figure 1 is a sub-model of Figure 14, but if ACDE(x) is bounded away from 0 then the $Z \rightarrow Y$ edge is present, and hence the exclusion restriction (1) is incompatible with the observed distribution.

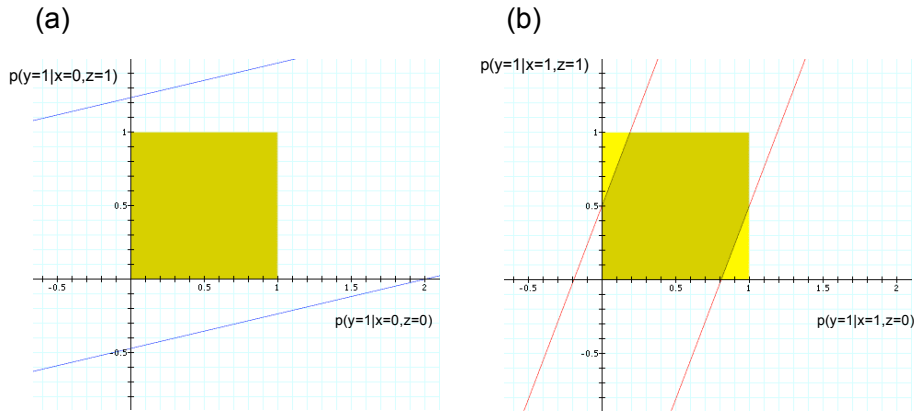


Figure 15. Illustration of the possible values for $p(y \mid x, z)$ compatible with the instrument inequalities, for a given distribution $p(x|z)$.The darker shaded region satisfies the inequalities: (a) $X = 0$, inequalities (a2); (b) $X = 1$, inequalities (b2). In this example $p(x = 1 \mid z = 0) = 0.84$, $p(x = 1 \mid z = 1) = 0.32$. Since $0.84/(1 - 0.32) > 1$, (a2) is trivially satisfied; see proof of Theorem 10.

6.2 How many instrument inequalities may be violated by a single distribution?

THEOREM 10. *For any distribution $p(x, y \mid z)$, at most one of the four instrument inequalities:*

(a2.1) $\sum_j p(y{=}j, x{=}0 \mid z{=}j) \leq 1$; (a2.2) $\sum_j p(y{=}j, x{=}0 \mid z{=}1{-}j) \leq 1$;

(b2.1) $\sum_j p(y{=}j, x{=}1 \mid z{=}j) \leq 1$; (b2.2) $\sum_j p(y{=}j, x{=}1 \mid z{=}1{-}j) \leq 1$;

is violated.

Proof: We first show that at most one of (a2.1) and (a2.2) may be violated. Letting $\theta_{ij} = p(y = 1 \mid x = j, z = i)$ we may express these inequalities as:

$$\begin{aligned} \theta_{10} \cdot p_{x_0|z_1} - \theta_{00} \cdot p_{x_0|z_0} &\leq p_{x_1|z_0}, & \text{(a2.1)} \\ \theta_{10} \cdot p_{x_0|z_1} - \theta_{00} \cdot p_{x_0|z_0} &\geq -p_{x_1|z_1}, & \text{(a2.2)} \end{aligned}$$

giving two half-planes in $(\theta_{00}, \theta_{10})$-space (see Figure 15(a)). Since the lines defining the half-planes are parallel, it is sufficient to show that the half-planes always intersect, and hence that the regions in which (a2.1) and (a2.2) are violated are disjoint. However, this is immediate since the (non-empty) set of points for which $\theta_{10} \cdot p_{x_0|z_1} - \theta_{00} \cdot p_{x_0|z_0} = 0$ always satisfy both inequalities.

The proof that at most one of (b2.1) and (b2.2) may be violated is symmetric.

We now show that the inequalities (a2.1) and (a2.2) place non-trivial restrictions on $(\theta_{00}, \theta_{10})$ iff (b2.1) and (b2.2) place trivial restrictions on $(\theta_{01}, \theta_{11})$. The line corresponding to (a2.1) passes through $(\theta_{00}, \theta_{10}) = (-p_{x_1|z_0}/p_{x_0|z_0}, 0)$ and

$(0, p_{x_1|z_0}/p_{x_0|z_1})$; since the slope of the line is non-negative, it has non-empty intersection with $[0,1]^2$ iff $p_{x_1|z_0}/p_{x_0|z_1} \leq 1$. Thus there are values of $(\theta_{01}, \theta_{11}) \in [0,1]^2$ which fail to satisfy (a2.1) iff $p_{x_1|z_0}/p_{x_0|z_1} < 1$. By a similar argument it may be shown that (a2.2) is non-trivial iff $p_{x_1|z_1}/p_{x_0|z_0} < 1$, which is equivalent to $p_{x_1|z_0}/p_{x_0|z_1} < 1$.

The proof is completed by showing that (b2.1) and (b2.2) are non-trivial if and only if $p_{x_1|z_0}/p_{x_0|z_1} > 1$. □

COROLLARY 11. *Every distribution $p(x, y \mid z)$ is consistent with randomization* (RX) *and* (2)*, and at least one of the exclusion restrictions* $\mathrm{E}_{X=0}$ *or* $\mathrm{E}_{X=1}$.

Flu Data Revisited

For the data in Table 3, all of the instrument inequalities hold. Consequently there is no evidence of a direct effect of Z on Y. (Again we emphasize that unlike [Hirano, Imbens, Rubin, and Zhou 2000], we are not using any information on baseline covariates in the analysis.) Finally we note that, since all of the instrumental inequalities hold, maximum likelihood estimates for the distribution $p(x, y \mid z)$ under the IV model are given by the empirical distribution. However, if one of the IV inequalities were to be violated then the MLE would not be equal to the empirical distribution, since the latter would not be a law within the IV model. In such a circumstance a fitting procedure would be required; see [Ramsahai 2008, Ch. 5].

7 Conclusion

We have built upon and extended the work of Pearl, displaying how the range of possible distributions over types compatible with a given observed distribution may be characterized and displayed geometrically. Pearl's bounds on the global ACE are sometimes objected to on the grounds that they are too extreme, since for example, the upper bound presupposes a 100% success rate among Never Takers if they were somehow to receive treatment, likewise a 100% failure rate among Always Takers were they not to receive treatment. Our analysis provides a framework for performing a sensitivity analysis. Lastly, our analysis relates the IV inequalities to the bounds on direct effects.

Acknowledgements

This research was supported by the U.S. National Science Foundation (CRI 0855230) and U.S. National Institutes of Health (R01 AI032475) and Jesus College, Oxford where Thomas Richardson was a Visiting Senior Research Fellow in 2008. The authors used Avitzur's *Graphing Calculator* software (`www.pacifict.com`) to construct two and three dimensional plots. We thank McDonald, Hiu and Tierney for giving us permission to use their flu vaccine data.

References

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*, 1171–1176.

Bonet, B. (2001). Instrumentality tests revisited. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pp. 48–55.

Cai, Z., M. Kuroki, J. Pearl, and J. Tian (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics 64*, 695–701.

Chickering, D. and J. Pearl (1996). A clinician's tool for analyzing non-compliance. In *AAAI-96 Proceedings*, pp. 1269–1276.

Erosheva, E. A. (2005). Comparing latent structures of the Grade of Membership, Rasch, and latent class models. *Psychometrika 70*, 619–628.

Fienberg, S. E. and J. P. Gilbert (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association 65*, 694–701.

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics 1*(1), 69–88.

Manski, C. (1990). Non-parametric bounds on treatment effects. *American Economic Review 80*, 351–374.

McDonald, C., S. Hiu, and W. Tierney (1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD Computing 9*, 304–312.

Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.

Ramsahai, R. (2008). *Causal Inference with Instruments and Other Supplementary Variables*. Ph. D. thesis, University of Oxford, Oxford, UK.

Robins, J. (1989). 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*. Washington, D.C.: U.S. Public Health Service.

Robins, J. and A. Rotnitzky (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika 91*(4), 763–783.

26

Pearl Causality and the Value of Control

Ross Shachter and David Heckerman

1 Introduction

We welcome this opportunity to acknowledge the significance of Judea Pearl's contributions to uncertain reasoning and in particular to his work on causality. In the decision analysis community causality had long been "taboo" even though it provides a natural framework to communicate with decision makers and experts [Shachter and Heckerman 1986]. Ironically, while many of the concepts and methods of causal reasoning are foundational to decision analysis, scholars went to great lengths to avoid causal terminology in their work. Judea Pearl's work is helping to break this barrier, allowing the exploration of some fundamental principles. We were inspired by his work to understand exactly what assumptions are being made in his causal models, and we would like to think that our subsequent insights have contributed to his and others' work as well.

In this paper, we revisit our previous work on how a decision analytic perspective helps to clarify some of Pearl's notions, such as those of the *do* operator and *atomic intervention*. In addition, we show how influence diagrams [Howard and Matheson 1984] provide a general graphical representation for cause. Decision analysis can be viewed simply as determining what interventions we want to make in the world to improve the prospects for us and those we care about, an inherently causal concept. As we shall discuss, causal models are naturally represented within the framework of decision analysis, although the causal aspects of issues about counterfactuals and causal mechanisms that arise in computing the value of clairvoyance [Howard 1990], were first presented by Heckerman and Shachter [1994, 1995]. We show how this perspective helps clarify decision-analytic measures of sensitivity, such as the value of control and the value of revelation [Matheson 1990; Matheson and Matheson 2005].

2 Decision-Theoretic Foundations

In this section we introduce the relevant concepts from [Heckerman and Shachter 1995], the framework for this paper, along with some extensions to those concepts.

Our approach rests on a simple but powerful primitive concept of *unresponsiveness*. An uncertain variable is unresponsive to a set of decisions if its value is unaffected by our choice for the decisions. It is unresponsive to those decisions in worlds limited by other variables if the decisions cannot affect the uncertain variable without also changing one of the other variables.

We can formalize this by introducing concepts based on Savage [1954]. We consider three different kinds of distinctions, which he called *acts*, *consequences*, and *possible states of the world*. We have complete control over the acts but no control over the uncertain state of the world. We might have some level of control over consequences, which are logically determined, after we act, by the state of the world. Therefore, a consequence can be represented as a deterministic function of acts and the state of the world, inheriting uncertainty from the state of the world while affected, more or less, by our choice of action.

In practice, it is convenient to represent acts and consequences with variables in our model. We call a variable describing a set of mutually exclusive and collectively exhaustive acts a *decision*, and we denote the set of decisions by D. We call a variable describing a consequence *uncertain*, and we denote the set of uncertain variables by U. At times we will distinguish between the uncertain variables that serve as our objectives or *value variables*, V, and the other uncertain variables which we call *chance variables*, $C = U \setminus V$. Finally, in this section we will use the variables S to represent the possible states of the world. As a convention we will refer to single variables with lower-case (x or d), sets of variables with upper-case (D or V), and particular instances of variables with bold ($\mathbf{x}$ or $\mathbf{D}$). In this notation, the set of uncertain variables X takes value $X[\mathbf{S}, \mathbf{D}]$ deterministically when $\mathbf{D}$ is chosen and $\mathbf{S}$ is the state of the world.

DEFINITION 1 (Unresponsiveness). Given a decision problem described by uncertain variables U, decision variables D, and state of the world S, and variable sets $X \subseteq U$ and $Y \subseteq D \cup U$, X is said to be *unresponsive to* D, denoted $X \not\rightsquigarrow D$, if we believe that

$$\forall \mathbf{S} \in S, \mathbf{D_1} \in D, \mathbf{D_2} \in D : X[\mathbf{S}, \mathbf{D_1}] = X[\mathbf{S}, \mathbf{D_2}]$$

and, if not, X is said to be *responsive to* D.
Furthermore, X is said to be *unresponsive to* D *in worlds limited by* Y, denoted $X \not\rightsquigarrow_Y D$, if we believe that

$$\forall \mathbf{S} \in S, \mathbf{D_1} \in D, \mathbf{D_2} \in D : Y[\mathbf{S}, \mathbf{D_1}] = Y[\mathbf{S}, \mathbf{D_2}] \Longrightarrow X[\mathbf{S}, \mathbf{D_1}] = X[\mathbf{S}, \mathbf{D_2}]$$

and, if not, X is said to be *responsive to* D *in worlds limited by* Y.

The distinctions of unresponsiveness and limited unresponsiveness seem natural for decision makers to consider. Unresponsiveness is related to independence, in that any uncertain variables X that are unresponsive to decisions D are independent of D. Although it is not necessarily the case that X independent of D is unresponsive to D, that implication is often assumed [Spirtes, Glymour, and Scheines 1993]. In contrast, there is no such general correspondence between limited unresponsiveness and conditional independence.

To illustrate these concepts graphically, we introduce influence diagrams [Howard and Matheson 1984]. An *influence diagram* is an acyclic directed graph G with

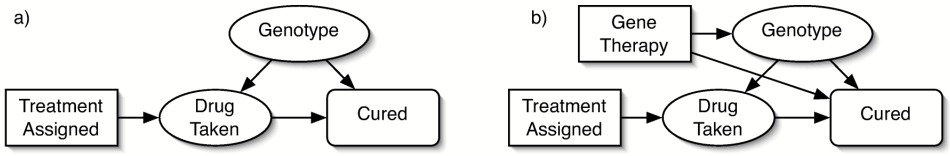


Figure 1. The treatment assignment only cures the patient if it affects whether the drug is taken, but genotype does not have a causal effect unless it is responsive to decisions.

nodes corresponding to the variables, rectangles for decisions, ovals for chance variables, and rounded rectangles for value variables. Arcs into chance and value nodes, are *conditional*. For each uncertain variable x there is a conditional probability distribution for x given its parents, $Pa(x)$. If the distribution is a deterministic function, we represent that in the graph by a double oval or double rounded rectangle. Arcs into decisions are *informational*, representing that the parent variables will be observed before the decision is made. Although there are significant issues involving informational arcs, we will focus primarily on models in which there are no informational arcs and all of the decisions could be made in any order, before any of the uncertain variables are observed.

We allow multiple value nodes, all with no children, assuming that their values will be summed. We assume that the criterion for making decisions is either the total value or an increasing exponential utility function of the total. This simplifies the valuation of a proposed change to a decision problem because the most a decision maker should be willing to pay for the change is the difference in the values of the diagrams with and without the proposed change.

Although we have defined unresponsiveness without regard to a graphical representation, there is an intuitive graphical interpretation (with some technical exceptions described in Heckerman and Shachter [1985]). The uncertain descendants of decisions are usually responsive to them, and the other uncertain variables are usually unresponsive. Also, X is usually unresponsive to D in worlds limited by Y if all of the directed paths from D to X include nodes in Y. When these rules of thumb are all satisfied, we say that an influence diagram is causal.

DEFINITION 2 (Causal Influence Diagram). An influence diagram with graph G and decision nodes D, chance nodes C, and value nodes V, is said to be *causal* if we believe that uncertain variables $X \subseteq C \cup V$ are unresponsive to decisions D, $X \not\rightharpoonup D$, whenever there is no directed path from D to X, and X is unresponsive to decisions D in worlds limited by Y, $X \not\rightharpoonup_Y D$, whenever every directed path from D to X includes a node from Y.

Consider the influence diagram shown in Figure 1a which we believe is causal. In this case, we believe that *Drug Taken* and *Cured* are responsive to *Treatment Assigned* while *Genotype* is unresponsive to *Treatment Assigned*. We also believe

that *Cured* is unresponsive to *Treatment Assigned* in worlds limited by *Drug Taken*. Note that *Treatnent Assigned* is *not* independent of *Genotype* or *Cured* given *Drug Taken*.

The concept of limited unresponsiveness allows us to define how one variable can cause another in a way that is natural for decision makers to understand.

DEFINITION 3 (Cause with Respect to Decisions). Given a decision problem described by uncertain variables U and decision variables D, and a variable $x \in U$, the set of variables $Y \subseteq D \cup U \setminus \{x\}$ is said to be a *cause for x with respect to D* if Y is a minimal set of variables such that $x \not\rightharpoonup_Y D$.

Defining cause with respect to a particular set of decisions adds clarity. Consider again the causal influence diagram shown in Figure 1a. With respect to the decision *Treatment Assigned*, the cause of *Cured* is either {*Treament Assigned*} or {*Drug Taken*}, while the cause of *Genotype* is {}. Because we believe that *Genotype* is unresponsive to *Treatment Assigned* it has no cause with respect to D. On the other hand, we believe that *Cured* is responsive to *Treatment Assigned* but not in worlds limited by *Drug Taken*, so {*Drug Taken*} is a cause of *Cured* with respect to D.

Consider now the causal influence diagram shown in Figure 1b, in which we have added the decision *Gene Therapy*. Because *Genotype* is now responsive to D, the cause of *Genotype* is {*Gene Therapy*} with respect to D. If the gene therapy has some side effect on whether the patient is cured, then {*Gene Therapy*, *Drug Taken*} but not {*Genotype*, *Drug Taken*} would be a cause of *Cured* with respect to the decisions, because *Cured* is unresponsive to D in worlds limited by the former but not the latter.

The concept of limited unresponsiveness also allows us to formally define direct and atomic interventions. A set of decision I is a direct intervention on a set of uncertain variables X if the effects of I on all other uncertain variables are mediated through their effects on X.

DEFINITION 4 (Direct Intervention). Given a decision problem described by uncertain variables U and decision variables D, a set of decisions $I \subseteq D$ is said to be a *direct intervention on $X \subseteq U$ with respect to D* if (1) $x \leftarrow\!\!\!\shortmid I$ for all $x \in X$, and (2) $y \not\rightharpoonup_X I$ for all $y \in U$.

In a causal influence diagram every node in I has children only in X and there is a directed path from I to every node in X. In the causal influence diagram shown in Figure 1b, *Treatment Assigned* is a direct intervention on *Drug Taken*, and the set of decisions is a direct intervention on all three uncertain variables. Note that whether a decision is a direct intervention depends on the underlying causal mechanism. If the gene therapy had no side effect then *Gene Therapy* would be a direct intervention on *Genotype*, but regardless whether there is a side effect, *Gene Therapy* is a direct intervention on {*Genotype*, *Cured*}.

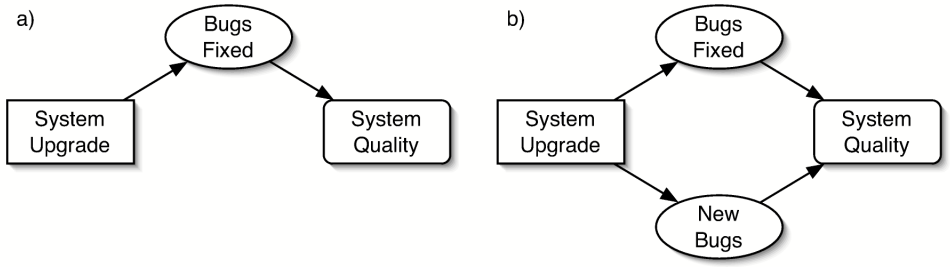


Figure 2. We believe that a system upgrade will affect system quality by fixing bugs unless new bugs are introduced in the process.

DEFINITION 5 (Atomic Intervention). Given a decision problem described by uncertain variables U and decision variables D, a decision $do(x) \in D$ is said to be a *atomic intervention on* $x \in U$ *with respect to* D if (1) $do(x)$ is a direct invention on x with respect to D, and (2) $do(x)$ has precisely the instances (a) **idle**, which corresponds to no intervention, and (b) $\mathbf{do(x)}$ for every instance $\mathbf{x}$ of x, where $x = \mathbf{x}$ whenever $do(x) = \mathbf{do(x)}$.

This is precisely the atomic intervention described without definition in Pearl [1993]. The assumptions underlying it are quite strong. The causal influence diagram shown in Figure 2a assumes that we can upgrade our system and improve the quality by fixing the bugs, but the diagram shown in (b) illustrates the all too familiar situation when new bugs are introduced in the process, compromising system quality. In that case, *System Upgrade* is not a direct intervention on *Bugs Fixed* and {*Bugs Fixed*} is not a cause of *System Quality* with respect to D. Although the system upgrade was intended to be an atomic intervention, it can have unintended and undesirable consequences.

We can now represent the relationship between an uncertain variable x and other variables Y, such as its parents in a causal influence diagram. We consider the uncertain function $x(Y)$ as a variable, and now x is a deterministic function of Y and $x(Y)$. In fact, if Y is a cause of x with respect to D then $x(Y)$ must be unresponsive to D.

DEFINITION 6 (Mapping Variable). Given a decision problem described by uncertain variables U and decision variables D, $x \in U$ and variables Y such that for every $y \in Y \cap U$ there exists an atomic intervention $do(y) \in D$, the *mapping variable* $x(Y)$ is the chance variable that represents all possible mappings from Y to x.

Finally, we have developed the machinery to characterize a *Pearl causal model* and structural equations [Pearl 1993]. Given uncertain variables U, suppose the decisions D comprise an atomic intervention $do(x)$ on every $x \in U$. Given a graph

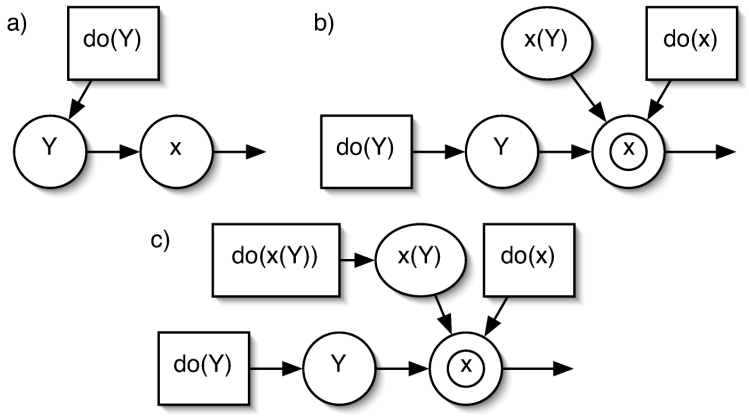


Figure 3. The partial influence diagram for x in a causal model, shown in (a) with parents Y, becomes the diagram shown in (b) explicitly representing the structural equation for x, and, when Y is nonempty, the diagram shown in (c) with an explicit atomic intervention on the mapping variable.

G with nodes U, such that $Pa(x) \cup \{do(x)\}$ is a cause for x with respect to D. Then

$$x = f_x(Pa(x), do(x), x(Pa(x)))$$

for all $x \in U$ where f_x is a deterministic function such that $x = \mathbf{x}$ if $do(x) = \mathbf{do}(\mathbf{x})$.

We can extend this to allow manipulation of a mapping variable for $x \in U$ with parents to obtain a *Pearl causal model with an atomic intervention for mapping variable* $x(Pa(x))$. The decisions D now also include a atomic intervention $do(x(Pa(x)))$. As a result, $Pa(x) \cup \{do(x), do(x(Pa(x)))\}$ is now a cause for x with respect to D and $x(Pa(x)) = \mathbf{x}(Pa(x))$ when $do(x(Pa(x))) = \mathbf{do}(\mathbf{x}(Pa(x)))$.

The causal model is represented by the partial influence diagrams shown in Figure 3 with $Y = Pa(x) \subseteq C$ in the graph G. We assume in (a) that there are atomic interventions $do(y)$ on each $y \in Y$ represented as $do(Y)$. The diagram shown in (b) explicitly represents the structural equation for x as a deterministic function of Y, an atomic intervention, $do(x)$, and the mapping variable, $x(Y)$. The influence diagram is causal, showing that $Y \cup \{do(x)\}$ is a cause for x with respect to D. We can extend the model by adding an atomic intervention for the mapping variable, $do(x(Y))$. If Y is empty then nothing needs to be added, as $do(x)$ is the same atomic intervention as $do(x())$, but otherwise we obtain the diagram shown in (c). Now $Y \cup \{do(x), do(x(Y))\}$ is a cause for x with respect to D.

An influence diagram is said to be in *canonical form* if each uncertain variable responsive to a decision is a descendant of that decision and represented as a deterministic node. Each decision, including atomic interventions, is explicit. Each uncertain variable that is responsive to D is a deterministic function of its parents,

including any decisions that are direct interventions on it, and a mapping variable. As an example, the influence diagram shown in Figure 3b is in canonical form.

In the next section we apply these concepts to define and contrast different measures for the value to a decision maker of manipulating (or observing) an uncertain variable.

3 Value of Control

When assisting a decision maker developing a model, sensitivity analysis measures help the decision maker to validate the model. One popular measure is the *value of clairvoyance*, the most a decision maker should be willing to pay to observe a set of uncertain variables before making particular decisions [Howard 1967]. Our focus of attention is another measure, the value of control (or wizardry), the most a decision maker should be willing to pay a hypothetical wizard to optimally control the distribution of an uncertain variable [Matheson 1990], [Matheson and Matheson 2005]. We consider and contrast the value of control with two other measures, the value of do, and the value of revelation, and we develop the conditions under which the different measures are equal.

In formalizing the value of control, it is natural to consider the value of an atomic intervention on uncertain variable x, in particular $\mathbf{do}(\mathbf{x}^*)$, that would set it to $\mathbf{x}^*$ the instance yielding the most valuable decision situation, rather than to **idle**. We call the most the decision maker should be willing to pay for such an intervention the *value of do* and compute it as the difference in the values of the diagrams.

DEFINITION 7 (Value of Do). Given a decision problem including an atomic intervention on uncertain variable $x \in U$, the *value of do for* x, denoted by $VoD(\mathbf{x}^*)$, is the most one should be willing to pay for an atomic intervention on uncertain variable x to the best possible deterministic instance, $\mathbf{do}(\mathbf{x}^*)$, instead of to **idle**.

Our goal in general is to value the optimal manipulation of the *conditional* distribution of a target uncertain variable x in a causal influence diagram, $P\{x|Y\}$, and the most we should be willing to pay for such an intervention is the *value of control*. The simplest case is when $\{do(x)\}$ is a cause of x with respect to D, $Y = \{\}$, so the optimal distribution is equivalent to an atomic intervention on x to $\mathbf{x}^*$, and control and *do* are the same intervention. Otherwise, the *do* operation effectively severs the arcs from Y to x and replaces the previous causal mechanism with the new atomic one. By contrast, the control operation is an atomic intervention on the mapping variable $x(Y)$ to its optimal value $\mathbf{do}(\mathbf{x}^*(Y))$ rather than to **idle**.

DEFINITION 8 (Value of Control). Given a decision problem including variables Y, a mapping variable $x(Y)$ for uncertain variable $x \in U$, and atomic interventions $do(x)$ and $do(x(Y))$ such that $Y \cup \{do(x), do(x(Y))\}$ is a cause of x with respect to D, the *value of control for* x, denoted by $VoC(\mathbf{x}^*(Y))$, is the most one should be willing to pay for an atomic intervention on the mapping variable for uncertain variable x to the best possible deterministic function of Y, $\mathbf{do}(\mathbf{x}^*(Y))$, instead of

to **idle**.

If $Y = \{\}$, then $do(x)$ is the same atomic intervention as $do(x(Y))$, and the values of do and control for x are equal, $VoD(\mathbf{x}^*) = VoC(\mathbf{x}^*())$.

In many cases, while it is tempting to assume atomic interventions, they can be cumbersome or implausible. In an attempt to avoid such issues, Ronald A. Howard has suggested an alternative passive measure, the *value of revelation*: how much better off the decision maker should be by observing that the uncertain variable in question obtained its most desirable value. This is only well-defined for variables unresponsive to D, except for those atomic interventions that are set to **idle**, because otherwise the observation would be made before decisions it might be responsive to. Under our assumptions this can be computed as the difference in value between two situations, but it is hard to describe it as a willingness to pay for this difference as it is more passive than intentional. (The value of revelation is in fact an intermediate term in the computation of the value of clairvoyance.)

DEFINITION 9 (Value of Revelation). Given a decision problem including uncertain variable $x \in U$ and a (possibly empty) set of atomic interventions, A, that is a cause for x with respect to D, the *value of revelation for* uncertain variable $x \in U$, denoted by $VoR(\mathbf{x}^*)$, is the increase in the value of the situation with $d =$ **idle** for all $d \in A$, if one observed that uncertain variable $x = \mathbf{x}^*$, the best possible deterministic instance, instead of not observing x.

To illustrate these three measures we, consider a partial causal influence diagram including x and its parents, Y, which we assume for this example are uncertain and nonempty, as shown in Figure 4a. There are atomic interventions $do(x)$ on x, $do(x(Y))$ on mapping variable $x(Y)$, and $do(y)$ on each $y \in Y$ represented as $do(Y)$. The variable x is a deterministic function of Y, $do(x)$ and $x(Y)$. In this model, $Y \cup \{do(x), do(x(Y))\}$ is a cause of x with respect to D. The dashed line from x to values V suggests that there might be some directed path from x to V. If not, V would be unresponsive to $do(x)$ and $do(x(Y))$ and the values of do and control would be zero.

To obtain the reference diagram for our proposed changes, we set all of the atomic interventions to **idle** as shown in Figure 4b1. We can compute the value of this diagram by eliminating the idle decisions and absorbing the mapping variable into x, yielding the simpler diagram shown in (b2). To compute the value of do for x, we can compute the value of the diagram with $\mathbf{do}(\mathbf{x}^*)$ by setting the other atomic interventions to **idle**, as shown in (c1). But since that is making the optimal choice for x with no interventions on Y or $x(Y)$, we can now think of x as a decision variable as indicated in the diagram shown in (c2). We shall use this shorthand in many of the examples that we consider. To compute the value of control for x, we can compute the value of the diagram with $\mathbf{do}(\mathbf{x}^*(Y))$ by setting the other atomic interventions to **idle**, as shown in (d1). But since that is making the optimal choice for $x(Y)$ with none of the other interventions, we can compute its value with

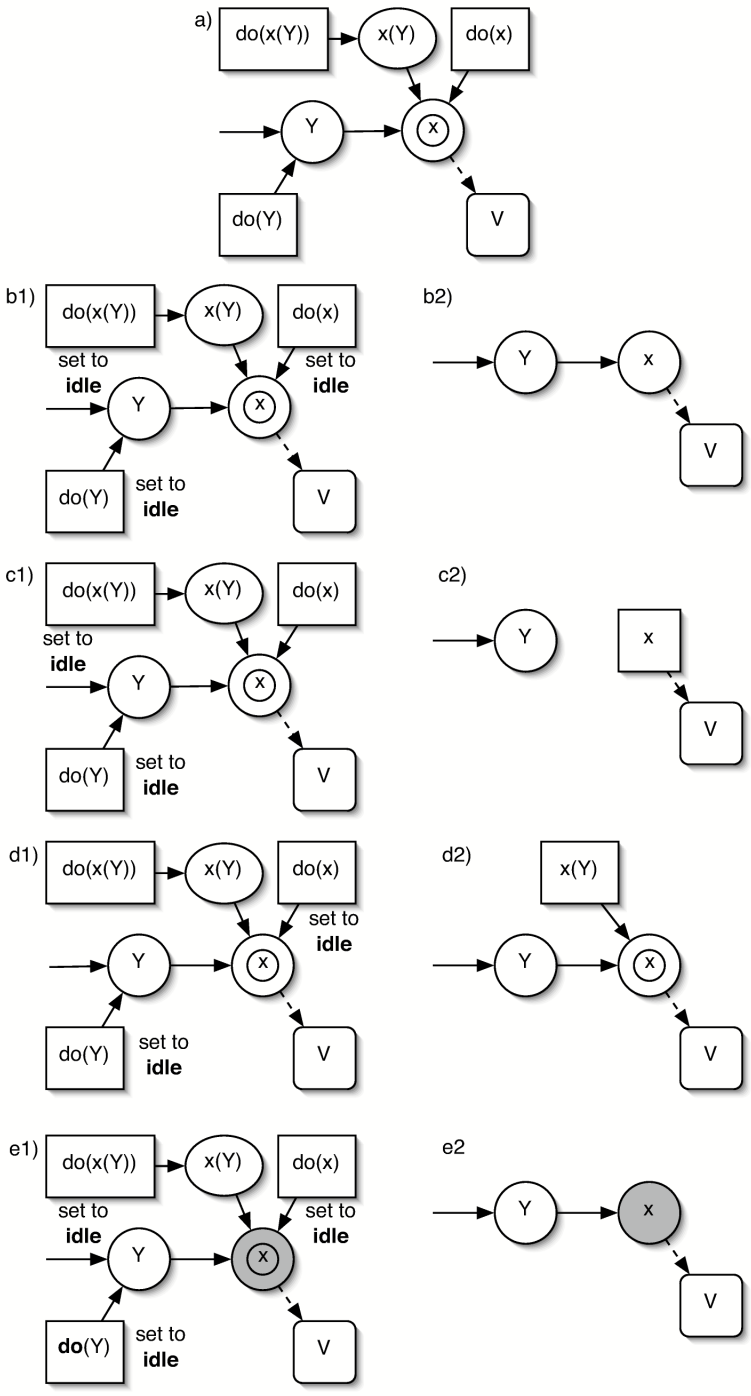


Figure 4. Partial causal influence diagrams to compute the values of do, control, and revelation for x when Y is nonempty.

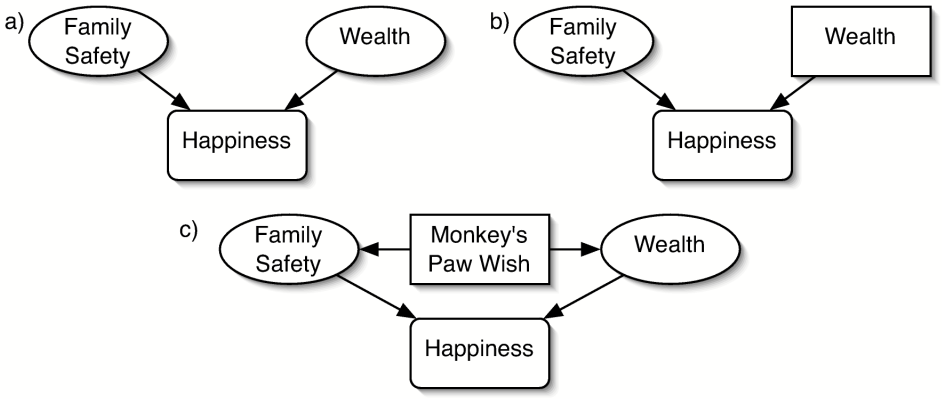


Figure 5. Unless the intervention is direct there can be disastrous side effects.

the simpler influence diagram shown in (d2), again using our shorthand. Finally, to compute the value of revelation for x, we can compute the value of the diagram with $x = x^*$ and all of the atomic interventions **idle**, as shown in (e1). The observation is well-defined because all of the interventions are **idle**, but that also means that we can compute its value with the simpler influence diagram shown in (e2).

Each of the three measures requires evaluation of two influence diagrams to determine its value, the reference diagram with all of the atomic interventions set to **idle** and a revised one, a diagram with either an atomic intervention or an observation. The values of these diagrams can be computed using simpler influence diagrams, with either one new decision, an atomic one made with no observations, or a new observation made before any decision, and the simpler diagram for the reference value has neither new decisions nor observations. These simpler diagrams are well-defined even if there are other decisions elsewhere and some observations prior to some of the other decisions [Shachter 1986]. Note that care must be taken in computing the value of control because there can be an exponential number of instances for the mapping variable.

The assumption of a direct intervention is crucial. Matheson and Matheson [2005](refer to it as "pure" and to an atomic intervention as "perfect".) There is a classic horror story of a man granted three wishes on a monkey's paw [Jacobs 1902]. He chooses to be wealthy and his wish is granted, tragically, through the death of his son. This corresponds to the causal influence diagrams shown in Figure 5. The value of his situation with no intervention is represented by the diagram in (a). The atomic intervention on *Wealth* he intends would yield the same value as a diagram in which *Wealth* is a decision as in (b), but the value with his intervention actually equals the value of the diagram shown in (c). The wish decision he actually made was not the direct intervention on *Wealth* he desired. The lesson is clear: in

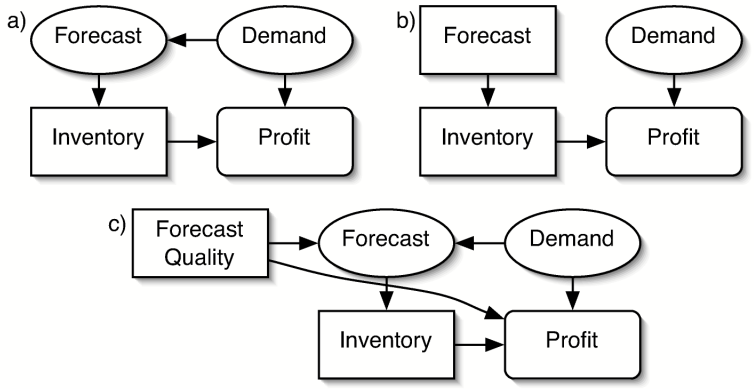


Figure 6. When we intervene on a forecast, we want to improve its quality, rather than to obtain a single most desirable instance.

manipulating our situation, we must beware of the unintended consequences.

Suppose the uncertain variable is being used to provide information, such as a forecast. Consider the causal influence diagram shown in Figure 6. This situation corresponds to one in which inventory decisions must be made before demand is observed, but a forecast relevant to demand will be observed before choosing inventory as shown in (a). Alas, an atomic intervention setting the forecast to our most desirable value ("highest demand") as in (b) does not improve profit since it tells us nothing about the real demand. What we would like to manipulate is the quality of the forecast, having it represent the best possible signal about demand as in (c). In this case, the value of do for *Forecast* is zero, but the value of control for *Forecast* should be positive. In fact, if there are as many instances for *Forecast* as there are for *Demand*, the highest quality forecast possible is clairvoyance on the demand, and the value of control would be equal to the value of clairvoyance. In the diagram *Forecast Quality* might not be an atomic intervention, both because there might only be a choice among imperfect information sources, and because there might be different costs associated with those different information sources.

Consider the causal influence diagrams shown in Figure 7, in which we believe that *Product Quality* is unresponsive to direct interventions (not shown) on *Sales* or *Profit*. We would like to understand how much we would improve our profit by manipulating our product quality. The diagram shown in (a) treats quality and sales as uncertain with its atomic interventions set to **idle**, and its value is the reference for any changes. The diagram shown in (b) has the same value as an atomic intervention on *Product Quality* to its optimal instance, and because that intervention is the cause of *Product Quality* with respect to D, the difference in values of this diagram relative to the one in (a) is both the value of do and the value of control for *Product Quality*. Alternatively, in (c) if we observed that *Product*

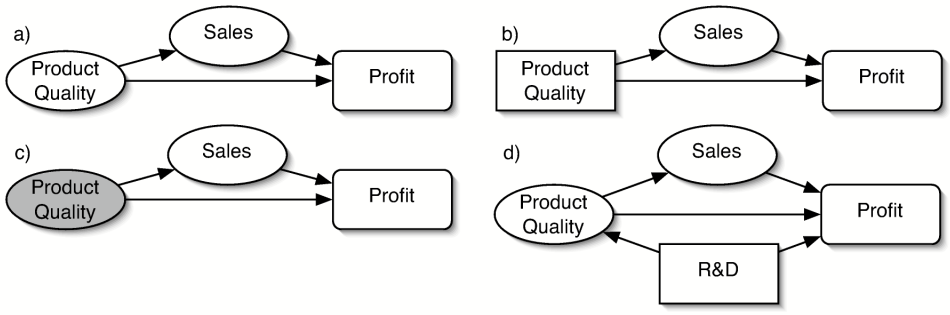


Figure 7. In this causal influence diagram the values of do, control, and revelation for *Product Quality* are equal.

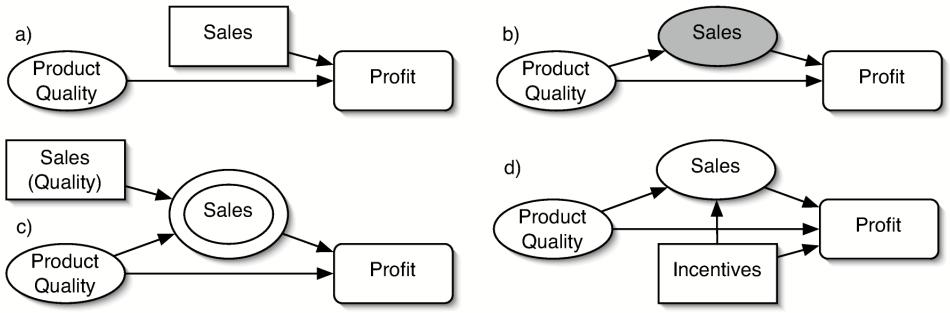


Figure 8. The values of do, control, and revelation for *Sales* might not be equal.

Quality takes the best possible value, this diagram has the same value as the one in (b). As a result, the value of revelation is equal to the other two values. Finally, in (d) we could contemplate a research and development effort that might lead to higher product quality. Because the diagram in (d) is causal, {*Product Quality*} is a cause of *Sales* with respect to D.

Now consider the causal influence diagrams shown in Figure 8, in which we are manipulating sales rather than product quality to improve our profit. We obtain the diagram shown in (a) by assuming that *Product Quality* is unresponsive to an atomic intervention on *Sales*. In (b) we could observe that *Sales* takes that same value, but this observation updates our belief about the *Product Quality*, and the value of this diagram might not be equal to the value of the diagram in (a). We obtain the diagram shown in (c) by an atomic intervention on the mapping variable for *Sales*, not determining sales but rather how it depends on quality (assuming that there is an atomic intervention on *Product Quality*). In this situation the values of do, control, and revelation could all be different! Finally, in (d) we consider offering incentives to boost sales, recognizing that it might affect our profits both directly

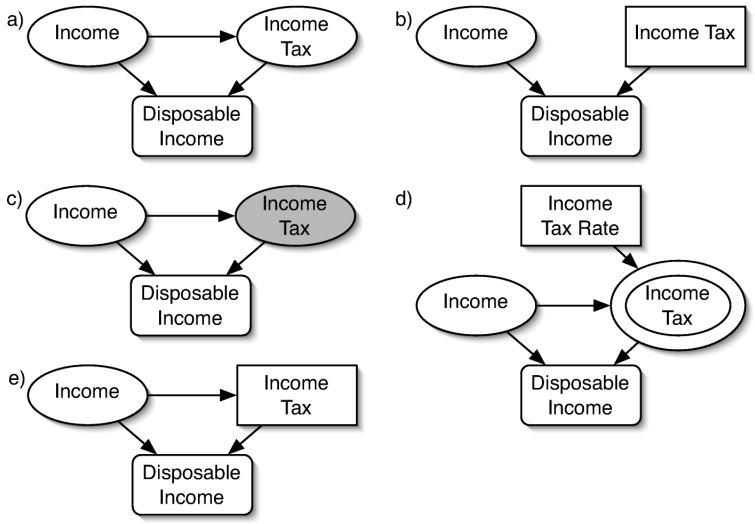


Figure 9. The values of do, control, and revelation are different for *Income Tax*.

and indirectly.

There can be a significant difference between passive observation of uncertain variable x and intervention on x. Consider the causal influence diagrams shown in Figure 9 representing disposable income after taxes. We believe that *Income* is unresponsive to a direct intervention on *Income Tax*, but *Income Tax* might be responsive to a direct intervention on *Income*. However, the value of do, the difference between the values of the diagrams in (b) and (a), is quite different from the value of revelation based on (c) and (a). Being able to choose not to pay any tax is quite different from learning that you will pay no tax, since it is more likely in the latter case that you have lost your job. Alternatively, we can consider setting the income tax rate as shown in (d), which would lead to the value of control. In this case, we can simplify the calculation in (d) that searches all possible mapping variable instances, to a simpler decision shown in (e), recognizing that in this case there is no interaction among the components of the mapping variable, and therefore we can independently search for the best possible instance for *Income Tax* for each possible instance of *Income*.

The correspondence between passive observation and intervention has been studied, primarily to identify causal effects from observational data [Robins 1986], [Pearl 1993] and [Spirtes, Glymour, and Scheines 1993]. In our framework, a set of variables Y is said to satisfy the *back door condition* for x if Y is unresponsive to $do(x)$ while $do(x)$ is d-separated from V by $\{x\} \cup Y$. When Y satisfies the back door condition, there is a correspondence among the values of do, control and revelation,

in that

$$P\{V|\mathbf{Y},\mathbf{x}^*\} = P\{V|\mathbf{Y},\mathbf{do}(\mathbf{x}^*)\} = P\{V|\mathbf{Y},\mathbf{do}(\mathbf{x}^*(Y))\}.$$

However, in valuing the decision situation we do not get to observe Y and thus $P\{V|\mathbf{x}^*\}$ might not be equal to $P\{V|\mathbf{do}(\mathbf{x}^*)\}$. Consider the diagrams shown in Figure 9. Because *Income* satisfies the back door criterion relative to *Income Tax*, the values of do, control and revelation on *Income Tax* would all be the same if we observed *Income*. But we do not know what our *Income* will be and the values of do, control, and revelation can all be different.

Nonetheless, if we make a stronger assumption, that Y is d-separated from V by x, the three measures will be equal. The atomic intervention on x or its mapping variable only affects the value V through the descendants of x in a causal model, and all other variables are unresponsive to the intervention in worlds limited by x. However, the atomic interventions might not be independent of V given x unless Y is d-separated from V by x. Otherwise, observing x or an atomic intervention on the mapping variable for x can lead to a different value for the diagram than an atomic intervention on x.

We establish this result in two steps for both general situations and for Pearl causal models. By assuming that $do(x)$ is independent of V given x, we first show that the values of do and revelation are equal. If we then assume that Y is d-separated from V by x, we show that the values of do and control are equal. The conditions under which these two different comparisons can be made are not identical either. To be able to compute the value of revelation for x we must set to **idle** all interventions that x is responsive to, while to compute the value of control for x we need to be ensure that we have an atomic intervention on a mapping variable for x.

THEOREM 10 (Equal Values of Do and Revelation). *Given a decision problem including uncertain variable $x \in U$, if there is a set of atomic interventions A, including $do(x)$, that is a cause of x with respect to D, and $do(x)$ is independent of V given x, then the values of do and revelation for x are equal, $VoD(\mathbf{x}^*) = VoR(\mathbf{x}^*)$.*

If $\{do(x)\}$ is a cause of x with respect to D, then they are also equal to the value of control for x, $VoC(\mathbf{x}^*()) = VoD(\mathbf{x}^*) = VoR(\mathbf{x}^*)$.

Proof. Consider the probability of V after the intervention $\mathbf{do}(\mathbf{x}^*)$ with all other interventions in A set to **idle**. Because x is determined by $\mathbf{do}(\mathbf{x}^*)$, and $do(x)$ is independent of V given x,

$$P\{V|\mathbf{do}(\mathbf{x}^*)\} = P\{V|\mathbf{x}^*,\mathbf{do}(\mathbf{x}^*)\} = P\{V|\mathbf{x}^*\} = P\{V|\mathbf{x}^*, do(x) = \mathbf{idle}\}.$$

If $\{do(x)\}$ is is a cause of x with respect to D then the values of do and control for x are equal. □

COROLLARY 11. *Given a decision problem described by a Pearl causal model including uncertain variable $x \in U$, if $Pa(x)$ is d-separated from V by x, then the*

values of do and revelation for x are equal, $VoD(\mathbf{x}^*) = VoR(\mathbf{x}^*)$. If x has no parents, then the the values of do, control, and revelation for x are equal,

$$VoD(\mathbf{x}^*) = VoC(\mathbf{x}^*()) = VoR(\mathbf{x}^*).$$

THEOREM 12 (Equal Values of Do and Control). *Given a decision problem described by an influence diagram including uncertain variable $x \in U$, and nonempty set of variables Y. If there are atomic interventions $do(x)$ for x, $do(y)$ for every $y \in Y \cap U$, and $do(x(Y)))$ for the mapping variable $x(Y)$, $Y \cup \{do(x), do(x(Y))\}$ is a cause of x with respect to D, and Y is d-separated from V by x, then the values of do and control are equal,*

$$VoD(\mathbf{x}^*) = VoC(\mathbf{x}^*(Y)).$$

Proof. We know that $Y \cup \{do(x), do(x(Y))\}$ is independent of V given x, because otherwise Y would not be d-separated from V by x. Because $do(x)$ is an atomic intervention on x and $do(x)$ is independent of V given x, as in Theorem 10, $P\{V|\mathbf{do}(\mathbf{x}^*)\} = P\{V|\mathbf{x}^*, \mathbf{do}(\mathbf{x}^*)\} = P\{V|\mathbf{x}^*\}$. Now consider the probability of V after the intervention $\mathbf{do}(\mathbf{x}^*(Y))$. Because $x = \mathbf{x}^*(\mathbf{Y})$ is determined by $\mathbf{do}(\mathbf{x}^*(Y))$ and $\mathbf{Y}$, and $Y \cup \{do(x(Y))\}$ is independent of V given x,

$$\begin{aligned} P\{V|\mathbf{do}(\mathbf{x}^*(Y)), \mathbf{Y}\} &= P\{V|x = \mathbf{x}^*(\mathbf{Y}), \mathbf{do}(\mathbf{x}^*(Y)), \mathbf{Y}\} \\ &= P\{V|x = \mathbf{x}^*(\mathbf{Y})\}, \end{aligned}$$

The optimal choice of $x(Y)$ does not depend on Y, $\mathbf{x}^*(Y) = \mathbf{x}^*$, yielding

$$P\{V|\mathbf{do}(\mathbf{x}^*(Y)), \mathbf{Y}\} = P\{V|\mathbf{x}^*\}.$$

As a result,

$$\begin{aligned} P\{V|\mathbf{do}(\mathbf{x}^*(Y))\} &= \sum_{\mathbf{Y}} P\{V, \mathbf{Y}|\mathbf{do}(\mathbf{x}^*(Y))\} \\ &= \sum_{\mathbf{Y}} P\{V|\mathbf{do}(\mathbf{x}^*(Y)), \mathbf{Y}\} P\{\mathbf{Y}|\mathbf{do}(\mathbf{x}^*(Y))\} \\ &= \sum_{\mathbf{Y}} P\{V|\mathbf{x}^*\} P\{\mathbf{Y}|\mathbf{do}(\mathbf{x}^*(Y))\} \\ &= P\{V|\mathbf{x}^*\} \sum_{\mathbf{Y}} P\{\mathbf{Y}|\mathbf{do}(\mathbf{x}^*(Y))\} \\ &= P\{V|\mathbf{x}^*\} \end{aligned}$$

□

COROLLARY 13. *Given an uncertain variable $x \in U$ with parents in a decision problem described by a Pearl causal model with an atomic intervention for mapping*

Figure 10. The values of do, control, and revelation are equal for each uncertain variable.

variable $x(Pa(x))$*, if* $Pa(x)$ *is d-separated from* V *by* x*, then the values of do, control, and revelation for* x *are equal,* $VoD(\mathbf{x}^*) = VoC(\mathbf{x}^*(Pa(x))) = VoR(\mathbf{x}^*)$.

Consider the causal influence diagrams shown in Figure 10, concerning a communicable disease, for which we believe that *Exposure* is unresponsive to any direct intervention on *Infection*, and both of them are unresponsive to any direct intervention on *Health*, but all of the uncertain variables might be responsive to a direction intervention on *Exposure*. Because *Exposure* has no parents, the values of do, control, and revelation for it will be equal. Furthermore, in this case, even though *Infection* has a parent, the values of do, control, and revelation for it will be also equal, because *Exposure* is independent of *Health* given *Infection*. Likewise, there will be equal values of do, control, and revelation for *Health*.

4 Conclusions

We have sharpened the distinctions underlying the value of control and related value of revelation and value of do, and shown that they are equivalent when the target variable x in a causal influence diagram either has no parents, or its parents, $Pa(x)$ are d-separated from the value V by x.

The general problem, which have only touched upon, permits multiple decisions and information sets at those other decisions. In that case, there is a question of how to recognize when $Pa(x)$ in d-separated from V by x. We can address this in general by either constructing the normal form diagram [Bhattacharjya and Shachter 2007] or by building a policy diagram, iteratively substituting deterministic policies for decisions starting with the latest decision [Shachter 1999]. These approaches exploit the causal structure and the separable value function represented in the influence diagram.

References

Bhattacharjya, D. and R. Shachter (2007). Evaluating influence diagrams with decision circuits. In R. Parr and L. van der Gaag (Eds.), *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 9–16. Oregon: AUAI Press.

Heckerman, D. and R. Shachter (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research 3*, 405–430.

Heckerman, D. E. and R. D. Shachter (1994). A decision-based view of causality.

In R. Lopez de Mantaras and D. Poole (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, pp. 302–310. San Mateo, CA: Morgan Kaufmann.

Howard, R. (1967). Value of information lotteries. *IEEE Transa. Systems Sci. Cybernetics SSC-3*(1), 54–60.

Howard, R. A. (1990). From influence to relevance to knowledge. In R. M. Oliver and J. Q. Smith (Eds.), *Influence Diagrams, Belief Nets, and Decision Analysis*, pp. 3–23. Chichester: Wiley.

Howard, R. A. and J. E. Matheson (1984). Influence diagrams. In R. A. Howard and J. E. Matheson (Eds.), *The Principles and Applications of Decision Analysis*, Volume II. Menlo Park, CA: Strategic Decisions Group.

Jacobs, W. W. (1902, September). The monkey's paw. *Harper's Monthly 105*, 634–639.

Matheson, D. and J. Matheson (2005). Describing and valuing interventions that observe or control decision situations. *Decision Analysis 2*(3), 165–181.

Matheson, J. E. (1990). Using influence diagrams to value information and control. In R. M. Oliver and J. Q. Smith (Eds.), *Influence Diagrams, Belief Nets, and Decision Analysis*, pp. 25–48. Chichester: Wiley.

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

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure results. *Mathematical Modeling 7*, 1393–1512.

Savage, L. (1954). *The Foundations o Statistics*. New York: Wiley.

Shachter, R. D. (1986). Evaluating influence diagrams. *Operations Research 34*(November-December), 871–882.

Shachter, R. D. (1999). Efficient value of information computation. In *Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference*, pp. 594–601. San Francisco, CA: Morgan Kaufmann.

Shachter, R. D. and D. E. Heckerman (1986). A backwards view for assessment. In *Workshop on Uncertainty in Artificial Intelligence*, University of Pennsylvania, Philadelphia, pp. 237–242.

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

27

Cause for Celebration, Cause for Concern

Yoav Shoham

It is truly a pleasure to contribute to this collection, celebrating Judea Pearl's scientific contributions. My focus, as well as that of several other contributors, is on his work in the area of causation and causal reasoning. Any student of these topics who ignores Judea's evolving contributions, culminating in the seminal [Pearl 2009], does so at his or her peril. In addition to the objective content of these contributions, Judea's unique energy and personality have led to his having unparalleled impact on the subject, in a diverse set of disciplines far transcending AI, his home turf. This body of work is truly a cause for celebration, and accounts for the first half of the title of this piece.

The second half of the title refers to a concern I have about the literature in AI regarding causation. As an early contributor to this literature I wade back into this area gingerly, aware of many of the complexities involved and difficulties encountered by earlier attempts to capture the notion formally. I am also aware of the fact that many developments have taken place in the past decade, indeed many associated with Judea himself, and only some of which I am familiar with. Still, it seems to me that the concern merits attention. The concern is not specific to Judea's work, and certainly applies to my own work in the area. It has to do with the yardsticks by which we judge this or that theory of causal representation or reasoning.

A number of years ago, the conference on Uncertainty in AI (UAI) held a panel on causation, chaired by Judea, in which I participated. In my remarks I listed a few requirements for a theory of causation in AI. One of the other panelists, whom I greatly respect, responded that he couldn't care less about such requirements; if the theory was useful that was good enough for him. In hindsight that was a discussion worth developing further then, and I believe it still is now.

Let us look at a specific publication, [Halpern and Pearl 2001]. This selection is arbitrary and I might as well have selected any number of other publications to illustrate my point, but it is useful to examine a concrete example. In this paper Halpern and Pearl present an account of "actual cause" (as opposed to "generic cause"; "the lighting last night caused the fire" versus "lightnings cause fire"). This account is also the basis for Chapter 10 of [Pearl 2009]. Without going into their specific (and worthwhile) account, let me focus on how they argue in its favor. In the third paragraph they say

> While it is hard to argue that our definition (or any other definition, for

> that matter) is the "right definition", we show that it deals well with the difficulties that have plagued other approaches in the past, especially those exemplified by the rather extensive compendium of [Hall 2004][1].

The reference is to a paper by a philosopher, and indeed of the thirteen references in the paper to work other than by the authors themselves, eight are to work by philosophers.

This orientation towards philosophy is evident throughout the paper, in particular in their relying strongly on particularly instructive examples that serve as test cases. This is an established philosophical tradition. The "morning star – evening star" example [Kripke 1980] catalyzed discussion of cross-world identity in first-order modal logic (you may have different beliefs regarding the star seen in the morning from those regarding the star seen in the evening, even though, unbeknownst to you, they are in fact the same star – Venus). Similarly, the example of believing that you will win the lottery and coincidentally later actually winning it served to disqualify the definition of knowledge as true belief, and a similar example argues against defining knowledge as *justified* true belief [Gettier 1963].

Such "intuition pumps" clearly guide the theory in [Halpern and Pearl 2001], as evidenced by the reference to [Hall 2004] mentioned earlier, and the fact that over four out of the paper's ten pages are devoted to examples. These examples can be highly instructive, but the question is what role they play. In philosophy they tend to serve as necessary but insufficient conditions for a theory. They are necessary in the sense that each of them is considered sufficient grounds for disqualifying a theory (namely, a theory which does not treat the example in an intuitively satisfactory manner). And they are insufficient since new examples can always be conjured up, subjecting the theory to ever-increasing demands.

This is understandable from the standpoint of philosophy, to the extent that it attempts to capture a complex, natural notion (be it knowledge or causation) it its full glory. But is this also the goal for such theories in AI? If not, what is the role of these test cases?

If taken seriously, the necessary-but-insufficient interpretation of the examples presents an impossible challenge to formal theory; a theoretician would never win in this game, in which new requirements may surface at any moment. Indeed, most of the philosophical literature is much less formal than the literature in AI, in particular [Halpern and Pearl 2001]. So where does this leave us?

This is not the first time computer scientists have faced this dilemma. Consider knowledge, for example. The S5 logic of knowledge [Fagin, Halpern, Moses, and Vardi 1994] captures well certain aspects of knowledge in idealized form, but the terms "certain" and "idealized" are important here. The logic has nothing to say about belief (as opposed to knowledge), nor about the dynamic aspects of knowledge (how it changes over time). Furthermore, even with regard to the static aspects of

[1] They actually refer to an earlier, unpublished version of Hall's paper from 1998.

knowledge, it is not hard to come up with everyday counterexamples to each of its axioms.

And yet, the logic proves useful to reason about certain aspects of distributed systems, and the mismatch between the properties of the modal operator K and the everyday word "know" does not get in the way, within these confines. All this changes as one switches the context. For example, if one wishes to consider cryptographic protocols, the K axiom $(Kp \wedge K(p \supset q) \supset Kq$, valid in any normal modal logic, and here representing logical omniscience) is blatantly inappropriate. Similarly, when one considers knowledge and belief together, axiom 5 of the logic ($\neg Kp \supset K\neg Kp$, representing negative introspection ability) seems impossible to reconcile with any reasonable notion of belief, and hence one is forced to retreat back from the S5 system to something weaker.

The upshot of all this is the following criterion for a formal theory of natural concepts: One should be explicit about the intended use of the theory, and within the scope of this intended use one should require that everyday intuition about the natural concepts be a useful guide in thinking about their formal counterparts.

A concrete interpretation of the above principle is what in [Shoham 2009] I called the *artifactual* perspective.[2] Artifactual theories attempt to shed light on the operation of a specific artifact, and use the natural notion almost as a mere visual aid. In such theories there is a precise interpretation of the natural notion, which presents a precise requirement for the formal theory. One example is indeed the use of "knowledge" to reason about protocols governing distributed systems. Another, discussed in [Shoham 2009], is the use of "intention" to reason about a database serving an AI planner.

Is there a way to instantiate the general criterion above, or more specifically the artifactual perspective, in the context of causation? I don't know the answer, but it seems to me worthy of investigation. If the answer is "yes" then we will be in a position to devise provably correct theories, and the various illustrative examples will be relegated to the secondary role of showing greater or lesser match with the everyday concept.

Acknowledgments: This work was supported by NSF grant IIS-0205633-001.

References

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1994). *Reasoning about Knowledge*. MIT Press.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis 23*, 121–123.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*. MIT Press.

[2] The discussion there is done in the context of formal models of intention, but the considerations apply here just as well.

Halpern, J. Y. and J. Pearl (2001). Causes and explanations: A structural-model approach. part I: Causes. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, San Francisco, CA, pp. 194–202. Morgan Kaufmann.

Kripke, S. A. (1980). *Naming and necessity* (Revised and enlarged ed.). Blackwell, Oxford.

Pearl, J. (2009). *Causality*. Cambridge University Press. Second edition.

Shoham, Y. (2009). Logics of intention and the database perspective. *Journal of Philosophical Logic 38*(6), 633–648.

28

Automated Search for Causal Relations – Theory and Practice

PETER SPIRTES, CLARK GLYMOUR, RICHARD SCHEINES, AND ROBERT TILLMAN

1 Introduction

The rapid spread of interest in the last two decades in principled methods of search or estimation of causal relations has been driven in part by technological developments, especially the changing nature of modern data collection and storage techniques, and the increases in the speed and storage capacities of computers. Statistics books from 30 years ago often presented examples with fewer than 10 variables, in domains where some background knowledge was plausible. In contrast, in new domains, such as climate research where satellite data now provide daily quantities of data unthinkable a few decades ago, fMRI brain imaging, and microarray measurements of gene expression, the number of variables can range into the tens of thousands, and there is often limited background knowledge to reduce the space of alternative causal hypotheses. In such domains, non-automated causal discovery techniques appear to be hopeless, while the availability of faster computers with larger memories and disc space allow for the practical implementation of computationally intensive automated search algorithms over large search spaces. Contemporary science is not your grandfather's science, or Karl Popper's.

Causal inference without experimental controls has long seemed as if it must somehow be capable of being cast as a kind of statistical inference involving estimators with some kind of convergence and accuracy properties under some kind of assumptions. Until recently, the statistical literature said *not.* While parameter estimation and experimental design for the effective use of data developed throughout the 20th century, as recently as 20 years ago the methodology of causal inference without experimental controls remained relatively primitive. Besides a cessation of hostilities from the majority of the statistical and philosophical communities (which has still only partially happened), several things were needed for theories of causal estimation to appear and to flower: well defined mathematical objects to represent causal relations; well defined connections between aspects of these objects and sample data; and a way to compute those connections. A sequence of studies beginning with Dempster's work on the factorization of probability distributions [Dempster 1972] and culminating with Kiiveri and Speed's [Kiiveri & Speed 1982] study of linear structural equation models, provided the first, in the form of directed acyclic graphs, and the second, in the form of the "local" Markov condition. Pearl and his students [Pearl 1988], and independently, Stefan

Lauritzen and his collaborators [Lauritzen, Dawid, Larsen, & Leimer 1990], provided the third, in the form of the "global" Markov condition, or d-separation in Pearl's formulation, and the assumption of its converse, which came to be known as "stability" or "faithfulness." Further fundamental conceptual and computational tools were needed, many of them provided by Pearl and his associates; for example, the characterization and representation of Markov equivalence classes and the idea of "inducing paths," essential to understanding the properties of models with unrecorded variables. Initially, most of these authors, including Pearl, did not connect directed graphical models with a causal interpretation (in the sense of representing outcomes of interventions). This connection between graphs and interventions was drawn from an earlier tradition in econometrics [Strotz & Wold 1960], and in our work [Spirtes, Glymour, & Scheines 1993]. With this connection, and the pieces Speed, Lauritzen, Pearl and others had established, a principled theory of causal estimation could, and did, begin around 1990, and Pearl and his students have made important contributions to it. Pearl has become the foremost advocate in the universe for reconceiving the relations between causality and statistics. Once begun for special cases, the understanding of search methods for causal relations has expanded to a variety of scientific and statistical settings, and in many scientific enterprises—neuroimaging for example—causal representations and search are treated as almost routine.

The theory of interventions also provided a coherent normative theory of inference using causal premises. That effort can also be traced back to Strotz and Wold [Strotz & Wold 1960], then to our own work [Spirtes, Glymour, & Scheines 1993] on prediction from classes of causal graphs, and then to the full development of a non-parametric theory of prediction for graphical models by Pearl and his collaborators [Shpitser & Pearl 2008]. Pearl brilliantly turned philosopher and developed the theory of interventions into a general account of counterfactual reasoning. Although we will not discuss it further, we think there remain interesting open problems about prediction algorithms for various parametric classes of graphical causal models.

The following paper surveys a broad range of causal estimation problems and algorithms, concentrating especially on those that can be illustrated with empirical examples that we and our students and collaborators have analyzed. This has naturally led to a concentration on the algorithms and tools that we have developed. The kinds of causal estimation problems and algorithms discussed are broadly representative of the most important developments in methods for estimating causal structure since 1990, but it is not a comprehensive survey. There have been so many improvements to the basic algorithms that we describe here there is not room to discuss them all. A good resource for a description of further research in this area is the Proceedings of the Conferences on Uncertainty in Artificial Intelligence, at http://uai.sis.pitt.edu.

The dimensions of the problems, as we have long understood them, are these:

1. Finding computationally and statistically feasible methods for discovering causal information for large numbers of variables, provably correct under standard sampling assumptions, assuming no confounding by unrecorded variables.

2. The same when the "no confounding" assumption is abandoned.
3. Finding methods for obtaining causal information when there is systematic sample selection bias — when values of some of the variables of interest are associated with sample membership.
4. Finding methods for establishing the existence of unobserved causes and estimating *their* causal relations with one another.
5. Finding methods for discovering causal relations in data produced by feedback systems.
6. Finding methods for discovering causal relations in time series data.
7. Finding methods for discovering causal relations in linear and in non-linear non-Gaussian systems with continuous variables.
8. Finding methods for discovering causal relations using distributed, multiple data sets.
9. Finding methods for merging the above with experimental design.

2 Assumptions

We assume the reader's familiarity with the standard notions used in discussions of graphical causal model search: conditional independence, Markov properties, d-separation, Markov equivalence, patterns, distribution equivalence, causal sufficiency, etc. The appendix gives a brief review of the essential definitions, assumptions and theorems required for known proofs of correctness of the algorithms we will discuss.

3 Model Search Assuming Causal Sufficiency

The assumption of causal sufficiency (roughly no unrecorded confounders) is often unrealistic, but it is useful in explicating search because the concepts and methods used in search algorithms that make more realistic assumptions are more complex versions of ideas that are used in searches that assume causal sufficiency.

3.1 The PC Algorithm

The PC algorithm is a constraint-based search that attempts to find the pattern that most closely entails all and only the conditional independence constraints judged to hold in the population. The SGS algorithm [Spirtes & Glymour 1991] and the IC algorithm [Verma & Pearl 1990] were early versions of this algorithm that were statistically and computationally feasible only on data sets with few variables because they required conditioning on all possible subsets of variables.) The PC algorithm solved both difficulties in typical cases.

The PC algorithm has an adjacency phase in which the adjacencies are determined, and an orientation phase in which as many edges as possible are oriented. The adjacency phase is stated below, and illustrated in Figure 1. Let **Adjacencies**(G,A) be the set of vertices adjacent to A in undirected graph G. (In the algorithm, the graph G is continually updated, so **Adjacencies**(G,A) may change as the algorithm progresses.)

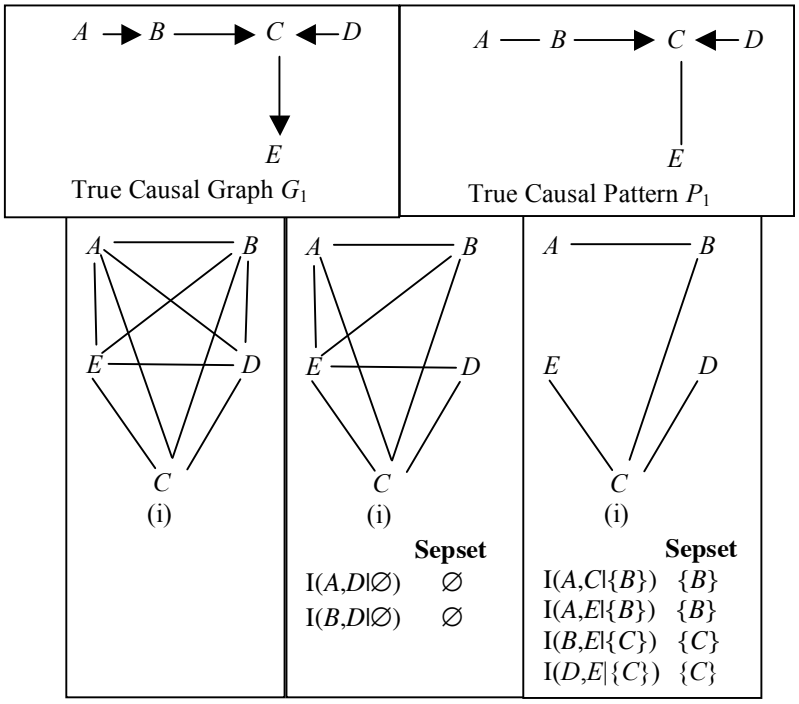


Figure 1: Constraint based search, where correct pattern is P_1

Adjacency Phase of PC Algorithm:

Form an undirected graph G in which every pair of vertices in **V** is adjacent.
$n := 0$.
repeat
  repeat
    Select an ordered pair of variables X and Y that are adjacent in G such that **Adjacencies**(G,X)\{Y} has cardinality greater than or equal to n, and a subset **S** of **Adjacencies**(G,X)\{Y} of cardinality n, and if X and Y are independent conditional on **S** delete edge X — Y from C and record **S** in **Sepset**(X,Y) and **Sepset**(Y,X);
  until all ordered pairs of adjacent variables X and Y such that **Adjacencies**(G,X)\{Y} has cardinality greater than or equal to n and all subsets **S** of **Adjacencies**(G,X)\{Y} of cardinality n have been tested for conditional independence;
  $n := n + 1$;
until for each ordered pair of adjacent vertices X, Y, **Adjacencies**(G,X)\{Y} is of cardinality less than n.

After the adjacency phase of the algorithm, the orientation phase of the algorithm is performed. The orientation phase of the algorithm is illustrated in Figure 2.

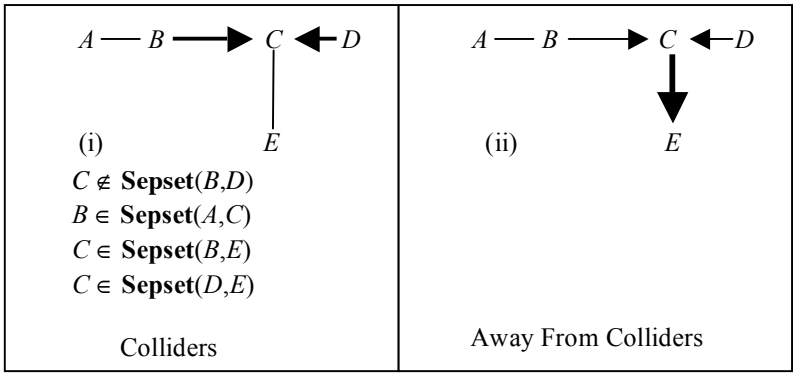


Figure 2: Orientation phase of PC algorithm, assuming true pattern is P_1

The orientation phase of the PC algorithm is stated more formally below. The last two orientation rules (Away from Cycles, and Double Triangle) are not used in the example, but are sound because if the edges were oriented in ways that violated the rules, there would be a directed cycle in the pattern, which would imply a directed cycle in the graph (which in this section is assumed to be impossible). The orientation rules are complete [Meek 1995], i.e. every edge that has the same orientation in every member of a DAG conditional independence equivalence class is oriented by these rules.

Orientation Phase of PC Algorithm

For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in graph G but the pair X, Z are not adjacent in G, orient $X — Y — Z$ as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in **Sepset**(X,Z).

repeat

Away from colliders: If $A \rightarrow B — C$, and A and C are not adjacent, then orient as $B \rightarrow C$.

Away from cycles: If $A \rightarrow B \rightarrow C$ and $A — C$, then orient as $A \rightarrow C$.

Double Triangle: If $A \rightarrow B \leftarrow C$, A and C are not adjacent, $A — D — C$, and there is an edge $B — D$, orient $B — D$ as $D \rightarrow B$.

until no more edges can be oriented.

The tests of conditional independence can be performed in the usual way. Conditional independence among discrete variables can be tested using the G^2 statistic; conditional independence among multivariate Gaussian variables can be tested using Fisher's Z-transformation of the partial correlations [Spirtes, Glymour, & Scheines 2001]. Section 3.4 describes more general tests of conditional independence. Such tests require specifying a significance level for the test, which is a user-specified parameter of the algorithm. Because the PC algorithm performs a sequence of tests without adjustment, the significance level does not represent any (easily calculable) statistical feature of the output, but should only be understood as a parameter used to guide the search.

Assuming that the causal relations can be represented by a directed acyclic graph, the Causal Markov Assumption, the Causal Faithfulness Assumption, and consistent tests of conditional independence, in the large sample (i.i.d.) limit for a causally sufficient set of variables, the PC algorithm outputs a pattern that represents the true causal graph.

The PC algorithm has been shown to apply to very high dimensional data sets (under a stronger version of the Causal Faithfulness Assumption), both for finding causal structure [Kalisch & Buhlmann 2007] and for classification [Aliferis, Tsamardinos, & Statnikov 2003]. A version of the algorithm controlling the false discovery rate is available [Junning & Wang 2009].

3.1.1 Example - Foreign Investment

This example illustrates how the PC algorithm can find plausible alternatives to a model built from domain knowledge. Timberlake and Williams used regression to claim foreign investment in third-world countries promotes dictatorship [Timberlake & Williams 1984]. They measured political exclusion (*PO*) (i.e., dictatorship), foreign investment penetration in 1973 (*FI*), energy development in 1975 (*EN*), and civil liberties (*CV*) for 72 countries. *CV* was measured on an ordered scale from 1 to 7, with lower values indicating greater civil liberties.

Their inference is unwarranted. Their model (with the relations between the regressors omitted) and the pattern obtained from the PC algorithm using a 0.12 significance level to test for vanishing partial correlations) are shown in Figure 3.[1] We typically run the algorithms at a variety of different significance levels, and compare the results to see if any of the features of the output are constant.

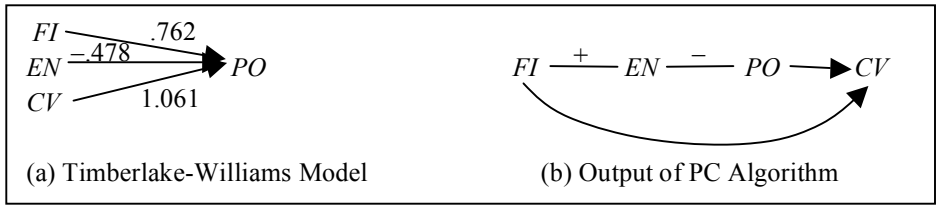


Figure 3: Two Models of Foreign Investment

The PC Algorithm will not orient the *FI – EN* and *EN – PO* edges, and assumes that the edges are not due to an unmeasured common cause. Maximum likelihood estimates of any linear, Gaussian parameterization of any DAG represented by the pattern output by the PC algorithm requires that the influence of *FI* on *PO* (if any) be negative, and the models easily pass a likelihood ratio test. If any of these SEMs is correct, Timberlake and William's regression model appears to be a case in which an effect of the outcome variable is taken as a regressor.

Given the small sample size, and the uncertainty about the distributional assumptions, we do not present the alternative models suggested by the PC algorithm as particularly well supported by the evidence. However, we do think that they are at least

[1] Searches at lower significance levels remove the adjacency between *FI* and *EN*.

as well supported as the regression model, and hence serve to cast doubt upon conclusions drawn from that model.

3.1.2 Example - Spartina Biomass

This example illustrates a case where the PC algorithm output received some experimental confirmation. A textbook on regression [Rawlings 1988] skillfully illustrates regression principles and techniques for a biological study from a dissertation [Linthurst 1979] in which it is reasonable to think there is a causal process at work relating the variables. The question at issue is plainly causal: among a set of 14 variables, which have the most influence on an outcome variable, the biomass of Spartina grass? Since the example is the principle application given for an entire textbook on regression, the reader who reaches the 13th chapter may be surprised to find that the methods yield almost no useful information about that question.

According to Rawlings, Linthurst obtained five samples of Spartina grass and soil from each of nine sites on the Cape Fear Estuary of North Carolina. Besides the mass of Spartina (*BIO*), fourteen variables were measured for each sample:

- Free Sulfide (H_2S)
- Salinity (*SAL*)
- Redox potentials at pH 7 (EH_7)
- Soil pH in water (*PH*)
- Buffer acidity at pH 6.6 (*BUF*)
- Phosphorus concentration (*P*)
- Potassium concentration (*K*)
- Calcium concentration (*CA*)
- Magnesium concentration (*MG*)
- Sodium concentration (*NA*)
- Manganese concentration (*MN*)
- Zinc concentration (*ZN*)
- Copper concentration (*CU*)
- Ammonium concentration (NH_4)

The aim of the data analysis was to determine for a later experimental study which of these variables most influenced the biomass of Spartina in the wild. Greenhouse experiments would then try to estimate causal dependencies out in the wild. In the best case one might hope that the statistical analyses of the observational study would correctly select variables that influence the growth of Spartina in the greenhouse. In the worst case, one supposes, the observational study would find the wrong causal structure, or would find variables that influence growth in the wild (e.g., by inhibiting or promoting growth of a competing species) but have no influence in the greenhouse.

Using the SAS statistical package, Rawlings analyzed the variable set with a multiple regression and then with two stepwise regression procedures from the SAS package. A search through all possible subsets of regressors was not carried out, presumably because the candidate set of regressors is too large. The results were as follows:

(i) a multiple regression of *BIO* on all other variables gives only *K* and *CU* significant regression coefficients;

(ii) two stepwise regression procedures[2] both yield a model with *PH*, *MG*, *CA* and *CU* as the only regressors, and multiple regression on these variables alone gives them all significant coefficients;

(iii) simple regressions one variable at a time give significant coefficients to *PH*, *BUF*, *CA*, *ZN* and *NH*4.

What is one to think? Rawling's reports that "None of the results was satisfying to the biologist; the inconsistencies of the results were confusing and variables expected to be biologically important were not showing significant effects." (p. 361).

This analysis is supplemented by a ridge regression, which increases the stability of the estimates of coefficients, but the results for the point at issue--identifying the important variables--are much the same as with least squares. Rawlings also provides a principal components factor analysis and various geometrical plots of the components. These calculations provide no information about which of the measured variables influence Spartina growth.

Noting that *PH*, for example, is highly correlated with *BUF*, and using *BUF* instead of *PH* along with *MG*, *CA* and *CU* would also result in significant coefficients, Rawlings effectively gives up on this use of the procedures his book is about:

> Ordinary least squares regression tends either to indicate that none of the variables in a correlated complex is important when all variables are in the model, or to arbitrarily choose one of the variables to represent the complex when an automated variable selection technique is used. A truly important variable may appear unimportant because its contribution is being usurped by variables with which it is correlated. Conversely, unimportant variables may appear important because of their associations with the real causal factors. It is particularly dangerous in the presence of collinearity to use the regression results to impart a "relative importance," whether in a causal sense or not, to the independent variables. (p. 362)

Rawling's conclusion is correct in spirit, but misleading and even wrong in detail. If we apply the PC algorithm to the Linthurst data then there is one robust conclusion*:* the only variable that may *directly* influence biomass in this population[3] is *PH*; *PH* is distinguished from all other variables by the fact that the correlation of every other variable (except *MG*) with *BIO* vanishes or vanishes when *PH* is conditioned on.[4] The relation is not symmetric; the correlation of *PH* and *BIO*, for example, does not vanish when *BUF* is controlled. The algorithm finds *PH* to be the only variable adjacent to *BIO*

[2]The "maximum R-square" and "stepwise" options in PROC REG in the SAS program.

[3]Although the definition of the population in this case is unclear, and must in any case be drawn quite narrowly.

[4]More exactly, at .05, with the exception of *MG* the partial correlation of every regressor with *BIO* vanishes when some set containing *PH* is controlled for; the correlation of *MG* with *BIO* vanishes when *CA* is controlled for.

no matter whether we use a significance level of .05 to test for vanishing partial correlations, or a level of 0.1, or a level of 0.2. In all of these cases, the PC algorithm (and the FCI algorithm, which allows for the possibility of latent variables in section 4.2) yields the result that *PH* and only *PH* can be directly connected with *BIO*. If the system is linear normal and the Causal Markov Assumption obtains, then in this population any influence of the other regressors on *BIO* would be blocked if *PH* were held constant. Of course, over a larger range of values of the variables there is little reason to think that *BIO* depends linearly on the regressors, or that factors that have no influence in producing variation within this sample would continue to have no influence.

Although the analysis cannot conclusively rule out possibility that *PH* and *BIO* are confounded by one or more unmeasured common causes, in this case the principles of the theory and the data argue against it. If *PH* and *BIO* have a common unmeasured cause *T*, say, and any other variable, Z_i, among the 13 others either causes *PH* or has a common unmeasured cause with *PH* (Figure 4, in which we do not show connections among the **Z** variables), then Z_i and *BIO* should be correlated conditional on *PH*, which is statistically not the case.

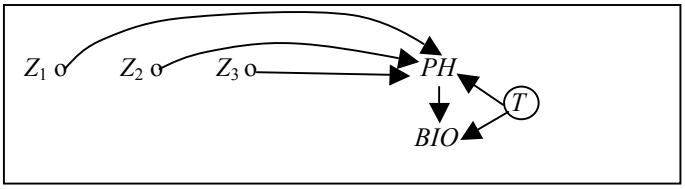


Figure 4 : *PH* and *BIO* Confounding?

The program and theory lead us to expect that if *PH* is forced to have values like those in the sample--which are almost all either below *PH* 5 or above *PH* 7-- then manipulations of other variables within the ranges evidenced in the sample will have no effect on the growth of Spartina. The inference is a little risky, since growing plants in a greenhouse under controlled conditions may not be a direct manipulation of the variables relevant to growth in the wild. If, for example, in the wild variations in *PH* affect Spartina growth chiefly through their influence on the growth of competing species not present in the greenhouse, a greenhouse experiment will not be a direct manipulation of *PH* for the system.

The fourth chapter of Linthurst's thesis partly confirms the PC algorithm's analysis. In the experiment Linthurst describes, samples of Spartina were collected from a salt marsh creek bank (presumably at a different site than those used in the observational study). Using a 3 x 4 x 2 (*PH* x *SAL* x *AERATION*) randomized complete block design with four blocks, after transplantation to a greenhouse the plants were given a common nutrient solution with varying values *PH* and *SAL* and *AERATION*. The *AERATION* variable turned out not to matter in this experiment. Acidity values were *PH* 4, 6 and 8. *SAL* for the nutrient solutions was adjusted to 15, 25, 35 and 45 ‰.

Linthurst found that growth varied with *SAL* at *PH* 6 but not at the other *PH* values, 4 and 8, while growth varied with *PH* at all values of *SAL* (p. 104). Each variable was

correlated with plant mineral levels. Linthurst considered a variety of mechanisms by which extreme *PH* values might control plant growth:

> At pH 4 and 8, salinity had little effect on the performance of the species. The pH appeared to be more dominant in determining the growth response. However, there appears to be no evidence for any causal effects of high or low tissue concentrations on plant performance unless the effects of pH and salinity are also accounted for. (p.108)
> The overall effect of pH at the two extremes is suggestive of damage to the root, thereby modifying its membrane permeability and subsequently its capacity for selective uptake. (p. 109).

A comparison of the observational and experimental data suggests that the PC Algorithm result was essentially correct and can be extrapolated through the variation in the populations sampled in the two procedures, but cannot be extrapolated through *PH* values that approach neutrality. The result of the PC search was that in the non-experimental sample, observed variations in aerial biomass were perhaps caused by variations in *PH*, but were not caused (at least not directly, relative to *PH*) by variations in other variables. In the observational data Rawlings reports (p. 358) almost all *SAL* measurements are around 30--the extremes are 24 and 38. Compared to the experimental study rather restricted variation was observed in the wild sample. The observed values of *PH* in the wild, however, are clustered at the two extremes; only four observations are within half a *PH* unit of 6, and no observations at all occurred at *PH* values between 5.6 and 7.1. For the observed values of *PH* and *SAL*, the experimental results appear to be in very good agreement with our results from the observational study: small variations in *SAL* have no effect on Spartina growth if the *PH* value is extreme.

3.1.3 College Plans

Sewell and Shah [Sewell & Shah 1968] studied five variables from a sample of 10,318 Wisconsin high school seniors.[5] The variables and their values are:

- *SEX* male = 0, female = 1
- *IQ* = Intelligence Quotient, lowest = 0, highest = 3
- *CP* = college plans yes = 0, no = 1
- *PE* = parental encouragement low = 0, high = 1
- *SES* = socioeconomic status lowest = 0, highest = 3

The question of interest is what the causes of college plans are. This data set is of interest because it has been used by a variety of different search algorithms that make different assumption. The different results illustrate the role that the different assumptions make in the output and are discussed in subsequent sections.

[5]Examples of the analysis of the Sewell and Shah data using Bayesian networks are given in Spirtes et al. (2001), and Heckerman (1998).

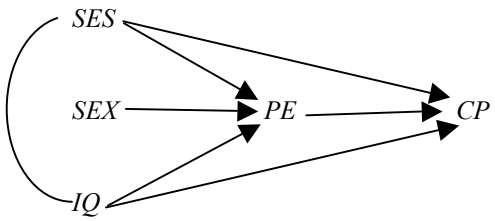


Figure 5: Model of Causes of College Plans

The pattern produced as the output of the PC algorithm is shown in Figure 5. The model predicts that *SEX* affects *CP* only indirectly via *PE*.

It is possible to predict the effects of some manipulations from the pattern, but not others. For example, because the pattern is compatible both with $SES \rightarrow IQ$ and with $SES \leftarrow IQ$, it is not possible to determine if *SES* is a cause or an effect of *IQ*, and hence it is not possible to predict the effect of manipulating *SES* on *IQ* from the pattern. On the other hand, it can be shown that all of the models in the conditional independence equivalence class represented by the pattern entail the same predictions about the quantitative effects of manipulating *PE* on *CP*. When *PE* is manipulated, in the manipulated distribution: $P(CP=0|PE=0) = .095$; $P(CP=1|PE=0) = .905$; $P(CP=0|PE=1) = .484$; $P(CP=1PE=1) = .516$ [Spirtes, Scheines, Glymour, & Meek 2004].

3.2 Greedy Equivalence Search Algorithm

Algorithms that maximize a score have certain advantages over constraint-based algorithms such as PC. When the data are not Gaussian, but the system is linear, extensive unpublished simulations find that at least one such algorithm, the Greedy Equivalence Search (GES) algorithm [Meek 1997] outperforms PC. GES can be used with a number of different scores for patterns, including posterior probabilities (for some parametric families and under some priors), and the Bayesian Information Criterion (BIC), which is an approximation of a class of posterior distributions in the large sample limit. The BIC score [Schwarz 1978] is: $-2 \ln(ML) + k \ln(n)$, where ML is the likelihood of the data at the maximum likelihood estimate of the parameters, k is the dimension of the model and n is the sample size. For uniform priors on models and smooth priors on the parameters, the posterior probability conditional on the data is a monotonic function of BIC in the large sample limit. In the forward stage of the search, starting with an initial (possibly empty) pattern, at each stage GES selects the pattern that is the one-edge addition compatible with the current pattern and has the highest score. The forward stage continues until no further additions improve the score. Then a reverse procedure is followed that removes edges according to the same criterion, until no improvement is found. The computational and convergence advantages of the algorithm depend on the fact that it searches over Markov equivalence classes of DAGs rather than individual DAGs, and that only one forward stage and one backward stage are required for an asymptotically correct search. In the large sample limit, GES identifies the Markov equivalence class of the true graph if the assumptions above are met [Chickering 2002].

GES has proved especially valuable in searches for latent structure (GESMIMBuild) and in searches with multiple data sets (IMaGES). Examples are discussed in sections 4.4 and 5.3 .

3.3 LiNGAM

Standard implementations of the constraint-based and score-based algorithms above usually assume that continuous variables have multivariate Gaussian distributions. This assumption is inappropriate in many contexts such as EEG analysis where variables are known to deviate from Gaussianity.

The LiNGAM (Linear Non-Gaussian Acyclic Model) algorithm [Shimizu, Hoyer, Hyvärinen, & Kerminen 2006] is appropriate specifically for cases where each variable in a set of measured variables can be written as a linear function of other measured variables plus an independent noise component, where at most one of the measured variables' noise components may be Gaussian. For example, consider the system with the causal graph shown in Figure 6 and assume X, Y, and Z are determined as follows, where a, b, and c are real-valued coefficients and ε_x , ε_y, and ε_z are independent noise components of which at least two are non-Gaussian.

(1) $X = \varepsilon_x$
(2) $Y = aX + \varepsilon_y$
(3) $Z = bX + cY + \varepsilon_z$

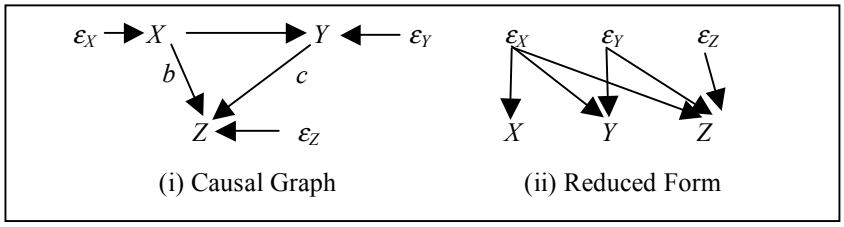


Figure 6: Causal Graph and Reduced Form

The equations can be rewritten in what economists called reduced form, also shown in Figure 6:

(4) $X = \varepsilon_X$
(5) $Y = a\varepsilon_X + \varepsilon_Y$
(6) $Z = b\varepsilon_X + ac\varepsilon_X + c\varepsilon_Y + \varepsilon_Z$

The standard Independent Components Analysis (ICA) procedure [Hyvärinen & Oja, 2000] can be used to recover a matrix containing the real-valued coefficients a, b, and c from an i.i.d. sample of data generated from the above system of equations. The LiNGAM algorithm finds the correct matching of coefficients in this ICA matrix to variables and prunes away any insignificant coefficients using statistical criteria.

The procedure yields correct values even if the coefficients were to perfectly cancel, and hence the variables such as *X, Z* above were to be uncorrelated. Since coefficients are determined for each variable, we can always reconstruct the true unique DAG, instead of its Markov equivalence class. The procedure converges (at least) pointwise to the true DAG and coefficients assuming: (1) there are no unmeasured common causes; (2) the dependencies among measured variables are linear; (3) none of the relations among measured variables are deterministic; (4) i.i.d. sampling; (5) the Markov Condition; (6) at most one error or disturbance term is Gaussian. We do not know its complexity properties.

The LiNGAM procedure can be generalized to estimate causal relations among observables when there are latent common causes [Hoyer, Shimizu, & Kerminen 2006], although the result is not in general a unique DAG, and LiNGAM has been combined [Shimizu, Hoyer, & Hyvarinen 2009] with Silva's clustering procedure (section 4.4) for locating latent variables to estimate a unique DAG among latent variables, and also with search for cyclic graphs [Lacerda, Spirtes, Ramsey, & Hoyer 2008], and combined with the PC and GES algorithms when more than one disturbance term is Gaussian [Hoyer et al. 2008].

3.4 The kPC Algorithm

The kPC algorithm [Tillman, Gretton, & Spirtes, 2009] relaxes distributional assumptions further, allowing not only non-Gaussian noise with continuous variables, but also nonlinear dependencies. In many cases, kPC will return a unique DAG (even when there is more than one DAG in the Markov equivalence class. However, unlike LiNGAM there is no requirement that a certain number of variables be non-Gaussian.

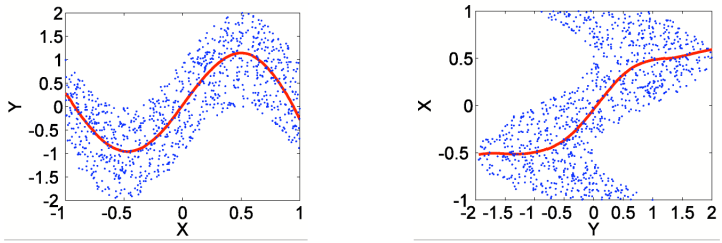
kPC consists of two stages. In the first stage of kPC, the standard PC algorithm is applied to the data using efficient implementations of the Hilbert-Schmidt Independence Criteria [Gretton, Fukumizu, Teo, Song, Scholkopf, & Smola, 2008], a nonparametric independence test and an extension of this test to the conditional cases based on the dependence measure given in [Fukumizu, Gretton, Sun, & Scholkopf, 2008]. This produces a pattern. Additional orientations are then possible if the true causal model, or a submodel (after removing some variables) of the true causal model is an *additive noise model* [Hoyer, Janzing, Mooij, Peters, & Scholkopf, 2009] that is *noninvertible*.

A set of variables is an additive noise model if (i) the function form of each variable can be expressed as a (possible nonlinear) smooth function of its parents in the true causal model plus an additive (Gaussian or non-Gaussian) noise component and (ii) the additive noise components are mutually independent. An additive noise model is noninvertible if we cannot reverse any edges in the model and still obtain smooth functional forms for each variable and mutually independent additive noise components that fit the data.

For example, consider the two variable case where $X \rightarrow Y$ is the true DAG and we have the following function forms and additive noise components for X and Y:

$$X = \varepsilon_X,\ Y = \sin(\pi X) + \varepsilon_X,\ \ \varepsilon_X \sim Uniform(-1,1),\ \varepsilon_Y \sim Uniform(-1,1)$$

If we fit a nonparametric regression model for Y regressed on X, the forward model, Figure 7a, and for X regressed on Y, the backward model, Figure 7b, we observe $I(\hat{\varepsilon}_Y, X)$ and $\neg I(\hat{\varepsilon}_X, Y)$ since this additive noise model is noninvertible.



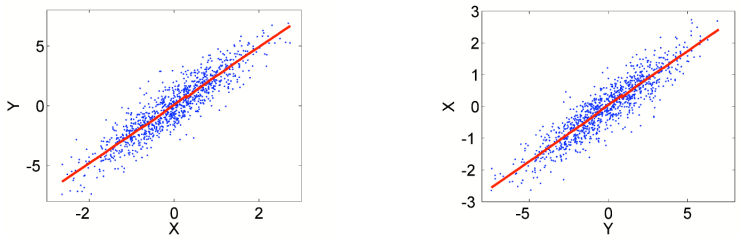
(a) **(b)**

Figure 7: Nonparametric regressions of (a) Y on X, and (b) X on Y with the data overlayed for nonlinear non-Gaussian case

Thus in this case, we can conclude that $X \to Y$ is the true DAG from the data since the additive noise model fits in only one direction, i.e. it is noninvertible. However, consider the following linear Gaussian case:

$$X = \varepsilon_X,\ Y = 2.4 \cdot X + \varepsilon_Y,\ \ \varepsilon_X \sim N(0,1),\ \varepsilon_Y \sim N(0,1)$$

After fitting nonparametric regression models for both directions, Figure 8, we find $I(\hat{\varepsilon}_Y, X)$ and $I(\hat{\varepsilon}_X, Y)$ so we cannot determine whether $X \to Y$ or $Y \to X$ is the correct DAG.



(a) **(b)**

Figure 8: Nonparametric regressions of (a) Y on X, and (b) X on Y with the data overlayed for linear Gaussian case

[Zhang and Hyvarinen, 2009] show that only a few special cases, other than the linear Gaussian case, exist where the additive noise model is invertible.

The second stage of kPC consists of searches for submodels that are consistent with the pattern learned in the first stage of kPC that may be noninvertible additive noise models. If such models are discovered, then further orientations of edges can be made resulting in an equivalence class of possible DAGs that is a proper subset of the Markov

equivalence class. In many cases, only a few variables need be nonlinear or non-Gaussian to obtain a unique DAG using kPC.

kPC requires the following additional assumption:

Weak Additivity Assumption: If the relationship between X and **Parents**(G,X) in the true DAG G cannot be expressed as a noninvertible additive noise model, there does not exist a Y in **Parents**(G,X) and alternative DAG G' such that Y and **Parents**(G',Y) can be expressed as a noninvertible additive noise model where X is included in **Parents**(G',Y).

This assumption does rule out invertible additive noise models or many cases where noise may not be additive, only the hypothetical case where we can fit an additive noise model to the data, but only in the incorrect direction. Weak additivity can be considered an extension of the simplicity intuitions underlying the causal faithfulness assumption, i.e. a complicated true model will not generate data resembling a different simpler model. Faithfulness can fail, but under a broad range of distributions, violations are Lebesgue measure zero [Spirtes, Glymour, & Scheines 2000]. Whether a similar justification can be given for the weak additivity assumption is an open question.

kPC is both correct and complete, i.e. it converges to the correct DAG or smallest possible equivalence class of DAGs in the limit under weak additivity and the assumptions of the PC algorithm.

3.4.1 Example - Auto MPG

Figure 9 shows the structures learned for the Auto MPG dataset, which records *MPG* fuel consumption of 398 automobiles in 1983 with 8 characteristics from the UCI database (Asuncion & Newman, 2007). The nominal variables *Year* and *Origin* were excluded.

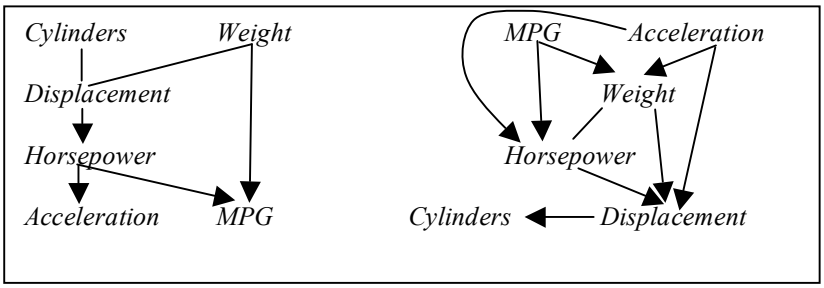


Figure 9: Automobile Models

The PC result indicates *MPG* causes *Weight* and *Horsepower*, and *Acceleration* causes *Weight*, *Horsepower*, and *Displacement*, which are clearly false. kPC finds the more plausible chain *Displacement* → *Horsepower* → *Acceleration* and finds *Horsepower* and *Weight* cause *MPG*.

3.4.2 Example - Forest Fires

The Forest Fires dataset contains 517 recordings of meteorological for forest fires observed in northeast Portugal and the total area burned (*Area*) [Asuncion & Newman 2007]. We again exclude nominal variables *Month* and *Year*. Figure 10 shows the

structures learned by PC and kPC for this dataset. kPC finds every variable other than *Area* is a cause of *Area*, which is sensible since each of these variables were included in the dataset by domain experts as predictors which influence the total area burned by forest fires.

The PC structure, however, indicates that Area is not associated with any of the variables, which are all assumed to be predictors by experts.

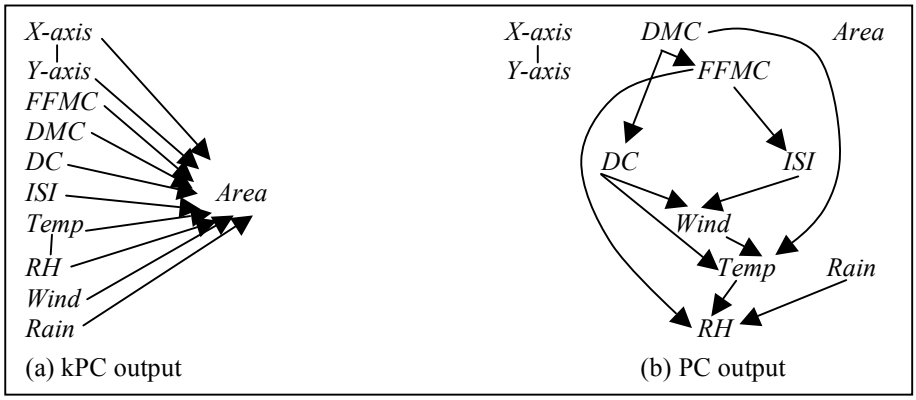


Figure 10: kPC and PC Forest Fires

4 Search For Latent Variable Models

The assumption that the observed variables are causally sufficient is usually unwarranted. In this section, we describe searches that do not make this assumption.

4.1 Distribution and Conditional Independence Equivalence

Let **O** be the set of observed variables, which may not be causally sufficient. If G_1 is a DAG over $\mathbf{V_1}$, G_2 is a DAG over $\mathbf{V_2}$, $\mathbf{O} \subseteq \mathbf{V_1}$, and $\mathbf{O} \subseteq \mathbf{V_2}$, G_1 and G_2 are **O**-conditional independence equivalent, if they both entail the same set of conditional independence relations among the variables in **O** (i.e. they have the same d-separation relations among the variables in **O**). $<G_1,\Theta_1>$ and $<G_2,\Theta_2>$ are **O-distribution equivalent** with respect to the parametric families Θ_1 and Θ_2 if and only if they represent the same set of marginal distributions over **O**.

It is possible that two directed graphs are conditional independence equivalent, or even distributionally equivalent (relative to given parametric families) but are not **O**-distributionally equivalent (relative to the same parametric families), as long as at least one of them contains a latent variable. Although there are algebraic techniques that determine when two Bayesian networks with latent variables are **O**-distributionally equivalent for some parametric families, or find features common to an **O**-distributional equivalence class, known algorithms to do so are not computationally feasible [Geiger & Meek 1999] for models with more than a few variables. In addition, if an unlimited number of latent variables are allowed, the number of DAGs that are **O**-distributionally equivalent may be infinite. Hence, instead of searching for **O**-distribution equivalence classes of models, we will describe how to search for **O**-conditional independence classes

of models. This is not as informative as the computationally infeasible strategy of searching for **O**-distribution equivalence classes, but is nevertheless correct.

It is often far from intuitive what constitutes a complete set of graphs **O**-conditional independence equivalent to a given graph although algorithms for deciding this now exist [Ali, Richardson, & Spirtes 2009].

4.2 The Fast Causal Inference Algorithm

The PC algorithm gives an asymptotically correct representation of the conditional independence equivalence class of a DAG without latent variables by outputting a pattern that represents all of the features that the DAGs in the equivalence class have in common. The same basic strategy can be used without assuming causal sufficiency, but the rules for detecting adjacencies and orientations are much more complicated, so we will not describe them in detail. The FCI algorithm[6] outputs an asymptotically correct representation of the **O**-conditional independence equivalence class of the true causal DAG (assuming the Causal Markov and Causal Faithfulness Principles), in the form of a graphical structure called a partial ancestral graph (PAG) that represents some of the features that the DAGs in the equivalence class have in common. The FCI algorithm takes as input a sample, distributional assumptions, optional background knowledge (e.g. time order), and a significance level, and outputs a partial ancestral graph. Because the algorithm uses only tests of conditional independence among sets of observed variables, it avoids the computational problems involved in calculating posterior probabilities or scores for latent variable models.

Just as the pattern can be used to predict the effects of some manipulations, a partial ancestral graph can also be used to predict the effects of some manipulations. Instead of calculating the effects of manipulations for which every member of the **O**-distribution equivalence class agree, we can calculate the effects only of those manipulations for which every member of the **O**-conditional independence equivalence agree. This will typically predict the effects of fewer manipulations than could be predicted given the **O**-distributional equivalence class (because a larger set of graphs have to make the same prediction), but the predictions made will still be correct.

Even though the set *S* of DAGs in an **O**-conditional independence equivalence class is infinite, it is still possible to extract the features that the members of *S* have in common. For example, every member of the conditional independence class over **O** that contains the DAG in Figure 11 has a directed path from *PE* to *CP* and no latent common cause of *PE* and *CP*. This is informative because even though the data do not help choose between members of the equivalence class, insofar as the data are evidence for the disjunction of the members in the equivalence class, they are evidence that *PE* is a cause of *CP*.

[6]The FCI algorithm is similar to Pearl's IC* algorithm [Pearl 2000] in many respects, and uses concepts bases on IC*; however IC* is computationally and statistically feasible only for a few variables.

A partial ancestral graph is analogous to a pattern, and represents the features common to an **O**-conditional independence class. Figure 11 shows an example of a DAG and the corresponding partial ancestral graph over **O** = {*IQ*, *SES*, *PE*, *CP*, *SEX*}. Two variables *A* and *B* are adjacent in a partial ancestral graph that represents an **O**-conditional independence class, when *A* and *B* are not entailed to be independent (i.e. they are d-connected) conditional on any subset of the variables in **O**\{*A*,*B*} for each DAG in the **O**-conditional independence class. The "–" endpoint of the $PE \rightarrow CP$ edge means that *PE* is an ancestor of *CP* in every DAG in the **O**-conditional independence class. The ">" endpoint of the $PE \rightarrow CP$ edges means that *CP* is not an ancestor of *PE* in any member of the **O**-conditional independence class. The "o" endpoint of the *SES* o–o *IQ* edge makes no claim about whether *SES* is an ancestor of *IQ* or not.

Applying the FCI algorithm to the Sewell and Shah data yields the PAG in Figure 11. The output predicts that when *PE* is manipulated, the following conditional probabilities hold: $P(CP=0|PE=0) = .063$; $P(CP=1|PE=0) = .937$; $P(CP=0|PE=1) = .572$; $P(CP=1PE=1) = .428$. These estimates are close to the estimates given by the output of the PC algorithm, although unlike the PC algorithm the output of the FCI algorithm posits the existence of latent variables. A bootstrap test of the output run at significance level 0.001 yielded the same results on 8 out of 10 samples. In the other two samples, the algorithm could not calculate the effect of the manipulation.

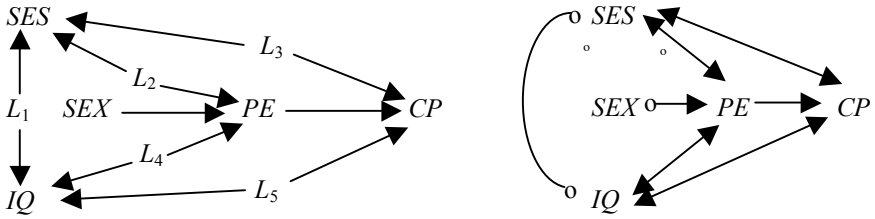


Figure 11: DAG and Partial Ancestral Graph

4.2.1 Online Course

Data from an online course provides an example where there was some experimental confirmation of the FCI causal model. Carnegie Mellon University offers a full semester online course that serves as a tutor on the subject of causal reasoning.[7] The course contains a number of different modules that contain both text and interactive online exercises that illustrate various concepts. Each module ends with a quiz that students must take. The interactive exercises are purely voluntary and play no role in calculating the student's final grade. It is possible to print the text from the online modules, but a student who studies from the printed text cannot use the online interactive exercises. The following variables were measured for each student:

- Pre-test (%)
- Print-outs (% modules printed)
- Quiz Scores (avg. %)

[7]See http://oli.web.cmu.edu/openlearning/forstudents/freecourses/csr

- Voluntary Exercises (% completed)
- Final Exam (%)
- 9 other variables

Using data from 2002, and some background knowledge about causal order, the output of the FCI algorithm was the PAG shown in Figure 12a. That model predicts that interventions that stops students from printing out the text and encourages students to use the online interactive exercises should raise the final grade in the class.

In 2003, students were advised that completing the voluntary exercises seemed to be important in helping grades, but that printing out the modules seemed to prevent completing the voluntary exercises. They were advised that, if they printed out the text they should make extra effort to go online and complete the interactive online exercises. Data on the same variables was gathered in 2003, and the output of the FCI algorithm is shown Figure 12b. The interventions to discourage printing and encourage the use of the online interactive exercises were largely successful, and the PAG output by the FCI algorithm from the 2003 data is exactly the PAG one would expect after intervening on the PAG output by the FCI algorithm from the 2002 data.

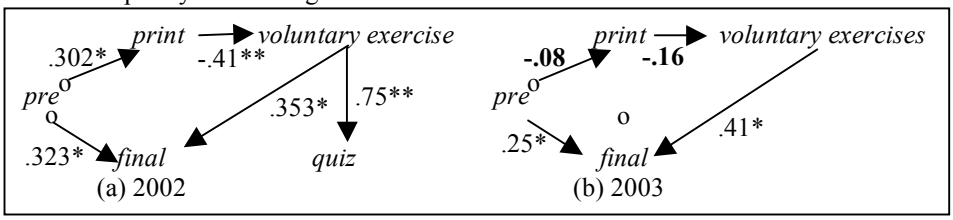


Figure 12: Online Course Printing

4.3 Errors in Variables: Combining Constraint Based Search and Bayesian Reasoning

In some cases the parameters of the output of the FCI algorithm are not identifiable or it is important to find not a particular latent variable model, but an equivalence class of latent variable models. In some of those cases the FCI algorithm can be combined with Bayesian methods.

4.3.1 Example - Lead and IQ

The next example shows how the FCI algorithm can be used to find a PAG, which can then be used as a starting point for a search for a latent variable DAG model and Bayesian estimation of parameters. It also illustrates how such a procedure produces different results than simply applying regression or using regression to generate more sophisticated models, such as errors-in-variables models.

By measuring the concentration of lead in a child's baby teeth, Herbert Needleman was the first epidemiologist to even approximate a reliable measure of cumulative lead exposure. His work helped convince the United States to eliminate lead from gasoline and most paint [Needleman 1979]. In their 1985 article in *Science* [Needleman, Geiger, & Frank 1985], Needleman, Geiger and Frank gave results for a multivariate linear regression of children's IQ on lead exposure. Having started their analysis with almost 40

covariates, they were faced with a variable selection problem to which they applied backwards-stepwise variable selection, arriving at a final regression model involving lead and five of the original 40 covariates. The covariates were measures of genetic contributions to the child's IQ (the parent's IQ), the amount of environmental stimulation in the child's early environment (the mother's education), physical factors that might compromise the child's cognitive endowment (the number of previous live births), and the parent's age at the birth of the child, which might be a proxy for many factors. The measured variables they used are as follows:

- *ciq* - child's verbal IQ score
- *piq* - parent's IQ scores
- *lead* - measured concentration in baby teeth
- *mab* - mother's age at child's birth
- *med* - mother's level of education in years
- *fab* - father's age at child's birth
- *nlb* - number of live births previous to the sampled child

The standardized regression solution[8] is as follows, with t-ratios in parentheses. Except for *fab*, which is significant at 0.1, all coefficients are significant at 0.05, and $R^2 = .271$.

$$\hat{ciq} = -.143\ lead + .219\ med + .247\ piq + .237\ mab - .204\ fab - .159\ nlb$$

(2.32) (3.08) (3.87) (1.97) (1.79) (2.30)

This analysis prompted criticism from Steve Klepper and Mark Kamlet, economists at Carnegie Mellon [Klepper, 1988/Klepper, Kamlet, & Frank 1993]. Klepper and Kamlet correctly argued that Needleman's statistical model (a linear regression) neglected to account for measurement error in the regressors. That is, Needleman's measured regressors were in fact imperfect proxies for the actual but latent causes of variations in IQ, and in these circumstances a regression analysis gives a biased estimate of the desired causal coefficients and their standard errors. Klepper and Kamlet constructed an errors-in-variables model to take into account the measurement error. See Figure 13, where the latent variables are in boxes, and the relations between the regressors are unconstrained.

Unfortunately, an errors-in-variables model that explicitly accounts for Needleman's measurement error is "underidentified," and thus cannot be estimated by classical techniques without making additional assumptions. Klepper, however, worked out an ingenious technique to bound the estimates, provided one could reasonably bound the amount of measurement error contaminating certain measured regressors [Klepper, 1988; Klepper et al. 1993]. The required measurement error bounds vary with each problem, however, and those required in order to bound the effect of actual lead exposure below 0 in Needleman's model seemed wholly unreasonable. Klepper concluded that the statistical evidence for Needleman's hypothesis was indeed weak. A Bayesian analysis, based on Gibbs sampling techniques, found that several posteriors corresponding to different priors lead to similar results. Although the size of the Bayesian point estimate

[8] The covariance data for this reanalysis was originally obtained from Needleman by Steve Klepper, who generously forwarded it. In this, and all subsequent analyses described, the correlation matrix was used.

for lead's influence on *IQ* moved up and down slightly, its sign and significance (the 95% central region in the posterior over the *lead-iq* connection always included zero) were robust.

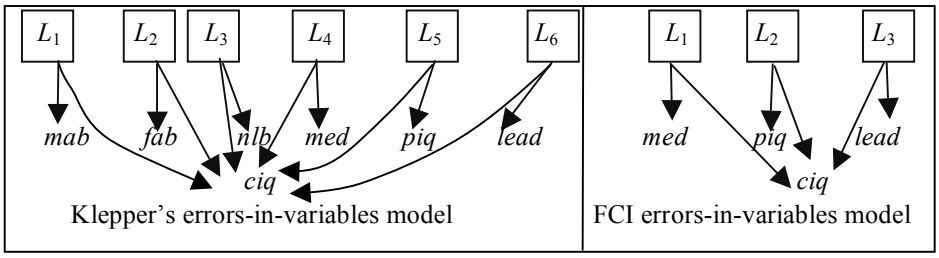


Figure 13: Errors-in-Variables Models

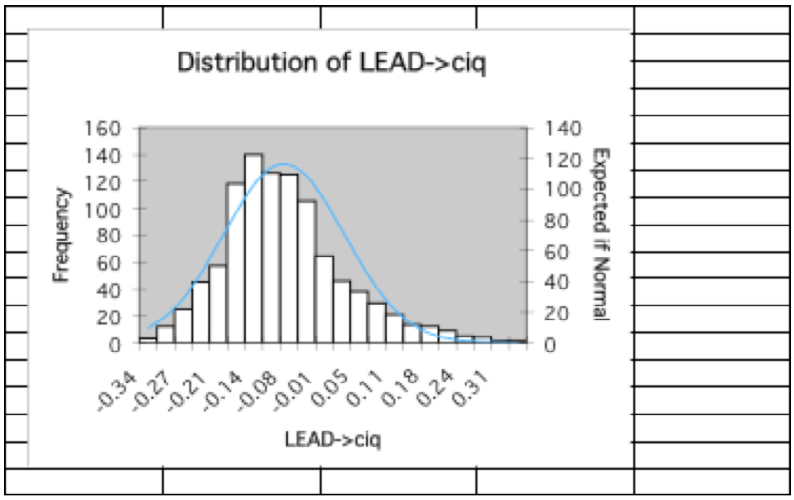


Figure 14: Posterior for Klepper's Model

A reanalysis using the FCI algorithm produced different results [Scheines 2000]. Scheines first used the FCI algorithm to generate a PAG, which was subsequently used as the basis for constructing an errors-in-variables model. The FCI algorithm produced a PAG that indicated that *mab*, *fab*, and *nlb* are *not* adjacent to *ciq*, contrary to Needleman's regression.[9] If we construct an errors-in-variables model compatible with the PAG produced by the FCI algorithm, the model does not contain *mab*, *fab*, or *nlb*. See Figure 13. (We emphasize that there are other models compatible with the PAG, which are not errors-in-variables models; the selection of an error-in-variables model from the

[9] The fact that *mab* had a significant regression coefficient indicates that *mab* and *ciq* are correlated conditional on the other variables; the FCI algorithm concluded that *mab* is not a cause of *ciq* because *mab* and *ciq* are unconditionally uncorrelated.

set of models represented by the PAG is an assumption.) In fact the variables that the FCI algorithm eliminated were precisely those, which required unreasonable measurement error assumptions in Klepper's analysis. With the remaining regressors, Scheines specified an errors-in-variables model to parameterize the effect of actual lead exposure on children's IQ. This model is still underidentified but under several priors, nearly all the mass in the posterior was over negative values for the effect of actual lead exposure (now a latent variable) on measured IQ. In addition, applying Klepper's bounds analysis to this model indicated that the effect of actual lead exposure on *ciq* was bounded below zero given reasonable assumptions about the degree of measurement error.

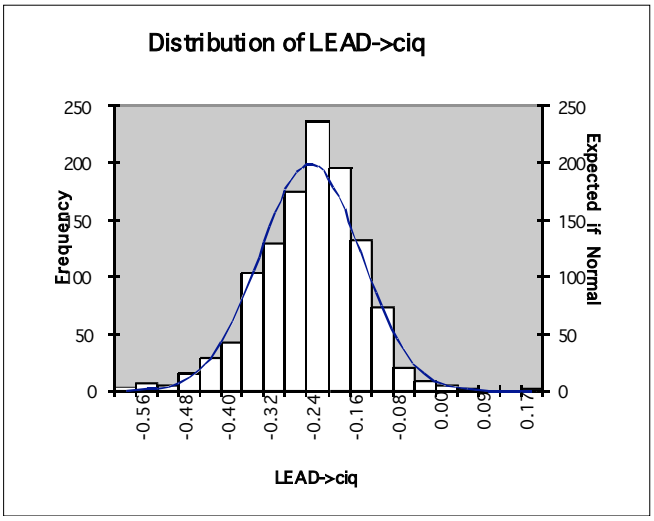


Figure 15: Posterior for FCI model

4.4 BuildPureClusters and MIMBuild

Searches using conditional independence constraints are correct, but completely uninformative for some common kinds of data sets. Consider the model *S* in Figure 16. The data comes from a survey of test anxiety indicators administered to 335 grade 12 male students in British Columbia [Gierl & Todd 1996]. The survey contains 20 measures of symptoms of anxiety under test conditions. Each question is about a symptom of anxiety. For example, question 8 is about how often one feels "jittery when taking tests". The answer is observed on a four-point approximately Likert scale (almost never, sometimes, often, or almost always). As in many such analyses, we will assume that the variables are approximately Gaussian.

Each *X* variable represents an answer to a question on the survey. For reasons to be explained later, not all of the questions on the test have been included in the model. There are three unobserved common causes in the model: *Emotionality*, *Care about achieving* (which will henceforth be referred to as *Care*) and *Self-defeating*. The test questions are of little interest in themselves; of more interest is what information they reveal about some unobserved psychological traits. If *S* is correct, there are no conditional

independence relations among the X variables alone - the only entailed conditional independencies require conditioning on an unobserved common cause. Hence the FCI algorithm would return a completely unoriented PAG in which every pair of variables in **X** is adjacent. Such a PAG makes no predictions at all about the effects of manipulations of the observed variables.

Furthermore, in this case, the effects of manipulating the observed variables (answers to test questions) are of no interest - the interesting questions are about the effects of manipulating the unobserved variables and the qualitative causal relationships between them.

Although PAGs can reveal the existence of latent common causes (as by the double-headed arrows in Figure 11 for example), before one could make a prediction about the effect of manipulating an unobserved variable(s), one would have to identify what the variable (or variables) is, which is never possible from a PAG.

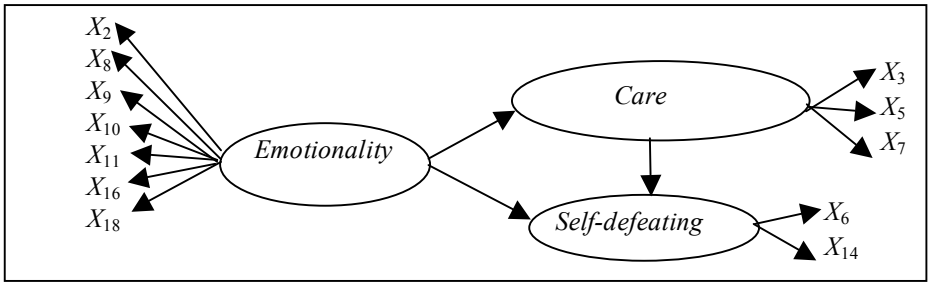


Figure 16: SEM S

Models such as S are *multiple indicator models*, and can be divided into two parts: the measurement model, which contains the edges between the unobserved variables and the observed variables (e.g. *Emotionality* $\rightarrow X_2$), and the structural model, which contains the edges between the unobserved variables (e.g. *Emotionality* $\rightarrow$ *Care*).

The **X** variables in S ($\{X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{14}, X_{16}, X_{18}\}$) were chosen with the idea that they indirectly measure some psychological trait that cannot be directly observed. Ideally, the **X** variables can be broken into clusters, where each variable in the cluster is caused by one unobserved cause common to the members of the cluster, and a unique error term uncorrelated with the other error terms, and nothing else. From the values of the variables in the cluster, it is then possible to make inferences about the value of the unobserved common cause. Such a measurement model is called *pure*. In psychometrics, pure measurement models satisfy the property of local independence: each measured variable is independent of all other variables, conditional on the unobserved variable it measures. In Figure 16, the measurement model of S is pure.

If the measurement model is impure (i.e. there are multiple common causes of a pair of variables in **X**, or some of the **X** variables cause each other) then drawing inferences about the values of the common causes is much more difficult. Consider the set **X'** = **X** $\cup$ $\{X_{15}\}$. If X_{15} indirectly measured (was a direct effect of) the unobserved variable *Care*, but X_{10} directly caused X_{15}, then the measurement model over the expanded set of

variables would not be pure. If a measurement model for a set **X'** of variables is not pure, it is nevertheless possible that some subset of **X'** (such as **X**) has a pure measurement model. If the only reason that the measurement model is impure is that X_{10} causes X_{15} then **X** = **X'**\{X_{15}} does have a pure measurement model, because all the "impurities" have been removed. *S* does not contain all of the questions on the survey precisely because various tests described below indicated that they some of them needed to be excluded in order to have a pure measurement model.

The task of searching for a multiple indicator model can then be broken into two parts: first finding clusters of variables so that the measurement model is pure; second, use the pure measurement model to make inferences about the structural model.

Factor analysis is often used to determine the number of unmeasured common causes in a multiple indicator model, but there are important theoretical and practical problems in using factor analysis in this way. Factor analysis constructs models with unobserved common causes (factors) of the observed **X** variables. However, factor analysis models typically connect each unobserved common cause (factor) to each **X** variable, so the measurement model is not pure. A major difficulty with giving a causal interpretation to factor analytic models is that the observed distribution does not determine the covariance matrix among the unobserved factors. Hence, a number of different factor analytic models are compatible with the same observed data [Harman 1976]. In order to reduce the underdetermination of the factor analysis model by the data, it is often assumed that the unobserved factors are independent of each other; however, this is clearly not an appropriate assumption for unobserved factors that are supposed to represent actual causes that may causally interact with each other. In addition, simulation studies indicate that factor analysis is not a reliable tool for estimating the correct number of unobserved common causes [Glymour 1998].

On this data set, factor analysis indicates that there are 2 unobserved direct common causes, rather than 3 unobserved direct common causes [Bartholomew, Steele, Moustaki, & Galbraith 2002]. If a pure measurement model is constructed from the factor analytic model by associating each observed *X* variable only with the factor that it is most strongly associated with, the resulting model fails a statistical test (has a p-value of zero) [Silva, Scheines, Glymour, & Spirtes 2006]. A search for pure measurement models that depends upon testing vanishing tetrad constraints is an alternative to factor analysis. Conceptually, the task of building a pure measurement model from the observed variables can be broken into 3 separate tasks:

1. Select a subset of the observed variables that form a pure measurement model.
2. Determine the number of clusters (i.e. the number of unobserved common causes) that the observed variables measure.
3. Cluster the observed variables into the proper groups (so each group has exactly one unobserved direct common cause.)

It is possible to construct pure measurement models using vanishing tetrad constraints as a guide [Silva et al. 2006]. A *vanishing tetrad constraint* holds among *X*, *Y*,

Z, W when $\text{cov}(X,Y) \cdot \text{cov}(Z,W) - \text{cov}(X,Z) \cdot \text{cov}(Y,W) = 0$. A pure measurement model entails that each X_i variables is independent of every other X_j variable conditional on its unobserved parent, e.g. S entails X_2 is independent of X_j conditional on *Emotionality*. These conditional independence relations cannot be directly tested, because *Emotionality* is not observed. However, together with the other conditional independence relations involving unobserved variables entailed by S, they imply vanishing tetrad constraints on the observed variables that reveal information about the measurement model that does not depend upon the structural model among the unobserved common causes. The basic idea extends back to Spearman's attempts to use vanishing tetrad constraints to show that there was a single unobserved factor of intelligence that explained a variety of observed competencies [Spearman 1904].

Because X_2 and X_8 have one unobserved direct common cause (*Emotionality*), and X_3 and X_5 have a different unobserved direct common cause (*Care*), S entails $\text{cov}_S(X_2, X_3) \cdot \text{cov}_S(X_5, X_8) = \text{cov}_S(X_2, X_5) \cdot \text{cov}_S(X_3, X_8) \neq \text{cov}_S(X_2, X_8) \cdot \text{cov}_S(X_3, X_5)$ for all values of the model's free parameters (here cov_S is the covariance matrix entailed by S).[10] On the other hand, because X_2, X_8, X_9, and X_{10} all have one unobserved common cause (*Emotionality*) as a direct common cause, the following vanishing tetrad constraints are entailed by S: $\text{cov}_S(X_2, X_8) \cdot \text{cov}_S(X_9, X_{10}) = \text{cov}_S(X_2, X_9) \cdot \text{cov}_S(X_8, X_{10}) = \text{cov}_S(X_2, X_{10}) \cdot \text{cov}_S(X_8, X_9)$ [Spirtes et al. 2001]. The BuildPureClusters algorithm uses the vanishing tetrad constraints as a guide to the construction of pure measurement models, and in the large sample limit reliably succeeds if there is a pure measurement model among a large enough subset of the observed variables [Silva et al. 2006].

In this example, BuildPureClusters automatically constructed the measurement model corresponding to the measurement model of S. The clustering on statistical grounds makes substantive sense, as indicated by the fact that it is similar to a prior theory-based clustering based on background knowledge about the content of the questions; however BuildPureClusters removes some questions, and splits one of the clusters of questions constructed from domain knowledge into two clusters.

Once a pure measurement model has been constructed, there are several algorithms for finding the structural model. One way is to estimate the covariances among the unobserved common causes, and then input the estimated covariances to the FCI algorithm. The output is then a PAG among the unobserved common causes. Alternative searches for the structural model include the MIMBuild and GESMIMBuild algorithms, which output patterns [Silva et al. 2006].

In this particular analysis, the MIMBuild algorithm, which also employs vanishing tetrad constraints, was used to construct a variety of output patterns corresponding to different values of the search parameters. The best pattern returned contains an undirected edge between every pair of unobserved common causes. (S is an example that is compatible with the pattern, but any other orientation of the edges among the three

[10] The inequality is based on an extension of the Causal Faithfulness Assumption that states that vanishing tetrad constraints that are not entailed for all values of the free parameters by the true causal graph are assumed not to hold.

unobserved common causes that does not create a cycle is also compatible with the pattern.) The resulting model (or set of models) passes a statistical test with a p-value of 0.47.

4.4.1 Example - Religion and Depression

Data relating religion and depression provides an example that shows how an automated causal search produces a model that is compatible with background knowledge, but fits much better than a model that was built from theories about the domain.

Bongjae Lee from the University of Pittsburgh organized a study to investigate religious/spiritual coping and stress in graduate students [Silva & Scheines 2004]. In December of 2003, 127 Masters in Social Works students answered a questionnaire intended to measure three main factors:

- *Stress*, measured with 21 items, each using a 7-point scale (from "not all stressful" to "extremely stressful") according to situations such as: "fulfilling responsibilities both at home and at school"; "meeting with faculty"; "writing papers"; "paying monthly expenses"; "fear of failing"; "arranging childcare";
- *Depression*, measured with 20 items, each using a 4-point scale (from "rarely or none" to "most or all the time") according to indicators as: "my appetite was poor"; "I felt fearful"; "I enjoyed life" "I felt that people disliked me"; "my sleep was restless";
- *Spiritual coping*, measured with 20 items, each using a 4-point scale (from "not at all" to "a great deal") according to indicators such as: "I think about how my life is part of a larger spiritual force"; "I look to God (high power) for strength in crises"; "I wonder whether God (high power) really exists"; "I pray to get my mind off of my problems";

The goal of the original study was to use graphical models to quantify how *Spiritual coping* moderates the association of *Stress* and *Depression*, and hypothesized that *Spiritual coping* reduces the association of *Stress* and *Depression*. The theoretical model (Figure 17) fails a chi-square test: $p = 0$. The measurement model produced by BuildPureClusters is shown in Figure 18. Note that the variables selected automatically are proper subsets of Lee's substantive clustering. The full model automatically produced with GESMIMBuild with the prior knowledge that *Stress* is not an effect of other latent variables is given in Figure 19. This model passes a chi square test, $p = 0.28$, even though the algorithm itself does not try to directly maximize the fit. Note that it supports the hypothesis that *Depression* causes *Spiritual Coping* rather than the other way around. Although this conclusion is not conclusive, the example does illustrate how the algorithm can find a theoretically plausible alternative model that fits the data well.

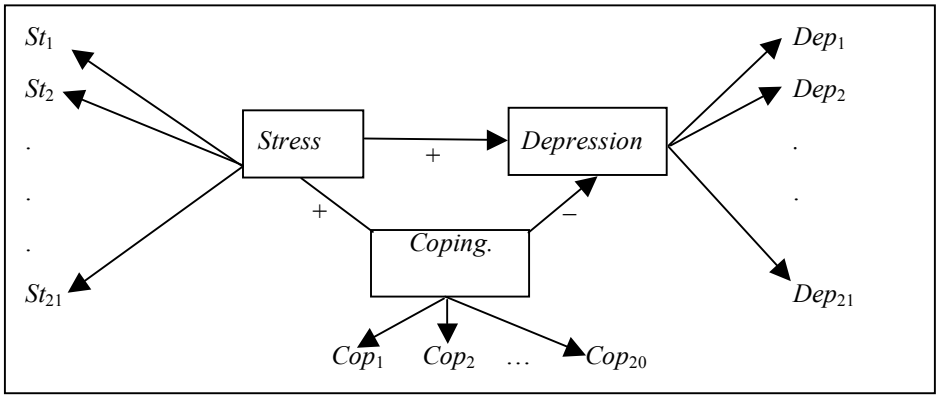


Figure 17: Model from Theory

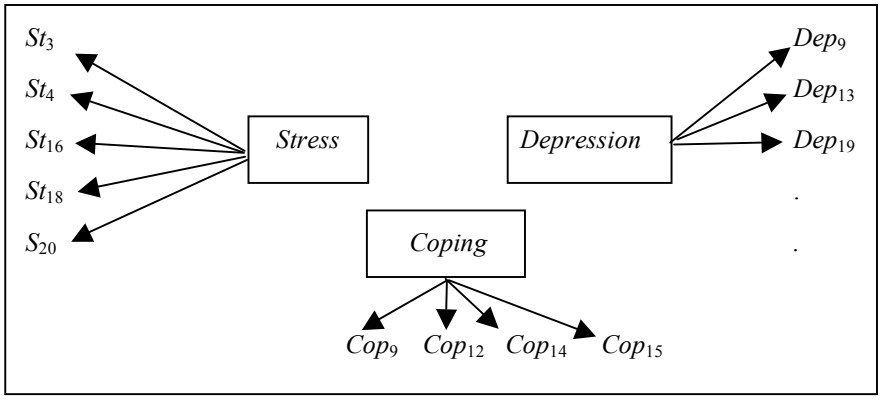


Figure 18: Output of BuildPureClusters

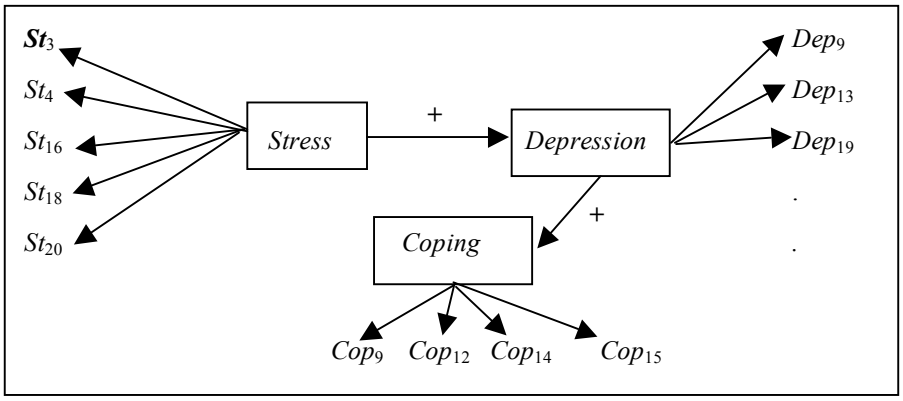


Figure 19: Output of GESMIMBuild

5 Time Series and Feedback

The models described so far are for "equilibrium." That is, they assume that an intervention fixes the values of a variable or variables, and that the causal process results in stable values of effect variables, so that time can be ignored. When time cannot be ignored, representation, interventions and search are all more complicated.

Time series models with a causal interpretation are naturally represented by directed acyclic graphs in at least three different forms: A graph whose variables are indexed by time, a "unit" graph giving a substructure that is repeated in the time indexed graph, and a finite graph that may be cyclic. Models of the first kind have been described as "Dynamical Causal Models" but the description does not address the difficulties of search. Pursuing a strategy of the PC or FCI kind, for example, requires a method of correctly estimating conditional independence relations.

5.1 Time series models

Chu and Glymour [2008] describe conditional independence tests for additive models, and use these tests in a slight modification of the PC and FCI algorithms. The series data is examined by standard methods to determine the requisite number of lags. The data are then replicated a number of times equal to the lags, delaying the first replicant by one time step, the second by two time steps, and so on, and conditional independence tests applied to the resulting sets of data. They illustrate the algorithm with climate data.

Climate teleconnections are associations of geospatially remote climate phenomena produced by atmospheric and oceanic processes. The most famous, and first established teleconnection, is the association of El Nino/Southern Oscillation (*ENSO*) with the failure of monsoons in India. A variety of associations have been documented among sea surface temperatures (*SST*), atmospheric pressure at sea level (*SLP*), land surface temperatures (*LST*) and precipitation over land areas. Since the 1970s data from a sequence of satellites have provided monthly (and now daily) measurements of such variables, at resolutions as small as 1 square kilometer. Measurements in particular spatial regions have been clustered into time-indexed indices for the regions, usually by principal components analysis, but also by other methods. Climate research has established that some of these phenomena are exogenous drivers of others, and has sought physical mechanisms for the teleconnections.

Chu and Glymour (2008) consider data from the following 6 ocean climate indices, recorded monthly from 1958 to 1999, each forming a time series of 504 time steps:

- *QBO* (*Quasi Biennial Oscillation*): Regular variation of zonal stratospheric winds above the equator
- *SOI* (*Southern Oscillation*): Sea Level Pressure (*SLP*) anomalies between Darwin and Tahiti
- *WP* (*Western Pacific*): Low frequency temporal function of the 'zonal dipole' *SLP* spatial pattern over the North Pacific.
- *PDO* (*Pacific Decadal Oscillation*): Leading principal component of monthly Sea Surface Temperature (*SST*) anomalies in the North Pacific Ocean, poleward of 20° N

- *AO* (*Arctic Oscillation*): First principal component of SLP poleward of 20° N
- *NAO* (*North Atlantic Oscillation*) Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland

Some connections among these variables are reasonably established, but are not assumed in the analysis that follows. In particular, *SO* and *NAO* are thought to be exogenous drivers.

After testing for stationarity, the PC algorithm yields the structure for the climate data shown in Figure 20. The double-headed arrows indicate the hypothesis of common unmeasured causes.[11] So far as the exogenous drivers are concerned, the algorithm output is in accord with expert opinion.

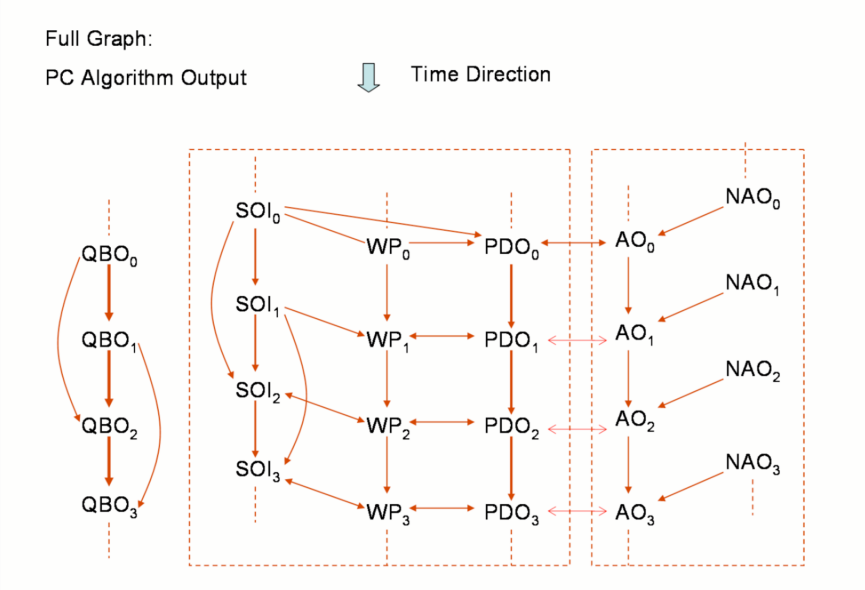


Figure 20: Climate Time Series

Monthly time series of temperatures and pressures at the sea surface present a case in which one might think that the causal processes take place more rapidly that the sampling rate. If so, then the causal structure in between time samples, the "contemporaneous" causal structure, should look much like a unit of the time series causal structure. When we sample at intervals of time as in economic, climate, and other time series, can we discover what goes on in the intervals between samples? Swanson and Granger suggested that an autoregression be used to remove the effects on each variable of variables at previous times, and a search could then be applied to the residual correlations [Swanson & Granger 1997]. The search they suggested was to assume a chain and to test

[11] When under the usual assumptions, the PC algorithm produces double headed arrows, they reliably indicate common unobserved causes as will FCI. But unlike FCI, PC is not complete in this respect.

it by methods described in [Glymour, Scheines, Spirtes, & Kelly 1987], some of the work whose aims and methods Cartwright previously sought to demonstrate is impossible [Cartwright 1994]. But a chain model of contemporaneous causes is far too special a case. Hoover & Demiralp, and later, Moneta & Spirtes, proposed applying PC to the residuals [Hoover & Demiralp 2003; Moneta & Spirtes 2006]. (Moneta also worked out the statistical corrections to the correlations required by the fact that they are obtained as residuals from regressions.) When that is done for model above, the result is the unit structure of the time series: $QBO \quad SOI \rightarrow WP \leftrightarrow PDO \leftrightarrow AO \leftarrow NA$.

5.2 Cyclic Graphs

Since the 1950s, the engineering literature has developed methods for analyzing the statistical properties of linear systems described by cyclic graphs. The literature on search is more recent. Spirtes showed that linear systems with independent noises satisfy a simple generalization of d-separation, and the idea of faithfulness is well-defined for such systems [Spirtes 1995]; Pearl & Dechter extended these results to discrete variable systems [Pearl & Dechter 1996]. Richardson proved some of the essential properties of such graphs, and developed a pointwise consistent PC style algorithm for search [Richardson 1996]. More recently, an extension of the LiNGAM algorithm for linear, cyclic, non-Gaussian models has been developed [Lacerda et al. 2008].

5.3 Distributed Multiple Data Sets: ION and IMaGES

Data mining has focused on learning from a single database, but inferences from multiple databases are often needed in social science, multiple subject time series in physiological and psychological experiments, and to exploit archived data in many subjects. Such data sets typically pose missing variable problems: some of what is measured in one study or for one subject, may not be measured in another. In many cases such multiple data sets cannot, for physical, sociological or statistical reasons, be merged into a single data set with missing variables. There are two strategies for this kind of problem: learn a structure or set of structures separately for each data set and then find the set of structures consistent with the several "marginal" structures, or learn a single set of structures by evaluating steps in a search procedure using all of the data sets. The first strategy could be carried out using PC, kPC GES, FCI, LiNGAM or other procedure on each data set, and then using an algorithm that returns a description of the set of all graphs, or mixed graphs, consistent with the results from each database [Tillman, Danks, & Glymour 2008]. Tillman, Danks and Glymour have used such a procedure in combination with GES and FCI. The result in some (surprising) cases is a unique partial ancestral graph, and in other cases a large set of alternatives collectively carrying little information. The second strategy has been implemented in the IMaGES algorithm [Ramsey et al. 2009]. The algorithm uses GES, but at each step in the evaluation of a candidate edge addition or removal, the candidate is scored separately by BIC on each data set and the average of the BIC scores is used by the algorithm in edge addition or deletion choices. The IMaGES strategy is more limited—no consistency proof is available when the samples are from mixed distributions, and a proof of convergence of

averages of BIC scores to a function of posteriors is only available when the sample sizes of several data sets are equal. Nonetheless, IMaGES has been applied to fMRI data from multiple subjects with remarkably good results. For example, an fMRI study of responses to visually presented rhyming and non-rhyming words and non-words should produce a left hemisphere cascade leading to right hemisphere effects, which is exactly what IMaGES finds, using only the prior knowledge that the input variable is not an effect of other variables.

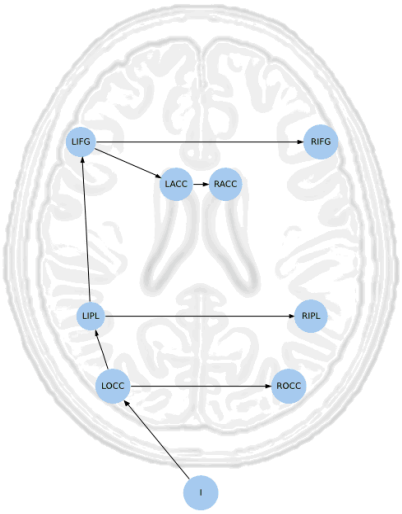


Figure 21: IMaGES Output for fMRI Data

6 Conclusion

The discovery of d-separation, and the development of several related notions, has made possible principled search for causal relations from observational and quasi-experimental data in a host of disciplines. New insights, algorithms and applications have appeared almost every year since 1990, and they continue. We are seeing a revolution in understanding of what is and is not possible to learn from data, but the insights and methods have seeped into statistics and applied science only slowly. We hope that pace will quicken.

7 Appendix

A *directed graph* (e.g. G_1 of Figure 22) consists of a set of vertices and a set of directed edges, where each edge is an ordered pair of vertices. In G_1, the vertices are $\{A,B,C,D,E\}$, and the edges are $\{B \rightarrow A, B \rightarrow C, D \rightarrow C, C \rightarrow E\}$. In G_1, B is a *parent* of A, A is a *child* of B, and A and B are *adjacent* because there is an edge $A \rightarrow B$. A *path* in a directed graph is a sequence of adjacent edges (i.e. edges that share a single common endpoint). A *directed path* in a directed graph is a sequence of adjacent edges all pointing in the same direction. For example, in G_1, $B \rightarrow C \rightarrow E$ is a directed path from B to E. In contrast, $B \rightarrow C \leftarrow D$ is a path, but not a directed path in G_1 because the two edges do not point in the same direction; in addition, C is a *collider* on the path because both edges on

the path are directed into *C*. A triple of vertices <*B*,*C*,*D*> is a *collider* if there are edges *B* → *C* ← *D* in G_1; <*B*,*C*,*D*> is an *unshielded collider* if in addition there is no edge between *B* and *D*. *E* is a *descendant* of *B* (and *B* is an *ancestor* of *E*) because there is a directed path from *B* to *E*; in addition, by convention, each vertex is a descendant (and ancestor) of itself. A directed graph is *acyclic* when there is no directed path from any vertex to itself: in that case the graph is a directed acyclic graph, or DAG for short.

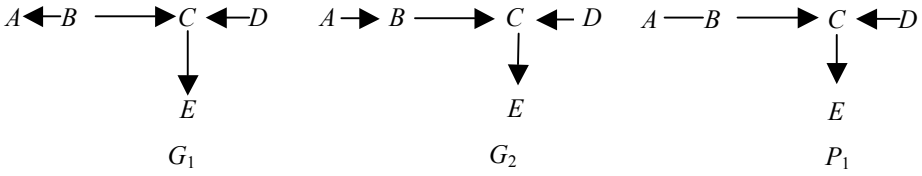


Figure 22: G_1, G_2, **and** P_1 **(pattern for** G_1 **and** G_2**)**

A probability distribution $P(\mathbf{V})$ satisfies the *local Markov condition* for a DAG G_1 with vertices **V** when each variable is independent of its non-parental non-descendants conditional on its parents. A *Bayesian network* is an ordered pair of a directed acyclic graph *G* and a set of probability distributions that satisfy the local Markov condition for *G*.

The graphical relationship among sets of variables in a DAG *G* called "d-separation" determines which conditional independence relations are entailed by satisfying the local directed Markov property). Following [Pearl 1988], in a DAG *G*, for disjoint variable sets **X**, **Y**, and **Z**, **X** and **Y** are *d-separated* conditional on **Z** in *G* if and only if there exists no path *U* between an $X \in \mathbf{X}$ and a $Y \in \mathbf{Y}$ such that (i) every collider on *U* has a descendent in **Z** and (ii) no other vertex on *U* is in **Z**. An important theorem in [Pearl 1988] is that a DAG *G* entails that **X** is independent of **Y** conditional on **Z** if and only if **X** is d-separated from **Y** conditional on **Z** in *G*.

A Bayesian network restricted to a parametric family <*G*,**Q**> where *G* is a DAG and **Q** is some parameterization of the DAG, e.g. multivariate Gaussian, has two distinct interpretations. First, it has a probabilistic interpretation as a distribution over the variables in *G*, for distributions that satisfy the local Markov condition for *G*. Under this interpretation, it is a useful tool for calculating conditional probabilities.

Second, it has a causal interpretation, and can be used to calculate the effects of manipulations. Intuitively, a manipulation of a variable is an exogenous action that *forces* a value (or a distribution over values) upon a variable in the system, e.g. as in a randomized experiment - if no exogenous action is taken on variable *X*, *X* is said to have undergone a null manipulation. An example of a manipulation is a randomized experiment, in which a distribution for some variables (e.g. ½ of the subjects take a given drug, and ½ of the subjects do not take the drug) is imposed from outside. The kinds of manipulations that we will consider are ideal in the sense that a manipulation of *X* directly affects only *X*.

X is a *direct cause* of *Y* relative to a set of variables **V** if there is a pair of manipulations (including possibly null manipulations, and including hypothetical manipulations in the many circumstances where no actual manipulations are feasible) of the values of the variables in **V**\{*Y*} that differ only in the value assigned to *X*, but that have different distributions for *Y*. This is in accord with the idea that the gold standard for determining causation is randomized experiments. (This is not a reduction of causality to non-causal concepts, because manipulation is itself a causal concept that we have taken as primitive.) Under the causal interpretation of DAGs, there is an edge $X \rightarrow Y$ when *X* is a direct cause of *Y* relative to the set of variables in the DAG. A set of variables **V** is *causally sufficient* if every direct cause (relative to **V**) of any pair of variables in **V**, is also in **V**. We will assume that causally interpreted DAGs are causally sufficient, although we will not generally all of the variables in a causally interpreted DAG are measured.

In automated causal search, the goal is to discover as much as possible about the true causal graph for a population from a sample from the joint probability distribution over the population, together with background knowledge (e.g. parametric assumptions, time order, etc.) This requires having some assumptions that link (samples from) probability distributions on the one hand, and causal graphs on the other hand. Extensive discussions of the following assumptions that we will make, including arguments for making the assumptions as well as limitations of the assumptions can be found in *Causation*, *Prediction*, & *Search* [Spirtes et al. 2001].

7.1 Causal Markov Assumption

The Causal Markov Assumption is a generalization of two commonly made assumptions: the immediate past screens off the present from the more distant past; and if *X* does not cause *Y* and *Y* does not cause *X*, then *X* and *Y* are independent conditional on their common causes. It presupposes that while the random variables of a unit in the population may causally interact, the units themselves are not causally interacting with each other.

Causal Markov Assumption: Let *G* be a causal graph with causally sufficient vertex set **V** and let *P* be a probability distribution over the vertices in **V** generated by the causal structure represented by *G*. *G* and *P* satisfy the Causal Markov Assumption if and only if for every *W* in **V**, *W* is independent of its non-parental non-descendants conditional on its parents in *G*.

In graphical terms, the Causal Markov Assumption states that in the population distribution over a causally sufficient set of variables, each variable is independent of its non-descendants and non-parents, conditional on its parents in the true causal graph.

While the Causal Markov Assumption allows for some causal conclusions from sample data, it only supports inferences that some causal connections exist - it does not support inferences that some causal connections do not exist. The following assumption does support the latter kind of inference.

7.2 Causal Faithfulness Assumption

Often the set of distributions that satisfy the local Markov condition for G is restricted to some parametric family Θ (e.g. Gaussian). In those cases, the set of distributions belonging to the Bayesian network will be denoted as $\boldsymbol{f}(<G,\Theta>)$, and $f(<G,\theta>)$ will denote a member of $\boldsymbol{f}(<G,\Theta>)$ for the particular value $\theta \in \Theta$ (and $f(<G,\theta>)$ is **represented** by $<G,\Theta>$). Let $I_f(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ denote that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in a distribution f.

If a DAG G does not entail that $I_{f(G,\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ for *all* $\theta \in \Theta$, nevertheless there may be *some* parameter values θ such that $I_{f(G,\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$. In that case say that $f(<G,\theta>)$ is **unfaithful** to G. In Pearl's terminology the distribution is *unstable* [Pearl 1988]. This would happen for example if taking birth control pills increased the probability of blood clots directly, but decreased the probability of pregnancy which in turn increased the probability of blood clots, and the two causal paths exactly cancelled each other. We will assume that such unfaithful distributions do not happen - that is there may be such canceling causal paths, but the causal paths do not exactly cancel each other.

Causal Faithfulness Assumption: For a true causal graph G over a causally sufficient set of variables $\mathbf{V}$, and probability distribution $P(\mathbf{V})$ generated by the causal structure represented by G, if G does not entail that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ then $\mathbf{X}$ is not independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in $P(\mathbf{V})$.

7.3 Conditional Independence Equivalence

Let $\mathbf{I}(<G,\Theta>)$ be the set of all conditional independence relations entailed by satisfying the local Markov condition. For any distribution that satisfies the local directed Markov property for G, all of the conditional independence relations in $\mathbf{I}(<G,\Theta>)$ hold. Since these independence relations don't depend upon the particular parameterization but only on the graphical structure and the local directed Markov property, they will henceforth be denoted by $\mathbf{I}(G)$.

G_1 and G_2 are *conditional independence equivalent* if and only if $\mathbf{I}(G_1) = \mathbf{I}(G_2)$. This occurs if and only if G_1 and G_2 have the same d-separation relations. A set of graphs that are all conditional independence equivalent to each other is a *conditional independence equivalence class*. If the graphs are all restricted to be DAGs, then they form a *DAG conditional independence equivalence class*. Two DAGs are conditional independence equivalent if and only if they have the same d-separation relations.

Theorem 1 (Pearl, 1988): Two directed acyclic graphs are conditional independence equivalent if and only if they contain the same vertices, the same adjacencies, and the same unshielded colliders.

For example, Theorem 1 entails that the set consisting of G_1 and G_2 in Figure 22 is a DAG conditional independence equivalence class. The fact that G_1 and G_2 are conditional independence equivalent, but are different causal models, indicates that in general any algorithm that relies only on conditional independence relations to discover the causal graph cannot (without stronger assumptions or more background knowledge) reliably

output a single DAG. A reliable algorithm could at best output the DAG conditional independence equivalence class, e.g. $\{G_1, G_2\}$.

Fortunately, Theorem 1 is also the basis of a simple representation called a pattern [Verma & Pearl 1990] of a DAG conditional independence equivalence class. Patterns can be used to determine which predicted effects of a manipulation are the same in every member of a DAG conditional independence equivalence class and which are not.

The adjacency phase of the PC algorithm is based on the following two theorems, where **Parents**(G,A) is the set of parents of A in G.

Theorem 2: If A and B are d-separated conditional on any subset **Z** in DAG G, then A and B are not adjacent in G.

Theorem 3: A and B are not adjacent in DAG G if and only if A and B are d-separated conditional on **Parents**(G,A) or **Parents**(G,B) in G.

The justification of the orientation phase of the PC algorithm is based on Theorem 4.

Theorem 4: If in a DAG G, A and B are adjacent, B and C are adjacent, but A and C are not adjacent, either B is in every subset of variables **Z** such that A and C are d-separated conditional on **Z**, in which case $<A,B,C>$ is not a collider, or B is in no subset of variables **Z** such A and C are d-separated conditional on Z, in which case $<A,B,C>$ is a collider.

A **pattern** (also known as a PDAG) P represents a DAG conditional independence equivalence class **X** if and only if:

1. P contains the same adjacencies as each of the DAGs in **X**;
2. each edge in P is oriented as $X \rightarrow Z$ if and only if the edge is oriented as $X \rightarrow Z$ in every DAG in **X**, and as X — Z otherwise.

There are simple algorithms for generating patterns from a DAG [Meek, 1995; Andersson, Madigan, & Perlman 1997; Chickering 1995]. The pattern P_1 for the DAG conditional independence equivalence class containing G_1 is shown in Figure 22. It contains the same adjacencies as G_1, and the edges are the same except that the edge between A and B is undirected in the pattern, because it is oriented as $A \leftarrow B$ in G_1, and oriented as $A \rightarrow B$ in G_2.

7.4 Distributional Equivalence

For multi-variate Gaussian distributions and for multinomial distributions, every distribution that satisfies the set of conditional independence relations in $\mathbf{I}(<G,\Theta>)$ is also a member of $\boldsymbol{f}(<G,\Theta>)$. However, for other families of distributions, it is possible that there are distributions that satisfy the conditional independence relations in $\mathbf{I}(<G,\Theta_\alpha>)$, but are not in $\boldsymbol{f}(<G,\Theta_\alpha>)$, i.e. the parameterization imposes constraints that are not conditional independence constraints [Lauritzen et al. 1990; Pearl 2000; Spirtes et al. 2001].

It can be shown that when restricted to multivariate Gaussian distributions, G_1 and G_2 in Figure 22 represent exactly the same set of probability distributions, i.e. $\boldsymbol{f}(<G_1,\Theta_1>)$

$= \boldsymbol{f}(<G_2,\Theta_2>)$. In that case say that $<G_1,\Theta_1>$ and $<G_2,\Theta_2>$ are *distributionally equivalent* (relative to the parametric family). Whether two models are distributionally equivalent depends not only on the graphs in the models, but also on the parameterization families of the models. A set of models that are all distributionally equivalent to each other is a *distributional equivalence class*. If the graphs are all restricted to be DAGs, then they form a *DAG distributional equivalence class*.

In contrast to conditional independence equivalence, distribution equivalence depends upon the parameterization families as well as the graphs. Conditional independence equivalence of G_1 and G_2 is a necessary, but not always sufficient condition for the distributional equivalence of $<G_1,\boldsymbol{\Theta}_A>$ and $<G_2,\boldsymbol{\Theta}_B>$.

Algorithms that rely on constraints beyond conditional independence may be able to output subsets of conditional independence equivalence classes, although without further background knowledge or stronger assumptions they could at best reliably output a DAG distribution equivalence class. In general, it would be preferable to take advantage of the non conditional independence constraints to output a subset of the conditional independence equivalence class, rather than simply outputting the conditional independence equivalence class. For some parametric families it is known how to take advantage of the non conditional independence constraints (sections 3.4 and 4.4); however in other parametric families, either there are no non conditional independence constraints, or it is not known how to take advantage of the non conditional independence constraints.

Acknowledgements: Clark Glymour and Robert Tillman thanks the James S. McDonnell Foundation for support of their research.

References

Ali, A. R., Richardson, T. S., & Spirtes, P. (2009). Markov Equivalence for Ancestral Graphs. *Annals of Statistics*, *37*(*5B*), 2808-2837.

Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. *Proceedings of the 2003 American Medical Informatics Association Annual Symposium*, Washington, DC, 21-25.

Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann Stat*, *25*(2), 505-541.

Asuncion, A. & Newman, D. J. (2007). UCI Machine Learning Repository.

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists (Texts in Statistical Science Series)*. Chapman & Hall/CRC.

Cartwright, N. (1994). *Nature's Capacities and Their Measurements (Clarendon Paperbacks)*. Oxford University Press, USA.

Chickering, D. M. (2002). Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, *3*, 507-554.

Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 87-98.

Chu, T., & Glymour, C. (2008). Search for Additive Nonlinear Time Series Causal Models. *Journal of Machine Learning Research*, 9(May):967-991.

Dempster, A. (1972). Covariance selection. *Biometrics*, *28*, 157-175.

Fukumizu, K., Gretton, A., Sun, X., & Scholkopf, B. (2008). Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems 20*.

Geiger, D. & Meek, C. (1999). Quantifier Elimination for Statistical Problems. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, 226-233.

Gierl, M. J. & Todd, R. W. (1996). A Confirmatory Factor Analysis of the Test Anxiety Inventory Using Canadian High School Students. *Educational and Psychological Measurement*, *56*(2), 315-324.

Glymour, C. (1998). What Went Wrong? Reflections on Science by Observation and the Bell Curve. *Philosophy of Science*, *65*(1), 1-32.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Scholkopf, B., & Smola, A. J. (2008) A kernel statistical test of independence, In *Advances in Neural Information Processing Systems 20*, 585-592.

Harman, H. H. (1976). *Modern Factor Analysis*. University Of Chicago Press.

Hoover, K. & Demiralp, S. (2003). Searching for the Causal Structure of a Vector Autoregression. *Oxford Bulletin of Economics and Statistics 65 (Supplement)*, *65*, 745-767.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Scholkopf, B. (2009). Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems 21*, 689-696.

Hoyer, P. O., Shimizu, S., & Kerminen, A. (2006). Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. *Third European Workshop on Probabilistic Graphical Models*, 155-162.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G. &Shimizu, S. (2008). Causal discovery of linear acyclic models with arbitrary distributions. *Proceedings of the Twentyfourth Annual Conference on Uncertainty in Artificial Intelligence*, 282-289.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5): 411 - 430.

Junning, L. & Wang, Z. (2009). Controlling the False Discovery Rate of the Association/Causality Structure Learned with the PC Algorithm. *Journal of Machine Learning Research,* 475 - 514.

Kalisch, M. & Buhlmann, P. (2007). Estimating high dimensional directed acyclic graphs with the PC algorithm. *Journal of Machine Learning Research*, *8*, 613-636.

Kiiveri, H. & Speed, T. (1982). Structural analysis of multivariate data: A review. In S. Leinhardt (Ed.), *Sociological Methodology 1982*. San Francisco: Jossey-Bass.

Klepper, S. (1988). Regressor Diagnostics for the Classical Errors-in-Variables Model. *J Econometrics*, *37*(2), 225-250.

Klepper, S., Kamlet, M., & Frank, R. (1993). Regressor Diagnostics for the Errors-in-Variables Model - An Application to the Health Effects of Pollution. *J Environ Econ Manag*, *24*(3), 190-211.

Lacerda, G., Spirtes, P., Ramsey, J., & Hoyer, P. O. (2008). Discovering Cyclic Causal Models by Independent Component Analysis. *Proceedings of the 24th Conference on Uncertainty In Artificial Intelligence*, 366-374.

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

Linthurst, R. A. (1979). Aeration, nitrogen, pH and salinity as factors affecting Spartina Alterniflora growth and dieback, Ph.D. dissertation, North Carolina State University.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 403-411.

Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 80-89.

Moneta, A. & Spirtes, P. (2006). Graphical Models for the Identification of Causal Structures in Multivariate Time Series Model. Paper presented at the *2006 Joint Conference on Information Sciences*.

Needleman, H. L. (1979). Deficits in psychologic and classroom performance of children with elevated dentine lead levels. N *Engl J Med*, *300(13)*, 689-695.

Needleman, H. L., Geiger, S. K., & Frank, R. (1985). Lead and IQ scores: a reanalysis. *Science*, *227(4688)*(4688), 701-2, 704.

Pearl, J. & Dechter, R. (1996). Identifying independencies in causal graphs with feedback. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, Portland, OR, 420-426.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., & Glymour, C. (2010). Six problems for causal inference from fMRI. *NeuroImage*, 49, 1545–1558.

Rawlings, J. (1988). *Applied Regression Analysis*. Belmont, CA: Wadsworth.

Richardson, T. S. (1996). A discovery algorithm for directed cyclic graphs. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, Portland, OR., 454-462.

Scheines, R. (2000). Estimating Latent Causal Influences: TETRAD III Variable Selection and Bayesian Parameter Estimation: the effect of Lead on IQ. In P. Hayes (Ed.), *Handbook of Data Mining*. Oxford University Press.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461-464.

Sewell, W. H. & Shah, V. P. (1968). Social Class, Parental Encouragement, and Educational Aspirations. *Am J Sociol*, *73*(5), 559-572.

Shimizu, S., Hoyer, P. O., & Hyvärinen, A. (2009). Estimation of linear non-Gaussian acyclic models for latent factors. *Neurocomputing*, *72*(7-9), 2024-2027.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, *7*, 2003-2030.

Shpitser, I. & Pearl, J. (2008). Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, *9*, 1941-1979.

Silva, R. & Scheines, R. (2004). Generalized Measurement Models. *reports-archive.adm.cs.cmu.edu*.

Silva, R., Scheines, R., Glymour, C., & Spirtes, P. (2006). Learning the structure of linear latent variable models. *J Mach Learn Res*, *7*, 191-246.

Spearman, C. (1904). General Intelligence objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 491-499.

Spirtes, P. & Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review*, *9*(1), 67-72.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Spring-Verlag Lectures in Statistics.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, Prediction, and Search, Second Edition (Adaptive Computation and Machine Learning)*. The MIT Press.

Spirtes, P., Scheines, R., Glymour, C., & Meek, C. (2004). Causal Inference. In D. Kaplan (Ed.), *SAGE Handbook of Quantitative Methodology*. (pp. 447-477). SAGE Publications.

Strotz, R. H. & Wold, H. O. A. (1960). Recursive VS Nonrecursive Systems- An Attempt At Synthesis. *Econometrica*, *28*(2), 417-427.

Sun, X. (2008). *Causal inference from statistical data*. MPI for Biological Cybernetics.

Swanson, N. R. & Granger, C. W. J. (1997). Impulse Response Function Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions. *Journal of the American Statistical Association*, *92*(437), 357-367.

Tillman, R. E., Danks, D., & Glymour, C. (2009). Integrating Locally Learned Causal Structures with Overlapping Variables. In *Advances in Neural Information Processing Systems 21*, 1665-1672.

Tillman, R. E., Gretton, A., & Spirtes, P. (2009). Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in Neural Information Processing Systems 22*.

Timberlake, M. & Williams, K. R. (1984). Dependence, Political Exclusion And Government Repression - Some Cross-National Evidence. *Am Sociol Rev*, *49*(1), 141-146.

Verma, T. S. & Pearl, J. (1990). Equivalence and Synthesis of Causal Models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 220-227.

Zhang, K. & Hyvarinen, A. (2009). On the Identifiability of the Post-Nonlinear Causal Model. *Proceedings of the 26th International Conference on Uncertainty in Artificial Intelligence*, 647-655.

29

The Structural Model and the Ranking Theoretic Approach to Causation: A Comparison

WOLFGANG SPOHN

1 Introduction

Large parts of Judea Pearl's very rich work lie outside philosophy; moreover, basically being a computer scientist, his natural interest was in computational efficiency, which, as such, is not a philosophical virtue. Still, the philosophical impact of Judea Pearl's work is tremendous and often immediate; for the philosopher of science and the formal epistemologist few writings are as relevant as his. Fully deservedly, this fact is reflected in some philosophical contributions to this Festschrift; I am glad I can contribute as well.

For decades, Judea Pearl and I were pondering some of the same topics. We both realized the importance of the Bayesian net structure and elaborated on it; his emphasis on the graphical part was crucial, though. We both saw the huge potential of this structure for causal theorizing, in particular for probabilistic causation. We both felt the need for underpinning the probabilistic account by a theory of deterministic causation; this is, after all, the primary notion. And we both came up with relevant proposals. Judea Pearl approached these topics from the Artificial Intelligence side, I from the philosophy side. Given our different proveniences, overlap and congruity are surprisingly large.

Nevertheless, it slowly dawned upon me that the glaring similarities are deceptive, and that we fill the same structure with quite different contents. It is odd how much divergence can hide underneath so much similarity. I have identified no less than fifteen different, though interrelated points of divergence, and, to be clear, I am referring here only to our accounts of deterministic causation, the structural model approach so richly developed by Judea Pearl and my (certainly idiosyncratic) ranking-theoretic approach. In this brief paper I just want to list the points of divergence in a more or less descriptive mood, without much argument. Still, the paper may serve as a succinct reference list of the many crucial points that are at issue when dealing with causation and may thus help future discussion.

At bottom, my comparison refers, on the one hand, to the momentous book of Pearl (2000), the origins of which reach back to the other momentous book of Pearl (1988) and many important papers in the 80's and 90's,s and, on the other hand, to the chapters 14 and 15 of Spohn (forthcoming) on causation, the origins of which reach back to Spohn (1978, sections 3.2 - 3, and 1983) and a bunch of subsequent papers. For ease of access,

though, I shall substantially refer to Halpern, Pearl (2005) and Spohn (2006) where the relevant accounts are presented in a more compact way. Let me start with reproducing the basic explications in section 2 and then proceed to my list of points of comparison in section 3. Section 4 concludes with a brief moral.

2 The Accounts to be Compared

For all those taking philosophical talk of events not too seriously (the vast majority among causal theorists) the starting point is a *frame*, a (non-empty, finite) set **U** of *variables*; X, Y, Z, W, etc. denote members of **U**, $\vec{X}$, $\vec{Y}$, $\vec{Z}$, $\vec{W}$, etc. subsets of **U**. Each variable $X \in U$ has a *range* Ω_X of values and is a function from some *possibility space* Ω into its range Ω_X. For simplicity, we may assume that Ω is the Cartesian product of all Ω_X and X the projection from Ω to Ω_X. For $x \in \Omega_X$ and $A \subseteq \Omega_X$, $\{X = x\} = \{\omega \in \Omega \mid X(\omega) = x\}$ and $\{X \in A\} = \{\omega \mid X(\omega) \in A\}$ are *propositions* (or events), and all those propositions generate a propositional algebra **A** over Ω. For $\vec{X} = \{X_1, \ldots, X_n\}$ and $\vec{x} = \langle x_1, \ldots, x_n\rangle$ $\{\vec{X} = \vec{x}\}$ is short for $\{X_1 = x_1$ and $\ldots$ and $X_n = x_n\}$. How a variable is to be precisely understood may be exemplified in the usual ways; however, we shall see that it is one of the issues still to be discussed.

The causal theorist may or may not presuppose a temporal order among variables; I shall. So, let $\prec$ be a linear order on the frame **U** representing temporal precedence. Linearity excludes simultaneous variables. The issue of simultaneous causation is pressing, but not one dividing us; therefore I put it to one side. Let, e.g., $\{\prec Y\}$ denote $\{Z \in \mathbf{U} \mid Z \prec Y\}$, that is, the set of variables preceding Y. So much for the algebraic groundwork.

A further ingredient is needed in order to explain causal relations. In the structural-model approach it is a set of structural equations, in the ranking-theoretic approach it is a ranking function.

A set **F** of *structural equations* is just a set of functions F_Y that specifies for each variable Y in some subset $\vec{V}$ of **U** how Y (essentially) functionally depends on some subset $\vec{X}$ of **U**; thus F_Y maps $\Omega_{\vec{X}}$ into Ω_Y. $\vec{V}$ is the set of *endogenous* variables, $\vec{U} = \mathbf{U} - \vec{V}$ the set of *exogenous* variables. The only condition on **F** is that no Y in $\vec{V}$ indirectly functionally depends on itself via the equations in **F**. Thus, **F** induces a DAG on **U** such that, if F_Y maps $\Omega_{\vec{X}}$ into Ω_Y, $\vec{X}$ is the set of parents of Y. (In their appendix A.4 Halpern, Pearl (2005) generalize their account by dropping the assumption of the acyclicity of the structural equations.) The idea is that **F** provides a set of laws that govern the variables in **U**, though, again, the precise interpretation of **F** will have to be discussed below. $\langle \mathbf{U}, \mathbf{F}\rangle$ is then called a *structural model* (*SM*). Note that a SM does not fix the values of any variables. However, once we fix the values $\vec{u}$ of all the exogenous variables in $\vec{U}$, the equations in **F** determine the values $\vec{v}$ of all the endogenous variables in $\vec{V}$. Let us call $\langle \mathbf{U}, \mathbf{F}, \vec{u}\rangle$ a *contextualized structural model* (*CSM*). Thus, each CSM determines a specific world or course of events $\omega = \langle \vec{u}, \vec{v}\rangle$ in Ω. Accordingly, each proposition A in **A** is *true* or *false* in a CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u}\rangle$, depending on whether or not $\omega \in A$ for the ω thereby determined.

For the structural model approach, causation is essentially related to intervention. Therefore we must first explain the latter notion. An *intervention* always intervenes on a CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$, more specifically, on a certain set $\vec{X} \subseteq \vec{V}$ of endogenous variables, thereby setting the values of $\vec{X}$ to some fixed values $\vec{x}$; that intervention or setting is denoted by $\vec{X} \leftarrow \vec{x}$. What such an intervention $\vec{X} \leftarrow \vec{x}$ does is to turn the CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$ into another CSM. The variables in $\vec{X}$ are turned into exogenous variables; i.e., the set $\mathbf{F}$ of structural equations is reduced to the set $\mathbf{F}^{\vec{X}}$, as I denote it, that consists of all the equations in $\mathbf{F}$ for the variables in $\vec{V} - \vec{X}$. Correspondingly, the context $\vec{u}$ of the original CSM is enriched by the chosen setting $\vec{x}$ for the new exogenous variables in $\vec{X}$. In short, the intervention $\vec{X} \leftarrow \vec{x}$ changes the CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$ into the CSM $\langle \mathbf{U}, \mathbf{F}^{\vec{X}}, \langle \vec{u}, \vec{x} \rangle \rangle$. Again, it will be an issue what this precisely means.

Now, we can proceed to Pearl's explication of actual causation; this is definition 3.1 of Halpern, Pearl (2005, p. 853) slightly adapted to the notation introduced so far (see also Halpern, Hitchcock (2010, Section 3)). Not every detail will be relevant to my further discussion below; I reproduce it here only for reasons of accuracy:

SM DEFINITION: $\{\vec{X} = \vec{x}\}$ is an *actual cause* of $\{Y = y\}$ in the CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$ iff the following three conditions hold:

(1) $\{\vec{X} = \vec{x}\}$ and $\{Y = y\}$ are true in $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$.

(2) There exists a partition $\langle \vec{Z}, \vec{W} \rangle$ of $\vec{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $\langle \vec{x}', \vec{w}' \rangle$ of the variables in $\vec{X}$ and $\vec{W}$ such that if $\{\vec{Z} = \vec{z}\}$ is true in $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$, then both of the following conditions hold:

   (a) $\{Y = y\}$ is false in the intervention $\langle \vec{X}, \vec{W} \rangle \leftarrow \langle \vec{x}', \vec{w}' \rangle$ on $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$, i.e., in $\langle \mathbf{U}, \mathbf{F}^{\vec{X},\vec{W}}, \langle \vec{u}, \vec{x}', \vec{w}' \rangle \rangle$. In other words, changing $\langle \vec{X}, \vec{W} \rangle$ from $\langle \vec{x}, \vec{w} \rangle$ to $\langle \vec{x}', \vec{w}' \rangle$ changes $\{Y = y\}$ from true to false.

   (b) $\{Y = y\}$ is true in $\langle \mathbf{U}, \mathbf{F}^{\vec{X},\vec{W}',\vec{Z}'}, \langle \vec{u}, \vec{x}, \vec{w}', \vec{z}' \rangle \rangle$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$, where $\vec{z}'$ is the subsequence of $\vec{z}$ pertaining to $\vec{Z}'$.

(3) $\vec{X}$ is minimal; i.e., no subset of $\vec{X}$ satisfies conditions (1) and (2).

This is not as complicated as it may look. Condition (1) says that the cause and the effect actually occur in the relevant CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$ and, indeed, had to occur given the structural equations in $\mathbf{F}$ and the context $\vec{u}$. Condition (2a) says that if the cause variables in $\vec{X}$ had been set differently, the effect $\{Y = y\}$ would not have occurred. It is indeed more liberal in allowing that also the variables in $\vec{W}$ outside $\vec{X}$ are set to different values, the reason being that the effect of $\vec{X}$ on Y may be hidden, as it were, by the actual values of $\vec{W}$, and uncovered only by setting $\vec{W}$ to different values. However, this alone would be too liberal; perhaps the failure of the effect $\{Y = y\}$ to occur is due only to the change of $\vec{W}$ rather than that of $\vec{X}$. Condition (2b) counteracts this permissiveness, and ensures that basically the change in $\vec{X}$ alone brings about the change of Y. Condition (3), finally, is to guarantee that the cause $\{\vec{X} = \vec{x}\}$ does not contain irrelevant parts; for the change described in (2a) all the variables in $\vec{X}$ are required. Note that $\vec{X}$ is a set of

variables so that $\{ \vec{X} = \vec{x} \}$ should be called a *total cause* of $\{Y = y\}$; its parts $\{X_i = x_i\}$ for $X_i \in \vec{X}$ may then be called *contributory causes*.

The details of the SM definition are mainly motivated by an adequate treatment of various troubling examples much discussed in the literature. It would take us too far to go into all of them. I should also mention that the SM definition is only preliminary in Halpern, Pearl (2005); but again, the details of their more refined definition presented on p. 870 will not be relevant for the present discussion.

The basics of the ranking-theoretic account may be explained in an equally brief way: A *negative ranking function* κ for Ω is just a function κ from Ω into $\mathbf{N} \cup \{\infty\}$ such that $\kappa(\omega) = 0$ for at least one $\omega \in \Omega$. It is extended to propositions in $\mathbf{A}$ by defining $\kappa(A) = \min\{\kappa(\omega) \mid \omega \in A\}$ and $\kappa(\varnothing) = \infty$; and it is extended to conditional ranks by defining $\kappa(B \mid A) = \kappa(A \cap B) - \kappa(A)$ for $\kappa(A) \neq \infty$. Negative ranks express degrees of disbelief: $\kappa(A) > 0$ says that A is disbelieved, so that $\kappa(\overline{A}) > 0$ expresses that A is believed in κ; however, we may well have $\kappa(A) = \kappa(\overline{A}) = 0$. It is useful to have both belief and disbelief represented in one function. Hence, we define the *two-sided rank* $\tau(A) = \kappa(\overline{A}) - \kappa(A)$, so that A is believed, disbelieved, or neither according to whether $\tau(A) > 0$, < 0, or $= 0$. Again, we have conditional two-sided ranks: $\tau(B \mid A) = \kappa(\overline{B} \mid A) - \kappa(B \mid A)$. The *positive relevance* of a proposition A to a proposition B is then defined by $\tau(B \mid A) > \tau(B \mid \overline{A})$, i.e., by the fact that B is more firmly believed or less firmly disbelieved given A than given $\overline{A}$; we might also say in this case that A *confirms* or *is a reason for* B. Similarly for negative relevance and irrelevance (= independence).

Like a set of structural equations, a ranking function κ induces a DAG on the frame $\mathbf{U}$ conforming with the given temporal order $\prec$. The procedure is the same as with probabilities: we simply define the set of parents of a variable Y as the unique minimal set $\vec{X} \subseteq \{\prec Y\}$ such that Y is independent of $\{\prec Y\} - \vec{X}$ given $\vec{X}$ relative to κ, i.e., such that Y is independent of all the other preceding variables given $\vec{X}$. If $\vec{X}$ is empty, Y is exogenous; if $\vec{X} \neq \varnothing$, Y is endogenous. The reading that Y directly causally depends on its parents will be justified later on.

Now, for me, being a cause is just being a special kind of conditional reason, i.e., being a reason given the past. In order to express this, for a subset $\vec{X}$ of $\mathbf{U}$ and a course of events $\omega \in \Omega$ let ${}^{\omega}[\vec{X}]$ denote the proposition that the variables in $\vec{X}$ behave as they do in ω. (So far, we could denote such a proposition by $\{\vec{X} = \vec{x}\}$, if $\vec{X}(\omega) = \vec{x}$, but we shall see in a moment that this notation is now impractical.) Then the basic definition of the ranking-theoretic account is this:

RT DEFINITION 1: For $A \subseteq W_X$ and $B \subseteq W_Y$ $\{X \in A\}$ is a *direct cause* of $\{Y \in B\}$ in $\omega \in \Omega$ relative to the ranking function κ (or the associated τ) iff

(a) $X \prec Y$,

(b) $X(\omega) \in A$ and $Y(\omega) \in B$, i.e., $\{X \in A\}$ and $\{Y \in B\}$ are facts in ω,

(c) $\tau(\{Y \in B\} \mid \{X \in A\} \cap {}^{\omega}[\{\prec Y\} - \{X\}]) > \tau(\{Y \in B\} \mid \{X \in \overline{A}\} \cap {}^{\omega}[\{\prec Y\} - \{X\}])$; i.e., $\{X \in A\}$ is a reason for $\{Y \in B\}$ given the rest of the past of Y as it is in ω.

It is obvious that the SM and the RT definition deal more or less with the same explicandum; both are after actual causes, where actuality is represented either by the context $\vec{u}$ of a CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$ in the SM definition or by the course of events ω in the RT definition. A noticeable difference is that in the RT definition the cause $\{X \in A\}$ refers only to a single variable X. Thus, the RT definition grasps what has been called a contributory cause, a total cause of $\{Y \in B\}$ then being something like the conjunction of its contributory causes. As mentioned, the SM definition proceeds the other way around.

Of course, the major differences lie in the explicantia; this will be discussed in the next section. A further noticeable difference in the definienda is that the RT definition 1 explains only direct causation; indeed, if $\{X \in A\}$ would be an indirect cause of $\{Y \in B\}$, we could not expect $\{X \in A\}$ to be positively relevant to $\{Y \in B\}$ conditional on the rest of the past of Y in ω, since that condition would not keep open the causal path from X to Y, but fix it to its actual state in ω. Hence, the RT definition 1 is restricted accordingly. As the required extension, I propose the following

RT DEFINITION 2: $\{X \in A\}$ is a (*direct or indirect*) *cause* of $\{Y \in B\}$ in $\omega \in \Omega$ relative to κ (or τ) iff there are $Z_i \in \mathbf{U}$ and $C_i \in \Omega_{Z_i}$ ($i = 1, \ldots, n \geq 2$) such that $X = Z_1$, $A = C_1$, $Y = Z_n$, $B = C_n$, and $\{Z_i \in C_i\}$ is a direct cause of $\{Z_{i+1} \in C_{i+1}\}$ in ω relative to κ for all $i = 1, \ldots, n - 1$.

In other words, causation in ω is just the transitive closure of direct causation in ω.

We may complete the ranking-theoretic account by explicating causal dependence between variables:

RT DEFINITION 3: $Y \in \mathbf{U}$ (*directly*) *causally depends* on $X \in \mathbf{U}$ relative κ iff there are $A \subseteq W_X$, $B \subseteq W_Y$, and $\omega \in \Omega$ such that $\{X \in A\}$ is a (direct) cause of $\{Y \in B\}$ in ω relative to κ.

One consequence of RT definition 3 is that the set of parents of Y in the DAG generated by κ and $\prec$ consists precisely of all the variables on which Y directly causally depends.

So much for the two accounts to be compared. There are all the differences that meet the eye. As we shall see, there are even more. Still, let me conclude this section by pointing out that there are also less differences than meet the eye. I have already mentioned that both accounts make use of the DAG structure of causal graphs. And when we supplement the probabilistic versions of the two accounts, they further converge. In the structural-model approach we would then replace the context $\vec{u}$ of a CSM $\langle \mathbf{U}, \mathbf{F}, \vec{u} \rangle$ by a probability distribution over the exogenous variables rendering them independent and extending via the structural equations to a distribution for the whole of **U**, thus forming a *pseudo-indeterministic system*, as Spirtes et al. (1993, pp. 38f.) call it, and hence a Bayesian net in which the probabilities agree with the causal graph. In the ranking-theoretic approach, we would replace the ranking function by a probability measure for **U** (or over **A**) that, together with the temporal order of the variables, would again induce a DAG or a

causal graph so as to form a Bayesian net. In this way, the basic ingredient of both accounts would become the same: a probability measure;the remaining differences appear to be of a merely technical nature.

Indeed, as I see the recent history of the theory of causation, this large agreement initially dominated the picture of probabilistic causation. However, the need for underpinning the probabilistic by a deterministic account was obvious; after all, the longer history of the notion was an almost entirely deterministic one up to the recent counterfactual accounts following Lewis (1973). And so the surprising ramification sketched above came about, both branches of which well agree with their probabilistic origins. The ramification is revealing since it makes explicit dividing lines that were hard to discern within the probabilistic harmony. Indeed, the points of divergence between the structural-model and the ranking-theoretic approach to be discussed in the next section apply to their probabilistic sisters as well, a claim that is quite suggestive, though I shall not elaborate on it.

3 Fifteen Points of Comparison

All in all, I shall come up with fifteen clearly distinguishable, though multiply connected points of comparison. The theorist of causation must take a stance towards all of them, and even more; my list is pertinent to the present comparison and certainly not exhaustive. Let us go through the list point for point:

(1) The most obvious instances provoking comparison and divergence are provided by *examples*, about preemption and prevention, overdetermination and switches, etc. The literature abounds in cases challenging all theories of causation and examples designed for discriminating among them, a huge bulk still awaiting systematic classification (though I attempted one in my (1983, ch. 3) as far as possible at that time). A theory of causation must do well with these examples in order to be acceptable. No theory, though, will reach a perfect score, all the more as many examples are contested by themselves, and do not provide a clear-cut criterion of adequacy. And what a 'good score' would be cannot be but vague. Therefore, I shall not even open this unending field of comparison regarding the two theories at hand.

(2) The main reason why examples provide only a soft criterion is that it is ultimately left to *intuition* to judge whether an example has been adequately treated. There are strong intuitions and weak ones. They often agree and often diverge. And they are often hard to compromise. Indeed, intuitions play an indispensable and important role in assessing theories of causation; they seem to provide the ultimate unquestionable grounds for that assessment.

Still, I have become cautious about the role of intuitions. Quite often I felt that the intuitions authors claim to have are guided by their theory; their intuitions seem to be what their theory suggests they should be. Indeed, the more I dig into theories of causation and develop my own, the harder it is for me to introspectively discern whether or not I share certain intuitions independently of any theorizing. So, again, the appeal to intui-

tions must be handled with care, and I shall not engage into a comparison of the relevant theories on an intuitive level.

(3) Another large field of comparison is the *proximity to* and the *applicability in scientific practice*. No doubt, the SM account fares much better in this respect than the RT approach. Structural modeling is something many scientists really do, whereas ranking theory is unknown in the sciences and it may be hard to say why it should be known outside epistemology. The point applies to other accounts as well. The regularity theory of causation seems close to the sciences, since they seem to state laws and regularities, whereas counterfactual analyses seem remote, since counterfactual claims are not an official part of scientific theories, even though, unofficially, counterfactual talk is ubiquitous. And probabilistic theories maintain their scientific appearance by ecumenically hiding disputes about the interpretation of probability.

Again, the importance of this criterion is undeniable; the causal theorist is well advised to appreciate the great expertise of the sciences, in general and specifically concerning causation. Still, I tend to downplay this criterion, not only in order to keep the RT account as a running candidate. The point is rather that the issue of causation is of a kind for which the sciences are not so well prepared. The counterfactual analysis is a case in point. If it should be basically correct, then the counterfactual idiom can no longer be treated as a second-rate vernacular (to use Quine's term), as the sciences do, but must be squarely faced in a systematic way, as, e.g., Pearl (2000, ch. 7) does, but qua philosopher, not qua scientist. Probabilities are a similar case. Mathematicians and statisticians by far know best how to deal with them. However, when it comes to say what probabilities mean, they are not in a privileged position.

The point of these three remarks is to claim primacy for theoretical issues about causation as such. External considerations are relevant and helpful, but they cannot release us from the task of taking some stance or other towards these theoretical issues. So, let us turn to them.

(4) Both, the SM and the RT account, are based on a *frame* providing a *framework of variables* and appertaining *facts*. I am not sure, however, whether we interpret it in the same way. A (random) variable is a function from some state space into some range of values, usually the reals; this is mathematical standard. That a variable takes a certain value is a proposition, and if the value is the true one (in some model), the proposition is a fact (in that model); so much is clear. However, the notion of a variable is ambiguous, and it is so since its statistic origins. A variable may vary over a given population as its state space and take on a certain value for each item in the population. E.g., size varies among Germans and takes (presently) the value 6' 0" for me. This is what I call a *generic variable*. Or a variable may vary over a set of possibilities as its state space and take values accordingly. For example, *my* (present) size is a variable in this sense and actually takes the value 6' 0", though it takes other values in other possibilities; I might (presently) have a different size. I call this a *singular variable* representing the possibility range of a

given single case. For each German (and time), size is such a singular variable. The generic variable of size, then, is formed by the actual values of all these singular variables.

The above RT account exclusively speaks about singular variables and their realizations; generic variables simply are out of the picture. By contrast, the ambiguity seems to afflict the SM account. I am sure everybody is fully clear about the ambiguity, but this clarity seems insufficiently reflected in the terminology. For instance, the equations of a SM represent laws or ceteris paribus laws or invariances in Woodward's (2003) terms or statistical laws, if supplemented by statistical 'error' terms, and thus state relations between generic variables. It is contextualization by which the model gets applied to a given single case; then, the variables should rather be taken as singular ones; their taking certain values then are specific facts. There is, however, no terminological distinction of the two interpretations; somehow, the notion of a variable seems to be intended to play both roles. In probabilistic extensions we find the same ambiguity, since probabilities may be interpreted as statistical distributions over populations or as realization propensities of the single case.

(5) I would not belabor the point if it did not extend to the causal relations we try to capture. We have causation among facts, as analyzed in the SM definition and the RT definitions 1 - 2; they are bound to apply to the single case. And we have causal relations among variables, i.e., causal dependence (though often and in my view confusingly the term "cause" is used here as well), and we find here the same ambiguity. Causal dependence between generic variables is a matter of causal laws or of *general causation*. However, there is also causal dependence between singular variables, something rarely made explicit, and it is a matter of *singular causation* applying to the single case just as much as causation between facts. Since its inception the discussion of probabilistic causality was caught in this ambiguity between singular and general causation; and I am wondering whether we can still observe the aftermath of that situation.

In any case, structural equations are intended to capture causal order, and the order among generic variables thus given pertains to general causation. Derivatively these equations may be interpreted as stating causal dependencies also between singular variables. In the SM account, though, singular causation is explicitly treated only as pertaining to facts. By contrast, the RT definition 3 explicates only causal dependence between singular variables. The RT account is so far silent about general causation and can grasp it only by generalizing over the causal relations in the single case. These remarks are not just pedantry; I think it is important to observe these differences for an adequate comparison of the accounts.

(6) I see these differences related to the issue of the role of *time* in an analysis of causation. The point is simply that generic variables as such are not temporally ordered, since their arguments, the items to which they apply, may have varying temporal positions; usually, statistical data do not come temporally ordered. By contrast, singular variables are temporally ordered, since their variable realizability across possibilities is tied to a fixed time. As a consequence, the SM definition makes no explicit reference to time,

whereas the RT definitions make free use of that reference. While I think that this point has indeed disposed Judea Pearl and me to our diverging perspectives on the relation between time and causation, it must be granted that the issue takes on much larger dimensions that open enough room for indecisive defenses of both perspectives.

Many points are involved: (i) Issues of analytic adequacy: while Pearl (2000, pp. 249ff.) argues that reference to time does not sufficiently further the analytic project and proposes ingenious alternatives (sections 2.3 - 4 + 8 - 9), I am much more optimistic about the analytic prospects of referring to time (see my 1990, section 3, and forthcoming, section 14.4). (ii) Issues of analytic policy (see also point 10 below): Is it legitimate to refer to time in an analysis of causation? I was never convinced by the objections. Or should the two notions be analytically decoupled? Or should the analytic order be even reversed by constructing a causal theory of time? Pearl (2000, section 2.8) shows sympathies for the latter project, although he suggests an evolutionary explanation, rather than Reichenbach's (1956) physical explanation for relating temporal direction with causal directionality. (iii) The issue of causal asymmetry: Is the explanation of causal asymmetry by temporal asymmetry illegitimate? Or incomplete? Or too uninformative, as far as it goes? If any of these, what is the alternative?

(7) Causation always is causation within given *circumstances*. What do the accounts say what the circumstances are? The RT definition 1 explicitly takes the entire past of the effect except the cause as the circumstances of a direct causal relationship, something apparently much too large and hence inadequate, but free of conceptual circularity, as I have continuously emphasized. In contrast, Pearl (2000, pp. 250ff.) endorses the circular explanation of Cartwright (1979) that those circumstances consist of the other causes of the effect and hence, in the case of direct causation, of the realizations of the other parents of the effect variable in the causal graph. Pearl thus accepts also Cartwright's conclusion that the reference to the obtaining circumstances does not help explicating causation; he thinks that this reference at best provides a kind of consistency test. I argue that the explicatory project is not doomed thereby, since Cartwright's circular explanation may be derived from my apparently inadequate definition (cf. Spohn 1990, section 4). As for the circumstances of indirect causation, the RT definition 2 is entirely silent, since it relies on transitivity; however, in Spohn (1990, Theorems 14 and 16) I explored how much I can say about them. In contrast, the SM definition contains an implicit account of the circumstances that applies to indirect causal relationships as well; it is hidden in the partition $\langle \vec{Z}, \vec{W} \rangle$ of the set $\vec{V}$ of endogenous variables. However, it still accepts Cartwright's circular explanation, since it presupposes the causal graph generated by the structural equations. So, this is a further respect in which our accounts are diametrically opposed.

(8) The preceding point contains two further issues. One concerns the distinction of *direct and indirect causation*. The SM approach explicates causation without attending to this distinction. Of course, it could account for it, but it does not acquire a basic importance. By contrast, the distinction receives analytic significance within the RT approach

that first defines direct causation and then, only on that basis, indirect causation. The reason is that, in this way, the RT approach hopes to reach a non-circular explication of causation, whereas the SM approach has given up on this hope (see also point 10 below) and thus sees no analytic rewards in this distinction.

(9) The other issue already alluded to in (7) is the issue of transitivity. This is a most vexed topic, and the community seems unable to find a stable attitude. Transitivity had to be given up, it seemed, within probabilistic causation (cf. Suppes 1970, p. 58), while it was derivable from a regularity account and was still defended by Lewis (1973) for deterministic causation. In the meantime the situation has reversed; transitivity has become more respectable within the probabilistic camp; e.g., Spirtes et al. (1993, p. 44) simply assume it in their definition of "indirect cause". By contrast, more and more tend to reject it for deterministic causation (cf., e.g., McDermott 1995 and Hitchcock 2001).

This uncertainty is also reflected in the present comparison. Pearl (2000, p. 237) rejects transitivity of causal dependence among variables, but, as the argument shows, only in the sense of what Woodward (2003, p. 51) calls "total cause". Still, Woodward (2003, p. 59), in his concluding explication **M**, accepts the transitivity of causal dependence among variables in the sense of "contributory cause", and I have not found any indication in Pearl (2000) or Halpern, Pearl (2005) that they would reject Woodward's account of contributory causation. However, all of them deny the transitivity of actual causation between facts.

I see it just the other way around. The RT definition 2 stipulates the transitivity of causation (with arguments, though; cf. Spohn 1990, p. 138, and forthcoming, section 14.12), whereas the RT definition 3 entails the transitivity of causal dependence among variables in the contributory sense only under (mild) additional assumptions. Another diametrical opposition.

(10) A much grander issue is looming behind the previous points, the issue of *analytic policy*. The RT approach starts defining direct causation between singular facts, proceeds to indirect causation and then to causal dependence between singular variables, and finally only hopes to thereby grasp general causation as well. It thus claims to give a non-circular explication or a reductive analysis of causation. The SM approach proceeds in the opposite direction. It presupposes an account of general causation that is contained in the structural equations, transfers this to causal dependence between singular variables (I mentioned in points 4 and 5 that this step is not fully explicit), and finally arrives at actual causation between facts. The claim is thereby to give an illuminating analysis of causation, but not a reductive one.

Now, one may have an argument about conceptual order: which causal notions to explicate on the basis of which? I admit I am bewildered by the SM order. The deeper issue, though, or perhaps the deepest, is the feasibility of reductive analysis. Nobody doubts that it would be most welcome to have one; therefore the history of the topic is full of attempts at such an analysis. Perhaps, though, they are motivated by wishful thinking. How to decide? One way of assessing the issue is by inspecting the proposals. The proponents

are certainly confident of their analyses, but their inspection revealed so many problems that doubts preponderate. However, this does not prove their failure. Also, one may advance principled arguments such as Cartwright's (1979) that one cannot avoid being entangled in conceptual circles. For such reasons, the majority, it seems, has acquiesced in non-reductive analysis; cf., e.g., Woodward (2003, pp. 104ff.) for an apology of non-reductivity or Glymour (2004) for a eulogy of the, as he calls it, Euclidean as opposed to the Socratic ideal.

Another way of assessing the issue is more philosophical. Are there any more basic features of reality to which causation may reduce? One may well say no, and thereby justify the rejection of reductive analysis. Or one may say yes. Laws may be such a more basic feature; this, however, threatens to result either in an inadequate regularity theory of causation or in an inability to say what laws are beyond regularities. Objective probabilities may be such a feature – if we only knew what they are. What else is there on offer? On the other hand, it is not so easy to simply accept causation as a basic phenomenon; after all, the point has deeply worried philosophers for centuries after Hume.

In any case, all these issue are involved in settling for a certain analytic policy. It will become clearer in the subsequent points why I nevertheless maintain the possibility of reductive analysis.

(11) The most conspicuous difference of the SM and the RT approach is a direct consequence of their different policies. The SM account bases its analysis on *structural models* or *equations*, whereas the RT account explicates causation in terms of *ranking functions*. These are entirely different things!

Prima facie, structural equations are easier to grasp. Despite its non-reductive procedure the SM approach incurs the obligation, though, to somehow explain how the structural equations can establish causal order among generic variables. They can do this, because Pearl (2000, pp. 157ff.) explicitly gives them an interventionistic interpretation that, in turn, is basically a counterfactual one, as is entirely clear to Pearl; most interventions are only counterfactual. Woodward (2003) repeatedly emphasizes the point that the interventionistic account clarifies the counterfactual approach by forcing a specific interpretation of the multiply ambiguous counterfactual idiom. Still, despite Woodward's (2003, pp. 121f.) claim to use counterfactuals only when they are clearly true of false, and despite Pearl's (2000, section 7.1) attempt to account for counterfactuals within structural models, the issue how counterfactuals acquire truth conditions remains a mystery in my view.

By contrast, it is quite bewildering to base an analysis of causation on ranking functions that are avowedly to be understood only as doxastic states, i.e., in a purely epistemological way. One of my reasons for doing so is that the closer inspection envisaged in (10) comes out, on the whole, more satisfactorily than for other accounts, that is, the overall score in dealing with examples is better. The other reason why I find ranking functions not so implausible a starting point lies in my profoundly Humean strategy in dealing with causation. There is no more basic feature of reality to which causation might reduce. The issue rather is how modal facts come into the world – where modal facts

pertain to lawhood, causation, counterfactuals, probabilities, etc. We do not find 'musts' and 'cans' in the world as we find apples and pears; this was Hume's crucial challenge. And his answer was what is now called Hume's projectivism (cf. Blackburn 1993, in particular the essays in part I). Ranking functions are well suited for laying out this projectivist answer in detail. This fundamental difference between the SM and the RT approach further unfolds in the final four points.

(12) A basic idea in our notion of causation between facts is, very roughly, that the cause does something for its effect, contributes to it, makes it possible or necessary or more likely, in short: that the cause is somehow positively relevant to its effect. One fact could also be negatively relevant to another, in which case the second obtains despite the first. As for causal dependence between variables, it is only required that the one is relevant for the other. What are the notions of *relevance* and *positive relevance* provided by the SM and the RT approach?

Ranking theory has a rich notion of positive and negative relevance, analogous and equivalent in formal behavior to the probabilistic notions. Its relevance notion is much richer and, I find, more adequate to the needs of causal theorizing than those provided by the key terms of other approaches to deterministic causation: laws, counterfactuals, interventions, structural equations, or whatever. This fact grounds my optimism that the RT approach is, on the whole, better able to cope with all the examples and problem cases.

I just said that the relevance notion provided by the SM approach is poorer. What is it? Clause (2b) of the SM definition says, in a way, that the effect $\{Y = y\}$ had to occur given the cause $\{\vec{X} = \vec{x}\}$ occurs, and clause (2a) says that the effect might not have occurred if the cause does not occur and, indeed, would not have occurred if the cause variable(s) $\vec{X}$ would have been realized in a suitable alternative way. In traditional terms, we could say that the cause is a necessary and sufficient condition of the effect provided the circumstances – where the subtleties of the SM approach lie in the proviso; that's the SM positive relevance notion. So, roughly, in SM terms, the only 'action' a cause can do is making its effect necessary, whereas ranking theory allows many more 'actions'. This is what I mean by the SM approach being poorer. For instance, it is not clear how a fact could be negatively relevant to another fact in the SM approach, or how one fact could be positively and another negatively relevant to a third one. And so forth.

(13) Let's take a closer look at what "action" could mean in the previous paragraph. In the RT approach it means comparing ranks conditional on the cause $\{X \in A\}$ and on its negation $\{X \in \overline{A}\}$; the rank raising showing up in that comparison is what the cause 'does'. In the SM approach we do not conditionalize on the cause $\{\vec{X} = \vec{x}\}$ and some alternative $\{\vec{X} = \vec{x}'\}$; rather, in clauses (2a-b) of the SM definition we look at the consequences of the interventions $\vec{X} \leftarrow \vec{x}$ and $\vec{X} \leftarrow \vec{x}'$, i.e., by replacing the structural equation(s) for $\vec{X}$ by the stipulation $\vec{X} = \vec{x}$ or, respectively, $= \vec{x}'$. The received view by now is that *intervention* is quite different from *conditionalization* (cf., e.g., Goldszmidt, Pearl 1992, and Meek, Glymour 1994), the suggestion being that interven-

tion is what causal theorizing requires, and that all approaches relying on conditionalization such as the RT approach therefore are misguided (cf. also Pearl 2000, section 3.2).

The difference looks compelling: intervention is a real activity, whereas conditionalization is only a mental, suppositional activity. But once we grant that intervention is mostly counterfactual (i.e., also suppositional), the difference shrinks. Indeed, I tend to say that there never is a real intervention in a given single case; after a real intervention we deal with a different single case than before. Hence, I think the difference the received view assumes is spurious; rather, interventions may be construed in terms of conditionalization:

Of course, the intervention $\vec{X} \leftarrow \vec{x}$ differs from conditioning on $\{ \vec{X} = \vec{x} \}$; in this, the received view is correct. However, the RT and other conditioning approaches do not simply conditionalize on the cause, but on much more. What the intervention $X_1 \leftarrow x_1$ on the single variable X_1 does is change the value of X_1 to x_1 while at the same time keeping fixed the values of all temporally preceding variables as they are in the given context, or, if only a causal graph and not temporal order is available, either of all ancestors of X_1 or of all non-descendants of X_1 (which comes to the same thing in structural models, and also in probabilistic terms given the common cause principle). Thus, the intervention is equivalent to conditioning on $\{X_1 = x_1\}$ *and* on the fixed values of those other variables.

Similarly for a double intervention $\langle X_1, X_2 \rangle \leftarrow \langle x_1, x_2 \rangle$. For assessing the behavior of the variables temporally between X_1 and X_2 (or being descendants of X_1, but not of X_2) under the double intervention, we have to look at the same conditionalization as in the single intervention $X_1 \leftarrow x_1$, whereas for the variables later than X_2 (or descending from both X_1 and X_2) we have to condition on $\{X_1 = x_1\}$, $\{X_2 = x_2\}$, the past of X_1 as it is in the given context, and on those intermediate variables taking the values as they are after the intervention $X_1 \leftarrow x_1$. And so forth for multiple interventions (that are so crucial for the SM approach).

Given this translation, this kind of difference between the SM and the RT approach vanishes, I think. Consider, e.g., the definition of direct causal dependence of Woodward (2003, p. 55): *Y* directly causally depends on *X* iff an intervention on *X* can make a difference to *Y*, provided the values of all other variables in the given frame **U** are somehow fixed by intervention. Translate this as proposed, and you arrive at the conditionalization I use in the above RT definitions to characterize direct causation.

(14) The preceding argument has a gap that emerges when we attend to another topic that I find crucial, but nowhere thoroughly discussed: the *frame-relativity* of causation. Everybody agrees that the distinction between direct and indirect causation is frame-relative; of course, a direct causal relationship relative to a coarse-grained frame may turn indirect under refinements. What about causation itself, though? One may try some moderate antirealism, e.g., general thoughts to the effect that science only produces models of reality and never truly represents reality as it really is; then causation would be model-relative, too.

However, this is not what I have in mind. The point is quite specific: The RT definition 1 refers, in a way I had explained in point 7, to the obtaining circumstances, however

only insofar as they are represented in the given frame **U**. This entails a genuine frame-relativity of causation as such; $\{X = x\}$ may be a (direct) cause of $\{Y = y\}$ within one frame, but not within another or more refined frame. As Halpern, Hitchock (2010, Section 4.1) argue, this phenomenon may also show up within the SM approach.

I do not think that this agrees with Pearl's intention in pursuing the SM account; an actual cause should not cease to be an actual cause simply by refining the frame. Perhaps, the intention was to arrive at a frame-independent notion of causation by assuming a frame-independent notion of intervention. My translation of the intervention $X_1 \leftarrow x_1$ into conditionalization referred to the past (or the ancestors or the non-descendants) of X_1 as far as they are represented in the given frame **U**, and thus reproduced only a frame-relative notion of intervention. However, the intention presumably is to refer to the entire past of X_1 absolutely, not leaving any hole for the supposition of $\{X_1 = x_1\}$ to backtrack. If so, there is another sharp difference between the SM and the RT approach with repercussions on the previous point.

Of course, I admit that our intuitive notion of causation is not frame-relative; we aim at an absolute notion. However, this aim bars us from having a reductive analysis of causation, since the analysis would have to refer then to the rest of the world, as it were, to many things outside the frame that are thus prevented from entering the analysis. In fact, any rigorous causal theorizing is thereby frustrated in my view. For, how can you theoretically deal with all those don't-know-what's? For this reason I always preferred to work with a fixed frame, to pretend that this frame is all there is, and then to say everything about causation that can be said within this frame. This procedure at least allows a reductive analysis of a frame-relative notion.

How, then, can we get rid of the frame-relativity? I propose, by ever more fine-graining and extending the frame, studying the frame-relative causal relations within all these well-defined frames, and finding out what remains stable across all these refinements; we may hope, then, that these stable features are preserved even in the maximally refined, universal frame (cf. Spohn forthcoming, section 14.9; for Halpern, Hitchcock (2010, Section 4.1) this stability is also crucial). I would not know how else to deal with the challenge posed by frame-relativity, and I suspect that considerable problems in causal theorizing result from not explicitly facing this challenge.

(15) The various points may be summarized in the final opposition: whether causation is to be *subjectivistically* or *objectivistically* conceived. Common sense, Judea Pearl, and many others are on the objectivistic side: "I now take causal relationships to be the fundamental building blocks both of physical reality and of human understanding of that reality" (Pearl 2000, pp. xiiif.). And insofar as structural equations are objective, the SM approach shares this objectivism. By contrast, frame-relativity is an element of subject-relativity; frames are chosen by us. And the use of only epistemically interpretable ranking functions involves a much deeper subjectivization of the topic of causation. (The issue of relevance, point 12, is related, by the way, since in my view only epistemic relevance is rich enough a concept.)

The motive of the subjectivistic RT approach was, I said, Hume's challenge. And the gain, I claimed, is the feasibility of a reductive analysis. Any objectivistic approach has to tell how else to cope with that challenge and how to make peace with non-reductivity. Still, we cannot simply acquiesce in subjectivism, since it flies in the face of everyone keeping some sense of reality. The general philosophical strategy to escape pure subjectivism has been aptly described by Blackburn (1993, part I) as Humean projectivism leading to so-called quasi-realism that is indistinguishable from 'real' realism.

This general strategy may be precisely explicated in the case of causation: I had indicated in the previous point how I propose to get rid of frame-relativity. And in Spohn (forthcoming, ch. 15) I develop an objectification theory for ranking functions, according to which some ranking functions, the objectifiable ones, may be said, to truly (or falsely) represent causal relations. No doubt, this objectification theory is disputable, but it shows that the subjectivistic starting point need not preclude us from objectivistic aims. Maybe, though, these aims are more convincingly served by approaching them in a more direct and realistic way, as the SM account does.

4 Conclusion

On none of the fifteen differences above could I seriously start discussion; obviously nothing below book length would do. Indeed, discussing these points was not my aim at all, let alone treating anyone conclusively (though, of course, I could not hide where my sympathies are). My first intention was simply to display the differences, not all of which are clearly seen in the literature; already the sheer number is surprising. And I expressed my second intention between point 3 and point 4: namely to show that there are many internal theoretical issues in the theory of causation. On all of them one must take and argue a stance, a most demanding requirement. My hunch is that those theoretical considerations will eventually override issues of exemplification and application. All the more important it is to take some stance; no less will do for reaching a considered judgment. Judea Pearl has paradigmatically shown how to do this. His brilliant theoretical developments have not closed, but tremendously advanced our understanding of all these issues pertaining to causation.

Acknowledgment: I am indebted to Joe Halpern for providing most useful comments and correcting my English.

References

Blackburn, S. (1993). *Essays in Quasi-Realism*, Oxford: Oxford University Press.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs 13*, 419-437.

Glymour, C. (2004). Critical notice on: James Woodward, Making Things Happen, *British Journal for the Philosophy of Science 55*, 779-790.

Goldszmidt, M., and J. Pearl (1992). Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In B. Nebel,

C. Rich, and W. Swartout (Eds.), *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, San Mateo, CA: Morgan Kaufmann, pp. 661-672.

Halpern, J. Y., and C. Hitchcock (2010). Actual causation and the art of modeling. This volume, chapter 22.

Halpern, J. Y., and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science 56*, 843-887.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy 98*, 273-299.

Lewis, D. (1973). Causation. *Journal of Philosophy 70*, 556-567.

McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science 46*, 523-544.

Meek, C., and C. Glymour (1994). Conditioning and intervening. *British Journal for the Philosophy of Science 45*, 1001-1021.

Reichenbach, H. (1956). *The Direction of Time*. Los Angeles: The University of California Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*. Berlin: Springer, 2nd ed. 2000.

Spohn, W. (1978). *Grundlagen der Entscheidungstheorie*, Kronberg/Ts.: Scriptor. Out of print, pdf-version at: http://www.uni-konstanz.de/FuF/Philo/Philosophie/philosophie/files/ge.buch.gesamt.pdf.

Spohn, W. (1983). *Eine Theorie der Kausalität*. Unpublished Habilitationsschrift, University of München, pdf-version at: http://www.uni-konstanz.de/FuF/Philo/Philosophie/philosophie/files/habilitation.pdf.

Spohn, W. (1990). Direct and indirect causes. *Topoi 9*, 125-145.

Spohn, W. (2006). Causation: An alternative. *British Journal for the Philosophy of Science 57*, 93-119.

Spohn, W. (forthcoming). *Ranking Theory. A Tool for Epistemology*.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*. Oxford: Oxford University Press.

30

On Identifying Causal Effects

Jin Tian and Ilya Shpitser

1 Introduction

This paper deals with the problem of inferring cause-effect relationships from a combination of data and theoretical assumptions. This problem arises in diverse fields such as artificial intelligence, statistics, cognitive science, economics, and the health and social sciences. For example, investigators in the health sciences are often interested in the effects of treatments on diseases; policymakers are concerned with the effects of policy decisions; AI research is concerned with effects of actions in order to design intelligent agents that can make effective plans under uncertainty; and so on.

To estimate causal effects, scientists normally perform randomized experiments where a sample of units drawn from the population of interest is subjected to the specified manipulation directly. In many cases, however, such a direct approach is not possible due to expense or ethical considerations. Instead, investigators have to rely on observational studies to infer effects. A fundamental question in causal analysis is to determine when effects can be inferred from statistical information, encoded as a joint probability distribution, obtained under normal, intervention-free behavior. A key point here is that it is not possible to make causal conclusions from purely probabilistic premises – it is necessary to make causal assumptions. This is because without any assumptions it is possible to construct multiple "causal stories" which can disagree wildly on what effect a given intervention can have, but agree precisely on all observables. For instance, smoking may be highly correlated with lung cancer either because it causes lung cancer, or because people who are genetically predisposed to smoke may also have a gene responsible for a higher cancer incidence rate. In the latter case there will be no effect of smoking on cancer.

In this paper, we assume that the causal assumptions will be represented by directed acyclic causal graphs [Pearl, 2000; Spirtes *et al.*, 2001] in which arrows represent the potential existence of direct causal relationships between the corresponding variables and some variables are presumed to be unobserved. Our task will be to decide whether the qualitative causal assumptions represented in any given graph are sufficient for assessing the strength of causal effects from nonexperimental data.

This problem of identifying causal effects has received considerable attention in the statistics, epidemiology, and causal inference communities [Robins, 1986;

Robins, 1987; Pearl, 1993; Robins, 1997; Kuroki and Miyakawa, 1999; Glymour and Cooper, 1999; Pearl, 2000; Spirtes *et al.*, 2001]. In particular Judea Pearl and his colleagues have made major contributions in solving the problem. In his seminal paper Pearl (1995) established a *calculus of interventions* known as *do-calculus* – three inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences, thus providing a syntactic method of deriving claims about interventions. Later, *do*-calculus was shown to be complete for identifying causal effects, that is, every causal effects that can be identified can be derived using the three *do*-calculus rules [Shpitser and Pearl, 2006a; Huang and Valtorta, 2006b]. Pearl (1995) also established the popular "back-door" and "front-door" criteria – sufficient graphical conditions for ensuring identification of causal effects. Using *do*-calculus as a guide, Pearl and his collaborators developed a number of sufficient graphical criteria: a criterion for identifying causal effects between singletons that combines and expands the front-door and back-door criteria [Galles and Pearl, 1995], a condition for evaluating the effects of plans in the presence of unmeasured variables, each plan consisting of several concurrent or sequential actions [Pearl and Robins, 1995]. More recently, an approach based on c-component factorization has been developed in [Tian and Pearl, 2002a; Tian and Pearl, 2003] and complete algorithms for identifying causal effects have been established [Tian and Pearl, 2003; Shpitser and Pearl, 2006b; Huang and Valtorta, 2006a]. Finally, a general algorithm for identifying arbitrary counterfactuals has been developed in [Shpitser and Pearl, 2007], while the special case of effects of treatment on the treated has been considered in [Shpitser and Pearl, 2009].

In this paper, we summarize the state of the art in identification of causal effects. The rest of the paper is organized as follows. Section 2 introduces causal models and gives formal definition for the identifiability problem. Section 3 presents Pearl's *do*-calculus and a number of easy to use graphical criteria. Section 4 presents the results on identifying (unconditional) causal effects. Section 5 shows how to identify conditional causal effects. Section 6 considers identification of counterfactual quantities which arise when we consider effects of relative interventions. Section 7 concludes the paper.

2 Notation, Definitions, and Problem Formulation

In this section we review the graphical causal models framework and introduce the problem of identifying causal effects.

2.1 Causal Bayesian Networks and Interventions

The use of graphical models for encoding distributional and causal assumptions is now fairly standard [Heckerman and Shachter, 1995; Lauritzen, 2000; Pearl, 2000; Spirtes *et al.*, 2001]. A *causal Bayesian network* consists of a DAG G over a set $V = \{V_1, \ldots, V_n\}$ of variables, called a *causal diagram*. The interpretation of such a graph has two components, probabilistic and causal. The probabilistic interpreta-

tion views G as representing conditional independence assertions: Each variable is independent of all its non-descendants given its direct parents in the graph.[1] These assertions imply that the joint probability function $P(v) = P(v_1, \ldots, v_n)$ factorizes according to the product [Pearl, 1988]

$$P(v) = \prod_i P(v_i|pa_i), \tag{1}$$

where pa_i are (values of) the parents of variable V_i in the graph. Here we use uppercase letters to represent variables or sets of variables, and use corresponding lowercase letters to represent their values (instantiations).

The set of conditional independences implied by the causal Bayesian network can be obtained from the causal diagram G according to the d-separation criterion [Pearl, 1988].

DEFINITION 1 (d-separation). A path [2] p is said to be *blocked* by a set of nodes Z if and only if

1. p contains a chain $V_i \rightarrow V_j \rightarrow V_k$ or a fork $V_i \leftarrow V_j \rightarrow V_k$ such that the node V_j is in Z, or
2. p contains an inverted fork $V_i \rightarrow V_j \leftarrow V_k$ such that V_j is not in Z and no descendant of V_j is in Z.

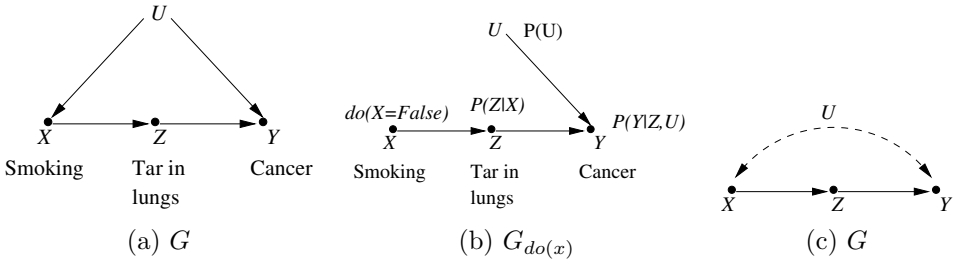
A path not blocked by Z is called *d-connecting* or *active*. A set Z is said to *d-separate* X from Y, denoted by $(X \perp\!\!\!\perp Y|Z)_G$, if and only if Z blocks every path from a node in X to a node in Y.

We have that if Z d-separates X from Y in the causal diagram G, then X is conditionally independent of Y given Z in the distribution $P(v)$ given in Eq. (1).

The causal interpretation views the arrows in G as representing causal influences between the corresponding variables. In this interpretation, the factorization of (1) still holds, but the factors are further assumed to represent autonomous data-generation processes, that is, each parents-child relationship characterized by a conditional probability $P(v_i|pa_i)$ represents a stochastic process by which the values of V_i are assigned in response to the values pa_i (previously chosen for V_i's parents), and the stochastic variation of this assignment is assumed independent of the variations in all other assignments in the model. Moreover, each assignment process remains invariant to possible changes in the assignment processes that govern other variables in the system. This modularity assumption enables us to infer the effects of interventions, such as policy decisions and actions, whenever interventions are described as specific modifications of some factors in the product of (1). The simplest such intervention, called *atomic*, involves fixing a set T of variables to some

[1] We use family relationships such as "parents," "children," and "ancestors" to describe the obvious graphical relationships.

[2] A path is a sequence of consecutive edges (of any directionality).



(a) G (b) $G_{do(x)}$ (c) G

Figure 1. A causal diagram illustrating the effect of smoking on lung cancer

constants $T = t$ denoted by $do(T = t)$ or $do(t)$, which yields the post-intervention distribution[3]

$$P_t(v) = \begin{cases} \prod_{\{i|V_i \notin T\}} P(v_i|pa_i) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \quad (2)$$

Eq. (2) represents a truncated factorization of (1), with factors corresponding to the manipulated variables removed. This truncation follows immediately from (1) since, assuming modularity, the post-intervention probabilities $P(v_i|pa_i)$ corresponding to variables in T are either 1 or 0, while those corresponding to unmanipulated variables remain unaltered. If T stands for a set of treatment variables and Y for an outcome variable in $V \setminus T$, then Eq. (2) permits us to calculate the probability $P_t(y)$ that event $Y = y$ would occur if treatment condition $T = t$ were enforced uniformly over the population. This quantity, often called the "causal effect" of T on Y, is what we normally assess in a controlled experiment with T randomized, in which the distribution of Y is estimated for each level t of T.

As an example, consider the model shown in Figure 1(a) from [Pearl, 2000] that concerns the relation between smoking (X) and lung cancer (Y), mediated by the amount of tar (Z) deposited in a person's lungs. The model makes qualitative causal assumptions that the amount of tar deposited in the lungs depends on the level of smoking (and external factors) and that the production of lung cancer depends on the amount of tar in the lungs but smoking has no effect on lung cancer except as mediated through tar deposits. There might be (unobserved) factors (say some unknown carcinogenic genotype) that affect both smoking and lung cancer, but the genotype nevertheless has no effect on the amount of tar in the lungs except indirectly (through smoking). Quantitatively, the model induces the joint distribution factorized as

$$P(u, x, z, y) = P(u)P(x|u)P(z|x)P(y|z, u). \quad (3)$$

[3] [Pearl, 1995; Pearl, 2000] used the notation $P(v|set(t))$, $P(v|do(t))$, or $P(v|\hat{t})$ for the post-intervention distribution, while [Lauritzen, 2000] used $P(v||t)$.

Assume that we could perform an ideal intervention on variable X by banning smoking[4], then the effect of this action is given by

$$P_{X=False}(u, z, y) = P(u)P(z|X = False)P(y|z, u), \tag{4}$$

which is represented by the model in Figure 1(b).

2.2 The Identifiability Problem

We see that, whenever all variables in V are observed, given the causal diagram G, all causal effects can be computed from the observed distribution $P(v)$ as given by Eq. (2). However, if some variables are not measured, or two or more variables in V are affected by unobserved confounders, then the question of *identifiability* arises. The presence of such confounders would not permit the decomposition of the observed distribution $P(v)$ in (1). For example, in the model shown in Figure 1(a), assume that the variable U (unknown genotype) is unobserved and we have collected a large amount of data summarized in the form of (an estimated) join distribution P over the observed variables (X, Y, Z). We wish to assess the causal effect $P_x(y)$ of smoking on lung cancer.

Let V and U stand for the sets of observed and unobserved variables, respectively. If each U variable is a root node with exactly two observed children, then the corresponding model is called a *semi-Markovian* model. In this paper, we will present results on semi-Markovian models as they allow for simpler treatment. However the results are general as it has been shown that causal effects in a model with arbitrary sets of unobserved variables can be identified by first projecting the model into a semi-Markovian model [Tian and Pearl, 2002b; Huang and Valtorta, 2006a].

In a semi-Markovian model, the observed probability distribution $P(v)$ becomes a mixture of products:

$$P(v) = \sum_u \prod_i P(v_i|pa_i, u^i)P(u) \tag{5}$$

where Pa_i and U^i stand for the sets of the observed and unobserved parents of V_i respectively, and the summation ranges over all the U variables. The post-intervention distribution, likewise, will be given as a mixture of truncated products

$$P_t(v) = \begin{cases} \sum_u \prod_{\{i|V_i \notin T\}} P(v_i|pa_i, u^i)P(u) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \tag{6}$$

And, the question of identifiability arises, i.e., whether it is possible to express some causal effect $P_t(s)$ as a function of the observed distribution $P(v)$, independent of the unknown quantities, $P(u)$ and $P(v_i|pa_i, u^i)$.

[4]Whether or not any actual action is an ideal manipulation of a variable (or is feasible at all) is not part of the theory - it is input to the theory.

It is convenient to represent a semi-Markovian model with a graph G that does not show the elements of U explicitly but, instead, represents the confounding effects of U variables using (dashed) bidirected edges. A bidirected edge between nodes V_i and V_j represents the presence of unobserved confounders that may influence both V_i and V_j. For example the model in Figure 1(a) will be represented by the graph in Figure 1(c).

In general we may be interested in identifying conditional causal effects $P_t(s|c)$, the causal effects of T on S conditioned on another set C of variables. This problem is important for evaluating *conditional plans* and stochastic plans [Pearl and Robins, 1995], where action T is taken to respond in a specified way to a set C of other variables – say, through a functional relationship $t = g(c)$. The effects of such actions may be evaluated through identifying conditional causal effects in the form of $P_t(s|c)$ [Pearl, 2000, chapter 4].

DEFINITION 2 (Causal-Effect Identifiability). The causal effect of a set of variables T on a disjoint set of variables S conditioned on another set C is said to be identifiable in a causal diagram G if the quantity $P_t(s|c)$ can be computed uniquely from any positive probability $P(v)$ of the observed variables—that is, if $P_t^{M_1}(s|c) = P_t^{M_2}(s|c)$ for every pair of models M_1 and M_2 with $P^{M_1}(v) = P^{M_2}(v) > 0$.

3 *Do*-calculus and Graphical Criteria

In general the identifiability of causal effects can be decided using Pearl's *do*-calculus – a set of inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences. A finite sequence of syntactic transformations, each applying one of the inference rules, may reduce expressions of the type $P_t(s)$ to subscript-free expressions involving observed quantities.

Let X, Y, and Z be arbitrary disjoint sets of nodes in G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. We denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. Similarly, $G_{\overline{X}\underline{Z}}$ will represent the deletion of both incoming and outgoing arrows.

THEOREM 3 (Rules of *do*-Calculus). *[Pearl, 1995] For any disjoint sets of variables X, Y, Z, and W we have the following rules.*

Rule 1 (*Insertion/deletion of observations*) :

$$P_x(y|z,w) = P_x(y|w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}}}. \tag{7}$$

Rule 2 (*Action/observation exchange*) :

$$P_{x,z}(y|w) = P_x(y|z,w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}}. \tag{8}$$

Rule 3 (*Insertion/deletion of actions*) :

$$P_{x,z}(y|w) = P_x(y|w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X},\overline{Z(W)}}}, \tag{9}$$

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

A key result about do-calculus is that any interventional distribution that is identifiable can be expressed in terms of the observational distribution by means of applying a sequence of do-calculus rules.

THEOREM 4. *[Shpitser and Pearl, 2006a] Do-calculus is complete for identifying causal effects of the form $P_x(y|z)$.*

In practice, do-calculus may be difficult to apply manually in complex causal diagrams, since, as stated, the rules give little guidance for chaining them together into a valid derivation.

Fortunately, a number of graphical criteria have been developed for quickly judging the identifiability by looking at the causal diagram G, of which the most influential are Pearl's back-door and front-door criteria. A path from X to Y is called *back-door* (relative to X) if it starts with an arrow pointing at X.

DEFINITION 5 (Back-Door). A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:

(i) no node in Z is a descendant of X_i; and

(ii) Z blocks every back-door path from X_i to X_j.

Similarly, if X and Y are two disjoint sets of nodes in G, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

THEOREM 6 (Back-Door Criterion). *[Pearl, 1995] If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula*

$$P_x(y) = \sum_z P(y|x, z)P(z). \tag{10}$$

For example, in Figure 1(c) X satisfies the back-door criterion relative to (Z, Y) and we have

$$P_z(y) = \sum_x P(y|x, z)P(x). \tag{11}$$

DEFINITION 7 (Front-Door). A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if:

(i) Z intercepts all directed paths from X to Y;

(ii) all back-door paths from X to Z are blocked (by empty set); and

(iii) all back-door paths from Z to Y are blocked by X.

THEOREM 8 (Front-Door Criterion). *[Pearl, 1995] If Z satisfies the front-door criterion relative to an ordered pair of variables (X, Y), then the causal effect of X on Y is identifiable and is given by the formula*

$$P_x(y) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x'). \tag{12}$$

For example, in Figure 1(c) Z satisfies the front-door criterion relative to (X, Y) and the causal effect $P_x(y)$ is given by Eq. (12).

There is a simple yet powerful graphical criterion for identifying the causal effects of a singleton. For any set S, let $An(S)$ denote the union of S and the set of ancestors of the variables in S. For any set C, let G_C denote the subgraph of G composed only of variables in C. Let a path composed entirely of bidirected edges be called a *bidirected path*.

THEOREM 9. *[Tian and Pearl, 2002a] The causal effect $P_x(s)$ of a variable X on a set of variables S is identifiable if there is no bidirected path connecting X to any of its children in $G_{An(S)}$.*

In fact, for X and S being singletons, this criterion covers both back-door and front-door criteria, and also the criterion in [Galles and Pearl, 1995].

These criteria are simple to use but are not necessary for identification. In the next sections we present complete systematic procedures for identification.

4 Identification of Causal Effects

In this section, we present a systematic procedure for identifying causal effects using so-called c-component decomposition.

4.1 C-component decomposition

The set of variables V in G can be partitioned into disjoint groups by assigning two variables to the same group if and only if they are connected by a bidirected path. Assuming that V is thus partitioned into k groups $S_1, \ldots, S_k$, each set S_j is called a *c-component* of V in G or a c-component of G. For example, the graph in Figure 1(c) consists of two c-components $\{X, Y\}$ and $\{Z\}$.

For any set $C \subseteq V$, define the quantity $Q[C](v)$ to denote the post-intervention distribution of C under an intervention to all other variables:[5]

$$Q[C](v) = P_{v \setminus c}(c) = \sum_u \prod_{\{i|V_i \in C\}} P(v_i|pa_i, u^i)P(u). \tag{13}$$

In particular, we have $Q[V](v) = P(v)$. If there is no bidirected edges connected with a variable V_i, then $U^i = \emptyset$ and $Q[\{V_i\}](v) = P(v_i|pa_i)$. For convenience, we will often write $Q[C](v)$ as $Q[C]$.

The importance of the c-component steps from the following lemma.

[5] Set $Q[\emptyset](v) = 1$ since $\sum_u P(u) = 1$.

LEMMA 10 (C-component Decomposition). *[Tian and Pearl, 2002a] Assuming that V is partitioned into c-components $S_1, \ldots, S_k$, we have*

(i) $P(v) = \prod_i Q[S_i]$.

(ii) Each $Q[S_i]$ is computable from $P(v)$. Let a topological order over V be $V_1 < \ldots < V_n$, and let $V^{(i)} = \{V_1, \ldots, V_i\}$, $i = 1, \ldots, n$, and $V^{(0)} = \emptyset$. Then each $Q[S_j]$, $j = 1, \ldots, k$, is given by

$$Q[S_j] = \prod_{\{i|V_i \in S_j\}} P(v_i|v^{(i-1)}) \tag{14}$$

The lemma says that for each c-component S_i the causal effect $Q[S_i] = P_{v \setminus s_i}(s_i)$ is identifiable. For example, in Figure 1(c), we have $P_{x,y}(z) = Q[\{Z\}] = P(z|x)$ and $P_z(x, y) = Q[\{X, Y\}] = P(y|x, z)P(x)$.

Lemma 10 can be generalized to the subgraphs of G as given in the following lemma.

LEMMA 11 (Generalized C-component Decomposition). *[Tian and Pearl, 2003] Let $H \subseteq V$, and assume that H is partitioned into c-components $H_1, \ldots, H_l$ in the subgraph G_H. Then we have*

(i) $Q[H]$ decomposes as

$$Q[H] = \prod_i Q[H_i]. \tag{15}$$

(ii) Each $Q[H_i]$ is computable from $Q[H]$. Let k be the number of variables in H, and let a topological order of the variables in H be $V_{m_1} < \cdots < V_{m_k}$ in G_H. Let $H^{(i)} = \{V_{m_1}, \ldots, V_{m_i}\}$ be the set of variables in H ordered before V_{m_i} (including V_{m_i}), $i = 1, \ldots, k$, and $H^{(0)} = \emptyset$. Then each $Q[H_j]$, $j = 1, \ldots, l$, is given by

$$Q[H_j] = \prod_{\{i|V_{m_i} \in H_j\}} \frac{Q[H^{(i)}]}{Q[H^{(i-1)}]}, \tag{16}$$

where each $Q[H^{(i)}]$, $i = 1, \ldots, k$, is given by

$$Q[H^{(i)}] = \sum_{h \setminus h^{(i)}} Q[H]. \tag{17}$$

Lemma 11 says that if the causal effect $Q[H] = P_{v \setminus h}(h)$ is identifiable, then for each c-component H_i of the subgraph G_H, the causal effect $Q[H_i] = P_{v \setminus h_i}(h_i)$ is identifiable.

Next, we show how to use the c-component decomposition to identify causal effects.

4.2 Computing causal effects

First we present a facility lemma. For $W \subseteq C \subseteq V$, the following lemma gives a condition under which $Q[W]$ can be computed from $Q[C]$ by summing over $C \setminus W$, like ordinary marginalization in probability theory.

LEMMA 12. *[Tian and Pearl, 2003] Let $W \subseteq C \subseteq V$, and $W' = C \setminus W$. If W contains its own ancestors in the subgraph G_C ($An(W)_{G_C} = W$), then*

$$\sum_{w'} Q[C] = Q[W]. \tag{18}$$

Note that we always have $\sum_c Q[C] = 1$.

Next, we show how to use Lemmas 10–12 to identify the causal effect $P_t(s)$ where S and T are arbitrary (disjoint) subsets of V. We have

$$P_t(s) = \sum_{(v \setminus t) \setminus s} P_t(v \setminus t) = \sum_{(v \setminus t) \setminus s} Q[V \setminus T]. \tag{19}$$

Let $D = An(S)_{G_{V \setminus T}}$. Then by Lemma 12, variables in $(V \setminus T) \setminus D$ can be summed out:

$$P_t(s) = \sum_{d \setminus s} \sum_{(v \setminus t) \setminus d} Q[V \setminus T] = \sum_{d \setminus s} Q[D]. \tag{20}$$

Assume that the subgraph G_D is partitioned into c-components $D_1, \ldots, D_l$. Then by Lemma 11, $Q[D]$ can be decomposed into products of $Q[D_i]$'s, and Eq. (20) can be rewritten as

$$P_t(s) = \sum_{d \setminus s} \prod_i Q[D_i]. \tag{21}$$

We obtain that $P_t(s)$ is identifiable if all $Q[D_i]$'s are identifiable.

Let G be partitioned into c-components $S_1, \ldots, S_k$. Then any D_i is a subset of certain S_j since if the variables in D_i are connected by a bidirected path in a subgraph of G then they must be connected by a bidirected path in G. Assuming $D_i \subseteq S_j$, $Q[D_i]$ is identifiable if it is computable from $Q[S_j]$. In general, for $C \subseteq T \subseteq V$, whether $Q[C]$ is computable from $Q[T]$ can be determined recursively by repeated applications of Lemmas 12 and 11, as given in the recursive algorithm shown in Figure 2. At each step of the algorithm, we either find an expression for $Q[C]$, find $Q[C]$ unidentifiable, or reduce the problem to a simpler one.

In summary, an algorithm for computing $P_t(s)$ is given in Figure 3, and the algorithm has been shown to be complete, that is, if the algorithm outputs FAIL, then $P_t(s)$ is not identifiable.

THEOREM 13. *[Shpitser and Pearl, 2006b; Huang and Valtorta, 2006a] The algorithm **ID** in Figure 3 is complete.*

5 Identification of Conditional Causal Effects

An important refinement to the problem of identifying causal effects $P_x(y)$ is concerned with identifying *conditional causal effects*, in other words causal effects in a particular subpopulation where variables Z are known to attain values z. These

Algorithm Identify(C, T, Q)
INPUT: $C \subseteq T \subseteq V$, $Q = Q[T]$. G_T and G_C are both composed of one single c-component.
OUTPUT: Expression for $Q[C]$ in terms of Q or FAIL.

Let $A = An(C)_{G_T}$.

- IF $A = C$, output $Q[C] = \sum_{t \setminus c} Q$.
- IF $A = T$, output FAIL.
- IF $C \subset A \subset T$
  1. Assume that in G_A, C is contained in a c-component T'.
  2. Compute $Q[T']$ from $Q[A] = \sum_{t \setminus a} Q$ by Lemma 11.
  3. Output Identify$(C, T', Q[T'])$.

Figure 2. An algorithm for determining if $Q[C]$ is computable from $Q[T]$.

Algorithm ID(s, t)
INPUT: two disjoint sets $S, T \subset V$.
OUTPUT: the expression for $P_t(s)$ or FAIL.
Phase-1:

1. Find the c-components of G: $S_1, \ldots, S_k$. Compute each $Q[S_i]$ by Lemma 10.
2. Let $D = An(S)_{G_{V \setminus T}}$ and the c-components of G_D be D_i, $i = 1, \ldots, l$.

Phase-2:
For each set D_i such that $D_i \subseteq S_j$:
Compute $Q[D_i]$ from $Q[S_j]$ by calling **Identify**$(D_i, S_j, Q[S_j])$ in Figure 2. If the function returns FAIL, then stop and output FAIL.

Phase-3: Output $P_t(s) = \sum_{d \setminus s} \prod_i Q[D_i]$.

Figure 3. A complete algorithm for computing $P_t(s)$.

conditional causal effects are written as $P_x(y|z)$, and defined just as regular conditional distributions as

$$P_x(y|z) = \frac{P_x(y,z)}{P_x(z)}$$

Complete closed form algorithms for identifying effects of this type have been developed. One approach [Tian, 2004] generalizes the algorithm for identifying *unconditional* causal effects $P_x(y)$ found in Section 4. There is, however, an easier approach which works.

The idea is to reduce the expression $P_x(y|z)$, which we don't know how to handle to something like $P_{x'}(y')$, which we do know how to handle via the algorithm already presented. This reduction would have to find a way to get rid of variables Z in the conditional effect expression.

Ridding ourselves of some variables in Z can be accomplished via rule 2 of do-calculus. Recall that applying rule 2 to an expression allows us to replace conditioning on some variable set $W \subseteq Z$ by fixing W instead. Rule 2 states that this is possible in the expression $P_x(y|z)$ whenever W contains no back-door paths to Y conditioned on the remaining variables in Z and X (that is $X \cup Z \setminus W$), in the graph where all incoming arrows to X have been cut.

It's not difficult to show the following uniqueness lemma.

LEMMA 14. *[Shpitser and Pearl, 2006a] For every conditional effect $P_x(y|z)$ there exists a unique maximal $W \subseteq Z$ such that $P_x(y|z)$ is equal to $P_{x,w}(y|z \setminus w)$ according to rule 2 of do-calculus.*

Lemma 14 states that we only need to apply rule 2 once to rid ourselves of as many conditioned variables as possible in the effect of interest. However, even after this is done, we may be left with some variables in $Z \setminus W$ past the conditioning bar in our effect expression. If we insist on using unconditional effect identification, we may try to identify the joint distribution $P_{x,w}(y, z \setminus w)$ to obtain an expression α, and obtain the conditional distribution $P_{x,w}(y|z \setminus w)$ by taking $\frac{\alpha}{\sum_y \alpha}$. But what if $P_{x,w}(y, z \setminus w)$ is not identifiable? Are there cases where $P_{x,w}(y, z \setminus w)$ is not identifiable, but $P_{x,w}(y|z \setminus w)$ is? Fortunately, it turns out the answer is no.

LEMMA 15. *[Shpitser and Pearl, 2006a] Let $P_x(y|z)$ be a conditional effect of interest, and $W \subseteq Z$ the unique maximal set such that $P_x(y|z)$ is equal to $P_{x,w}(y|z \setminus w)$. Then $P_x(y|z)$ is identifiable if and only if $P_{x,w}(y, z \setminus w)$ is identifiable.*

Lemma 15 gives us a simple algorithm for identifying arbitrary conditional effects by first reducing the problem into one of identifying an unconditional effect – and then invoking the complete algorithm **ID** in Figure 3. This simple algorithm is actually complete since the statement in Lemma 15 is if and only if. The algorithm itself is shown in Fig. 4. The algorithm as shown picks elements of W one at a time, although the set it picks as it iterates will equal the maximal set W due to the following lemma.

Algorithm **IDC**(y, x, z)
INPUT: disjoint sets $X, Y, Z \subset V$.
OUTPUT: Expression for $P_x(y|z)$ in terms of P or FAIL.

1 if $(\exists W \in Z)(Y \perp\!\!\!\perp W | X, Z \setminus \{W\})_{G_{\overline{x},\underline{w}}}$,
return **IDC**$(y, x \cup \{w\}, z \setminus \{w\})$.

2 else let $P' =$ **ID**$(y \cup z, x)$.
return $P' / \sum_y P'$.

Figure 4. A complete identification algorithm for conditional effects.

LEMMA 16. *Let $P_x(y|z)$ be a conditional effect of interest in a causal model inducing G, and $W \subseteq Z$ the unique maximal set such that $P_x(y|z)$ is equal to $P_{x,w}(y|z \setminus w)$. Then $W = \{W' | P_x(y|z) = P_{x,w'}(y|z \setminus \{w'\})\}$.*

Completeness of the algorithm easily follows from the results we presented.

THEOREM 17. *[Shpitser and Pearl, 2006a] The algorithm **IDC** is complete.*

We note that the procedures **ID** and **IDC** served as a means to prove the completeness of do-calculus (Theorem 4). The proof [Shpitser and Pearl, 2006b] proceeds by reducing the steps in these procedures to sequences of do-calculus derivations.

6 Relative Interventions and the Effect of Treatment on the Treated

Interventions considered in the previous sections are what we term "absolute," since the values x to which variables are set by $do(x)$ bear no relationship to whatever natural values were assumed by variables X prior to an intervention. Such absolute interventions correspond to clamping a wire in a circuit to ground, or performing a randomized clinical trial for a drug which does not naturally occur in the body.

By contrast, many interventions are *relative*, in other words, the precise level x to which the variable X is set depends on the values X naturally attains. A typical relative intervention is the addition of insulin to the bloodstream. Since insulin is naturally synthesized by the human body, the effect of such an intervention depends on the initial, pre-intervention concentration of insulin in the blood, even if a constant amount is added for every patient. The insulin intervention can be denoted by $do(i + X)$, where i is the amount of insulin added, and X denotes the random variable representing pre-intervention insulin concentration in the blood. More generally, a relative intervention on a variable X takes the form of $do(f(X))$ for some function f.

How are we to make sense of a relative intervention $do(f(X))$ on X applied to a given population where the values of X are not known? Can relative interventions

be reduced to absolute interventions? It appears that in general the answer is "no." Consider: if we knew that X attained the value x for a given unit, then the effect of an intervention in question on the outcome variable Y is really $P(y|do(f(x)), x)$. This expression is almost like the (absolute) conditional causal effect of $do(f(x))$ on y, except the evidence that is being conditioned on is on the same variable that is being intervened. Since x and $f(x)$ are not in general the same, it appears that this expression contains a kind of value conflict. Are these kinds of probabilities always 0? Are they even well defined?

In fact, expressions of this sort are a special case of a more general notion of a *counterfactual distribution*, which can be derived from *functional causal models* [Pearl, 2000, Chapter 7]. Such models consist of two sets of variables, the observable set V representing the domain of interest, and the unobservable set U representing the background to the model that we are ignorant of. Associated with each observable variable V_i in V is a function f_i which determines the value of V_i in terms of values of other variables in $V \cup U$. Finally, there is a joint probability distribution $P(u)$ over the unobservable variables, signifying our ignorance of the background conditions of the model.

The causal relationships in functional causal models are represented, naturally, by the functions f_i; each function causally determines the corresponding V_i in terms of its inputs. Causal relationships entailed by a given model have an intuitive visual representation using a causal diagram. Causal diagrams contain two kinds of edges. Directed edges are drawn from a variable X to a variable V_i if X appears as an input of f_i. Directed edges from the same unobservable U_i to two observables V_j, V_k can be replaced by a bidirected edge between V_j to V_k. We will consider semi-Markovian models which induce acyclic graphs where $P(u) = \prod_i P(u_i)$, and each U_i has at most two observable children. A graph obtained in this way from a model is said to be induced by said model.

Unlike causal Bayesian networks introduced in Section 2, functional causal models represent fundamentally deterministic causal relationships which only appear stochastic due to our ignorance of background variables. This inherent determinism allows us to define counterfactual distributions which span multiple worlds under different interventions regimes. Formally, a joint counterfactual distribution is a distribution over events of the form Y_x where Y is a post-intervention random variable in a causal model (the intervention in question being $do(x)$). A single joint distribution can contain multiple such events, with different, possibly conflicting interventions.

Such joint distributions are defined as follows:

$$P(Y^1_{x^1} = y^1, ..., Y^k_{x^k} = y^k) = \sum_{\{u|Y^1_{x^1}(u)=y^1 \wedge ... \wedge Y^k_{x^k}(u)=y^k\}} P(u), \tag{22}$$

where U is the set of unobserved variables in the model. In other words, a joint counterfactual probability is obtained by adding up the probabilities of every setting

of unobserved variables in the model that results in the observed values of each counterfactual event Y_x in the expression. The query with the conflict we considered above can then be expressed as a conditional distribution derived from such a joint, specifically $P(Y_{f(x)} = y|X = x) = \frac{P(Y_{f(x)}=y, X=x)}{P(X=x)}$. Queries of this form are well known in the epidemiology literature as the effect of treatment on the treated (ETT) [Heckman, 1992; Robins *et al.*, 2006].

In fact, relative interventions aren't quite the same as ETT since we don't actually know the original levels of X. To obtain effects of relative interventions, we simply average over possible values of X, weighted by the prior distribution $P(x)$ of X. In other words, the relative causal effect $P(y|do(f(X)))$ is equal to $\sum_x P(Y_{f(x)} = y|X = x)P(X = x)$.

Since relative interventions reduce to ETT, and because ETT questions are of independent interest, identification of ETT is an important problem. If interventions are performed over multiple variables, it turns out that identifying ETT questions is almost as intricate as general counterfactual identification [Shpitser and Pearl, 2009; Shpitser and Pearl, 2007]. However, in the case of a singleton intervention, there is a formulation which bypasses most of the complexity of counterfactual identification. This formulation is the subject of this section.

We want to approach identification of ETT in the same way we approached identification of causal effects in the previous sections, namely by providing a graphical representation of conditional independences in joint distributions of interest, and then expressing the identification algorithm in terms of this graphical representation. In the case of causal effects, we were given as input the causal diagram representing the original, pre-intervention world, and we were asking questions about the post-intervention world where arrows pointing to intervened variables were cut. In the case of counterfactuals we are interested in joint distributions that span multiple worlds each with its own intervention. We want to construct a graph for these distributions.

The intuition is that each interventional world is represented by a copy of the original causal diagram, with the appropriate incoming arrows cut to represent the changes in the causal structure due to the intervention. All worlds are assumed to share history up to the moment of divergence due to differing interventions. This is represented by all worlds sharing unobserved variables U. In the special case of two interventional worlds the resulting graph is known as the *twin network graph* [Balke and Pearl, 1994b; Balke and Pearl, 1994a].

In the general case, a refinement of the resulting graph (to account for the possibility of duplicate random variables) is known as the *counterfactual graph* [Shpitser and Pearl, 2007]. The counterfactual graph represents conditional independences in the corresponding counterfactual distribution via the d-separation criterion just as the causal diagram represents conditional independences in the observed distribution of the original world. The graph in Figure 5(b) is a counterfactual graph for the query $P(Y_x = y|X = x')$ obtained from the original causal diagram shown in

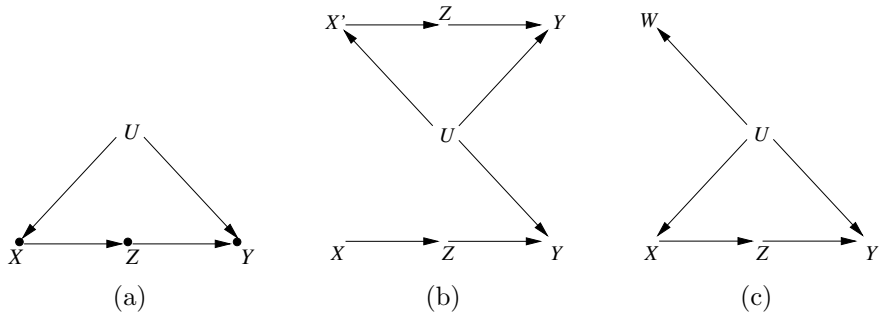


Figure 5. (a) A causal diagram G. (b) The counterfactual graph for $P(Y_x = y|x')$ in G. (c) The graph G' from Theorem 18.

Figure 5(a).

There exists a rather complicated general algorithm for identifying arbitrary counterfactual distributions from either interventional or observational data [Shpitser and Pearl, 2007; Shpitser and Pearl, 2008], based on ideas from the causal effect identification algorithms given in the previous sections, only applied to the counterfactual graph, rather than the causal diagram. It turns out that while identifying ETT of a single variable X can be represented as an identification problem of ordinary causal effects, ETT of multiple variables is significantly more complex [Shpitser and Pearl, 2009]. In this paper, we will concentrate on single variable ETT with multiple outcome variables Y.

What makes single variable ETT $P(Y_x = y|X = x')$ particularly simple is the form of its counterfactual graph. For the case of all ETTs, this graph will have variables from two worlds – the "natural" world where X is observed to have taken the value x' and the interventional world, where X is fixed to assume the value x. There are two key points that simplify matters. The first is that no descendant of X (including variables in Y) is of interest in the "natural" world, since we are only interested in the outcome Y in the interventional world. The second is that all non-descendants of X behave the same in both worlds (since interventions do not affect non-descendants). Thus, when constructing the counterfactual graph we don't need to make copies of non-descendants of X, and we can ignore descendants of X in the "natural" world. But this means the only variable in the "natural" world we will construct is a copy of X itself.

What this implies is that a problem of identifying the ETT $P(Y_x = y|X = x')$ can be rephrased as a problem of identifying a certain conditional causal effect.

THEOREM 18. *[Shpitser and Pearl, 2009] For a singleton variable X, and a set Y, $P(Y_x = y|X = x')$ is identifiable in G if and only if $P_x(y|w)$ is identifiable in G', where G' is obtained from G by adding a new node W with the same set of parents (both observed and unobserved) as X, and no children. Moreover, the estimand for*

$P(Y_x = y|X = x')$ is equal to that of $P_x(y|w)$ with all occurrences of w replaced by x'.

We illustrate the application of Theorem 18 by considering the graph G in Fig. 5(a). The query $P(Y_x = y|X = x')$ is identifiable by considering $P_x(y|w)$ in the graph G' shown in Fig. 5(c), while the counterfactual graph for $P(Y_x = y|x')$ is shown in Fig. 5(b). Identifying $P_x(y|w)$ in G' using the algorithm **IDC** in the previous section leads to $\sum_z P(z|x) \sum_x P(y|z, w, x)P(w, x)/P(w)$. Replacing w by x' yields the expression $\sum_z P(z|x) \sum_{x''} P(y|z, x', x'')P(x', x'')/P(x')$.

Ordinarily, we know that $P(y|z, x', x'')$ is undefined if x' is not equal to x''. However, in our case, we know that observing $X = x'$ in the natural world implies $X = x'$ in any other interventional world which shares ancestors of X with the natural world. This implies the expression $\sum_{x''} P(y|z, x', x'')P(x', x'')/P(x')$ is equivalent to $P(y|z, x')$, thus our query $P(Y_x = y|X = x')$ is equal to $\sum_z P(y|z, x')P(z|x)$.

It is possible to use Theorem 18 to derive analogues of the back-door and front-door criteria for ETT.

COROLLARY 19 (Back-door Criterion for ETT). *If a set Z satisfies the back-door criterion relative to (X, Y), where X is a singleton variable, then $P(Y_x = y|X = x')$ is identifiable and equal to $\sum_z P(y|z, x)P(z|x')$.*

The intuition for the back-door criterion for ETT is that Z, by assumption, screens X and Y from observed values of X in other counterfactual worlds. Thus, the first term in the back-door expression does not change. The second term changes in an obvious way since Z depends on observing $X = x'$.

COROLLARY 20 (Front-door Criterion for ETT). *If a set Z satisfies the front-door criterion relative to (X, Y), where X, Y are singleton variables, then $P(Y_x = y|X = x')$ is identifiable and equal to $\sum_z P(y|z, x')P(z|x)$.*

Proof. We will be using a number of graphs in this proof. G is the original graph. G^w is the graph obtained from G by adding a copy of X called W with the same parents (including unobserved parents) as X and no children. G' is a graph representing independences in $P(X, Y, Z)$. It is obtained from G by removing all nodes other than X, Y, Z, by adding a directed arrow between any remaining A and B in X, Y, Z if there is a d-connected path containing only nodes not in X, Y, Z which starts with a directed arrow pointing away from A and ends with any arrow pointing to B. Similarly, a bidirected arrow is added between any A and B in X, Y, Z if there is a d-connected path containing only nodes not in X, Y, Z which starts with any arrow pointing to A and ends with any arrow pointing to B. (This graph is known as a latent projection [Pearl, 2000]). The graphs $G'^w, G'^w_{\overline{x}}$ are defined similarly as above.

We want to identify $P_x(y, z, w)$ in G'^w. First, we want to show that no node in Z shares a c-component with W or any node in Y in $G'^w_{\overline{x}}$. This can only happen if a node in Z and W or a node in Y share a bidirected arc in $G'^w_{\overline{x}}$. But this means that

either there is a back-door d-connected path from Z to Y in $G_{\overline{x}}$, or there is a back-door d-connected path from X to Z in G. Both of these claims are contradicted by our assumption that Z satisfies the front-door criterion for (X, Y).

This implies $P_x(y, z, w) = P_{z,x}(y, w)P_{x,w}(z)$ in G^w.

By construction of G^w and the front-door criterion, $P_{x,w}(z) = P_x(z) = P(z|x)$. Furthermore, since no nodes in Z and Y share a c-component in G'^w, no node in Z has a bidirected path to Y in G'^w. This implies, by Lemma 1 in [Shpitser *et al.*, 2009], that $P_z(y, w, x) = P(y|z, w, x)P(w, x)$.

Since Z intercepts all directed paths from X to Y (by the front-door criterion), $P_{z,x}(y, w) = P_z(y, w) = \sum_x P(y|z, w, x)P(w, x)$.

We conclude that $P_x(y, w)$ is equal to $\sum_z P(z|x) \sum_x P(y|z, w, x)P(w, x)$. Since $P_x(w) = P(w)$ in G'^w, $P_x(y|w) = \sum_z P(z|x) \sum_x P(y|z, w, x)P(x|w)$.

Finally, recall that W is just a copy of X, and X is observed to attain value x' in the "natural" world. This implies that our expression simplifies to $\sum_z P(z|x)P(y|z, x')$, which proves our result. □

If neither the back-door nor the front-door criteria hold, we must invoke general causal effect identification algorithms from the previous sections. However, in the case of ETT of a single variable, there is a simple complete graphical criterion which works.

THEOREM 21. *[Shpitser and Pearl, 2009] For a singleton variable X, and a set Y, $P(Y_x = y|X = x')$ is identifiable in G if and only if there is no bidirected path from X to a child of X in $G_{an(y)}$. Moreover, if there is no such bidirected path, the estimand for $P(Y_x = y|X = x')$ is obtained by multiplying the estimand for $\sum_{an(y)\setminus(y\cup\{x\})} P_x(an(y) \setminus x)$ (which exists by Theorem 9) by $\frac{Q[S^x]'}{P(x')\sum_x Q[S^x]}$, where S^x is the c-component in G containing X, and $Q[S^x]'$ is obtained from the expression for $Q[S^x]$ by replacing all occurrences of x with x'.*

7 Conclusion

In this paper we described the state of the art in identification of causal effects and related quantities in the framework of graphical causal models. We have shown how this framework, developed over the period of two decades by Judea Pearl and his collaborators, and presented in Pearl's seminal work [Pearl, 2000], can sharpen causal intuition into mathematical precision for a variety of causal problems faced by scientists.

Acknowledgments: Jin Tian was partly supported by NSF grant IIS-0347846. Ilya Shpitser was partly supported by AFOSR grant #F49620-01-1-0055, NSF grant #IIS-0535223, MURI grant #N00014-00-1-0617, and NIH grant #R37AI032475.

References

A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds, and applications. In R. Lopez de Mantaras and D. Poole, editors,

Uncertainty in Artificial Intelligence 10, pages 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume I, pages 230–237. MIT Press, Menlo Park, CA, 1994.

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

C. Glymour and G. Cooper, editors. *Computation, Causation, and Discovery*. MIT Press, Cambridge, MA, 1999.

D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

J.J. Heckman. Randomization and social policy evaluation. In C. Manski and I. Garfinkle, editors, *Evaluations: Welfare and Training Programs*, pages 201–230. Harvard University Press, 1992.

Y. Huang and M. Valtorta. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1149–1154, Menlo Park, CA, July 2006. AAAI Press.

Y. Huang and M. Valtorta. Pearl's calculus of interventions is complete. In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. AUAI Press, July 2006.

M. Kuroki and M. Miyakawa. Identifiability criteria for causal effects of joint interventions. *Journal of the Japan Statistical Society*, 29(2):105–117, 1999.

S. Lauritzen. Graphical models for causal inference. In O.E. Barndorff-Nielsen, D. Cox, and C. Kluppelberg, editors, *Complex Stochastic Systems*, chapter 2, pages 67–112. Chapman and Hall/CRC Press, London/Boca Raton, 2000.

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

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

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

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–710, December 1995.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, NY, 2000.

James M. Robins, VanderWeele Tyler J., and Thomas S. Richardson. Comment on causal effects in the presence of non compliance: a latent variable interpretation by antonio forcina. *METRON*, LXIV(3):288–298, 2006.

J.M. Robins. 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*, 7:1393–1512, 1986.

J.M. Robins. 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.

J.M. Robins. Causal inference from complex longitudinal data. In *Latent Variable Modeling with Applications to Causality*, pages 69–117. Springer-Verlag, New York, 1997.

I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444. AUAI Press, July 2006.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226, Menlo Park, CA, July 2006. AAAI Press.

Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. In *Twenty Third Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2007.

I. Shpitser and J. Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Ilya Shpitser and Judea Pearl. Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 25, 2009.

Ilya Shpitser, Thomas S. Richardson, and James M. Robins. Testing edges by truncations. In *International Joint Conference on Artificial Intelligence*, volume 21, pages 1957–1963, 2009.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search (2nd Edition)*. MIT Press, Cambridge, MA, 2001.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI)*, pages 567–573, Menlo Park, CA, 2002. AAAI Press/The MIT Press.

J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.

J. Tian and J. Pearl. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles, 2003.

J. Tian. Identifying conditional causal effects. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.