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

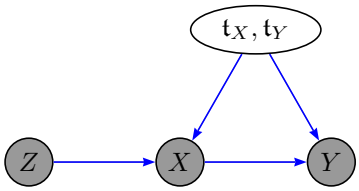


Figure 1. Graphical representation of the IV model given by assumptions (1) and (2).The shaded nodes are observed.

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:

$$\left\{ \begin{array}{lcll} p(1,1) &=& t & \\ p(1,0) &=& c_1 - t & \\ p(0,1) &=& c_2 - t & t \in \left[\max\{0, (c_1+c_2)-1\}, \min\{c_1, c_2\}\right] \\ p(0,0) &=& 1 - c_1 - c_2 + t & \end{array} \right\}. \tag{3}$$

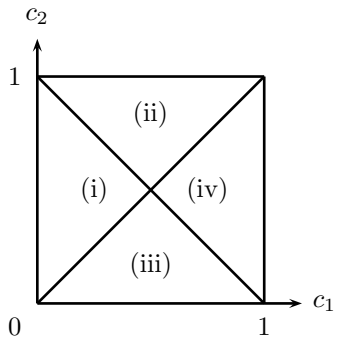


Figure 2. The four regions corresponding to different supports for t in (3); see Table 1.

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

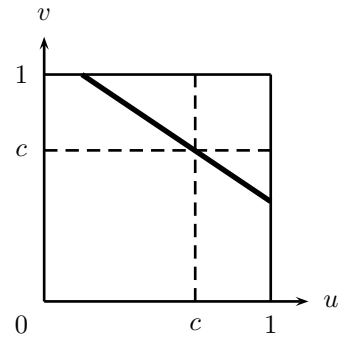


Figure 3. Illustration of $\mathcal{Q}_{c,\alpha}$.

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:

$$\left\{ \begin{array}{lcll} u & = & (c_1 - t(1-\alpha_1))/\alpha_1, & \\ v & = & (c_2 - t(1-\alpha_2))/\alpha_2, & t \in [t_l, t_u] \\ w & = & t, & \end{array} \right\},$$

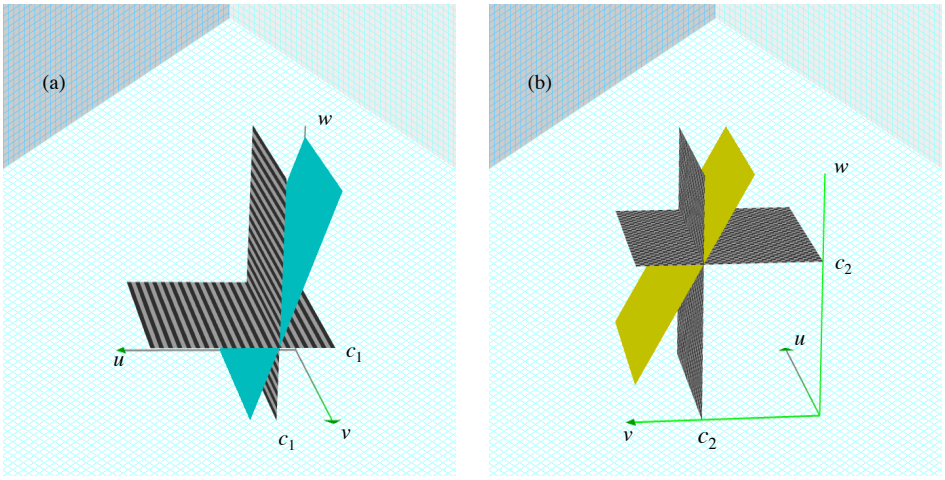


Figure 4. (a) The plane without stripes is $\alpha_1 u + (1-\alpha_1)w = c_1$. (b) The plane without checks is $\alpha_2 v + (1-\alpha_2)w = c_2$.

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:

$$([0,c_1] \times [0,c_2] \times [c_2,1]) \cup ([0,c_1] \times [c_2,1] \times [c_1,c_2]) \cup ([c_1,1] \times [c_2,1] \times [0,c_1]).$$

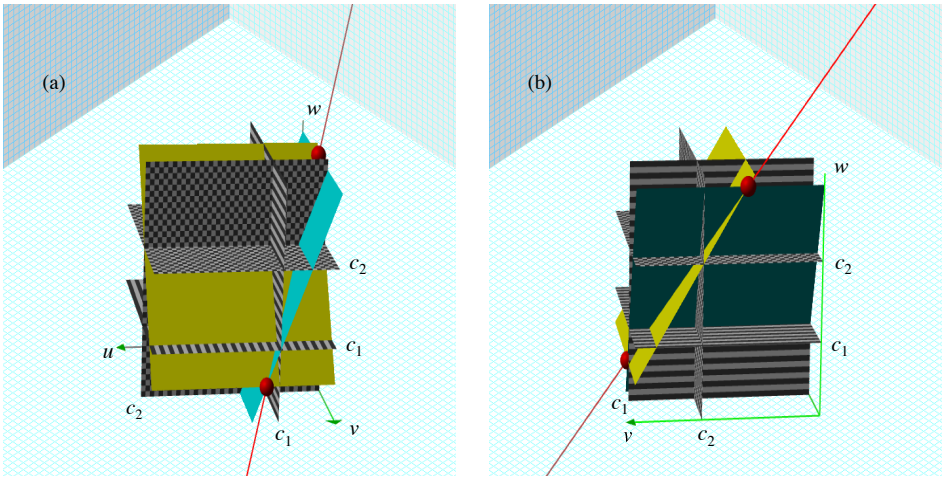


Figure 5. $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$ corresponds to the section of the line between the two marked points; (a) view towards u-w plane; (b) view from v-w plane. (Here $c_1 < c_2$.)

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-

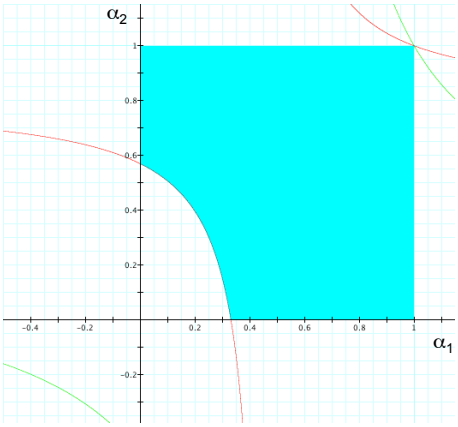


Figure 6. $\mathcal{R}_{c_1,c_2}$ corresponds to the shaded region. The hyperbola of which one arm forms a boundary of this region corresponds to the active constraint; the other hyperbola to the inactive constraint.

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.

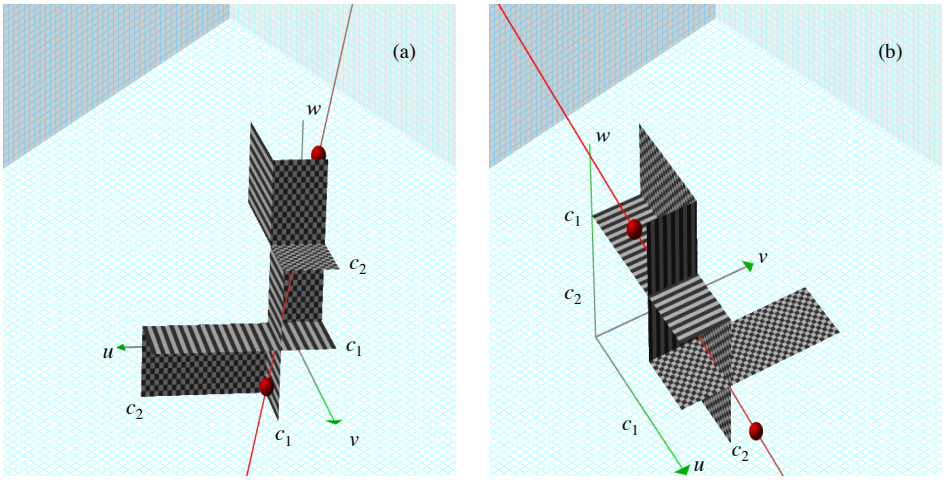


Figure 7. 'Staircases' of three boxes illustrating the possible support for $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$; (a) $c_1 < c_2$; (b) $c_2 < c_1$. Sides of the boxes that are formed by (subsets of) faces of the unit cube are not shown. The line segments shown are illustrative; in general they 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:

$$\gamma^i_{\mathbf{t}_X} \equiv p(y_{x=i} = 1 \mid \mathbf{t}_X), \text{ for } i \in \{0,1\} \text{ and } \mathbf{t}_X \in \mathbb{D}_X,$$

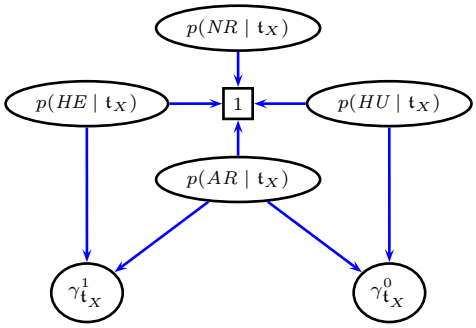


Figure 8. A graph representing the functional dependencies used in the reduction step in §3.2. The rectangular node indicates that the probabilities are required to sum to 1.

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

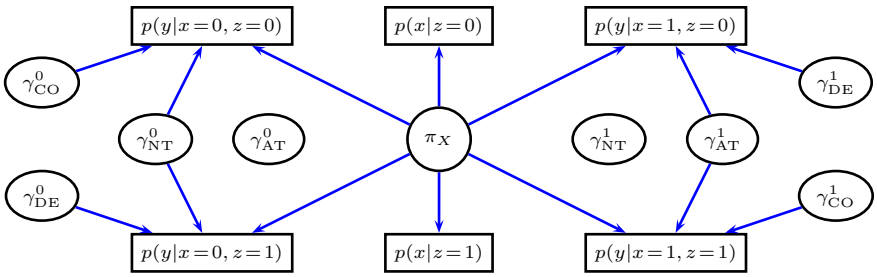


Figure 9. A graph representing the functional dependencies in the analysis of the binary IV model. Rectangular nodes are observed; oval nodes are unknown parameters. See text for further explanation.

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

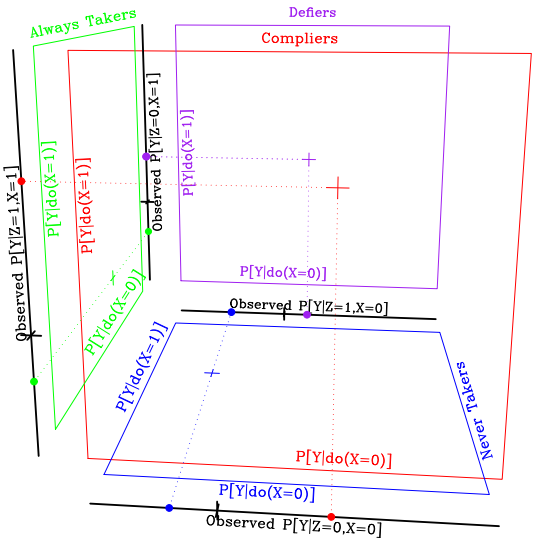


Figure 10. Geometric picture illustrating the relation between the $\gamma^i_{\mathbf{t}_X}$ parameters and $p(y \mid x, z)$. See also Figure 9.

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.

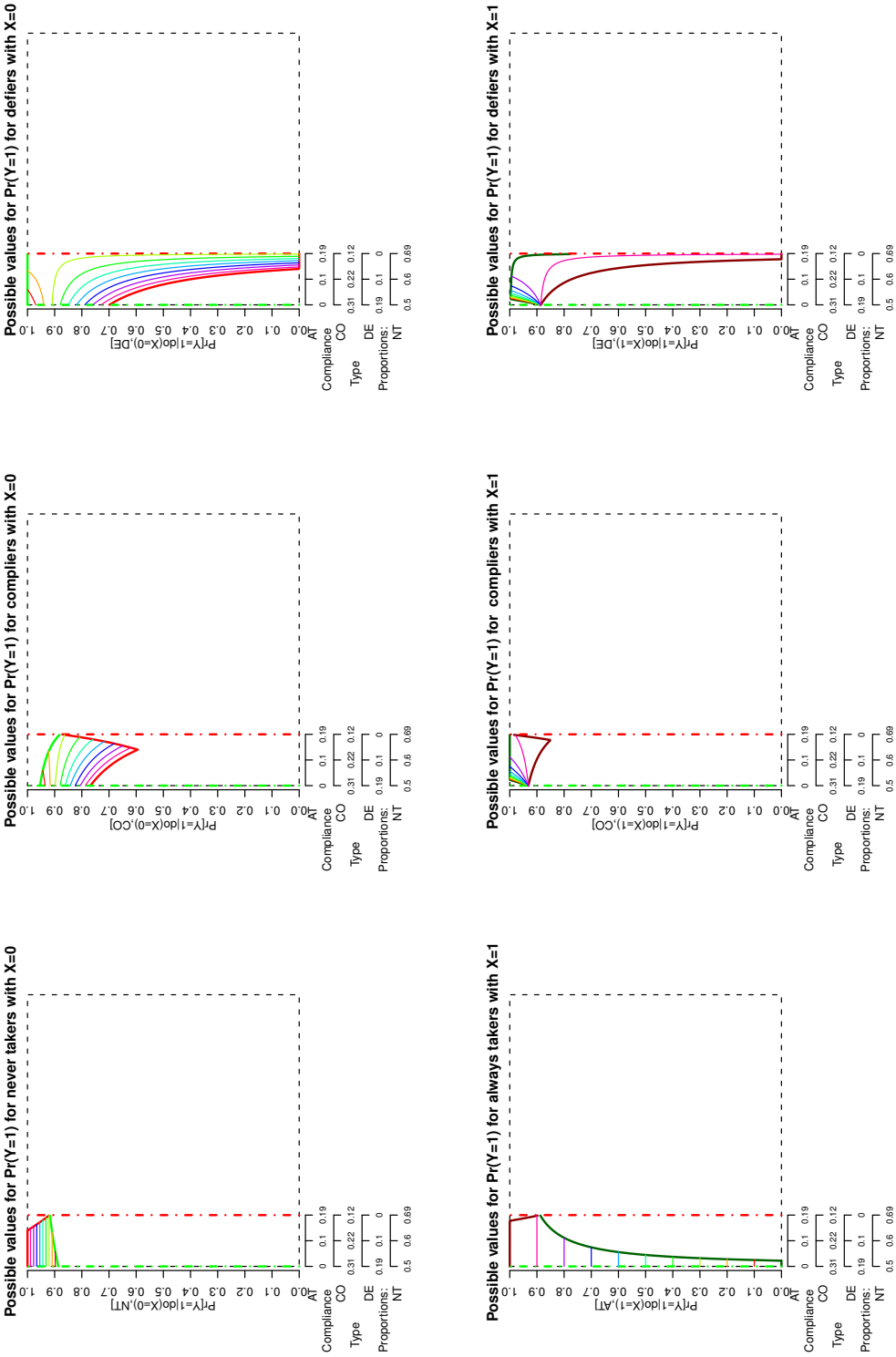


Figure 11. Depiction of the set of values for π_X vs. $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ (upper row), and π_X vs. $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$ for the flu data.

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.

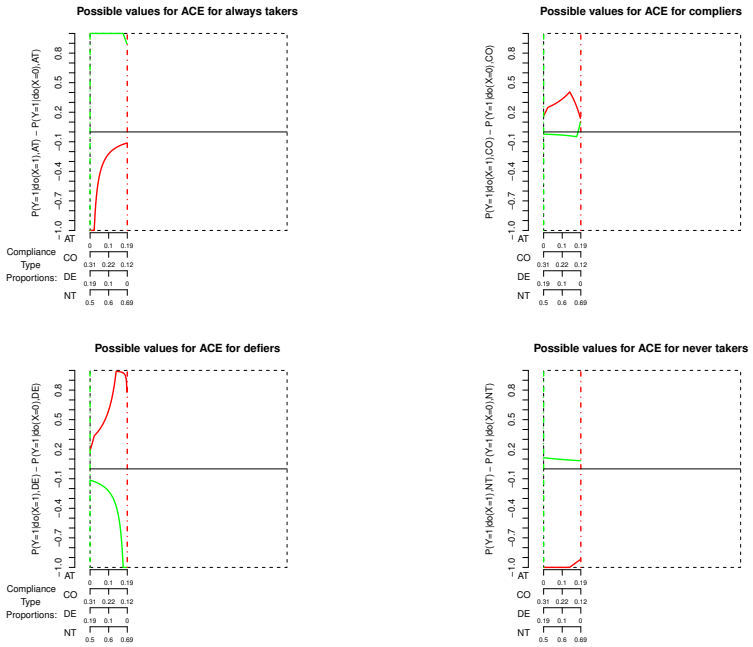


Figure 12. Depiction of the set of values for π_X vs. $\mathrm{ACE}_{\mathbf{t}_X}(\pi_X)$ for $\mathbf{t}_X \in \mathbb{D}_X$ for the flu data.

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) $\max\left\{\sum_j p(y{=}j, x{=}1 \mid z{=}j),\ \sum_k p(y{=}k, x{=}1 \mid z{=}1-k)\right\} \leq 1$.

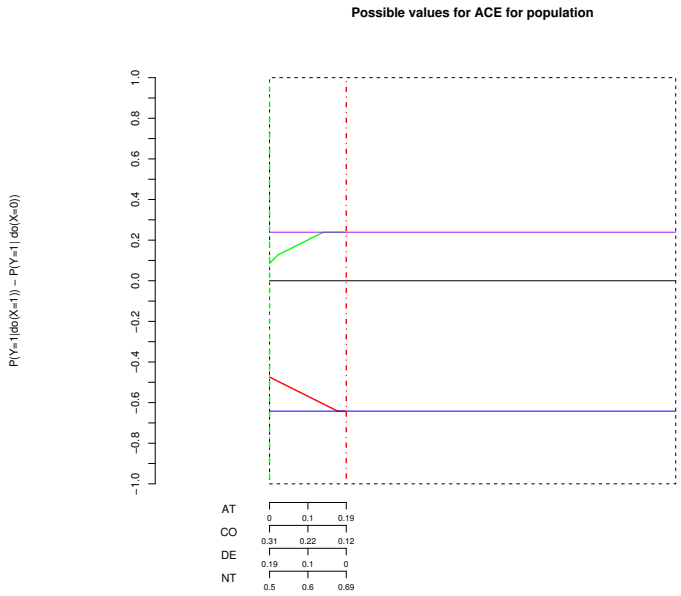


Figure 13. Depiction of the set of values for π_X vs. the global ACE for the flu data. The horizontal lines represent the overall bounds on the global ACE due to Pearl.

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. □

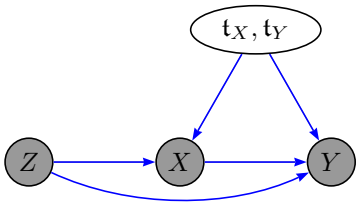


Figure 14. Graphical representation of the model given by the randomization assumption (2) alone. It is no longer assumed that Z does not have a direct effect on Y.

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.

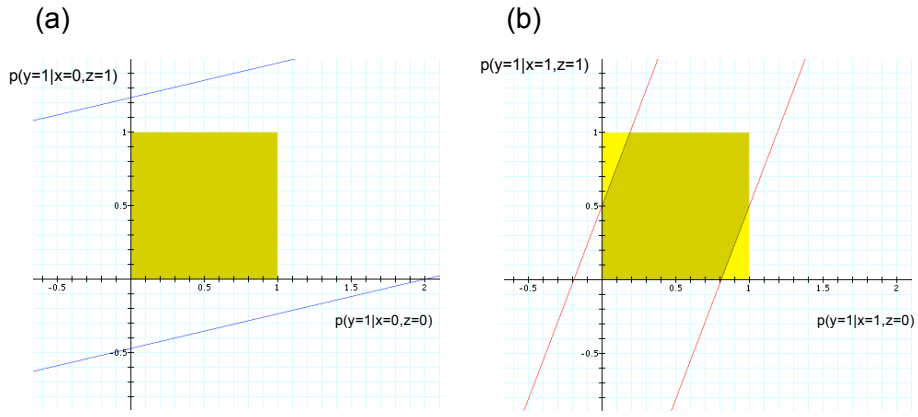


Figure 15. Illustration of the possible values for $p(y \mid x, z)$ compatible with the instrument inequalities, for a given distribution $p(x|z)$.The darker shaded region satisfies the inequalities: (a) $X = 0$, inequalities (a2); (b) $X = 1$, inequalities (b2). In this example $p(x = 1 \mid z = 0) = 0.84$, $p(x = 1 \mid z = 1) = 0.32$. 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

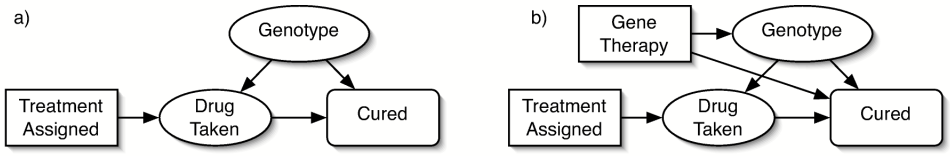


Figure 1. The treatment assignment only cures the patient if it affects whether the drug is taken, but genotype does not have a causal effect unless it is responsive to decisions.

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*}.

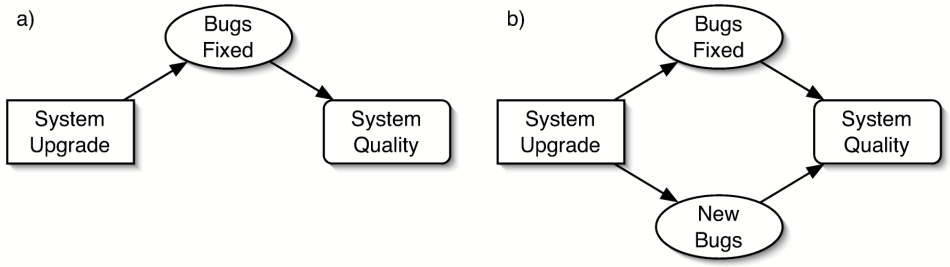


Figure 2. We believe that a system upgrade will affect system quality by fixing bugs unless new bugs are introduced in the process.

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

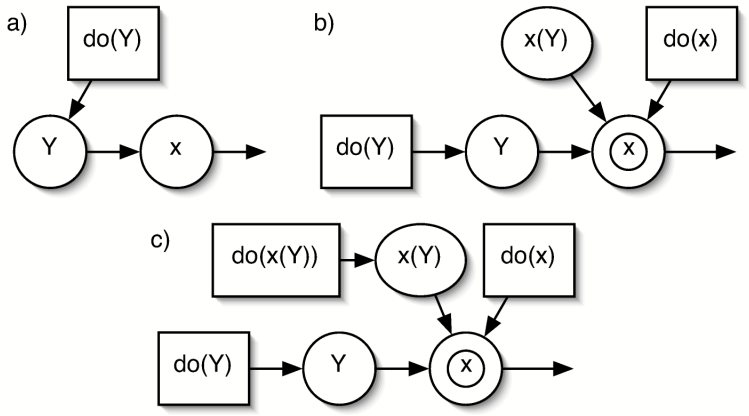


Figure 3. The partial influence diagram for x in a causal model, shown in (a) with parents Y, becomes the diagram shown in (b) explicitly representing the structural equation for x, and, when Y is nonempty, the diagram shown in (c) with an explicit atomic intervention on the mapping variable.

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

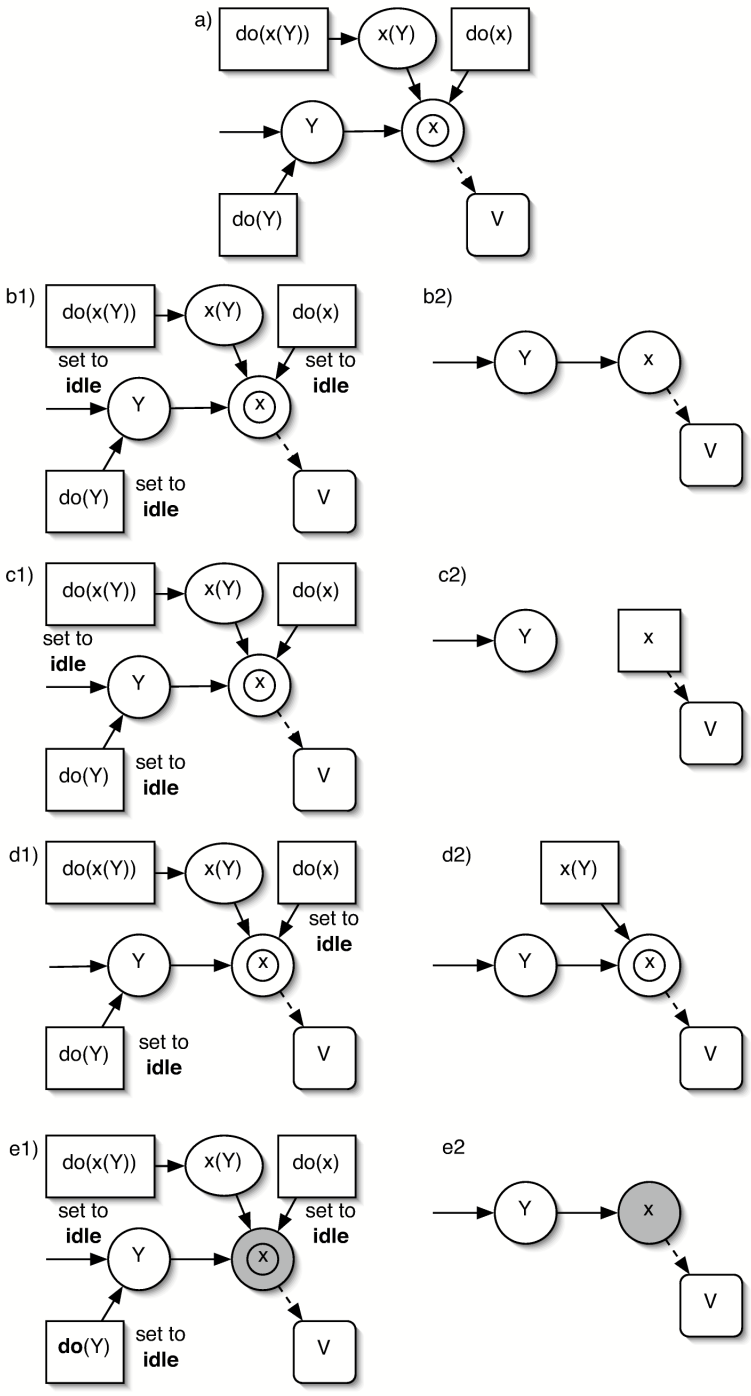


Figure 4. Partial causal influence diagrams to compute the values of do, control, and revelation for x when Y is nonempty.

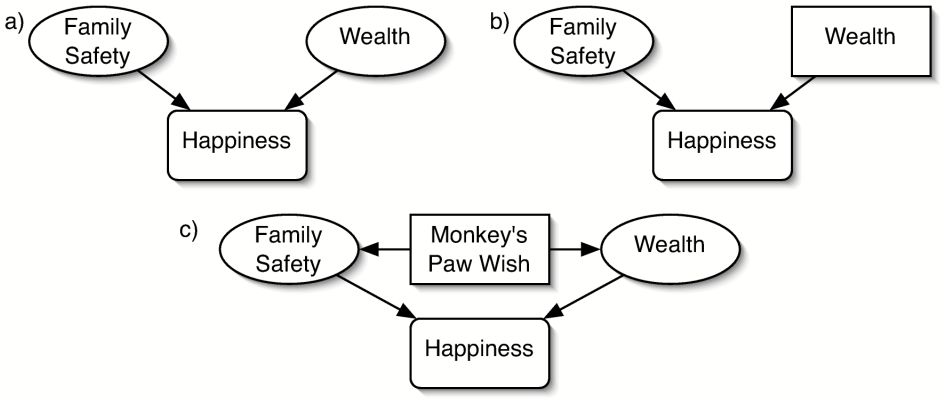


Figure 5. Unless the intervention is direct there can be disastrous side effects.

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

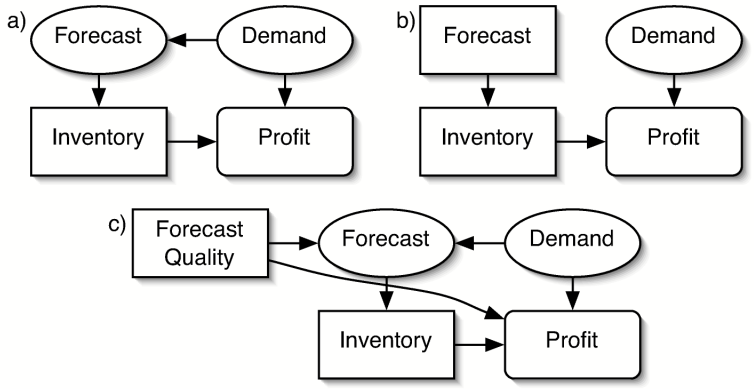


Figure 6. When we intervene on a forecast, we want to improve its quality, rather than to obtain a single most desirable instance.

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*

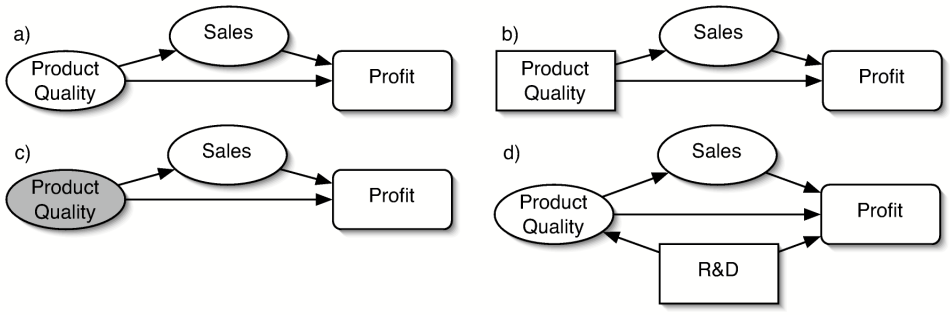


Figure 7. In this causal influence diagram the values of do, control, and revelation for *Product Quality* are equal.

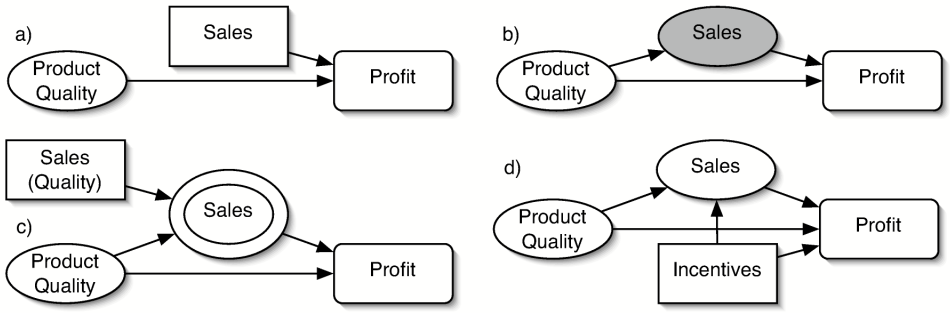


Figure 8. The values of do, control, and revelation for *Sales* might not be equal.

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

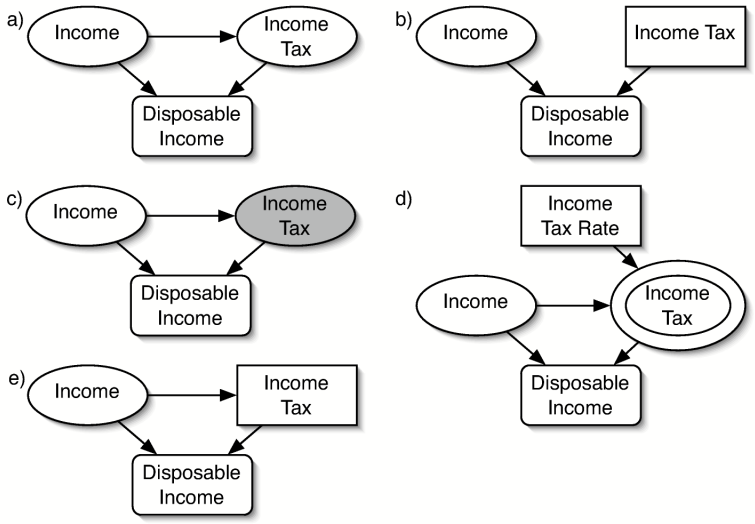


Figure 9. The values of do, control, and revelation are different for *Income Tax*.

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.)

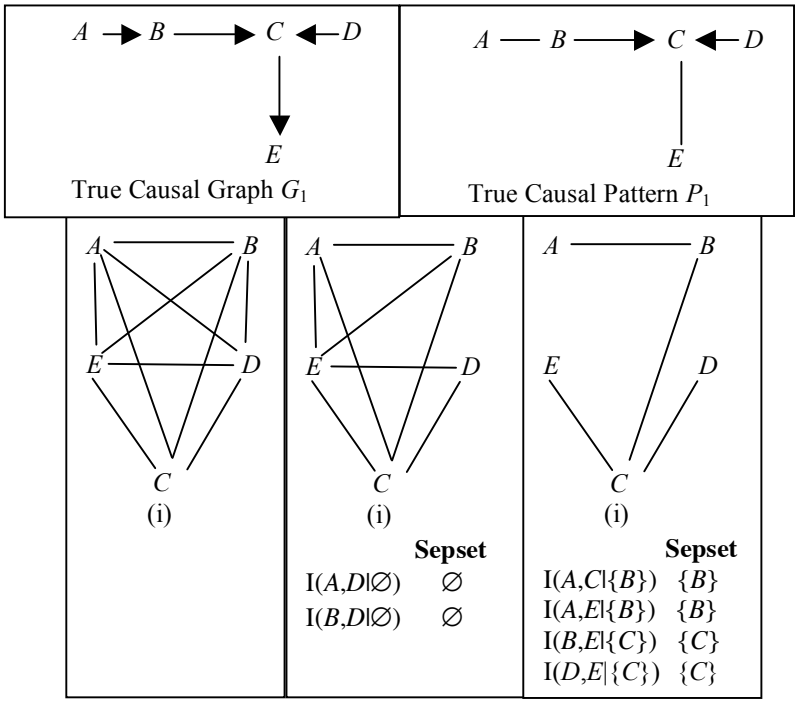


Figure 1: Constraint based search, where correct pattern is P_1

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.

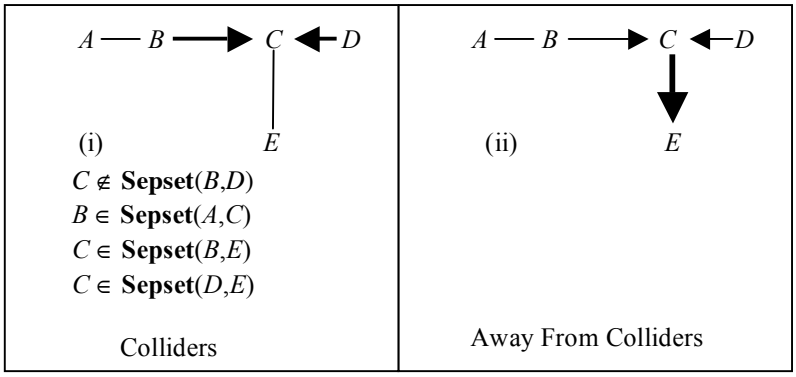


Figure 2: Orientation phase of PC algorithm, assuming true pattern is P_1

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.

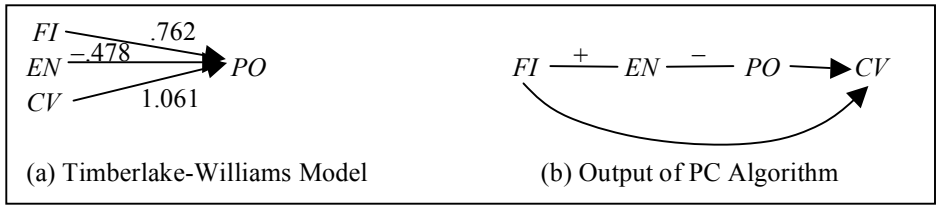


Figure 3: Two Models of Foreign Investment

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.

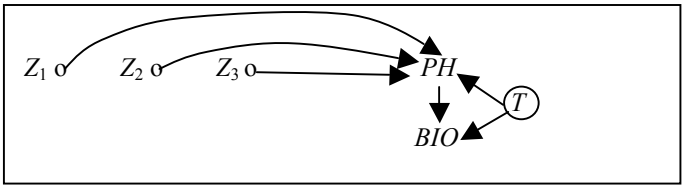


Figure 4 : *PH* and *BIO* Confounding?

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).

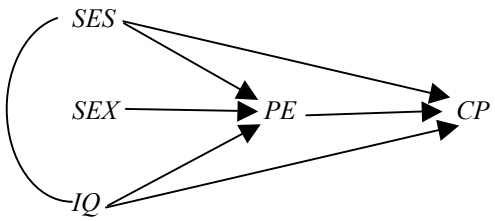


Figure 5: Model of Causes of College Plans

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) $Z = bX + cY + \varepsilon_z$

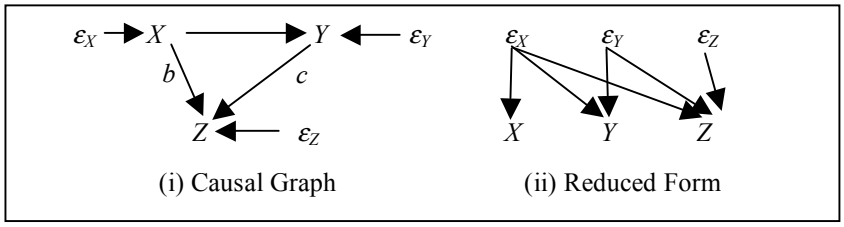


Figure 6: Causal Graph and Reduced Form

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.

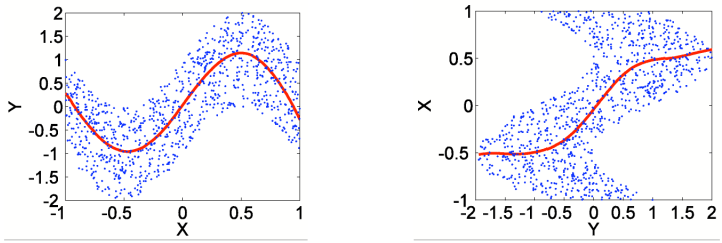
kPC consists of two stages. In the first stage of kPC, the standard PC algorithm is applied to the data using efficient implementations of the Hilbert-Schmidt Independence Criteria [Gretton, Fukumizu, Teo, Song, Scholkopf, & Smola, 2008], a nonparametric independence test and an extension of this test 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.



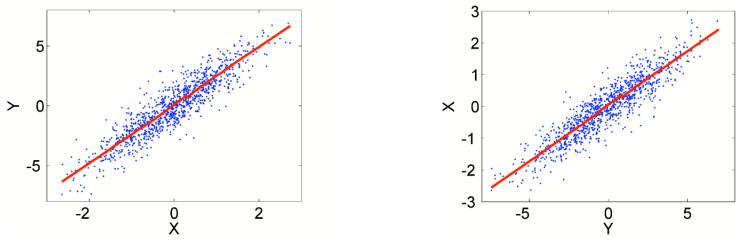
(a) **(b)**

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.

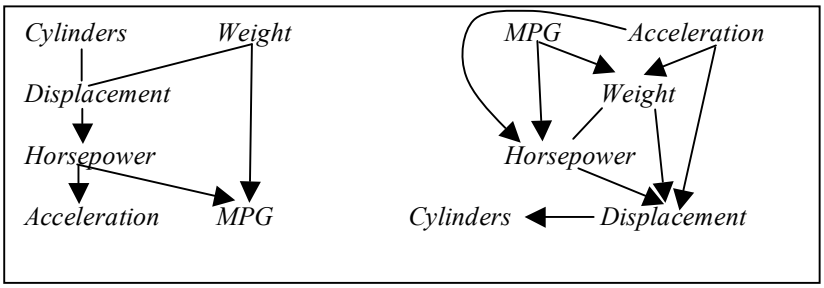


Figure 9: Automobile Models

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.

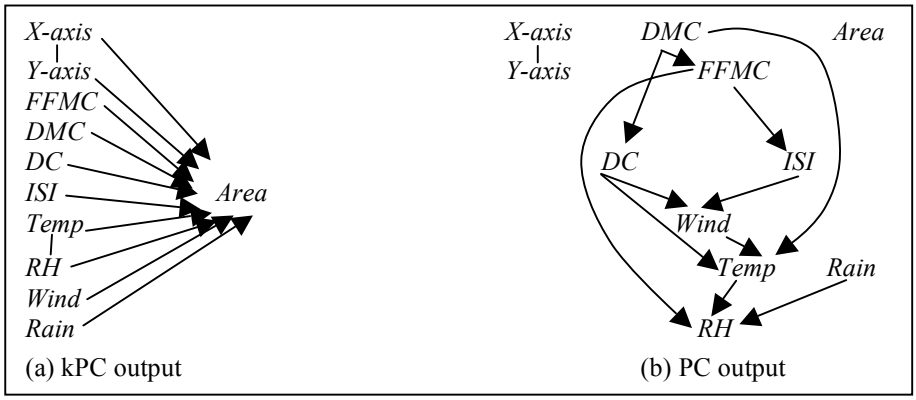


Figure 10: kPC and PC Forest Fires

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.

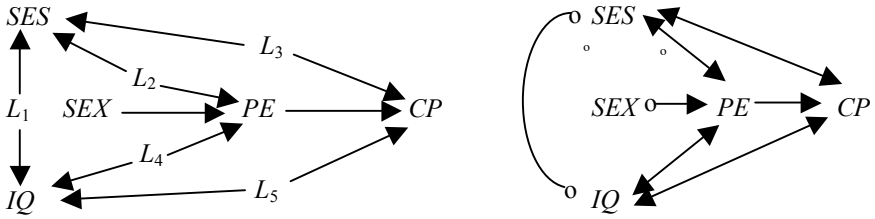


Figure 11: DAG and Partial Ancestral Graph

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.

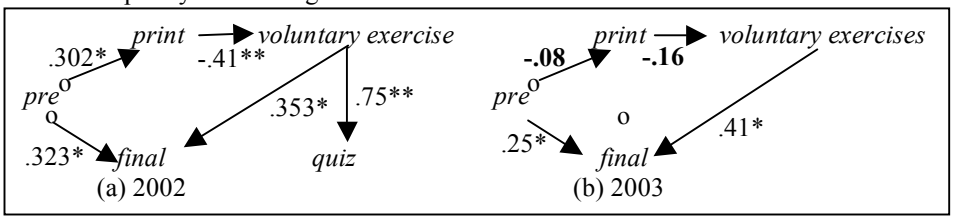


Figure 12: Online Course Printing

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.

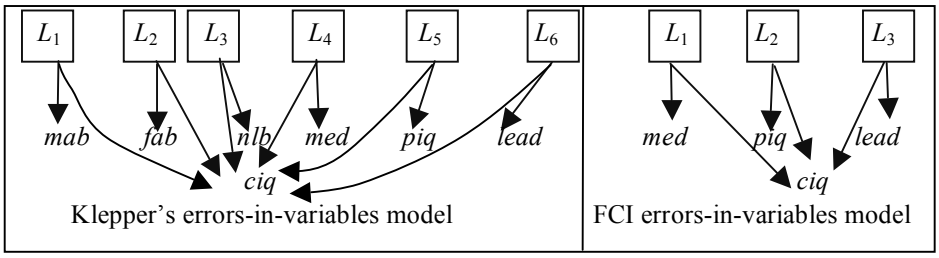


Figure 13: Errors-in-Variables Models

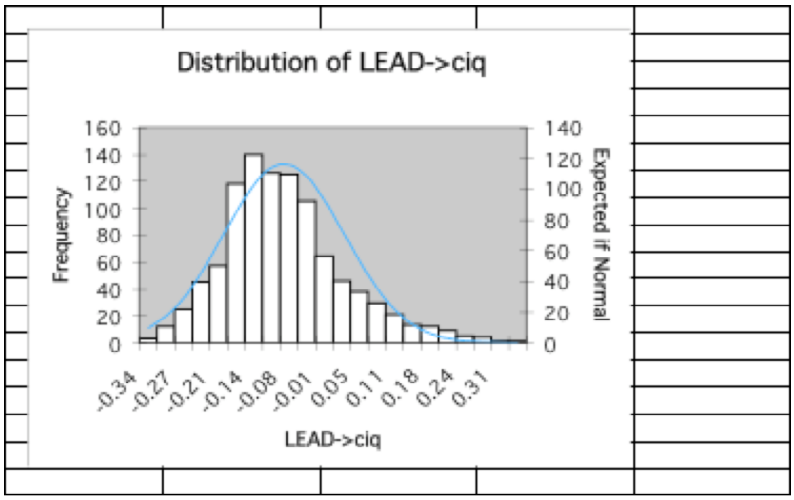


Figure 14: Posterior for Klepper's Model

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.

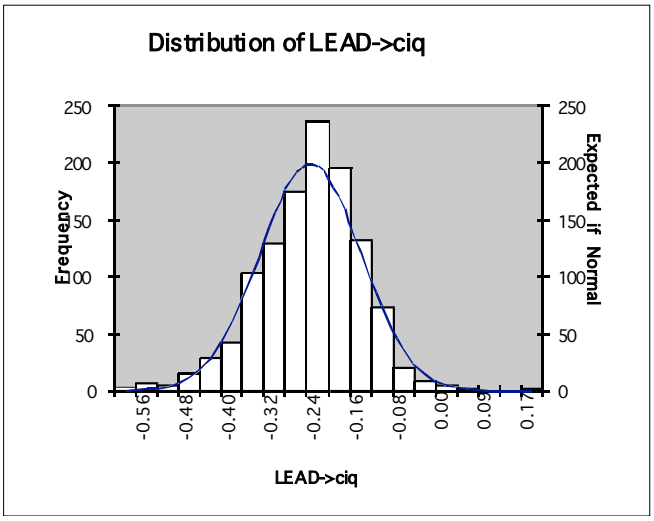


Figure 15: Posterior for FCI model

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.

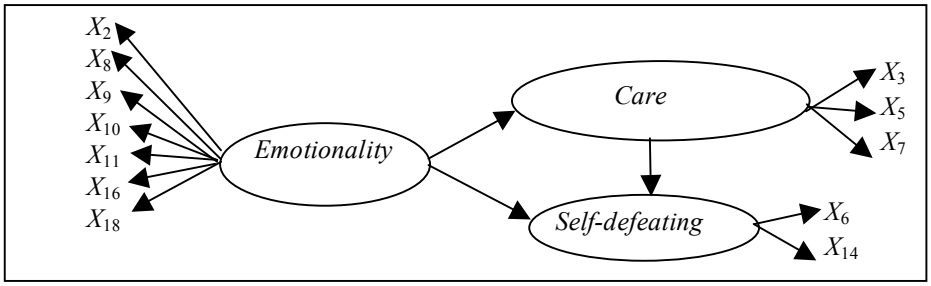


Figure 16: SEM S

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.

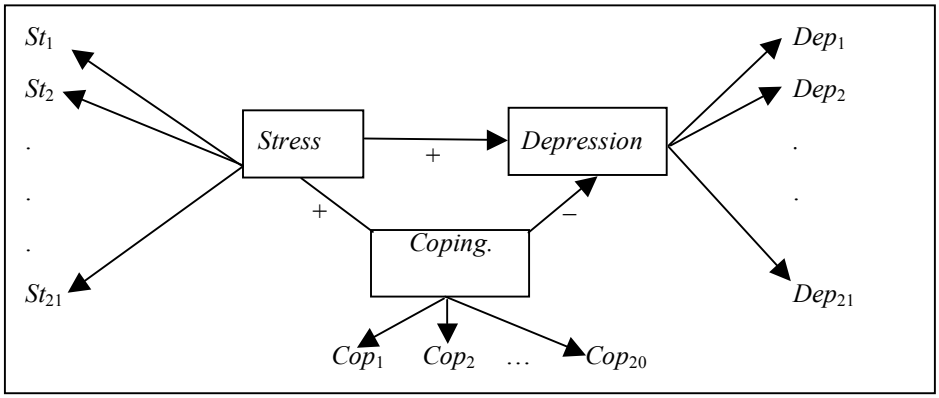


Figure 17: Model from Theory

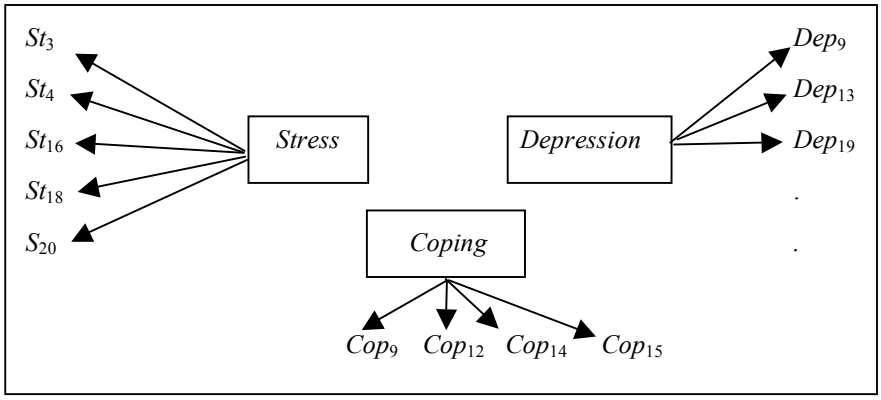


Figure 18: Output of BuildPureClusters

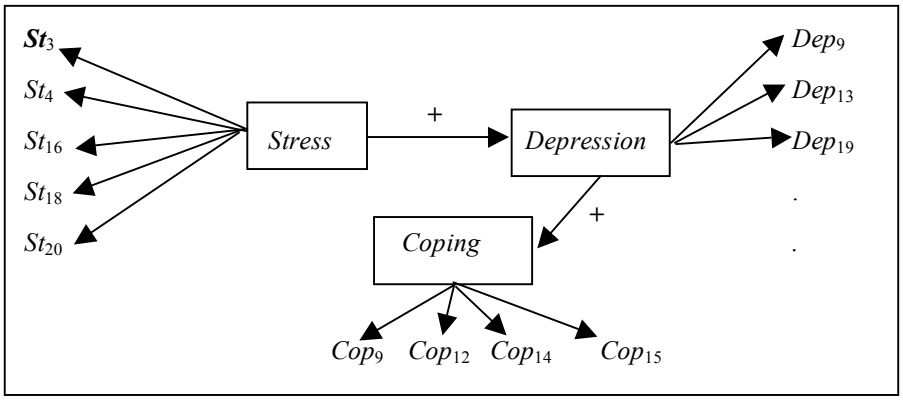


Figure 19: Output of GESMIMBuild

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.

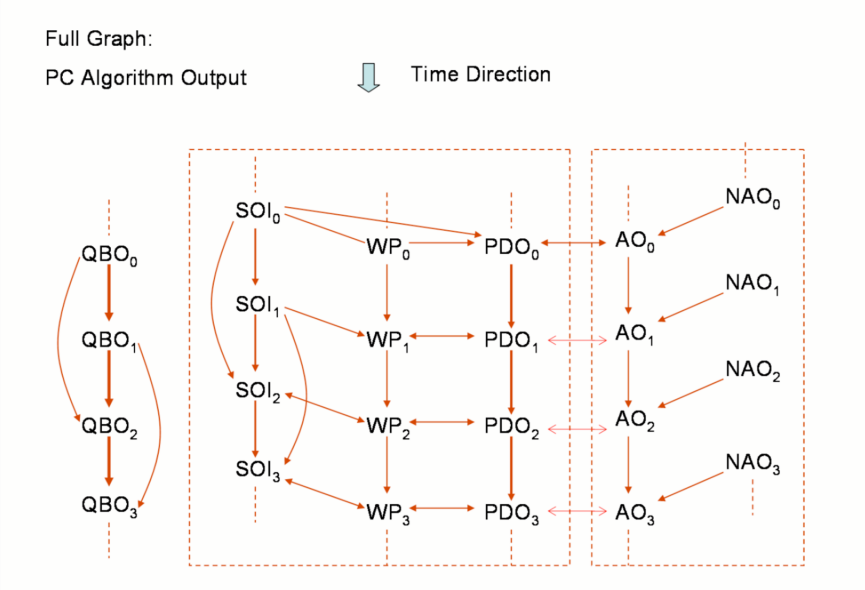


Figure 20: Climate Time Series

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.

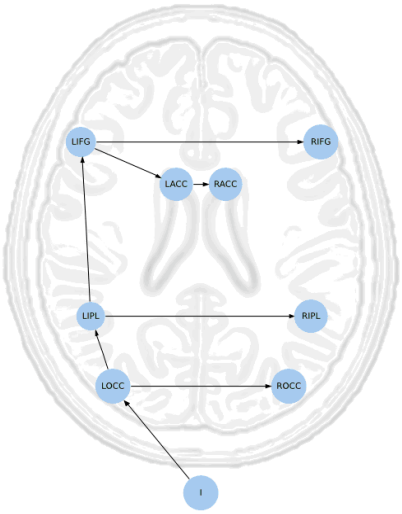


Figure 21: IMaGES Output for fMRI Data

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.

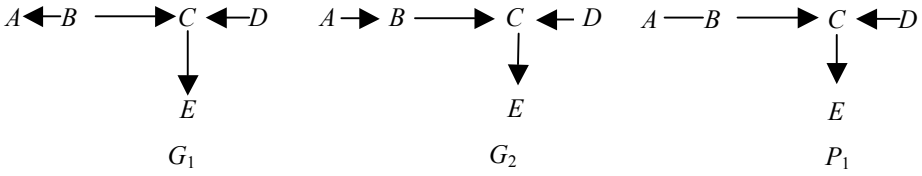


Figure 22: G_1, G_2, **and** P_1 **(pattern for** G_1 **and** G_2**)**

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.

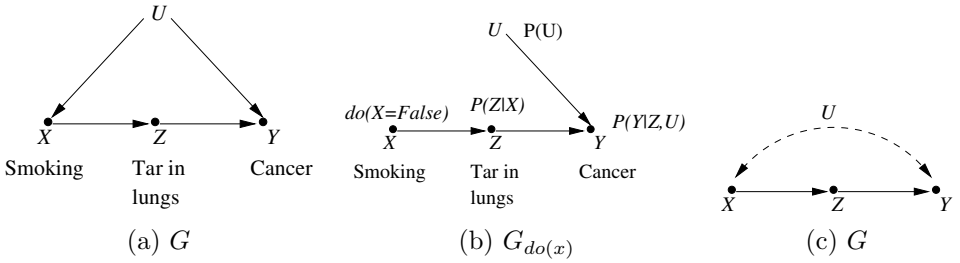
A path not blocked by Z is called *d-connecting* or *active*. A set Z is said to *d-separate* X from Y, denoted by $(X \perp\!\!\!\perp Y|Z)_G$, if and only if Z blocks every path from a node in X to a node in Y.

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

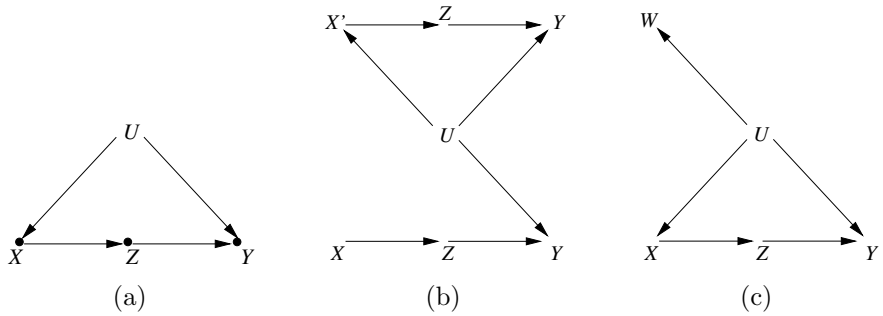


Figure 5. (a) A causal diagram G. (b) The counterfactual graph for $P(Y_x = y|x')$ in G. (c) The graph G' from Theorem 18.

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.