JMLR: Workshop and Conference Proceedings 6: 59–86 NIPS 2008 Workshop on Causality

Beware of the DAG!

A. Philip Dawid APD@STATSLAB.CAM.AC.UK
Statistical Laboratory
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB, UK

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Directed acyclic graph (DAG) models are popular tools for describing causal relationships and for guiding attempts to learn them from data. They appear to supply a means of extracting causal conclusions from probabilistic conditional independence properties inferred from purely observational data. I take a critical look at this enterprise, and suggest that it is in need of more, and more explicit, methodological and philosophical justification than it typically receives. In particular, I argue for the value of a clean separation between formal causal language and intuitive causal assumptions.

Keywords: Directed acyclic graph, conditional independence, probabilistic causality, statistical causality, causal DAG, augmented DAG, Pearlian DAG, causal discovery, causal Markov condition, reification fallacy, instrumental variable

1. Introduction

This article is based on a talk given at the 2008 NIPS Workshop *Causality: Objectives and Assessment*, where I was commissioned to play "devil's advocate" in relation to the enterprise of *causal discovery*, which—as evidenced by the other contributions to the workshop—has become an important and vibrant strand of modern machine learning. Like Cartwright (2007, Chapter II), I take a sceptical attitude to the widespread view that we can learn about causal processes by constructing DAG models of observational data.

In taking on this sceptical rôle I would not wish to be thought entirely negative and destructive: on the contrary, I am impressed by the overall quality of work in this area, be it fundamental methodology, algorithmic development, or scientific application. Moreover, many of the cautions I shall raise have been clearly identified, appreciated and enunciated by the major players in the field, and will be at the back, if not the forefront, of the minds of many of those who use the techniques and algorithms. I do feel, however, that there is still a useful contribution to be made by reiterating and to some extent reframing these cautions. A companion paper (Dawid, 2010) makes similar points, with emphasis on the variety of concepts of causality rolled up in the statistical methodology.

My principal concern here is to clarify and emphasise the strong assumptions that have to be made in order to make progress with causal modelling and causal discovery, and to argue that these should never be accepted glibly or automatically, but deserve careful attention and context-specific discussion and justification whenever the methods are applied. And, in a more positive vein, I describe a formal language that can assist in expressing such assumptions in an unambiguous way, thereby facilitating this process of discussion and justification.

1.1 DAG models

A unifying feature of the discussion is the use of directed acyclic graph (DAG) representations. These can interpreted and applied in a number of very different ways, which I attempt to elucidate and contrast. Here I give a very brief preview of these different interpretations.

Consider for example the simple DAG

(a) $X \leftarrow Z \rightarrow Y$.

One interpretation of this is as a *probabilistic DAG*, which is just a graphical way of describing the probabilistic conditional independence (CI) property $X \perp\!\!\!\perp Y \mid Z$—and is thus interchangeable with the entirely equivalent descriptions of this CI property by means of the DAGs

(b) $X \rightarrow Z \rightarrow Y$; or

(c) $X \leftarrow Z \leftarrow Y$.

A totally different interpretation of **(a)** is as a *causal DAG*, saying that Z is (in some sense) a "common cause" of both X and Y, which are otherwise causally unrelated. Under this causal interpretation the DAGs **(a)**, **(b)** and **(c)** are *not* interchangeable.

Although these interpretations (probabilistic and causal) have absolutely nothing in common, it is often assumed that a single DAG can fulfil both these interpretative functions simultaneously. When this is so, it follows that any variable will be independent of its non-effects, given its direct causes—the *causal Markov* property that forms the basis of "causal discovery" algorithms that attempt to infer causal relationships from observationally discovered probabilistic conditional independencies. Unfortunately there is no clear way of deciding when (if ever) it is appropriate to endow a DAG with this dual interpretation.

Pearlian DAGs aim to clarify this connexion, using interventions to define causal relationships, and making strong assumptions to relate the non-interventional probabilistic regime with various interventional causal regimes. For example, this interpretation of DAG **(a)** would require that the observational joint conditional distribution of (X,Y) given $Z = z$ (under which X and Y are in fact conditionally independent) is the same as the joint distribution of (X,Y) that would ensue when we intervene on Z to set its value to z. Such Pearlian assumptions, which are testable (at least in principle), support a rich causal calculus. There are also valuable variations on this approach that require fewer assumptions (*e.g.*, we envisage intervention at some, but not all, of the nodes in the DAG).

This abundance of different interpretations of the same DAG is rich in possibilities, but at the same time a potential source of confusion.

1.2 Outline

In § 2 I recall the importance of distinguishing between passive observation ("seeing") and intervention ("doing"). Section 3 introduces the algebraic theory of conditional independence (CI), relevant to the "seeing" context, while graphical representations of CI are described and discussed in § 4. In § 5 I switch to considering causal models and their graphical representations, as relevant to the "doing" context, while § 6 discusses possible relationships that might be assumed to hold between graphical representations in the two contexts. Section 7 treats the more specific assumptions underlying Judea Pearl's use of DAGs to represent and manipulate causality. In § 8, I comment on the

strong assumptions that are implicit in these causal models. Section 9 then presents an approach to modelling causality that does not require any such assumptions (though these can be represented when desired), and this is further illustrated, and contrasted with other approaches, in §§ 10 and 11. The need for (and possibilities for) contextual justification of causal assumptions is highlighted in § 12, while § 13 summarises the arguments presented and considers what might be an appropriate rôle for causal discovery.

2. Seeing and doing

Spirtes et al. (2000) and Pearl (2009), among others, have stressed the fundamental importance of distinguishing between the activities of *Seeing* and *Doing*. *Seeing* involves passive observation of a system in its natural state. *Doing*, on the other hand, relates to the behaviour of the system in a disturbed state, typically brought about by some external intervention. For statistical applications a strong case can be made (Dawid, 2000, 2002b) for regarding the philosophically problematic concept of *causation* as simply describing how the system responds to external intervention—a stripped-down "agency" or "manipulationist" interpretation of causality (Hausman, 1998; Woodward, 2003). *Causal inference* then refers to the problem of drawing conclusions, from available data, about such responses to interventions.

The cleanest case is when the data were collected under the very interventional regime in which we are interested. "To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)" (Box, 1966). This is the credo underlying the whole discipline of experimental design and inference, as exemplified by the most important medical advance of the 20th century: the controlled clinical trial.

Often, however, for reasons of cost, practicality, ethics, *etc.*, we can not experiment, but are confined to passive observation of the undisturbed system. Now it is a logically trivial but fundamentally important point that there is no necessary connexion between the different regimes of seeing and doing: a system may very well behave entirely differently when it is kicked than when it is left alone. So any understanding one might achieve by observation of the system's undisturbed behaviour is at best indirectly relevant to its disturbed behaviour, and thus to causal inference. We might attempt to proceed by *assuming* connexions between the different regimes, which—if valid—would allow us to transfer knowledge gained from *seeing* to inferences about the effects of *doing*. But it is important to be entirely explicit about such assumptions; to attempt, so far as is possible, to justify them; and to be fully aware of the sensitivity of any conclusions drawn to their validity.

In recent years there has grown up a body of methodology, broadly described as *causal discovery*, that purports to extract causal (doing) conclusions from observational (seeing) data in fairly automatic fashion (Spirtes et al., 2000; Glymour and Cooper, 1999; Neapolitan, 2003). This approach largely revolves around directed acyclic graph (DAG) models, which have interpretations in both the seeing and the doing contexts, so that a DAG model identified from observational (seeing) data can be imbued with causal (doing) content. However, these two interpretations of DAGs, while related, are logically distinct, and have no necessary connexion. Hence it is important to clearly identify, understand, and provide contextual justification for, the assumptions that are needed to support replacement of one interpretation by another. There can be nothing fully automatic about causal discovery.

I will survey various different interpretations of DAG models, and their relationships with conditional independence.

3. Seeing: Conditional independence

We start by concentrating on the behaviour, under a single stable regime, of a collection of variables of interest. We assume that this behaviour will be modelled by means of a fixed joint probability distribution P.[1] If we can obtain and record repeated observations under the same regime, we might hope to estimate P. Here we largely ignore problems of inference, and restrict attention to purely probabilistic properties.

One of the most important of such properties is that of *conditional independence*, CI (Dawid, 1979a, 1980). We write $X \perp\!\!\!\perp Y \mid Z\ [P]$ to denote that, under the distribution P, variables X and Y are probabilistically independent given $Z = z$, for any observable value z of Z. When P can be understood we write simply $X \perp\!\!\!\perp Y \mid Z$. This can be interpreted in various equivalent ways, but for our purposes the most useful is the following:[2]

$$P(X = x \mid Y = y, Z = z) \text{ depends only on } z, \text{ and not further on } y. \tag{1}$$

Universal[3] qualitative properties of probabilistic CI include (Dawid, 1979a; Spohn, 1980; Pearl and Paz, 1986):

$$\begin{aligned}
&X \perp\!\!\!\perp Y \mid X \\
&X \perp\!\!\!\perp Y \mid Z && \Rightarrow \quad Y \perp\!\!\!\perp X \mid Z \\
&X \perp\!\!\!\perp Y \mid Z, \quad W \le Y && \Rightarrow \quad X \perp\!\!\!\perp W \mid Z \\
&X \perp\!\!\!\perp Y \mid Z, \quad W \le Y && \Rightarrow \quad X \perp\!\!\!\perp Y \mid (W,Z) \\
&\left.\begin{array}{l} X \perp\!\!\!\perp Y \mid Z \\ \text{and} \\ X \perp\!\!\!\perp W \mid (Y,Z) \end{array}\right\} && \Rightarrow \quad X \perp\!\!\!\perp (Y,W) \mid Z
\end{aligned} \tag{2}$$

(where $W \le Y$ denotes that W is a function of Y).

There is another useful property, which is however valid not universally, but only under additional conditions (Dawid, 1979b, 1980):[4]

$$X \perp\!\!\!\perp Y \mid (Z,W) \text{ and } X \perp\!\!\!\perp Z \mid (Y,W) \Rightarrow X \perp\!\!\!\perp (Y,Z) \mid W. \tag{3}$$

While (2) (and, where appropriate, (3)) do not exhaust all the general properties of probabilistic CI (Studený, 1992), they are adequate for most statistical purposes.

4. Graphical representation

It can be helpful to use mathematical constructions of various kinds to represent and manipulate CI (Dawid, 2001a). This involves making formal analogies between properties of probabilistic

1. There are of course many interpretations of probability (Galavotti, 2005). For present purposes a naïve frequentist view, which can also be given a subjective Bayesian intepretation in terms of exchangeability (de Finetti, 1975), will suffice. Williamson (2005) argues for an "objective Bayesian" interpretation as most appropriate for causal inference. The formal mathematical framework is the same in all cases.
2. Purely for simplicity, we may here suppose the variables are discrete, and all combinations of logically possible values have positive probability. For a rigorous definition in the general case, see Dawid (1980).
3. *i.e.* holding for any distribution P and any variables $X, Y, \ldots$
4. For example, when the sample space is discrete and each elementary outcome has positive probability.

CI and non-probabilistic properties of the representations we use. The representations themselves can look very different from probability distributions, and we need to be very clear as to how we are to interpret properties of such a representation as "saying something about" properties of CI. As with any use of representations to assist understanding and construct arguments, the *semantics* (or *meaning*) of a representation—describing exactly just how it is to be taken as relating to the external "reality" it is intended to represent—is at least as important as its *syntax*—describing its internal grammar.

One of the most popular and useful of such representations is the *directed acyclic graph* (DAG). A DAG $\mathscr{D}$ has a set $\mathscr{V}$ of nodes, and arrows joining them, with no loops or directed cycles. A full description and analysis of the formal semantics of the relationship between DAGs and the collections of CI properties they represent, together with the associated notation and terminology, can be found in Cowell et al. (2007). Although this theory will be familiar to many readers, I repeat here the specific features I wish to emphasise—more to clarify what is *not* being said than what is.

4.1 *d*-separation

Given node-sets $S,T,U \subseteq \mathscr{V}$, we say U *d-separates* S *from* T *in* $\mathscr{D}$, and write $S \perp_d T \mid U\ [\mathscr{D}]$, if the following somewhat complex geometric property[5] is satisfied. First we delete all nodes that are not "ancestors" of some node in $S \cup T \cup U$, as well as all their incoming arrows; then we add undirected edges between any two nodes that are "parents" of a common "child" node, if they are not already joined by an arrow; next we delete all arrowheads, so obtaining an undirected graph, the relevant *moralized ancestral graph*. Finally, in this graph we look for paths joining S and T that do not intersect U. If there are none such, then S and T are d-separated by U in $\mathscr{D}$.

It turns out (Lauritzen et al., 1990) that this graph-theoretic separation property also obeys the formal rules (2) (with $\leq$ interpreted as $\subseteq$), and is thus potentially able to represent some collections of probabilistic conditional independence properties. Specifically, when the nodes of $\mathscr{D}$ represent random variables, we say that $\mathscr{D}$ *represents* a collection $\mathscr{C}$ of conditional independence relations between sets of variables if the graph-theoretic property $S \perp_d T \mid U\ [\mathscr{D}]$ holds exactly when the CI relation $S \perp\!\!\!\perp T \mid U$ either belongs to $\mathscr{C}$, or can be logically deduced from $\mathscr{C}$ by application of the rules in (2). For a probability distribution P over $\mathscr{V}$, we say $\mathscr{D}$ *represents* P if (the *Markov condition)*:

$$S \perp_d T \mid U\ [\mathscr{D}] \Rightarrow S \perp\!\!\!\perp T \mid U\ [P]. \tag{4}$$

This will be so if and only if, under P, for each $V \in \mathscr{V}$, V is conditionally independent of its parents in $\mathscr{V}$, $\mathrm{pa}(V)$, given its non-descendants in $\mathscr{V}$, $\mathrm{nd}(V)$. Such a representation is termed (probabilistically) *faithful* when the converse implication to (4) also holds, *i.e.* the *only* conditional independence properties holding in P between the variables in $\mathscr{V}$ are those represented by $\mathscr{D}$. These relationships between the d-separation properties of a DAG and a collection of CI properties, or a joint distribution P, constitute the *semantic interpretation* of the DAG.

As a simple example, Figure 1 shows the unique DAG over four variables (Z,U,X,Y) that represents the following pair of CI properties:

$$U \perp\!\!\!\perp Z \tag{5}$$

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

5. We here describe the "moralisation" version of this property (Lauritzen et al., 1990). This is logically equivalent to the d-separation property as described by Pearl (1986); Verma and Pearl (1990).

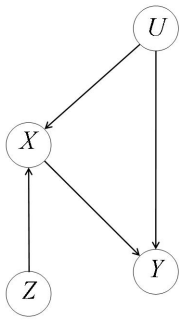

Figure 1: Simple DAG

It is important to note that, for given variable set $\mathscr{V}$, the collections of CI properties $\mathscr{C}$ that can be represented by a DAG are very special.[6] Thus with $\mathscr{V} = \{X,Y,Z\}$, the pair of properties $\{X \perp\!\!\!\perp Y,\ X \perp\!\!\!\perp Y \mid Z\}$ has no DAG representation, and this is indeed the typical state of affairs. Conversely, when a DAG representation is available, it need not be unique. Distinct DAGs on $\mathscr{V}$ are termed *Markov equivalent* when they represent the same collection of CI relations: this will be so if and only if they have the same *skeleton* (undirected version) and *immoralities* (configurations of the form $A \to C \leftarrow B$ with no arrow between A and B) (Frydenberg, 1990; Verma and Pearl, 1991). Thus the three DAGs **(a)**, **(b)** and **(c)** of § 1.1 are Markov equivalent, all representing the same single CI property $X \perp\!\!\!\perp Y \mid Z$, and are all equally valid for this purpose. This representational flexibility is extended further when we allow the set $\mathscr{V}$ of variables considered to vary: thus both DAGs of Figure 2 represent the single CI property $A \perp\!\!\!\perp B$.

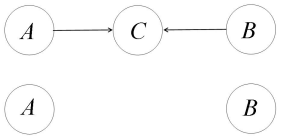

Figure 2: Two DAGs representing $A \perp\!\!\!\perp B$

4.2 What do the arrows mean?

According to the theory presented above, the purpose of a DAG representation is to mirror, *via* the d-separation semantics described in § 4, the probabilistic relationship of conditional independence—a relationship that, it is worth emphasising, is entirely symmetrical, as captured by the second line of (2). However, it is in the very nature, and indeed name, of a directed acyclic graph that it contains *directed* arrows between variables, so that this particular graphical representation embodies a non-symmetrical relationship between nodes. But this is a pure artifact: thus Figure 1, although

6. They are exactly those that are logically equivalent (using (2)) to a collection of the form $V_i \perp\!\!\!\perp \{V_1,\ldots,V_{i-1}\} \mid S_i$ for $i = 1,\ldots,N$, where $V_1,\ldots,V_N$ is an ordering of $\mathscr{V}$, and $S_i \subseteq \{V_1,\ldots,V_{i-1}\}$. In this case the associated DAG has an arrow into each V_i from each node in S_i.

composed of directed arrows, is nothing but an alternative way of representing the symmetrical CI relationships (5) and (6). The rôle of an arrow in a DAG model is much like that of a construction line in an architect's drawing: although it plays an important rôle in the formal syntax of the model, it has no direct counterpart in the world, and contributes only indirectly to the semantic interpretation of the model.

4.3 Reification

Nevertheless, having built a DAG representation of a probability distribution, it is hard to resist the temptation to interpret an arrow from node X to node Y in the DAG as representing something meaningful in the real-world system that the DAG is modelling—for example, as embodying some conception of the non-symmetrical relation of *cause and effect*: that X is, in some sense, a "direct cause" of Y. Likewise, we might be tempted to read off from the Figure 1 such intuitive properties as "X lies on the causal pathway between Z and Y". But no such inferences are justified from the formal semantics relating DAG representations to conditional independence. Such interpretation of an incidental formal attribute of a mathematical representation of the world as corresponding to something real in the external (physical or mental) world[7] may be termed "reification". While reification can often be indicative and fruitful, it is important to be very clear as to when we are reaching beyond the formal semantics by which the representation has been supposed to encode real-world properties, and in that case to consider very carefully whether, when and how this might be justifiable.

5. Causal DAGs

An entirely different use of a DAG representation is to model causal relations directly. Unlike conditional independence, which is a clearly defined property of a probability distribution, causality is a slippery and ambiguous concept. In dealing with causal relations, we can either regard them as fundamental undefined primitives in themselves, or as defined in terms of still more basic ingredients, such as the effect of interventions. In either case the important thing, if a representation is to be used for meaningful communication, is that all parties have the same (explicit or implicit) understanding of the things it is supposed to be representing, and of the nature and mechanics of the representation.

A common causal interpretation of a DAG is along the following lines, quoted from Hernán and Robins (2006) (their Figure 2 is redrawn here as our Figure 3), in discussion of a certain problem relating to the use of "instrumental variables" (see § 10 below):

> "A `causal DAG` is a DAG in which:
>
> (i). the lack of an arrow from V_j to V_m can be interpreted as the absence of a `direct causal effect` of V_j on V_m (`relative to` the other variables on the graph)[8]
>
> (ii). all `common causes`, even if unmeasured, of any pair of variables on the graph are themselves on the graph. In Figure 2... the inclusion of the measured variables (Z, X, Y) implies that the causal DAG must also include their unmeasured common causes (U, U^*)."

7. Bourdieu (1977, *p.* 29) speaks of "sliding from the model of reality to the reality of the model"
8. A stronger and potentially more useful requirement is that an arrow be present *if and only if* there is such a direct causal effect.

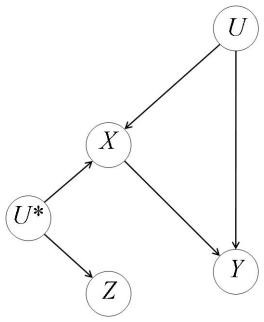

Figure 3: Figure 2 of Hernán and Robins (2006)

Here we have used `teletype font` (not in the original) to highlight non-mathematical causal concepts.[9] Only when we have pre-existing understanding and interpretation of these concepts will it be possible to say whether or not a given DAG is indeed a *causal* DAG for the problem it is intended to represent. Such a question can never be addressed solely in formal terms, by reference only to the DAG. In particular, we can not define concepts such as "direct causal effect" or "common cause" by reference to a putative DAG model unless that model has previously been justified as "causal" by other, necessarily non-graphical, considerations not involving these terms.

However we may choose to understand the causal terms involved, it is clear that the semantics whereby a DAG represents causal properties are qualitatively totally different from those whereby it represents conditional independence properties. In the former case, the arrows are supposed to have a direct interpretation in terms of cause and effect; whereas, as emphasised in § 4.2, for conditional independence the arrows are nothing but incidental construction features supporting the d-separation semantics. A related distinction is that conditional independence is an externally determined all-or-nothing affair, whose validity is unaffected by which other variables and properties we may choose to represent in our DAG: this means, in particular, we can not interpret the presence or absence of an arrow from X to Y in some probabilistic DAG representation as representing a fundamental CI property, since such arrows can come and go as we vary the set of variables represented. In contrast, the very meaning of the properties (such as "direct effect") represented by a causal DAG may be dependent on the specification of the variable set $\mathscr{V}$, and may change (with corresponding changes in the relevant representation) as we vary $\mathscr{V}$. Correspondingly an arrow in a causal DAG *can* be considered as having independent meaning (*e.g.* as representing a "direct effect")—albeit only in terms of causal concepts defined *relative* to the specific variables represented.

Contrasting conditional independence and causality, we see that both in their subject matter and in their graphical representation they differ markedly. We could simply keep them in entirely different pockets, with nothing whatsoever to do with each other. However, there is a long-standing tradition of attempting to forge connexions between the two. Indeed, some thinkers (Shafer, 1996; Spohn, 2001) regard the very concept of causality as entirely supervenient on that of conditional independence (for an appropriate collection of variables).

9. Note that (i) involves a causal concept that is not regarded as absolute, but rather as relative to a specific collection of variables under consideration.

6. Probabilistic causality

One approach, going back at least to Reichenbach (1956) and pursued by Suppes (1970) among others, essentially proceeds by relating the *probabilistic conditional independence* of two variables X and Y given some third variable Z to the *causal independence* of X and Y, in the sense that neither of these variables causally affects the other. However it is not easy to make this precise. In one direction, it apparently implies that, if X and Y are completely independent probabilistically (so that we can take Z to be vacuous), then neither can causally affect the other. However, this degree of implication is not usually claimed, since such independence could be an accidental result of numerical cancellation between two or more non-null causal probabilistic relationships involving these and other variables.[10] Likewise, if we find that X and Y are not independent given any variable Z currently under consideration, we can not immediately deduce causal dependence between X and Y, since we can not rule out the possibility that we have simply not examined enough Zs.

In the converse direction, it might be claimed (the "weak causal Markov assumption", Scheines and Spirtes, 2008) that, if X and Y are "causally disconnected", in the sense that neither X nor Y causally affects the other and they have no other common cause Z, then they should be probabilistically independent.

6.1 Causal Markov condition

A still more thoroughgoing approach is based on the "Causal Markov condition", CMC (Spohn, 1980; Spirtes et al., 2000). Essentially, this supposes that, when we do have a causal DAG representation[11] of a system, the identical DAG will also represent its CI properties. Equivalently,[12] CMC requires that any variable be probabilistically independent of its non-effects,[13] conditional on its direct causes—all understood relative to the set of variables in the causal DAG. When valid, CMC allows us to infer conditional independence properties from causal assumptions (so long as these can be represented by a causal DAG).

6.2 Other interpretations of probabilistic causality

Recently other ideas have been suggested for relating causal relationships between variables to properties of their joint probability distribution. For example, Janzing and Schölkopf (2008b,a) distinguish between the two decompositions of a joint distribution, $p(x,y) = p(x)\,p(y\,|\,x)$ and $p(x,y) = p(y)\,p(x\,|\,y)$, in terms of their algorithmic complexity: if, say, the former is simpler by this criterion, one might regard this as indicating a causal effect of X on Y. Similarly (Zhang and Hyvärinen, 2010), if one can reasonably describe $p(y\,|\,x)$, but not $p(x\,|\,y)$, in terms of an implicit additive error structure, one might again interpret that as implying that X is a cause of Y. It is clear that the assumptions underlying such claims are very different from those of "probabilistic causality" above. The extent to which they might be appropriate, and indeed whether they even relate to the same conception of causality, deserves deeper attention.

10. Such a state of affairs is sometimes dismissed as being due to a "non-faithful" DAG representation of the problem. But at this level of generality we do not have a DAG.
11. *e.g.*, as described in §5.
12. At any rate, with the stronger interpretation of footnote 8.
13. The "effect" relation here is the transitive closure of the "direct effect" relation.

6.3 Causal discovery

"Causal discovery" aims to deduce causal properties of a system from its CI properties, themselves typically inferred (with a consequent degree of uncertainty) from an analysis of data generated by the system. There are many variations and algorithms, but all share the same basic philosophy. The fundamental assumptions[14] needed to validate this enterprise in any particular application are:

Assumption 6.1 (Causal representation) *There exists* some *DAG $\mathscr{D}$ that is a causal DAG representation of the system.*

Assumption 6.2 (Causal Markov condition) *The identical DAG $\mathscr{D}$ also represents (by means of the Markov condition (4)) the probabilistic conditional independence properties of the system.*

The more sophisticated causal discovery methods appreciate that (especially in the light of (ii) of § 5) it would not generally be reasonable to expect the causal DAG $\mathscr{D}$ to involve only the variables that happen to have been measured, and so will allow for the inclusion of additional unobserved variables.

Some putative causal DAG representations might be eliminated directly on *a priori* grounds, *e.g.* taking into account temporal order. Under Assumption 6.2, any remaining putative causal DAG representation will have implications for the probability distribution over the observed variables—either directly in terms of conditional independencies when there are no unobserved variables in the causal DAG, or more subtle consequences of "latent conditional independence" when there are. Consequently, if those implications are not supported by the data, then the hypothesised causal DAG may be eliminated on empirical grounds. In this way, and under the assumptions made, we can gain partial knowledge of causal structure from a combination of *a priori* reasoning and observational data.

To make further progress, it is common to strengthen Assumption 6.2 as follows:

Assumption 6.3 (Causal faithfulness) *The causal DAG $\mathscr{D}$ is a probabilistically faithful representation of the system.*

In this case, the *only* conditional independence properties enjoyed by the variables in $\mathscr{D}$ will be those represented by the causal DAG $\mathscr{D}$. Under Assumption 6.1 and Assumption 6.3, knowledge of which (patent or latent) conditional independencies do or do not hold between the observed variables allows us to eliminate still more putative causal DAG representations of the problem. However, even if we knew which variables were to be included in the causal DAG $\mathscr{D}$, and all the conditional independence properties they possess, we might not be able to identify $\mathscr{D}$ uniquely, since we could never distinguish observationally[15] between distinct DAGs that are Markov (but not causal) equivalent. Even with this remaining ambiguity, it may be possible to make some causal inferences: thus if every uneliminated causal DAG description of the problem involves the same variable set $\mathscr{V}$, and all contain an arrow from X to Y, we can infer that X is a direct cause of Y relative to $\mathscr{V}$.

Zhang and Spirtes (2008) point out that certain implications of the combination of Assumptions 6.1 and 6.3 can be tested empirically. For it follows from Assumption 6.3 that the conditional

14. There are also variations using different graphical representations of causal and CI properties, such as partial ancestral graphs (Richardson and Spirtes, 2002; Zhang, 2008).
15. although we may be able to do so on *a priori* grounds

independence properties of the observational distribution P are faithfully represented by *some* DAG, and this property has testable consequences. But this does not make much progress towards *causal* inference without the additional strong Assumption 6.1.

7. Pearlian DAGs

Judea Pearl, through his book (Pearl, 2009) and many other works, has popularised a particular use of DAGs to represent both CI and causal properties simultaneously—the latter understood as describing the effects of interventions. We shall refer to a DAG imbued with such an interpretation as a *Pearlian DAG*.[16]

Such a representation applies to a collection of variables measured on some system, such that we can intervene (or at least can conceive of the possibility of intervening) on any one variable or collection of variables, so as to "set" the value(s) of the associated variable(s) in a way that is determined entirely externally. This gives rise to a wide variety of interventional regimes, while the observational regime arises as the special case that no variables are set. A DAG $\mathscr{D}$ is then a Pearlian representation of the system when the following properties hold:

Property 7.1 (Locality) *Under regimes in which all variables other than V are set, at arbitrary values, the associated distribution of V depends only on the settings of its parents,* $\mathrm{pa}(V)$*, in $\mathscr{D}$.*

This can be interpreted as requiring that only the DAG parents of V have a direct effect on V, relative to the other variables in the DAG.

Property 7.2 (CMC) *Under any regime, $\mathscr{D}$ represents (by means of the Markov condition (4)) the probabilistic conditional independence properties of the associated joint distribution.*

Under any interventional regime that sets the value of $V \in \mathscr{V}$, there trivially can be no dependence of the (one-point) distribution of V on $\mathrm{pa}(V)$: the arrows into V could thus be removed while retaining a DAG representation of this regime.

Under Property 7.1, $\mathscr{D}$ can plausibly be interpreted as a causal DAG representation of the problem: property (i) of §5 is incorporated in Property 7.1, while property (ii), though not directly interpreted or represented, might be regarded as implicit in Property 7.2. With this interpretation, Property 7.2, applied to the observational regime, implies the causal Markov condition.

However, a Pearlian DAG representation must also satisfy an additional "modularity" (or "invariance") condition:

Property 7.3 (Modularity) *For any node $V \in \mathscr{V}$, its conditional distribution, given its DAG parents* $\mathrm{pa}(V)$*, is the same, no matter which variables in the system (other than V itself) are intervened on.*

16. We in fact shall deal only with Pearl's initial, fully stochastic, theory. More recently (see the second-half of Pearl (2009), starting with Chapter 7), he has moved to an interpretation of DAG models based on deterministic functional relationships, with stochasticity deriving solely from unobserved exogenous variables. That interpretation does however imply all the properties of the stochastic theory, and can be regarded as a alternative description of it. (This is however not so when we move from DAG models to more general representations, when such deterministic models have restricted generality: see Example 11.2 below.)

(Note that Property 7.1 follows from Property 7.2 combined with Property 7.3).

Property 7.3 extends CMC: not only can we relate the *qualitative* conditional independence properties and causal properties represented by $\mathscr{D}$ (as embodied in CMC), but we can further relate the various *quantitative* distributional behaviours of the system when subjected to different interventions. In particular, from purely observational data on $\mathscr{V}$ we could estimate the modular parent-child distributions, and piece these together to deduce the joint distribution for the system under any set of interventions: a fully quantitative solution to the problem of inferring causality from observational data.

We see that a Pearlian DAG representation embodies CMC (for its particular interpretation of causality), but much more besides. When we can assume that a system is represented by some Pearlian DAG, we can attempt "quantitative causal discovery", in which we attempt to learn the quantitative as well as the qualitative causal structure of the problem, as embodied in the underlying Pearlian DAG.

8. How do we get started?

As is brilliantly attested by the work of Pearl, an extensive and fruitful theory of causality can be erected upon the foundation of a Pearlian DAG. So, when we can assume that a certain DAG is indeed a Pearlian DAG representation of a system, we can apply that theory to further our causal understanding of the system. But this leaves entirely untouched the vital questions: when is a Pearlian DAG representation of a system appropriate at all?; and, when it is, when can a specific DAG $\mathscr{D}$ be regarded as filling this rôle? As we have seen, Pearlian representability requires many strong relationships to hold between the behaviours of the system under various kinds of interventions.

Causal discovery algorithms, as described in § 6.3, similarly rely on strong assumptions, such as Assumption 6.1 and Assumption 6.2, about the behaviour of the system. The need for such assumptions chimes with Cartwright's maxim "No causes in, no causes out" (Cartwright, 1994, Chapter 2), and goes to refute the apparently widespread belief that we are in possession of a soundly-based technology for drawing causal conclusions from purely observational data, without further assumptions.[17] This belief perhaps arises because every DAG model can be given both a probabilistic and a causal interpretation, so it is easy to conclude that, once we have derived a DAG model to describe observational conditional independencies, it must necessarily also be interpretable according to more sophisticated causal semantics (*e.g.*, as a Pearlian DAG). While this is evidently untrue (in particular, distinct but Markov equivalent DAG models, representing identical observational CI properties, will always have different implications when interpreted causally), such reification of a DAG CI representation can be very tempting.

In my view, the strong assumptions needed even to get started with causal interpretation of a DAG are far from self-evident as a matter of course,[18] and whenever such an interpretation is proposed in a real-world context these assumptions should be carefully considered and justified. Without such justification, why should we have any faith at all in, say, the application of Pearl's causal theory, or in the output of causal discovery algorithms?

17. See Geneletti (2005) for further discussion of the hidden assumptions made in this enterprise.
18. It is commonly recognised (Scheines and Spirtes, 2008) that there are cases where such assumptions should *not* be expected to hold, such as in the presence of measurement error or coarsening (which might however be rehabilitated by including the original variables in the DAG), and, more fundamentally, when dealing with dynamic processes in equilibrium (Dash, 2005).

But what would count as justification? We return to this important question in § 12. For the moment we merely remark that it cannot be conducted entirely within a model, but must, as a matter of logic, involve consideration of the interpretation of the terms in the model in the real world.

9. A formal language for causality

Another difference between a DAG representation of CI and a DAG representation of causality is that the former is always available, while the latter is not. In particular, a complete DAG over a collection of variables is totally non-committal as to their CI properties, and so (vacuously) correct. However, interpreted causally, even a complete DAG makes strong assertions. If we do not wish to make any such assertions, we can not even begin to consider using a causal DAG representation.

A less restrictive approach to causal modelling (Didelez and Sheehan, 2007a) is to develop a formal framework, with clear semantics relating mathematical properties of a putative representation to causal properties of the external system it is intended to represent, but without any commitment as to what properties the system should have: such properties should be expressible within the system, but not imposed by it. In particular, no rigid assumptions about how causality relates to probability need be made. Rather, the aim is to present a completely general language, in terms of which we can clearly express and manipulate whatever tentative causal assumptions we may wish to entertain in a specific context (in particular, it should be possible to make no such assumptions whatsoever). In these respects the rôle of such a theory would be similar to that of the theory of probabilistic conditional independence, as described in Sections 3 and 4.

One way of proceeding involves extending that same CI theory into the causal domain, using a manipulationist conception of causality (similar to that underlying the approach of Pearl). The basic ingredients are of two kinds, intended to represent, respectively, the variables ("domain variables") in the system, and the "regimes" under which those variables are generated. For application to modelling a particular external system, we must fully understand what real-world variables are supposed represented by the domain variables in the model, and what real-world regimes by the regime variables in the model. To accommodate our manipulationist stance, at least one of the regimes modelled should result from an external intervention.

The kind of causal property that will be expressible in this theory will concern relationships between the probabilistic behaviours of the domain variables, across the various regimes. Specifically, we are able (but are not obliged!) to postulate the identity, across two or more regimes, of the *conditional distribution* for one set of domain variables given another set of domain variables. When this holds we can regard that conditional distribution as a stable "modular component",[19] transferable across regimes.

This invariance or (stochastic) "stability" concept, in addition to being fundamental to my interpretation of causality, has other useful applications, arguably outside "causal inference", which can be modelled and analysed in essentially the same way. Thus we might consider the differing

19. Modularity—though more typically conceived in terms of transferable *deterministic* relationships between variables—has often been taken as an essential or defining property of causality, though this view has been challenged (Cartwright, 2007, Chapter II-3). While I make no metaphysical commitment to modularity as essential to the understanding of causality, nor even to the expression of modularity solely in terms of invariant conditional distributions, I consider that this particular approach covers a very great deal of ground, and is able to handle most aspects of "statistical" causality. A similar approach, regarding causality as residing in the "structural stability" of random variation, is taken by Russo (2008).

probabilistic behaviours of some collection of random variables in various different hospitals. We could then introduce a non-random regime indicator (but now without an interventional interpretation) to index which hospital we are looking at: this would allow us to express an assumption that a certain conditional distribution is the same in all hospitals. Or (see Example 9.1 below), we could express the property that a certain imperfect diagnostic test has the same error probabilities, no matter who it is used on. Such "reusable invariant modules" can be conveniently implemented in "object-oriented" software such as HUGIN 6[20] (Dawid et al., 2007), and have been found useful in generic schemes for handling and interpreting evidence (Hepler et al., 2007).

We observe that Property 7.3 of a Pearlian DAG representation is of just this modular form. The essential difference between Pearl's approach and that described here is that, in a Pearlian DAG model, Property 7.3 requires many modularity properties to hold—for each $V \in \mathscr{V}$, under many different observational-interventional regimes—in a way that is fully determined by the form of the DAG. In contrast, we do not seek to impose any particular modularity requirements, nor do we require that the problem be representable by a DAG. We simply provide a language for expressing and manipulating any modularity properties that we might think it appropriate, on the basis of subject matter understanding, to impose or hyothesise. As we shall see in §9.2, in some (special) cases such more limited assumptions can themselves be usefully represented by DAG-type models, but these will be non-prescriptive, and will make explicit exactly what modularity assumptions it has been considered appropriate to incorporate.

9.1 Extended conditional independence

Suppose then that there is a collection of domain variables that together describe relevant aspects of the behaviour of a system under each regime of interest. Under any one of these regimes, these variables will have a joint distribution. Any conditional independence properties that distribution may have could be expressed algebraically as in §3 or—where appropriate—graphically as in §4. We now indicate how to extend such mathematical representations to incorporate any relationships, as described above in terms of invariant conditional distributions, that might be assumed to hold between the various different regimes.

Example 9.1 As a simple example, let X be a patient's actual systolic blood pressure, and Y the value of this as recorded on a certain sphygnomanometer. The same sphygnomanometer might be used on different patients at different times, but it might be reasonable to assume that the distribution of Y given X is stable, irrespective of the circumstances of use. We could introduce a regime indicator F, whose values specify the conditions, environment, kind of patient, *etc.*. Note that whereas the domain variables (X,Y) are random, F is not: rather, it has the status of a *statistical parameter*, indexing the probabilistic regime under consideration. In particular, any probability or independence statements must, explicitly or implicitly, be conditioned on the value of F.

The stability assumption is just that the conditional density $p(y \mid F = f, X = x)$ for Y, given $X = x$, in regime $F = f$, is in fact the same for all values of f. In the light of (1), we see that this can be expressed in the form of a conditional independence property:

$$Y \perp\!\!\!\perp F \mid X. \tag{7}$$

□

20. <http://www.hugin.com/>

It is important to note that expression (7) makes sense, even though F is not a random variable. In general, for the expression $X \perp\!\!\!\perp Y \mid Z$ in (1) to be meaningful, while X must be random there is no requirement that the conditioning variables Y and Z be random: either or both could be a parameter variable or regime indicator (see Dawid (1979a, 2002b) for further details). This language of *extended conditional independence* (ECI) thus provides a natural way of expressing stability across regimes of modular conditional distributions. In particular, ECI supplies an appropriate formal language (syntax and semantics) for describing and handling causality in our modular manipulationist understanding of the term.

Example 9.2 Consider a system involving domain variables Z, U, X, Y. We wish to model the effect of an intervention that sets the value of X. To this end we introduce an *intervention variable*, F_X, a special regime indicator with values corresponding to the different regimes that arise on intervening to set the value of X in various ways (Spohn, 1976; Spirtes et al., 2000; Pearl, 2009). If X is binary, then F_X might have values $\emptyset$, 0 and 1, the interpretation being that, when $F_X = \emptyset$ (the *idle* regime), the domain variables arise from the undisturbed system; whereas when $F_X = 0$ [resp., 1] they arise from the system disturbed by an external intervention that forces X to take the value 0 [resp., 1].

In general, the joint distributions of (Z, U, X, Y) under the three different regimes (*i.e.*, given $F_X = \emptyset$, 0 or 1) could be entirely arbitrary,[21] and unrelated to each other. But *should* we wish to specify or describe connexions between them, we can usefully do so using ECI. This programme can be effected, in great generality, in entirely algebraic fashion: we can use the general properties (2) and (with due care) (3) to manipulate ECI properties, almost exactly as for probabilistic CI. We just have to ensure that no non-random variable occurs as the first term in any ECI relation in either our assumptions or our conclusions.

Again, these manipulations are most conveniently described and conducted in graphical terms—though we once again warn that by no means every problem that can be manipulated algebraically can be modelled graphically. □

9.2 Augmented DAGs

Just as for regular CI it is sometimes possible, and then is helpful, to represent a collection of ECI properties by means of a DAG[22]—but now extended to include nodes to represent non-random regime variables (generally drawn as square), in addition to nodes representing domain variables (generally drawn as round). Indeed, this can be done with essentially the identical constructions and interpretations as for regular DAGs. Such a DAG is termed an *influence diagram* (ID) (Dawid, 2002b).

Many of the IDs considered in a causal context have a specific form, as "*augmented DAGs*"(Pearl, 1993). Figure 4 shows an augmented DAG, a variation on the simple, purely probabilistic, DAG of Figure 1, that also incorporates, in a particular way, an *intervention node* F_X, interpreted as in Example 9.2.

What does it mean to say that a particular system is modelled by this augmented DAG? To address this question, we apply the "d-separation semantics" described in §4—but now ignoring the

21. Except that, to express our intended interpretation of F_X, under $F_X = 0$ [resp., 1] we should require $X = 0$ [resp., 1] with probability 1. There is however no immediate implication for the distribution of any other variables.
22. Other kinds of graphical CI representations can be similarly extended to include intervention variables (Dawid, 2002a; Zhang, 2008; Eichler and Didelez, 2009).

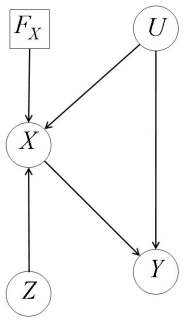

Figure 4: Augmented DAG

distinction between domain and regime variables. The DAG thus represents the following (algebraically expressed) conditional independence properties:

$$(U,Z) \perp\!\!\!\perp F_X \tag{8}$$
$$U \perp\!\!\!\perp Z \mid F_X \tag{9}$$
$$Y \perp\!\!\!\perp F_X \mid (X,U) \tag{10}$$
$$Y \perp\!\!\!\perp Z \mid (X,U;F_X). \tag{11}$$

Using (1), property (8) is to be interpreted as saying that the joint distribution of (U,Z) is independent of the regime F_X: *i.e.*, it is the same in all three regimes. In particular, it is unaffected by whether, and if so how, we intervene to set the value of X. The identity of this joint distribution across the two interventional regimes $F_X = 0$ and $F_X = 1$ could be interpreted as expressing a causal property: manipulating X has no (probabilistic) effect on the pair of variables (U,Z). Furthermore, since this common joint distribution is also supposed the same in the idle regime, $F_X = \emptyset$, we could in principle use observational data to estimate it—thus opening up the possibility of causal inference.

Property (9) asserts that, in their (common) joint distribution in any regime, U and Z are independent: this however is a purely probabilistic, not a causal, property.

Property (10) says that the conditional distribution of Y given (X,U) is the same in both interventional regimes, as well as in the observational regime, and can thus be considered as a modular component, fully transferable between the three regimes — again, I regard this as expressing a causal property.

Finally, property (11) asserts that this common conditional distribution is unaffected by further conditioning on Z (not in itself a causal property).

Just as for regular CI, it is possible for a collection of ECI properties to have more than one representation as an augmented DAG. This is the case for Figure 5, where the direction of the arrow between U and V is not determined.

We see that the ingredients required for at least some causal assertions and inferences—namely that certain marginal or conditional distributions be unaffected by whether or how certain interventions are made—are readily expressible using the familiar language of conditional independence (specifically, they arise when the second argument of an ECI relation is a regime variable). They are just as readily manipulated by means of the rules embodied in (2). And in those special cases that it

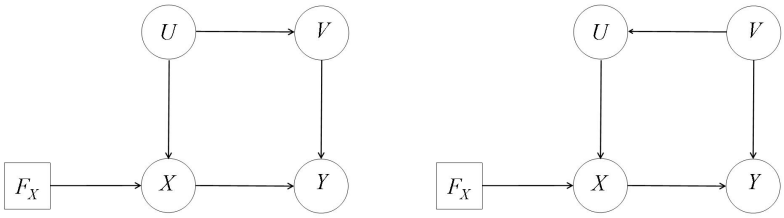

Figure 5: Two Markov-equivalent augmented DAGs

is possible to express all the causal and probabilistic assumptions made in a form that can be represented by an augmented DAG, we can use the d-separation semantics of §4 as a "theorem-proving machine" to discover their logical implications.

9.3 Pearlian DAGs as augmented DAGs

A Pearlian DAG is readily represented as a special kind of augmented DAG. To do this, we elaborate the given DAG by including, for *every* domain variable V in it, an intervention node F_V, and an arrow pointing from F_V to V. Properties 7.1, 7.2 and 7.3 are then explicitly represented by the d-separation semantics. Correspondingly all the implications of a Pearlian representation can be deduced from this augmented DAG and d-separation.

In Pearl's earlier work (Pearl, 1993, 1995) he moved backwards and forwards between explicit and implicit representation of the intervention variables in the DAG. More recently he, and most of those following him, have been using only the implicit version, in which the intervention variables F_V are not explicitly included in the diagram, but (to comply with the Pearlian interpretation) the DAG is nevertheless to be interpreted as if they were. I regard this demotion of the intervention indicators as a retrograde move, since the resulting graphical representation, while imbued with Pearlian causal semantics, is visually indistinguishable from a DAG used to describe purely probabilistic CI. Consequently, great care is needed to be clear just what a given DAG is intended to represent, and to avoid slipping unthinkingly from one interpretation to another. Explicit representation of intervention nodes helps to guard against such confusion, as well as simplifying interpretation and manipulation.[23]

10. Instrumental variables

To clarify the similarities and differences between augmented DAG representations and other causal DAG representations, we revisit the example of §5. Hernán and Robins (2006) present the causal DAG of Figure 3 as a counterexample to the supposition (Martens et al., 2006) that the following conditions are necessary for a variable Z to qualify as an "instrumental variable" for estimating the causal effect of an "exposure" X on a "response" Y, in the presence of an additional, unmeasured, variable U (a confounder), that affects both X and Y, when we can not directly manipulate X:

23. For example, Pearl (1995) derives his "do-calculus" rules using an explicit augmented DAG representation, but then re-expresses them in terms of the unaugmented graph—when they become considerably more complex. It is not clear what is gained to compensate for this loss of transparencqy.

(i). Z has a causal effect on X

(ii). Z affects the outcome Y only through X (*i.e.*, no direct effect of Z on Y)

(iii). Z does not share common causes with the outcome Y (*i.e.*, no confounding for the effect of Z on Y).

The causal DAG presented by Hernán and Robins (2006) as embodying these assumptions is essentially the same as our Figure 1. This is contrasted with the causal DAG of Figure 3, which is not regarded as embodying condition (i), since Z has no direct causal effect on X, but is merely associated with it through sharing a common cause U^*.

Note that the descriptions of both problems employ intuitive causal terms, and that these are associated with the presence and directionality of the arrows in the causal DAG representations.

DAG representations of the ECI versions of these stories are presented in Figure 4 and Figure 6. In each case an intervention node F_X associated with X has been added, describing three regimes of interest: the idle regime $F_X = \emptyset$ corresponding to pure observation, and the two interventional regimes $F_X = 0$ and 1, corresponding to an intervention in which X is externally manipulated to take values 0 and 1, respectively. While data can be gathered only under the idle regime, which is thus all that can be directly estimated, our interest is nevertheless in estimating (if possible), and, especially, comparing, the distributions of the response Y under the interventional regimes, $F_X = 0$ and $F_X = 1$.

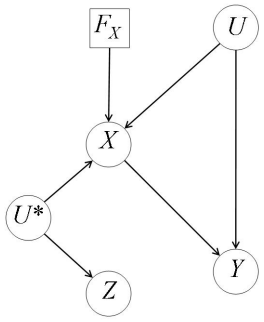

Figure 6: Augmented DAG corresponding to Figure 3

Now in the story represented by Figure 3 or Figure 6, the variable U^*, while apparently required for a full causal specification of the structure of the problem, plays no rôle in the analysis of Z as an instrumental variable. So we can restrict attention to the joint distribution, under the various regimes, of the variables (U,X,Y,Z), and their independence properties. We then find that the augmented DAG of Figure 6 embodies the identical conditional independence properties (8)–(11) as the alternative augmented DAG of Figure 4, in which U^* does not figure at all. Consequently, for our purposes this is just as good an augmented DAG representation of the problem as Figure 6. This equivalence should be contrasted with the apparent attitude of Hernán and Robins (2006), that if Figure 3 is a "true" causal DAG representation of the problem, then Figure 1 is not. We would likewise have to distinguish these two DAG representations under a Pearlian interpretation, which would be equivalent to attaching intervention nodes to *every* domain variable. But this would

involve making many additional and possibly questionable assumptions, none of which is needed for the analysis.

An important interpretive difference between causal DAGs as described in § 5 and as represented by augmented DAGs is that, in the former, causal meaning is understood as carried by the arrows, whereas, in the latter, it is entirely carried by extended conditional independence properties, involving intervention variables, which are represented only indirectly in the DAG, *via* d-separation. In particular, in Figure 4 (and in contrast to the causal interpetation of Figure 1) the arrow from Z to X is *not* to be construed as representing a relationship of cause and effect between Z and X (see Didelez and Sheehan (2007b) for more on this in the context of Mendelian randomization).

The ECI properties (8)–(9) are "core conditions" for a variable Z to be an *instrument* for the effect of X on Y.[24] Once so characterised, these properties can be manipulated algebraically using the rules of (2) (together with properties such as $F_X = 0 \Rightarrow X = 0$), without reference to any graphical representation: the "theorem-proving" properties of DAG representations, while immensely useful, are logically inessential. But if we do want to use graphical representations to help us, there is no point in arguing whether it is Figure 4 or Figure 6 that is "correct"—since each of them embodies (8)–(9) equally well.

11. Non-DAG modularity

Bertrand Russell (Russell, 1913) famously opined "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm". Many scientific laws are symmetric, and thus inappropriate for representation in terms of directional causal properties. Nevertheless, they can still be described by means of modularity properties, and these can frequently be expressed using ECI.[25]

Example 11.1 Ideal gas

Consider a system involving a fixed number N of molecules of a monatomic ideal gas in an impermeable container, whose volume can be adjusted by manipulating a piston. The container is immersed in a heat bath of constant absolute temperature t^*. Let V, P denote, respectively, the volume and pressure of the gas.

Suppose first that the manipulations of the piston are *isothermal*, *i.e.* slow enough that, by heat transfer through the walls of the container, the gas always remains at the external temperature t^*. Then P and V are functionally related by Boyle's law:

$$PV = c \tag{12}$$

where the constant c is kNt^*, k being Boltzmann's constant.

24. There is one more core condition, expressible in terms of ECI though not graphically representable: $X \not\perp\!\!\!\perp Z \mid F_X = \emptyset$. In addition to these core conditions, precise identification of a causal effect by means of an instrumental variable requires further modelling assumptions, such as linear regressions (Didelez and Sheehan, 2007b).

25. Alternative modular descriptions can also be used in this more general context, for example, based on non-recursive systems of simultaneous structural equations, such as in Pearl (2009, Chapter 7) (see Example 11.2 below). An immediate advantage of the ECI description over representations in terms of equations is that, because of its relationship with conditional distributions, each ECI property, of the form $X \perp\!\!\!\perp Y \mid Z$, automatically comes with an associated directionality: from its conditioning variables (Y, Z) to its response variables X.

Let the regime indicator F describe various isothermal manipulations of the piston, in either fixed or random ways. In all cases the relation (12) will hold. In particular, given either V or P, the other is determined, irrespective of the regime that brought the situation about. We shall thus have, simultaneously, the ECI properties

$$P \perp\!\!\!\perp F \mid V \tag{13}$$
$$V \perp\!\!\!\perp F \mid P \tag{14}$$

where, for example, the modular conditional distribution of P given $V = v$ associated with (13) is a 1-point distribution at c/v. We note that there is no DAG representation of the pair of ECI properties (13) and (14).

We could alternatively consider *adiabatic* manipulations, which proceed sufficiently fast that no heat can transfer through the walls of the container (but not so fast as to add energy to the gas). Then (12) is replaced by

$$PV^{\frac{5}{3}} = \text{ constant.} \tag{15}$$

Again (13) and (14) will hold, but now with different specifications for the modular conditional distributions.

Finally, consider arbitrary manipulations of the piston. There are now no modular relationships holding between P and V, but modularity can be restored by introducing the additional variable T, the absolute temperature of the gas. The (symmetrical) invariant relationship is

$$PV = kNT. \tag{16}$$

In terms of ECI we have

$$P \perp\!\!\!\perp F \mid (V, T) \tag{17}$$
$$V \perp\!\!\!\perp F \mid (P, T) \tag{18}$$
$$T \perp\!\!\!\perp F \mid (P, V) \tag{19}$$

where, for example, the modular conditional distribution of P given $V = v, T = t$ associated with (17) is a 1-point distribution at kNt/v. Again, there is no DAG representation of this collection of ECI properties. □

Example 11.2 Price and demand

A simple econometric model relates price, P, and quantity demanded, Q, for some good. It is supposed possible to manipulate either of these to any given value. There are additional unobserved explanatory variables U_P, U_Q, which are supposed unaffected by such manipulations, having a given joint distribution. Let F_P, F_Q denote the indicators for interventions at P, Q respectively. We suppose we have specified the interventional conditional distribution of Q, given $(P = p, U_P, U_Q; F_P = p, F_Q = \emptyset)$, and that this does not in fact depend on U_P; and similarly we have specified the conditional distribution for P, given $(Q = q, U_P, U_Q; F_P = \emptyset, F_Q = q)$, which is independent of U_Q.

The idle regime, when $F_P = F_Q = \emptyset$, is taken as referring to the joint distribution "in equilibrium". On the basis of economic theory it is supposed—constituting our "modular assumptions"—that all the above specified marginal and conditional distributions continue to apply, simultaneously,

in this equilibrium regime. (Note that consistency conditions then constrain the possible specifications of the conditional distributions for Q and P).

The modular assumptions made are encapsulated in the following ECI properties:

$$(U_P, U_Q) \quad \perp\!\!\!\perp \quad (F_P, F_Q) \tag{20}$$

$$Q \quad \perp\!\!\!\perp \quad (F_P, U_P) \mid (P, F_Q, U_Q) \tag{21}$$

$$P \quad \perp\!\!\!\perp \quad (F_Q, U_Q) \mid (Q, F_P, U_P). \tag{22}$$

There is no DAG representation of this collection of properties, but they can be represented using the more general graphical semantics of *chain-graphs* (Cowell et al., 2007; Dawid, 2002a), involving undirected as well as directed links: the relevant diagram is shown in Figure 7.[26]

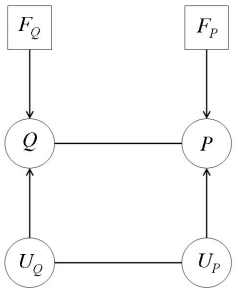

Figure 7: Chain-graph for price-demand relationship

When each of the modular distributions of Q given (P, U_Q), and of P given (Q, U_P), is concentrated on a single point, so that they represent deterministic functional relationships, and moreover those relationships are linear and (U_P, U_Q) are bivariate normal, our model is isomorphic to the structural equation model considered by Pearl (2009, §7.2.1).[27] However—and in sharp contrast to the analogous case for DAG models—even if we consider only the case that U_P and U_Q are independent, this model is *not* equivalent, in terms of the properties of the observables (Q, P) under the different regimes, to a case where the latent variables (U_P, U_Q) are absent (equivalently, taken as trivial), but we allow genuine stochasticity in the above conditional distributions. In particular, the appropriate generalisation of d-separation (Frydenberg, 1990) applied to the chain-graph of Figure 7 (even without the link U_Q—U_P) does not yield $Q \perp\!\!\!\perp F_P \mid (P, F_Q)$, although (by (21) with the U's absent) this does hold in the stochastic model. A confirmation of this non-equivalence is that, in the stochastic model without U's, the observational and interventional distributions of Q given P are identical, as follows immediately from the above relation; whereas Pearl's own analysis shows that this is typically not the case for the deterministic model incorporating U's.

We can entertain more general models, in which the U's are non-trivial and the specified distributions are genuinely stochastic. Again, the observational and interventional and distributions of Q given P will differ, but the relationship between them will be different from both the cases considered above.

26. In reality we can not vary F_P and F_Q independently: at least one of them must be idle. This "variation non-independence" (Dawid, 2001a,b) could be represented in Figure 7 by a further undirected link between F_P and F_Q; however this is of no real consequence here.
27. We have omitted Pearl's observable explanatory variables I, W for simplicity.

Our general ECI model (20)–(22) thus incorporates Pearl's deterministic structural model, but allows other cases too—which, it could be argued, are no less obviously appropriate as descriptions of the problem. A moral of this analysis is that we should not be cavalier in setting out the ingredients (variables and modular conditional distributions) of such a model, but need to think very carefully about them in the context of the problem we are modelling and any relevant theory. And in order for us to be able to approach this task in a meaningful way, we must be able to identify the unobserved explanatory variables U_P and U_Q as real-world quantities. We can not just treat them as convenient mathematical fictions ("error terms"), for then how are we to decide whether our model should be deterministic or stochastic?—a choice that will make a difference to our analysis and conclusions.

□

12. Justifying assumptions

Perhaps the most important characteristic of my suggested approach to causality, using extended conditional independence and (where appropriate) augmented DAGs or other graphical representations, is that it is descriptive, not prescriptive. It makes no assumptions as to how causality ought to behave or be represented; rather, it supplies a language by which we are able clearly to express and manipulate any such assumptions we might wish to make in any given context. In this respect it differs from other theories of "probabilistic causality"[28] in much the same way as Kolmogorov's purely formal theory of probability differs from other theories such as the "classical theory" based on the assumption that intuitively "equally possible" outcomes should be assigned equal probabilities, or von Mises's theory of collectives, which sought to represent assumed empirical properties of probability, such as the existence and stability of limiting relative frequencies, directly within the formal theory. This strict separation of the formal general-purpose language from any special assumptions that might be made in specific contexts allows for much greater clarity and flexibility. It also protects against the ever-present danger of unthinking reification of incidental formal properties of our representations. In particular, it does not in itself support causal interpretation of a probabilistic DAG. If we wish to represent this, we have very explicitly to introduce (using ECI) whatever additional assertions we are making about effects of interventions. ECI is a purely mechanical tool for manipulating causal properties, not a philosophical foundation for defining them.

This purely formal approach does, of necessity, leave entirely untouched such essential questions as "Where do we get our causal assumptions from?" and "How can they be justified?" It is at this point, entirely removed from representational issues, that we might find a place for more informal arguments, based on intuitive understandings of cause and effect.

In principle, the meaning of ECI assumptions such as (8)–(11) is straightforward; and they could indeed all be tested empirically if we had access to data collected on (U,Z,X,Y) under the various regimes. In practice, however, we will usually not have such data (and it may not even be clear which unobserved external variable or variables are represented by the symbol U). Then

28. By this term I do not mean to include general theories of " statistical causality," such as that of Rubin (1978), which likewise make no prescriptive assumptions. See Dawid (2000, 2002b) for comparisons and contrasts between my own approach and other approaches to statistical causality. The general points I have made could have been developed from the viewpoint of those other theories, though these mostly do not focus, as I do, on modularity at the level of conditional distributions, which supplies a natural point of contact with the intuitive concepts of "probabilistic causality".

the appropriateness of the assumptions made requires and deserves further, necessarily context-dependent, argument.

For example, physical *randomization* of a treatment T in the "idle" regime is generally agreed to provide a convincing reason for believing that the observational distribution of a response Y, given $T = t$, is the same as its distribution would be under an intervention to set T to t (formally: $Y \perp\!\!\!\perp F_T \mid T$), thus justifying causal interpretation of these conditional distributions. Although this property of randomization is usually taken as intuitively obvious, I am not aware of any argument for it based on deeper principles. One such argument could be based on the assumed existence of some *sufficient covariate* U, such that (a) $U \perp\!\!\!\perp F_T$ and (b) $Y \perp\!\!\!\perp F_T \mid (T,U)$ (Dawid, 2002b). Here, (a) says that the distribution of U is unaffected by which regime is operating—typically believable if U is a "pre-treatment" variable; while (b) says that, conditional on U and *which* treatment T is applied, the response Y of the system is unaffected by *how* (*i.e.*, in which regime) it is applied. While it may not be easy to identify a specific pre-treatment variable U with this property, one might be willing to accept that some such variable does exist. Randomization, and the pretreatment status of U, now gives good cause to accept $T \perp\!\!\!\perp U \mid F_T = \emptyset$, whence (since T is in any case non-random in any interventional regime) (c) $T \perp\!\!\!\perp U \mid F_T$. Using the rules of (2), it is straightforward to deduce, from the three CI properties (a), (c), (b), the desired conclusion $Y \perp\!\!\!\perp F_T \mid T$. Alternatively, these CI properties can be represented by the augmented DAG of Figure 8, from which we can readily read off $Y \perp\!\!\!\perp F_T \mid T$. Similar arguments can be made to justify suitably expressed causal interpretations

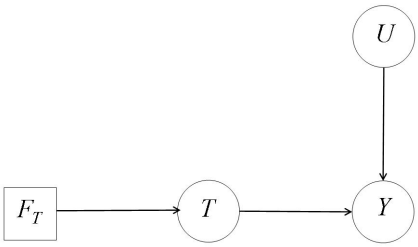

Figure 8: Augmented DAG for randomization

of data generated under more complex randomization schemes. But the appropriateness of any such argument needs to be carefully considered, not just taken for granted.[29]

When physical randomization is not possible, it will be necessary to attempt to justify causal CI assumptions on other grounds. For example, in the instrumental variable problem of § 10, we need to argue for the appropriateness of the assumptions (8)–(11). (Once again, it is enough that there exist *some* variable U, which we need not however specify in detail, for which the conditions can be assumed to hold.)

29. Indeed, even in a randomized double-blind clinical trial—the "gold standard" of evidence-based medicine—one could argue that the very artificiality of the trial negates assumption (b) above: we would not expect the same response process to operate for future treated patients as for those in the trial. To make progress we might make weaker assumptions, such as transferability from the clinical trial into general practice of the "specific causal effect": $\mathrm{E}(Y \mid T = 1, U = u, F_T = \emptyset) - \mathrm{E}(Y \mid T = 0, U = u, F_T = \emptyset) = \mathrm{E}(Y \mid F_T = 1, U = u) - \mathrm{E}(Y \mid F_T = 0, U = u)$. While not expressible in terms of ECI, such an assumption still relates to the invariance of probabilistic properties across different regimes. Again, it should be made explicit, and justified (ideally empirically).

Property (8) essentially requires that both U and Z be pre-treatment variables, and then (10) implies that U must be a sufficient covariate.

Properties (9) and (11) are more problematic. Property (9) could be plausible if Z is itself determined by randomization: a scenario in which this occurs is that of "incomplete compliance" (Dawid, 2003), where patients are randomized to treatment, with randomization indicator Z, but the treatment X actually taken might not be the same as that assigned. Alternatively, in "Mendelian randomization" (Didelez and Sheehan, 2007b), Z might be a gene that naturally affects X: property (9) might then be justified, for suitable U, on the basis of the random assortment of genes under Mendelian genetics. As described by Didelez and Sheehan (2002): "If we think of U as some behavioural pattern or life style, this independence condition can be justified as long as we are reasonably certain that any possible genetic factors influencing the behavioural pattern are unrelated to this particular gene".

Finally, (11) requires that the distribution of Y given (X,U) (which has been assumed the same in all regimes) is unaffected, in any regime, by further conditioning on Z—intuitively expressed as "no direct effect of Z on Y". This might be plausible in the imperfect compliance context, where we could believe that behaviour of the response Y could depend on the treatment X actually taken and further pre-existing individual characteristics U, but not further on the treatment Z that the individual was supposed to take. In the context of Mendelian randomization, we require that "there is no association between the genotype and the disease status given the intermediate phenotype and the life style" (Didelez and Sheehan, 2002). (However, core conditions (9) and (11) can be violated in the presence of various complications, such as linkage disequilibrium, pleiotropy, genetic heterogeneity or population stratification (Didelez and Sheehan, 2007b).)

When attempting to justify the core conditions in a specific context, it is plausible that thinking about the problem in terms of further unobserved variables, such as U^* in Figure 6, can play a valuable rôle in the process. However, once these conditions have been settled on as the assumptions we wish to introduce, there is no need to make irrelevant distinctions between alternative, equally valid representations of them, such as Figure 4 and Figure 6.

In the ECI framework, attention is clearly drawn to any assumptions we may choose to make, since these have to be clearly expressed as explicit ingredients added to our model, and justified in the context of the real-world application under consideration. In other approaches the assumptions are often hidden, and it is easy to be misled into believing that they are not in need of justification. For example, the weak causal Markov assumption (§ 6) rules out certain ECI representations purely on the basis of ordinary CI properties in the observational regime; but there is no logical reason why this should be so, and its validity should be carefully considered in every intended application.

13. Conclusion

We have contrasted various approaches to the interpretation of graphical models of probabilistic causal processes. Each of these purports to relate properties of the mathematical model and properties of the process.

The most common approach, "probabilistic causality" (see § 6), works with intuitive understandings of causal terms, which are often taken as undefined and self-evident primitives, although they can also be regarded as deriving from an underlying manipulationist conception. Its most important feature is that it assumes links (*via e.g.* the "Causal Markov Condition") between such causal

concepts and certain probabilistic conditional independence properties—links that, however, there is no reason to believe hold in complete generality.

In contrast the approach described in § 9, based on the algebraic theory of extended conditional independence and its graphical representations, is based on a clearly defined internal mathematical structure (syntax), and clearly described rules of interpretation (semantics). In these respects it is similar to Pearl's approach. However, unlike both that approach and that of probabilistic causality, it does not suppose any special relationship between causality and conditional independence. It merely supplies a formal language by means of which we can express and explore interesting causal conjectures, phrased as the identity of certain conditional distributions across a variety of different regimes (typically encompassing both intervention and pure observation). This surgical separation of the formal language from *ad hoc* causal assumptions enforces clear and unambiguous articulation of those assumptions, allows us to develop the logical implications of our assumptions, and clarifies exactly what needs to be justified in any particular context. That justification is itself, however, an entirely separate task, that can not rely on formal representations of any kind but must relate to the real-world context of the problem. Perhaps the most important contribution of modelling "causality" in terms of ECI is to highlight the vital need for such external justification.

13.1 What rôle for "causal discovery"?

The enterprise of "causal discovery" aims to extract causal conclusions from observationally inferred conditional independencies. However it can not do so without making (explicitly or, more often, implicitly) strong causal assumptions—which may rest unjustified, so invalidating the process. Such methods can nevertheless be useful in suggesting interesting causal conjectures for further investigation. Ideally we should then gather data from appropriate interventional studies, to investigate—and if necessary revise—the validity of conjectures, made purely on the basis of observational data, about the effects of interventions. Williamson (2005), among others, has argued for such a "hybrid hypothetico-deductive/inductive" approach.

Alternatively, when we can collect data under a variety of regimes, including interventional studies, we could directly apply variations of causal discovery techniques, to uncover genuinely causal properties. Thus, if we had data on variables (U,Z,X,Y) under all three regimes $F_X = \emptyset$, $F_X = 0$, $F_X = 1$, we could empirically test the ECI properties (8) and (10), by (for example) simple χ^2-tests (which are equally valid for testing homogeneity of conditional distributions as they are for testing conditional independence); alternatively, Bayesian techniques could be used (Cooper and Yoo, 1999). Only with such experimental data could we hope to obtain genuine empirical evidence in favour of a causal DAG representation such as Figure 4.

Acknowledgments

My thanks to the Editors for their encouragement to prepare both the NIPS talk and this paper very loosely based on it. I am grateful to Nancy Cartwright, Vanessa Didelez, Sara Geneletti, Paul Rosenbaum, Federica Russo and Jon Williamson, the referees, and many contributors to the "Causality and Machine Learning Reading Group" http://www.afia-france.org/tiki-index.php?page=Groupe+de+lecture, for valuable feedback on an earlier draft.

References

Pierre Bourdieu. *Outline of a Theory of Practice*. Cambridge University Press, 1977.

George E. P. Box. Use and abuse of regression. *Technometrics*, 8:625–629, 1966.

Nancy Cartwright. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford, 1994.

Nancy Cartwright. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, 2007.

Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 116–125, San Francisco, CA, 1999. Morgan Kaufmann.

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer, New York, 2007.

Denver Dash. Restructuring dynamic causal systems in equilibrium. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005. URL http://www.gatsby.ucl.ac.uk/aistats/fullpapers/264.pdf.

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

A. Philip Dawid. Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society, Series B*, 41:249–52, 1979b.

A. Philip Dawid. Conditional independence for statistical operations. *Annals of Statistics*, 8:598–617, 1980.

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

A. Philip Dawid. Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32:335–372, 2001a.

A. Philip Dawid. Some variations on variation independence. In Tommi Jaakkola and Thomas S. Richardson, editors, *Artificial Intelligence and Statistics 2001*, pages 187–191, San Francisco, California, 2001b. Morgan Kaufmann Publishers.

A. Philip Dawid. In discussion of Lauritzen and Richardson (2002). *Journal of the Royal Statistical Society, Series B*, 64:348–351, 2002a.

A. Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002b. Corrigenda, *ibid*., 437.

A. Philip Dawid. Causal inference using influence diagrams: The problem of partial compliance (with Discussion). In Peter J. Green, Nils L. Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems*, pages 45–81. Oxford University Press, 2003.

A. Philip Dawid. Seeing and doing: The Pearlian synthesis. In Rina Dechter, Hector Geffner, and Joseph Y. Halpern, editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, chapter 18, pages 309–325. College Publications, London, 2010.

A. Philip Dawid and Vanessa Didelez. Identifying the consequences of dynamic treatment strategies. Research Report 262, Department of Statistical Science, University College London, 2005. URL http://www.ucl.ac.uk/Stats/research/reports/abs05.html#262.

A. Philip Dawid, Julia Mortera, and Paola Vicard. Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International*, 169:195–205, 2007.

Bruno de Finetti. *Theory of Probability (Volumes 1 and 2)*. John Wiley and Sons, New York, 1975. (Italian original Einaudi, 1970).

Vanessa Didelez and Nuala A. Sheehan. Mendelian randomisation and instrumental variables: What can and what can't be done. Technical Report 05-02, University of Leicester Department of Health Sciences, 2002.

Vanessa Didelez and Nuala A. Sheehan. Mendelian randomisation: Why epidemiology needs a formal language for causality. In Federica Russo and Jon Williamson, editors, *Causality and Probability in the Sciences*, volume 5 of *Texts In Philosophy Series*, pages 263–292. College Publications, London, 2007a.

Vanessa Didelez and Nuala A. Sheehan. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16:309–330, 2007b.

Michael Eichler and Vanessa Didelez. On Granger-causality and the effect of interventions in time series. Research Report 09:01, Statistics Group, University of Bristol, 2009.

Morten Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17: 333–353, 1990.

Maria Carla Galavotti. *Philosophical Introduction to Probability*. CSLI Publications, Stanford, 2005.

Sara G. Geneletti. *Aspects of Causal Inference in a Non-Counterfactual Framework*. PhD thesis, Department of Statistical Science, University College London, 2005.

Clark Glymour and Gregory F. Cooper, editors. *Computation, Causation and Discovery*. AAAI Press, Menlo Park, CA, 1999.

Daniel Hausman. *Causal Asymmetries*. Cambridge University Press, Cambridge, 1998.

Amanda B. Hepler, A. Philip Dawid, and Valentina Leucari. Object-oriented graphical representations of complex patterns of evidence. *Law, Probability & Risk*, 6:275–293, 2007. doi: 10.1093/lpr/mgm005.

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

Dominik Janzing and Bernhard Schölkopf. Distinguishing between cause and effect via the algorithmic Markov condition. Paper presented at NIPS 2008 Workshop "Causality: Objectives and Assessment", Whistler, Canada, 2008a.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition, 2008b. URL http://arxiv.org/abs/0804.3678.

Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations (with Discussion). *Journal of the Royal Statistical Society, Series B*, 64:321–361, 2002.

Steffen L. Lauritzen, A. Philip Dawid, Birgitte N. Larsen, and Hanns-Georg Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–505, 1990.

Edwin P. Martens, Wiebe R. Pestman, Anthonius de Boer, Svetlana V. Belitser, and Olaf H. Klungel. Instrumental variables: Applications and limitations. *Epidemiology*, 17:260–267, 2006.

Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, New Jersey, 2003.

Judea Pearl. A constraint–propagation approach to probabilistic reasoning. In Laveen N. Kanal and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 357–370, Amsterdam, 1986. North-Holland.

Judea Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8:266–269, 1993.

Judea Pearl. Causal diagrams for empirical research (with Discussion). *Biometrika*, 82:669–710, 1995.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, second edition, 2009.

Judea Pearl and Azaria Paz. Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In *ECAI*, pages 357–363, 1986.

Hans Reichenbach. *The Direction of Time*. University of Los Angeles Press, Berkeley, 1956.

Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30: 962–1030, 2002.

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

Bertrand Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26, 1913.

Federica Russo. *Causality and Causal Modelling in the Social Sciences: Measuring Variations*, volume 5 of *Methodos Series*. Springer, 2008.

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

Glenn Shafer. *The Art of Causal Conjecture*. MIT Press, Cambridge, Mass, 1996.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. Springer-Verlag, New York, Second edition, 2000.

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

Wolfgang Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, 9:73–99, 1980.

Wolfgang Spohn. Bayesian nets are all there is to causal dependence. In Maria Carla Galavotti, Patrick Suppes, and Domenico Costantini, editors, *Stochastic Dependence and Causality*, chapter 9, pages 157–172. University of Chicago Press, Chicago, 2001.

Milan Studený. Conditional independence relations have no finite complete characterization. In *Transactions of the Eleventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, volume B, pages 377–396, Prague, 1992. Academia.

Patrick Suppes. *A Probabilistic Theory of Causality*. North Holland, Amsterdam, 1970.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In Ross D. Shachter, Tod S. Levitt, Laveen N. Kanal, and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence 4*, pages 69–76, Amsterdam, 1990. North-Holland.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In Piero P. Bonissone, Max Henrion, Laveen N. Kanal, and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 255–268. North-Holland, Amsterdam, 1991.

Jon Williamson. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford University Press, Oxford, 2005.

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

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9: 1437–1474, 2008.

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.

Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 6:157–164, 2010.