International Statistical Review (2002), **70**, 2, 161–189, Printed in The Netherlands

Influence Diagrams for Causal Modelling and Inference

A.P. Dawid

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK. E-mail: dawid@stats.ucl.ac.uk

Summary

We consider a variety of ways in which probabilistic and causal models can be represented in graphical form. By adding nodes to our graphs to represent parameters, decisions, *etc.*, we obtain a generalisation of influence diagrams that supports meaningful causal modelling and inference, and only requires concepts and methods that are already standard in the purely probabilistic case. We relate our representations to others, particularly functional models, and present arguments and examples in favour of their superiority.

Key words: Augmented DAG; Causal inference; Confounder; Counterfactual; Directed acyclic graph; Graphical model; Intervention; Functional model.

1 Introduction

The use of graphical models to represent and manipulate complex multivariate probability distributions is by now well-established. Such representations are also increasingly being applied to problems of causal modelling. However there are several ways in which we can describe a problem graphically, and these may differ in their representational power. In particular, Judea Pearl has introduced at least three distinct (though related) frameworks for causal modelling based on directed acyclic graphs (DAGs). His original approach used two variant graphical representations of purely probabilistic causal relations (Pearl, 1993a, 1993b), while his later papers (Pearl, 1995; Balke & Pearl, 1994) replaced these with a formulation based on *functional models*. The latter are closely related to *counterfactual models*, as discussed and criticised by Dawid (2000). Pearl's recent book (Pearl, 2000) gives a thorough account of both approaches, the first half (Chapters 1–6) being largely probability-based, while the second half is largely function-based.

In this paper we review these and other graphical representations, and attempt to clarify the connections between them. In particular, we demonstrate the advantages of a straightforward elaboration of probability-based graphical modelling, based on influence diagrams, extended where appropriate to include graphical specification of relevant probability distributions and/or decision strategies. These have an appealing and expressive semantics, are straightforward to manipulate, and support meaningful causal modelling and analysis. This particular graphical framework has just the right degree of expressive power: all meaningful questions relating to 'effects of causes' can be formulated and analysed within it, while, by shunning arbitrary and unnecessary extraneous ingredients, it protects us from asking (and extracting purported answers to) questions that appear well-formulated but are in fact scientifically meaningless. We also examine in detail other models, based on introducing deterministic functional relations or counterfactual variables into our representations, and their relationship to influence diagram representations. We argue that these extensions are both unnecessary and undesirable.

Our approach is illustrated by several examples. In particular, we show how it can be used to clarify both the definition and the analysis of problems having 'no unobserved confounders'.

2 Probabilistic DAGs

We start by reviewing the use of directed acyclic graphs to represent independence properties of joint probability distributions (Cowell *et al.*, 1999). Figure 1 gives a simple example involving five variables, A, B, C, D and E. To say that a distribution for these five variables is represented by this

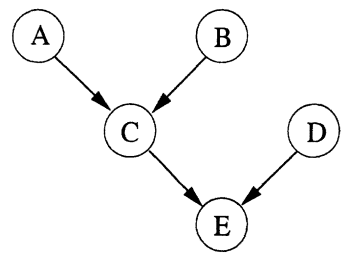

Figure 1. *A DAG describing probabilistic structure.*

DAG is to assert that the joint probability or density function factorises in the form

$$p(a, b, c, d, e) = p(a)p(b)p(c|\, a, b)p(d)p(e|\, c, d), \tag{1}$$

where each term on the right hand side is the density of one of the variables, conditional on the values of its 'parents' in the graph. This factorisation property also implies, and is in fact equivalent to, a collection of conditional independence properties for the joint distribution, namely that each variable is independent of its 'non-descendants', conditional on its 'parents'. For example

$$E \perp\!\!\!\perp (A, B) |\, (C, D),$$

read as "E is independent of (A, B) given C and D"; and

$$D \perp\!\!\!\perp (A, B, C),$$

i.e. D is (marginally) independent of (A, B, C).

Other, more complex, conditional independence properties are also implied by the structure of the graph, and may be determined by inspection of it, using either the 'd-separation' criterion of Verma & Pearl (1990), or the equivalent 'moralisation criterion' of Lauritzen *et al.* (1990)—see Cowell *et al.* (1999), §5.3. The latter, which we shall use repeatedly throughout this paper, proceeds as follows. Suppose we wish to ascertain whether or not $X \perp\!\!\!\perp Y |\, Z$, where X, Y, Z are collections of variables. We perform the following steps:

Ancestral graph Remove from the DAG any node which is neither in $X \cup Y \cup Z$ nor an ancestor of a node in this set, together with any edges in or out of such nodes.

Moralisation Add a line between any two remaining nodes which have a common child, but are not already connected by an arrow. Then remove remaining arrowheads.

Separation In the undirected graph so constructed, look for a path which joins a node in X to one in Y but does not intersect Z. If there is no such path, deduce that $X \perp\!\!\!\perp Y |\, Z$. If there is, then there is some set of variables, and some joint distribution over them which is represented by the DAG, for which it is not true that $X \perp\!\!\!\perp Y |\, Z$.

A graph such as that of Figure 1 is in itself an incomplete description of a full probability model: to

complete the description we need to specify, in addition, each of the conditional probability terms on the right-hand side of (1). Thus the full model is specified partly by the graphical structure, and partly by the requisite associated numerical probability values. We have already seen how the graphical part of the representation, even though incomplete, can convey a great deal of information, and suffice for many investigations and manipulations (such as for describing conditional independence structure, as detailed above).

2.1 Non-random Nodes

Notwithstanding the above, a guiding principle of the graphical modelling approach might be phrased as follows:

> "Seek to represent and manipulate as much as possible of the relevant structure and details of the model by purely graphical means, keeping any additional external information required to a minimum."

In this paper we shall try to follow this precept. (Of course, what is relevant detail for one purpose may be irrelevant clutter for another.)

With the above in mind, we might wish to represent the details of the specified probability distribution internally in the DAG. We can do so by including additional 'parameter nodes'. Thus a fuller description of the problem represented by Figure 1 would be as in Figure 2. Here, for each

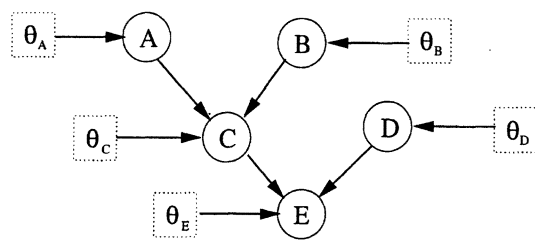

Figure 2. *DAG with parameter nodes.*

original 'domain node' ν, we have added an additional parameter node θ_ν (drawn as a dotted box) as a new parent of ν: θ_ν specifies (explicitly or implicitly) the conditional distributions $p(\nu|\,\mathrm{pa}^0(\nu))$ for ν, given the various possible configurations of its parent domain nodes (where we are here using pa^0 to denote domain parents, i.e. parents in the original graph of Figure 1, rather than in the new graph of Figure 2). We regard the parameters (θ_ν) as non-random, and logically free to vary independently over their respective domains (functional relationships between the parameters could be introduced, but will not be considered here); it is this logical independence, formally expressible as *variation independence* (Dawid, 2001), that is now represented by the absence of arrows between parameter nodes.

If we were to ignore the distinctions between domain and parameter nodes, we might apply the standard DAG moralisation semantics to Figure 2, to deduce properties such as

$$B \perp\!\!\!\perp (E, \theta_A, \theta_D, \theta_E) |\, (A, C, \theta_B, \theta_C). \tag{2}$$

Moreover, even though (2) contains non-random quantities, it can be interpreted in a way that makes sense: *viz.* as asserting that the conditional distribution of B given A, C and E, under a joint distribution θ (specified by $(\theta_A, \theta_B, \theta_C, \theta_D, \theta_E)$) does not depend on the value of E, and moreover only depends on θ through the values of the parameters θ_B and θ_C. In order to justify such a conclusion, we can imagine a fully probabilistic 'Bayesian' DAG, in which the parameters

are modelled as arising randomly and independently according to some joint distribution Π. The conclusion then follows—at least with probability one under Π. To make this argument rigorous would require more abstract considerations, along the lines of the treatment in Dawid (1980). We shall not go into the matter any more deeply here, but merely remark that, whenever we have non-random nodes in a DAG, any conditional independence property that follows from standard DAG moralisation semantics, and still has a meaningful interpretation (as in the example above) in the non-random setting, is fully justified.

3 Causal Interpretation?

When DAG models are constructed, it is often with an explicit or implicit understanding that they represent some sort of 'causal relationships' between the variables. For example, in Figure 1 we might regard A, B and D as externally determined 'exogenous' variables; C as being 'caused' (in a non-deterministic fashion) by A and B jointly; and E as being caused by C and D jointly. However, there is absolutely nothing in the probabilistic semantics by which such graphs are supposed to be interpreted that is relevant to such causal intuitions. For example, a completely general distribution for two variables (A, B) can be equally well represented, either by the DAG of Figure 3, corresponding to the factorisation $p(a, b) = p(a)p(b|a)$, or by that of Figure 4, corresponding

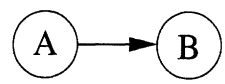

Figure 3. *A before B.*

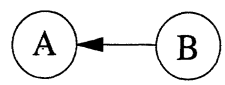

Figure 4. *B before A.*

to $p(a, b) = p(b)p(a|b)$. Although we might feel tempted to regard the former as representing a situation in which A is exogenously determined, and in turn causes B, there is in fact nothing to distinguish it from the latter, in which the rôles of the variables are interchanged. So far as the probabilistic assertions made by the two graphs are concerned, they are identical. The terms in the associated factorisations will of course differ, but can be related to each other using general laws of probability: specifically, Bayes's theorem and the law of total probability.

Example 1. Suppose that A is a binary variable, with

$$\text{pr}(A = 1) = 1 - \text{pr}(A = 2) = p, \tag{3}$$

and that, conditionally on $A = a$ $(a = 1, 2)$, B has the normal distribution

$$B \sim \mathcal{N}(\mu_a, 1) \tag{4}$$

with known means μ_1, μ_2.

Because we have specified $p(a)$ and $p(b|a)$, we might regard the joint probability structure determined by (3) and (4) as naturally associated with the DAG of Figure 3. However, it can just as well be associated with that of Figure 4, by re-expressing it in terms of the marginal distribution of

B, which is the mixture

$$B \sim p\mathcal{N}(\mu_1, 1) + (1-p)\mathcal{N}(\mu_2, 1); \tag{5}$$

and the conditional distributions of A given B, which have the logistic form

$$\text{logit}\{\text{pr}(A = 1 | B = b)\} = \alpha + \beta b, \tag{6}$$

where

$$\begin{aligned} \alpha &= \text{logit}\, p + \mu_2^2 - \mu_1^2, \\ \beta &= \mu_1 - \mu_2. \end{aligned}$$

Neither order of considering the variables is intrinsically more basic than the other (although, for purely pragmatic reasons, we might be more likely to use this particular joint distribution if we felt that "A causes B" rather than "B causes A"; in the latter case, while we might well retain the conditional model (6) for A given B, we would be more likely to have chosen a simpler normal model for the marginal distribution of B, rather than the mixture model (5)). □

Various authors (Verma & Pearl, 1991; Frydenberg, 1990) have developed a theory for deciding when two different DAGs (or, more generally, 'chain graphs', which we shall not consider here) are equivalent, in the sense of representing the same probabilistic or conditional independence structure over the observed variables. Define the *skeleton* of a DAG as its undirected version, arrowheads being omitted; and an *immorality* in a DAG as a configuration of three nodes, say A, B, and C, such that C is a child of each of A and B, but neither of A and B is a child of the other. Then two DAGs represent the same set of probability models if and only if they have the same skeleton and the same set of immoralities. This condition is trivially seen to hold for Figures 3 and 4.

Such equivalences have been used (Spirtes *et al.*, 1993; Glymour *et al.*, 1999) in an attempt to extract purportedly causal properties out of purely probabilistic structure (in particular, conditional independence structure), which can in principle be learned from data. For example, if all equivalent DAGs representing the discovered probabilistic structure contain an arrow pointing from variable X to variable Z, we might consider that X is a (partial) cause of Z, and/or that Z does not contribute to causing X. However, there seems to be no good logical reason why this should be the case, especially since DAG and related graphical models, though often useful, provide a very limited framework for representing conditional independence properties (Studený, 2001). Causal modelling is an intrinsically more complicated enterprise than probability modelling, and it is naïve to believe it can be accomplished so painlessly. If we are to base causal understandings on the structure of graphical representations, these must be supplied with a richer semantics, already embodying causal properties.

4 Influence Diagrams

As in Dawid (2000), our interpretation of the nature of causal understanding will be a decision-theoretic one, essentially reducing our task to the description of the differential effects of various possible external actions or *interventions* on some system. In many problems, such a description can be most clearly represented by means of an *influence diagram*, as originally developed as a concise means of representing and solving a Bayesian decision problem (Howard & Matheson, 1984; Shachter, 1986; Oliver & Smith, 1990). We shall base our development on the more general 'limited memory influence diagrams' (Nilsson & Lauritzen, 2000; Lauritzen & Nilsson, 2001).

An example is given in Figure 5. In this representation, nodes A and C are *decision nodes*, drawn as rectangles, while B, D and E are *random nodes*. The value of a decision variable is determined by outside intervention, rather than being left to arise naturally. As with any such graphical representation, additional information is needed to complete the description of the model.

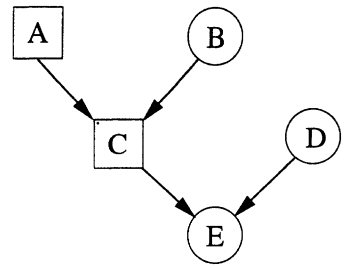

Figure 5. *Influence diagram.*

In an influence diagram, each random node is supposed supplied with its conditional distribution given its parent nodes, just as in a fully probabilistic DAG. So far as a decision node is concerned, its parents represent the information that is supposed available to the decision maker at the point at which that decision has to be made. A full description for such a node would thus require specification of a function of the values of the parent nodes, describing exactly which of the available decisions is to be selected for any particular set of information. We can extend this to allow randomised decision-making, described by specifying conditional probability distributions over the available decisions at a node, given its parents. We then recover exactly the structure of a probabilistic DAG—and this applies, albeit in degenerate form, also in the non-randomised case.

In our example, we thus suppose that we have specified externally the densities $p(b)$, $p(d)$, and $p(e|\, c, d)$; the choice of a value for (or distribution over) the decision at A; and, for each pair of values of A and B, a value for (or distribution over) the decision at C.

Traditionally the extra information required for an influence diagram has been left only partially specified: conditional distributions are given for random nodes, but no description is supplied of the functions or distributions involved at the decision nodes, which are left arbitrary and at the choice of the decision maker. These inputs then determine what we may term the *partial* distribution of random nodes, given decision nodes. In our example, this is

$$p(b, d, e : a, c) = p(b)p(d)p(e|\, c, d).$$

Any given specification, π say, of the functions or distributions at decision nodes constitutes a *decision strategy*. This too specifies a partial distribution, now for the decision nodes given the random nodes: in our example we get

$$\pi(a, c : b, d, e) = \pi(a)\pi(c|\, a, b),$$

where the terms are to be interpreted as probabilities or densities. Once a strategy π has been specified we readily obtain the full joint distribution for all the variables:

$$p_\pi(a, b, c, d, e) = p(b, d, e : a, c)\pi(a, c : b, d, e). \tag{7}$$

It is important to realise that the partial distributions are not, in general, the same as the associated conditional distributions calculated from the full joint distribution.

Most work to date on influence diagrams has focused on ways of manipulating the graph and the probability inputs so as to calculate the overall expected utility $E_{p_\pi}(U)$ of any given strategy π, and, in particular, to identify the optimal strategy, π_{opt}, maximising this value. Here the utility function U is an additional external input, representing the overall value of any configuration of decisions and outcomes. (Certain structural aspects of the utility function may also be represented in the graph, by means of *value nodes*—these will not be needed for our present purposes). Our purpose here however is different: to make maximal use of the powerful and flexible representational ability of an influence diagram, to encode fundamental structural assumptions and to reveal their consequences.

In particular, an influence diagram will typically encode certain assumptions of *stability*. Thus for the example given, no matter how the values at the decision nodes A and C are chosen, whether deterministically or randomly, the distribution of E given C and D is always supposed described by the same given input conditional density $p(e|c,d)$. (Of course, this is merely what our model *says*; whether or not it is a good representation of reality is an entirely separate question, which can only be addressed by performing experiments in which a variety of interventions are made to set the values of the decision nodes).

A related approach to representing causal relationships through influence diagrams has been taken by Heckerman & Shachter (1995). However, their modelling and analysis, based on the axiomatic framework of Savage (1954), is closer to the 'counterfactual' approach that we here take pains to avoid (see Section 10 below).

4.1 Semantics of Influence Diagrams

An influence diagram can be regarded as representing the (partial) probabilistic structure of all the random variables, given the decision variables, with essentially the same semantics as described in Section 2.1. In particular, we can continue to use the moralisation criterion to discover conditional independencies between the variables, both random and decision, which hold universally. We can further enquire as to what features of the joint distribution are unaffected by the choice of decision strategy π. This can be examined by means of a diagram such as Figure 6, which includes additional non-random 'strategy nodes' indicating the choices made for the various components of π: thus π_C, for example, specifies the functional dependence of C on A and B, or more generally the conditional distributions $\pi(c|a,b)$ of C, to be used in the strategy. It is easy to see that, in the moral

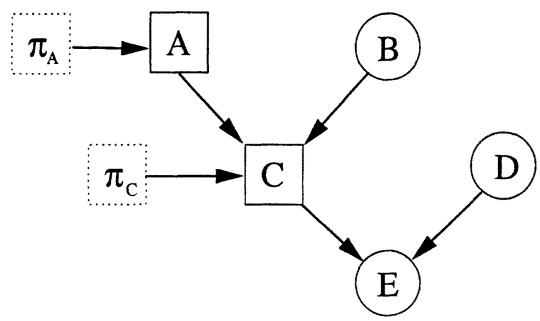

Figure 6. *Influence diagram with strategy nodes.*

graph constructed from Figure 6, the strategy nodes are separated from all other nodes by the set $\Delta := \delta \cup \mathrm{pa}^0(\delta)$, where δ denotes the set of decision nodes (here (A, C)), and $\mathrm{pa}^0(\delta)$ is the set of their domain parents (here B). Applying standard DAG semantics, as justified in Section 2.1, we deduce that

$$S \perp\!\!\!\perp \pi \mid T \cup \Delta, \tag{8}$$

for any sets of random nodes S and T. That is to say, the conditional distribution of S, given T and Δ, will not depend on how decisions are taken in the light of the observed variables. This argument works in general. In our example, $\Delta = (A, B, C)$, and we obtain $(D, E) \perp\!\!\!\perp \pi \mid (A, B, C)$: the conditional distribution of (D, E) given (A, B, C), which, from the original influence diagram, is the same as that of (D, E) given C alone (since $(D, E) \perp\!\!\!\perp (A, B) \mid C$), is unaffected by the strategy π used, and is thus determined solely by the (externally specified) initial probabilistic inputs p—an obvious property in this case.

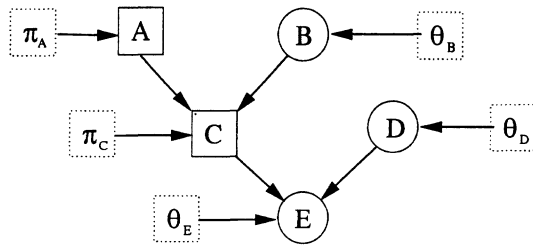

Figure 7. *Extended influence diagram.*

A still more complete graphical representation of this problem is given by the *extended influence diagram* of Figure 7, incorporating domain variables, parameter nodes, and strategy nodes. From this we can deduce, e.g.:

$$B \perp\!\!\!\perp (E, \pi_A, \theta_D, \theta_E) \,|\, (A, C, \theta_B, \pi_C), \tag{9}$$

which says that the conditional distribution of B, given A, C and E, does not depend on the value of E, nor on the way in which A was chosen, nor on the distributions $p(D)$ and $p(E|\,C, D)$; although it may depend on the distribution $p(B)$ and the strategy used to select C in the light of A and B. (Note that this is essentially the same as (2), since the associated DAGs have the same structure, but with different interpretations of the added nodes).

4.2 *Causal Enquiries*

Armed with an influence diagram representation of a causal problem, we can interrogate it to extract answers to questions about the 'effects of causes' (Dawid, 2000). By this we mean queries of the form: "If I were to intervene in the system in such and such a way"—described by a suitable specification, deterministic or randomised, for the overall strategy π— "and observe that $A = a$, what uncertainty would I then have about the ensuing values of the random variables B?". Such a query can be answered by extracting, from the joint distribution p_π over all variables specified by (7), the conditional distribution for B given $A = a$.

For example, consider the influence diagram of Figure 8, where L represents a pre-treatment characteristic of a patient, T represents the treatment applied (*active*, coded 1, or *control*, coded 0), and Y is the response variable.

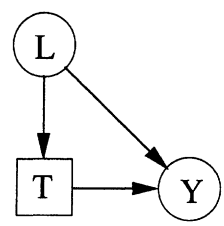

Figure 8. *Treatment comparison.*

The characteristic L has a specified distribution $p(l)$. The arrow from L to T means that L can be observed before treatment is assigned. The experimenter is free to choose an arbitrary strategy for such treatment assignment, represented by a probability distribution $\pi(t|\,l)$. The other two arrows indicate an external specification for the conditional distribution $p(y|\,l, t)$ of the response Y, this being supposed to depend on both the patient characteristic L and the treatment T applied. The

overall distribution of (L, T, Y), when strategy π is applied, is then given by

$$p_\pi(l, t, y) = p(l)\,\pi(t|l)\,p(y|t, l).$$

In the simplest case, we could just consider the two trivial treatment strategies, π_1 and π_0, where π_i involves always giving treatment $T = i$, ignoring the patient characteristic L. Then (taking account of the fact that T is constrained to equal i) we obtain: $p_i(l, y) := p_{\pi_i}(l, y) = p(l)\,p(y|i, l)$. We could then calculate, in either case, the ensuing marginal distribution $p_i(y)$ for the response Y when treatment $T = i$ is given. Measuring the '(overall) effect of treatment' essentially consists in making a suitable comparison of the two distributions $p_1(y)$ and $p_0(y)$. We might also look more closely at a subgroup of patients, say those having $L \in W$ for some set W: the appropriate comparison would then be between $p_1(y|\, L \in W)$ and $p_0(y|\, L \in W)$, each of these terms being easily calculable from $p_i(l, y)$.

We could similarly calculate and compare the effects of other, more complex, treatment strategies π, each specifying a way of assigning T in the light of L. Again, this could be done on an overall population basis (focusing on the ensuing marginal response distribution $p_\pi(y)$), or within subpopulations (focusing on $p_\pi(y|\, L \in W)$).

5 Intervention DAGs

Pearl's original development (Pearl, 1995) of a causal semantics for DAGs was motivated by the important realisation (following Spirtes *et al.* (1993)) that, as well as modelling the joint distribution of a system of variables as these might arise naturally in the world, we must also, quite independently, consider how this distribution might be affected by outside interventions. We are particularly interested in interventions which force one or more of the variables in the system to take on some externally assigned value or values, rather than arising naturally from within the system. If we can say how the distribution of the remaining variables is affected by such 'setting' of the values for some of them, we have essentially described the causal influence of the 'set' variables on the others. In full generality there is no reason to expect any relationship whatsoever, let alone a simple one, between the distributions describing 'natural' and interventional scenarios, but it will often seem appropriate to make links between these. Pearl thus extends the semantics by which a DAG is to be interpreted so as to express, simultaneously, the joint distributions of the variables under a variety of scenarios for setting the values for some of them. For example, in Figure 1 we might set the value of C, to c_0 say. The effect of this on the joint distribution, as expressed by means of the factorisation (1), is supposed to be confined to the replacement of the term $p(c|\, a, b)$, associated with the set variable C, by the one-point distribution on the assigned value c_0. In particular, the distributions of variables which are non-descendants of C (namely A, B and D) are entirely unaffected; while that of E given D becomes $p(e|\, c_0, d)$. Thus the joint distribution of (A, B, D, E) when C is set to c_0 is supposed given by

$$p(a, b, d, e\|\, c_0) = p(a)p(b)p(d)p(e|\, c_0, d). \tag{10}$$

Setting of several variables simultaneously is handled similarly. The use of the symbol $\|$, rather than the normal conditioning symbol $|$, to represent this 'conditioning by intervention' follows Lauritzen (2000); alternative notations that have been used, by Pearl and others, are $p(a, b, d|\hat{c}, \hat{e})$ and $p(a, b, d\|\, \mathrm{do}(C = c), \mathrm{do}(E = e))$.

In Pearl's account, it is supposed that, for each variable in the system, an intervention is possible that would set that variable to any of its possible values, and that in all such cases the effect of that intervention would be as exemplified above. We call a DAG equipped with these modified semantics an *intervention DAG* (*caution:* Lauritzen (2000) uses this description for what we term an *augmented DAG* below).

Example 2. With this new causal interpretation, Figures 3 and 4, considered as intervention DAGs, no longer represent the same model. Suppose we use Figure 3, with the distributional assumptions (3) and (4). Then, on setting $A = 1$, the implied distribution for B is its conditional, $B \sim \mathcal{N}(\mu_1, 1)$. However, if we use Figure 4, with the probabilistically equivalent structure (5) and (6), when we set $A = 1$ we find that B retains its marginal distribution (5). In the former model A 'affects' B, but not in the latter.

In particular, we will never be able to decide which of these distinct causal models is appropriate merely by examining observational data: we need to perform experiments in which we do in fact intervene to set the value of A, and see what effect that intervention has on B. (And it may of course turn out that neither of the above choices represents the actual distribution of B in the experiment, in which case the intervention DAG semantics are simply inadequate to describe the actual causal relationships). □

In general, an intervention DAG provides a concise and handy way of encapsulating and displaying certain assumptions about the effects of setting some variables on the others in the system. But it is as always a matter of experiment, beyond mere passive observation, to decide whether any such assumptions about the behaviour of the real-world system are appropriate. Furthermore, there may be various real-world mechanisms by which the value of a variable may be set: setting a patient's treatment to 'none' by (a) withholding it from him, (b) wiring his jaw shut, or (c) killing him are all very different interventions, with different effects, and we must be very clear as to which mechanism we are considering. An intervention DAG model can be justified only to the extent that it fits the behaviour of the world in the setting to which it is intended to apply.

5.1 *Conditioning by Observation and Conditioning by Intervention*

If, without making any interventions, we were to observe the (naturally arising) values of C and E, we could calculate the resulting uncertainty about the remaining unobserved variables, A, B and D, by the usual rules of the probability calculus applied to the joint distribution $p(a, b, c, d, e)$, yielding

$$p(a, b, d \mid c, e) = \frac{p(a, b, c, d, e)}{\sum_{a', b', d'} p(a', b', c, d', e)}, \tag{11}$$

with $p(a, b, c, d, e)$ specified by (1). This classical operation is called 'conditioning by observation'. The conditional distribution (11) could be calculated using algorithms and software for probability propagation (Cowell *et al.*, 1999), based on the DAG structure of Figure 1, on entering the 'evidence' $C = c$, $E = e$. Although barely necessary in this simple case, these methods can greatly streamline the relevant calculations in more complex problems.

Alternatively, suppose that we intervened in the system, setting $C = c$ and $E = e$. This conditioning by intervention would result in a different joint distribution for (A, B, D), given, according to the intervention semantics, by

$$\begin{aligned} p(a, b, d \parallel c, e) &= p(a)\, p(b)\, p(d) \\ &= \frac{p(a, b, c, d, e)}{p(c \mid a, b)\, p(e \mid c, d)}. \end{aligned} \tag{12}$$

It is important to notice that (in contradistinction to (11)) the appropriate formula (here (12)) expressing the results of such conditioning by intervention depends not only on the joint density p, but also on the assumed structure of the relevant intervention DAG. In particular, DAGs that are equivalent as probabilistic DAGs but not as intervention DAGs will yield different intervention formulae.

In Section 6.1 below we shall consider an alternative approach to expressing and conducting these two kinds of conditioning, which makes both the distinctions and the connections between them very clear, while at the same time avoiding the need to introduce any new concepts or notation.

6 Augmented DAGs

Let X be a random variable naturally taking values in $\mathcal{X}$, say. Suppose that it is possible to intervene somehow, so forcing X to take on some chosen value $x \in \mathcal{X}$. We now represent this explicitly by introducing an *intervention variable* F_X, which is a special kind of decision variable. The state space of F_X is constructed by augmenting $\mathcal{X}$ with an additional state $\emptyset$. Intuitively, if $F_X = \emptyset$, X is allowed to arise 'naturally'; while any value $x \in \mathcal{X}$ for F_X indicates an intervention to set the value of X to x. In particular, the conditional distribution of X, given $F_X = x \in \mathcal{X}$ (and any other information) will be degenerate at the value x. The conditional distribution of X given $F_X = \emptyset$, and other information H, will be just its 'natural' distribution, given H.

In cases of interest there will be other variables in the problem. Again, formal conditioning on $F_X = \emptyset$ is deemed to have no effect. However, the global effect of an intervention, $F_X = x$ (beyond its immediate effect at setting $X = x$) needs further specification. There can be no magic solution to this problem: what is appropriate must depend on the context and meaning of the variables (including the way in which intervention is effected), and on what assumptions appear reasonable in the circumstances.

The semantics of an intervention DAG, as described in Section 5 above, offer one (out of many) possible ways of modelling global response to intervention. An alternative, and more versatile, explicit representation of the assumptions embodied in such a model is by means of an 'augmented DAG': a special kind of influence diagram, which, for each existing domain node X of the original probabilistic DAG, explicitly adds a new intervention node F_X, as described above, as a decision node parent of X (Spirtes *et al.* (1993); Pearl (2000), §3.2.2; Lauritzen (2000)). Formally, let $\text{pa}^0(X)$ be the set of domain parents of X. In the augmented DAG, the conditional distribution of X, given any configuration $\boldsymbol{y}$ of $\text{pa}^0(X)$ and the value x of F_X, is taken to be the same as the original distribution for X given $\text{pa}^0(X) = \boldsymbol{y}$ if $x = \emptyset$; but otherwise puts all its probability mass on the value x for X. The augmented DAGs corresponding to Figures 3 and 4 (regarded as intervention DAGs) are given by the influence diagrams of Figures 9 and 10. An immediate payoff of the explicit display of the intervention variables in the DAGS is that we can see directly that these graphs do not represent equivalent causal assumptions since, although they have the same skeleton, they have different immoralities.

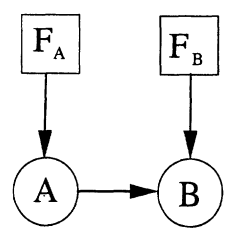

Figure 9. *Augmented DAG: A causes B.*

The augmented DAG corresponding to Figure 1 is given by Figure 11. We can read off the graph that, for example, conditional on its parents (A, B, F_C), C is conditionally independent of its other non-descendant nodes (D, F_A, F_B, F_D, F_E). In particular, if we fix F_C at $\emptyset$ (and drop this conditioning from the notation), the 'natural' conditional distribution of C given A and B is not

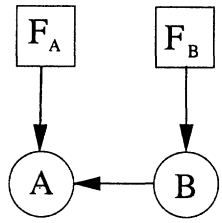

Figure 10. *Augmented DAG: B causes A.*

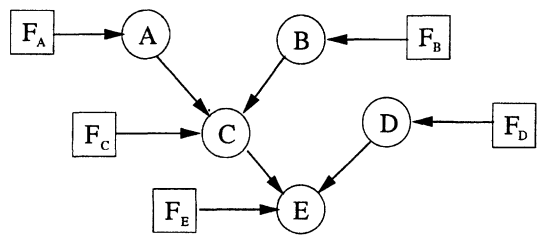

Figure 11. *Augmented DAG.*

further affected by additional conditioning on the value of D, *nor by whether or not any or all of A, B, D or E arose naturally or by intervention.* Similar properties hold for any other domain node in place of C. In particular the conditional distribution for a node, given its domain parents, when it is allowed to 'arise naturally', remains unchanged when its parents are set by intervention. That is, the augmented DAG *explicitly* encodes the assumptions that are only implicit in the intervention interpretation of a probabilistic DAG; and further makes it possible to read off their implications directly.

An augmented DAG model is thus completely equivalent to an intervention DAG model; but the augmented DAG representation is greatly to be preferred, since the assumed structure is explicitly represented in the diagram, its implications can easily be read off the graph, and equivalence of different structures is easy to check. (Note however that the requirement that $p(X|F_X = x, \mathrm{pa}^0(X))$ be degenerate at x ($x \neq \emptyset$) is not explicitly displayed in the graph, and still has to be introduced as an implicit, externally specified, constraint). Such a representation also easily allows for the incorporation of additional flexibility, by varying specific details while retaining the general structure of an influence diagram. For example, we could restrict the possibility of intervention to a subset of the nodes, or to a subset of the state-space of a node; or modify the effect of intervention, for example so as to determine the value of the associated node in some externally decided random way; or allow dependence between the decision nodes, or have a decision node dependent on ancestral nodes in the graph, or make some domain nodes pure decision nodes. All such modifications are easily made explicit and manipulated by means of suitable influence diagrams.

6.1 Observation and Intervention: Reprise

Because an augmented DAG represents each variable by two nodes, X and F_X, it makes it possible to make very clear the distinction between observation and intervention.

In the augmented DAG model of Figure 11, conditioning on the observation $C = c$, $E = e$ can be effected by ordinary conditioning on $C = c$, $E = e$, $F_v = \emptyset$ (all v)—in a Probabilistic Expert System shell such as HUGIN, one would enter this as evidence into a model whose graphical structure is given by the augmented DAG (and whose probability tables at intervention nodes are

arbitrary, for example being set to be identically 1). In contrast, conditioning by intervention on $C = c$, $E = e$ is equivalent to ordinary conditioning on ('inserting evidence') $F_C = c$, $F_E = e$, $F_v = \emptyset$ (all $v \neq C, E$). (Because of the way we have modelled the dependence of C on (F_C, A, B), this automatically implies that $C = c$, and similarly $E = e$—it is not important whether or not this redundant information is explicitly regarded as part of the evidence). Whether we are in the observational or the interventional case, entering and propagating the requisite evidence, to find the associated ordinary conditional distribution for the remaining variables, automatically leads to the correct answer.

The augmented DAG makes it very clear why we get different distributions depending on whether we condition or intervene at a node, since we are asking different questions. But both questions are clearly represented (and just as easily distinguished) in the augmented DAG, and each can be answered by exactly the same machinery—no new concepts or methods are required.

There is no problem in combining the two types of conditioning: the uncertainty resulting from setting $A = a$ and observing $E = e$ is obtained by entering evidence $A = a$, $F_E = e$, $F_v = \emptyset$ (all $v \neq E$). To be more precise, this represents the uncertainty of an observer (who may or may not be the experimenter) who has observed that an intervention was made to set $A = a$, and further observes that $E = e$—we continue to suppose that, as represented in Figure 11, when the experimenter sets A no information about the other variables is available to him. We remark that, according to the initial formulation of Lauritzen (2000), the order in which one performs observation and intervention might matter—something that clearly cannot hold in our interpretation. Lauritzen avoids this ambiguity by deeming that intervention always precedes observation, which brings his formulation into agreement with ours—and so with his own graphical elaboration of the problem, which is essentially the same as ours.

We can further interrogate the augmented DAG to investigate what aspects of an interventional conditional distribution are the same as in the observational case. Thus applying the standard moralisation criterion to Figure 11, we can read off the property:

$$B \perp\!\!\!\perp (D, E, F_A, F_C, F_D, F_E) \mid (A, C, F_B), \tag{13}$$

and so deduce that the conditional distribution of B, given all the other variables, will be the same no matter whether those other variables arose naturally or by intervention (and in addition will not depend on the realised values at D and E).

More generally let S and T be any sets of domain variables in an augmented DAG. Then from the augmented graph we readily obtain

$$S \perp\!\!\!\perp F_T \mid T \cup \mathrm{pa}^0(T), \tag{14}$$

where we have defined $F_T := \{F_v : v \in T\}$ and $\mathrm{pa}^0(T) := \bigcup\{\mathrm{pa}^0(v) : v \in T\}$. That is, intervening to set the values for the variables in T has no effect on the conditional distribution of S given $T \cup \mathrm{pa}^0(T)$. (Note that there is a close formal correspondence with the similar use of the extended influence diagram of Figure 7 to derive properties such as (8) and (9).)

6.2 The Back-door Criterion

Let T, R, C be sets of domain variable, which might be interpreted respectively as Treatments, Responses, and Covariates. We also have intervention variables F_T associated with the domain variables in T, as described above (any other intervention variables will be ignored, being implicitly fixed at $\emptyset$).

Suppose we are able to assert the following two properties:

$$C \perp\!\!\!\perp F_T \tag{15}$$

$$R \perp\!\!\!\perp F_T \mid C \cup T. \tag{16}$$

Condition (15) looks simple, but care is needed in its interpretation. According to (15), uncertainty about C should be the same, whether we let the variables in T arise naturally, or intervene to set them at arbitrary values. In particular, conditional on intervening to set T, the distribution of C is unaffected by the specific values set. Intuitively, T has 'no causal effect' on C (although we should interpret all such phrases with circumspection). A typical example is when C consists of pre-treatment measurements. However, this is not enough to ensure (15). For example, if the choice of the value to set for T were allowed to depend on the observed values for C, or for precursors of C, (15) would typically fail. Over and above this required irrelevance of the specific interventional setting for T, (15) also incorporates the important additional requirement that this common interventional distribution for C be the same as its distribution when T is allowed to 'arise naturally'.

As for (16), this requires that, for any specific values of the treatments T, the conditional distribution of the responses R, given the covariates C, will be the same, no matter whether the treatments were determined naturally or by intervention. In this case we may call C a *sufficient set of covariates* (for the effect of T on R). This definition of sufficiency is weaker than that of Lauritzen (2000), which is expressed in terms of additional unobserved variables.

Typically (though not necessarily), in order for condition (16) to be reasonable we would expect the set C to be large. Intuitively we want C to account for the totality of those individual characteristics that are relevant to the process whereby T generates R. However, (15) acts as a constraint on just how much we can throw into C. In any case, whatever intuitions we may use to select C, the essential point is to ensure that we are satisfied as to the appropriateness of (15) and (16), as interpreted above.

When both (15) and (16) hold, we have, for interventions t at T:

$$\begin{aligned} p(r|\, F_T = t) &= \sum_c p(r|\, F_T = t, c)\, p(c|\, F_T = t) \\ &= \sum_c p(r|\, F_T = \emptyset, T = t, c)\, p(c|\, F_T = \emptyset). \end{aligned} \tag{17}$$

To obtain (17), we have used $p(r|\, F_T = t, c) = p(r|\, F_T = t, T = t, c)$ (since $F_T = t \Rightarrow T = t$), which is the same as $p(r|\, F_T = \emptyset, T = t, c)$ by (16); and $p(c|\, F_T = t) = p(c|\, F_T = \emptyset)$, as follows immediately from (15). It is noteworthy that (17) expresses the effect on R of intervention at T purely in terms of ingredients of the observational regime, in which $F_T = \emptyset$. That is, in the presence of (15) and (16), we can estimate the 'causal effect', on R, of setting T to t, from an observational study in which (R, C, T) are all observed. (To be more accurate, for such estimability we also require an additional 'positivity condition': under the observational regime $F_T = \emptyset$, we want $p(T = t|\, C = c) > 0$ whenever $p(C = c) > 0$).

Thus far we have made no use of graphical representations—merely conditional independence properties. However, if the variables involved in (15) and (16) can be represented in a suitable augmented DAG, then it will generally be straightforward to check whether those conditions are satisfied, using standard moralisation semantics.

In this case it is also possible to re-express these conditions in terms of separation conditions on the (unaugmented) intervention DAG. Thus by considering the moralisation criterion in the augmented DAG it is easily seen that (15) will hold if and only if

$$\mathrm{an}^0(C) \cap T = \emptyset; \tag{18}$$

while (16) is equivalent to

$$R \perp\!\!\!\perp \mathrm{pa}^0(T) | C \cup T, \tag{19}$$

where an^0, pa^0 denote ancestors and parents in the unaugmented DAG. However, there is no particular advantage to such re-expression, which also obscures the clear and simple meaning expressed in (15) and (16). It is easier and more appropriate to query the augmented graph directly to check (15) and (16).

Pearl (2000), Theorem 3.3.2, derives (17) using (18), together with a re-expression of (16), in the unaugmented DAG, in terms of a variation on his 'd-separation' criterion that only examines trails leaving T by way of $\mathrm{pa}^0(T)$ (hence his description of these graphical conditions as 'the back-door criterion'). For an analysis of this problem much in line with our own development, see also Lauritzen (2000). But again, this reformulation only replaces a simple condition, based on standard DAG semantics, with a more complex (and less readily generalisable) one. It is better to make *explicit* use of augmented DAGs (or, more generally, influence diagrams), so avoiding completely the need for any special calculus for describing and handling interventions.

6.3 Estimability

Properties (15) and (16) are by no means necessary for it to be possible for one to estimate interventional response distributions from observational data—although under other conditions (17) need not be the appropriate formula. The 'front-door' formula (Pearl, 1995) and the 'G-computation formula' (Robins, 1986, 1987, 1997) supply alternative estimation formulae, that are appropriate under different circumstances—Pearl & Robins (1995) give some helpful graphical procedures for identifying when and how the G-computation formula may be applied. Fairly general sufficient conditions for estimability are given by Galles & Pearl (1995), but it is not known whether or not these are also necessary. In all cases, verification of the appropriate formula can be performed by manipulations similar to those above whereby (17) was derived from (15) and (16).

The general estimability problem can be expressed as follows. Suppose that we can observe (T, C, R) under natural conditions, and can also apply interventions F_T to set T. There may be other, unobserved, variables U in the problem. Let the joint distributions of (T, C, R, U), whether under observational or experimental conditions, be fully determined by a parameter θ, and the observational joint distribution for the observable variables (T, C, R) by θ^*, a function of θ. Then only θ^* is estimable from the observational data, and we thus require that the interventional distribution for R, when T is set, is in fact fully determined by θ^*.

This problem can also be expressed in terms of conditional independence. The subparameter θ^* is defined by the property that, for any function ψ of θ,

$$(T, C, R) \perp\!\!\!\perp \theta \mid \psi, F_T = \emptyset \tag{20}$$

if and only if θ^* is a function of ψ. From this, together with whatever other structure we impose, we wish to derive:

$$R \perp\!\!\!\perp \theta \mid \theta^*, F_T. \tag{21}$$

(Note: When $F_T = \emptyset$, (21) follows automatically from (20). This requirement thus has significance only for the interventional case $F_T = t, t \neq \emptyset$). In particular, whenever the interventional distribution is fully determined by the observational distribution θ (as holds if we have an augmented DAG model, for example), and there are no additional variables U, (21) must hold.

The above conditional independence formulation might be thought to open up an approach to finding necessary and sufficient conditions for estimability in, say, augmented DAG models with added parameter nodes. However, its usefulness is limited because typically it will not be easy to represent θ^* naturally in such a graph. There may nonetheless be scope for further investigations along these lines.

7 'No Unobserved Confounders'

The simple augmented DAG of Figure 12 describes a situation in which

$$Y \perp\!\!\!\perp F_T \mid T. \tag{22}$$

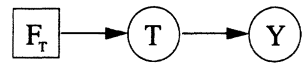

Figure 12. *No unobserved confounders.*

Regarding T as a treatment and Y as a response, this asserts that the distribution of Y given T will be the same, whether the variables arise naturally or T is set by intervention. In particular, causal enquiries about the 'effect of T on Y', which we regard as relating to (comparisons between) the distributions of Y given $F_T = t$ for various settings t, can be addressed directly from observational data in this situation.

Now we may not initially be willing to make the bold assumption (22). However, we might be happy to assume that, for a certain additional variable U, the joint structure of (F_T, T, U, Y) is described by the DAG of Figure 13. The properties represented in this DAG are:

$$U \perp\!\!\!\perp F_T \tag{23}$$

$$Y \perp\!\!\!\perp F_T \,|\, (U, T). \tag{24}$$

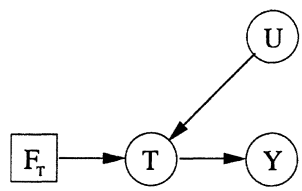

Figure 13. *A potential confounder.*

Note that these are the same as (15) and (16), with U in place of C and Y in place of R, and are subject to the same comments on their interpretation. In particular, condition (24) requires U be a sufficient covariate for the effect of T on Y. If we could observe U, then we could apply the 'back-door formula' (17) to estimate the causal effect of T on Y from observational data. However, without observing U we can in general no longer do this. In particular, on marginalising out over U it will generally *not* be the case that (22) holds (it is *not* a consequence of the moralisation criterion in Figure 13). When we have (23) and (24), we may call U a *potential confounder*, becoming a *non-confounder* if in fact (22) does hold, and otherwise a *confounder*. It is a great advantage of our language of intervention variables and conditional independence relations (whether or not these are expressed graphically) that it enables such properties to be described with a much greater degree of clarity than is usually attained. Moreover, our definition (22) of non-confounding is absolute, rather than relative to a specific model (as is the case, for example, for Definition 6.2.1 of Pearl (2000)).

When will a potential confounder in fact be a non-confounder? Two separate sufficient conditions are given in the following.

LEMMA 7.1. *Suppose that, in addition to (23) and (24),* either *of the following conditions holds:*

$$Y \perp\!\!\!\perp U \,|\, T \tag{25}$$

$$T \perp\!\!\!\perp U \,|\, F_T. \tag{26}$$

Then $Y \perp\!\!\!\perp F_T \,|\, T$.

Proof. These results can be shown algebraically, by manipulating conditional independence properties as in Dawid (1979). Alternatively we can use graphical properties. Thus under (25) the arrow b in Figure 13 is absent, and we can represent the problem by Figure 14. The property $Y \perp\!\!\!\perp F_T \,|\, T$ is

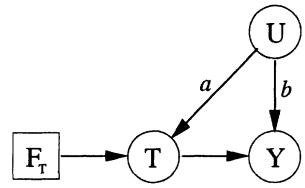

Figure 14. *Irrelevance.*

then easily verified using the moralisation criterion. If, instead, we assume (26), we can drop arrow a from Figure 13, thus obtaining Figure 15. We can again read off the desired property $Y \perp\!\!\!\perp F_T \mid T$. □

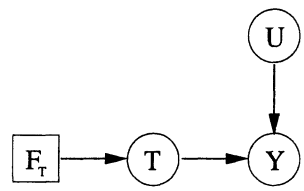

Figure 15. *Randomisation.*

Condition (25) says that the conditional distribution of response Y given treatment T and covariate U (which, by (24) has already been assumed unaffected by whether treatment arose naturally or by intervention), is not in fact affected by the value of U. It is indeed obvious then that we can entirely ignore U, *even if we were to observe it*; and that the conditional distribution of Y given treatment T (which is the same as that given both T and U) is unaffected by whether treatment arose naturally or by intervention. In this case we could call U *irrelevant* for Y. In a sense condition (25) represents a rather trivial way of obtaining (22), and in cases where it holds we might have been just as willing to assume (22) directly.

More interesting is (26). Because the conditional distribution of T given F_T is degenerate whenever $F_T \neq \emptyset$, this condition is only non-trivial for the case $F_T = \emptyset$. It is thus equivalent to requiring that T and U be independent when both arise naturally—in which case we can say that U is *unassociated* with treatment. Since this would occur in an experiment with completely randomised assignment of T (at any rate when U is a pre-treatment quantity), we may describe (26) as the 'randomisation condition'. Under this condition, even though U might not be irrelevant, and hence would be useful were it to be observed, when it is not available it does not bias the assessment of the distribution of response given only treatment, whose observational and interventional versions will still be identical.

To justify proceeding without taking any account of confounding, it is necessary that, for *any* potential confounder U, it in fact turns out to be a non-confounder—the situation described by the phrase 'no unobserved confounders'. Now this initially might appear to require explicit consideration of all possible potential confounders—a task that would generally be impossible in practice (and perhaps even in principle). However, our analysis shows that this is not required. So long as we can identify *just one* potential confounder U that is a non-confounder, we can deduce (22), and hence will be justified in interpreting the observational distributions of Y given T as equally meaningful under intervention. In particular, it will be enough to have either (25) or (26) hold for *some* potential confounder. Then any other potential confounder must also be a non-confounder (although it might not satisfy either (25) or (26), since although these conditions are sufficient they are not necessary).

We recall here the discussion after (15) and (16), which is equally relevant to (23) and (24). In particular, for it to be reasonable to assume (24) we will typically want U to contain a large amount

of information. Notwithstanding this, we emphasise once again that it is not required that U should comprise *all* unobserved potential confounders (whatever this phrase might mean).

Finally we note that it may be possible to demonstrate (22) under other conditions than those considered above. As a simple example, consider the 'combination' structure displayed in Figure 16. Here again, $U = (U_1, U_2)$ is a potential confounder that is in fact a non-confounder, although now neither (25) nor (26) holds. More generally, given a suitable model involving specified relations between Y, T, F_T and some additional variables, it will sometimes be possible to demonstrate (22), either directly, through manipulation of conditional independence properties, or indirectly, by applying the moralisation criterion to a suitable graphical representation.

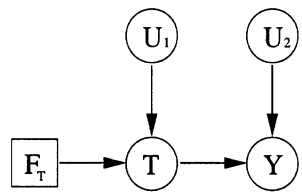

Figure 16. *Combination.*

8 A More Complex Problem

Here we demonstrate how the general approach above can be applied to a more elaborate version of the problem of confounding, taking us beyond simple augmented DAGs to more general influence diagram representations. See Dawid (2002) for a related analysis of the problem of 'partial compliance'. Still more complex cases arise in problems that involve progression through a number of stages of observation and action—these are treated in Dawid *et al.* (2002).

Consider a situation represented by the extended influence diagram of Figure 17. The observable

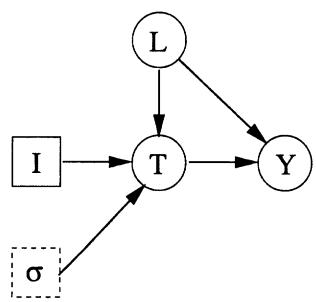

Figure 17. *No unobserved confounders.*

variables T, Y, L denote, respectively, 'treatment', 'response' and '*observed* covariate' (possibly multivariate). Node I is a 'switch', with values N and E, indicating whether the data are generated under 'Natural' or 'Experimental' conditions. When Experimenter is in control, he first observes L, and then chooses (possibly with randomisation) the treatment to be assigned. Node σ holds details of this treatment assignment strategy, in the form of a specified distribution $p(t \mid l)$—more fully described as $p(t \mid I = E, l, \sigma)$; σ is irrelevant to T when $I = N$. We have omitted from the diagram nodes θ_L and θ_Y, describing the distributions (under either regime) of L and of Y given (T, L), and θ_T, describing the distribution of T given L when Nature is in control. These are unknown, but can be estimated from observational studies.

The significant missing arrows in Figure 17 encode the following assumptions (compare (23) and (24)):

$$L \perp\!\!\!\perp (I, \sigma). \tag{27}$$

$$Y \perp\!\!\!\perp (I, \sigma) | (L, T). \tag{28}$$

Assumption (27) says that the observed covariate L does not affect, and is not affected by, either the choice of Natural or Experimental conditions, or the experimental treatment assignment strategy. Assumption (28) says that L is a sufficient covariate for the effect of T on Y, in that the distribution of Y given T and L is the same under both Experimental and Natural conditions.

8.1 Causal Queries

Here we show that, when we can assume that (27) and (28) hold, we are in a position to answer causal queries from purely observational data on Y, T and L. We can regard (27) and (28) as defining the appropriate property of 'no confounding' for this problem.

Suppose then that an experiment has been performed to observe Y, T and L under Natural conditions, so that all aspects of the joint distribution $p(y, t, l | I = N)$ are estimable. However, we wish to assess the distributions of responses under experimental conditions, for some specified treatment assignment strategy σ. We could, if wished, further specialise our query to patients with particular values for the covariate L, or, as an intermediate stage, to patients with a specified value for some function (itself possibly multivariate) B of L. That is, the *target* of our inference is of the form

$$p(y | I = E, \sigma, B = b). \tag{29}$$

Now (29) can be expressed as:

$$\sum_l \sum_t p(y | l, t, I = E, \sigma) \times p(t | I = E, l, \sigma) \times p(l | B = b, I = E, \sigma). \tag{30}$$

By (28), the first term in (30) is $p(y | l, t) = p(y | l, t, I = N)$, and is thus estimable from the data. Since B is a function of L, the third term can be calculated from the distribution of L given $(I = E, \sigma)$, which, by (27), is just the 'natural' distribution of L, given $I = N$, and is likewise estimable. As for the second term, this is fully determined by the specified treatment allocation strategy σ. It follows that (29) is estimable from the data, being given by the formula

$$p(y | I = E, \sigma, B = b) = \sum_l \sum_t p(y | l, t) \times p(t | I = E, l, \sigma) \times p(l | B = b). \tag{31}$$

In fact, the above analysis would still hold if the assumption (27) were weakened to:

$$L \perp\!\!\!\perp (I, \sigma) | B, \tag{32}$$

thus allowing a selection process whereby the inclusion of a case into the observational study is allowed to depend on its values for the reduced covariates B appearing in (29).

Equation (31) can be regarded as a simple version of the 'G-computation formula' (Robins 1986, 1987). A general treatment of that formula from the point of view of the theory outlined here can be found in Dawid *et al.* (2002).

Important special cases of formula (31) arise:

(a) When σ is non-randomised.
(b) When $T \perp\!\!\!\perp L | I = E, \sigma$.
(c) When L is trivial.
(d) When B is trivial.
(e) When $B = L$.

In particular, when (a) and (b) hold, we are considering a 'uniform' strategy σ that always sets T to some fixed value t_0. In this case (31) reduces to:

$$\sum_l p(y|l, t_0) \times p(l|B = b). \tag{33}$$

If, further, (d) holds, so that we are looking at the 'average effect', in the population, of setting $T = t_0$, we get

$$\sum_l p(y|l, t_0) \times p(l), \tag{34}$$

which is just the 'back-door formula' (17).

8.2 *A Potential Confounder*

Whereas the validity of assumption (27), or its weaker version (32), will usually be easy to assess, the sufficiency assumption (28), for the observed covariate L, may require more courage. Once again, suppose that we are not initially willing to take this step, but are instead willing to assume that, for some particular set U of further covariates, not fully observed, we could recover (27) and (28) if we expanded L to (L, U):

$$(L, U) \perp\!\!\!\perp (I, \sigma). \tag{35}$$

$$Y \perp\!\!\!\perp (I, \sigma)|\,(L, T, U). \tag{36}$$

In particular, (L, U) are supposed jointly sufficient.

Properties (35) and (36) are embodied in Figure 18. They do not, in themselves, imply the validity

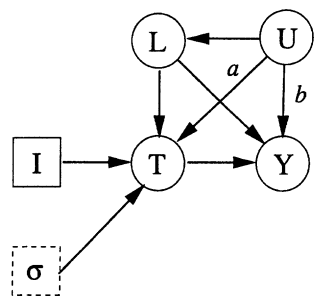

Figure 18. *A potential confounder.*

of (27) and (28), so that U is a potential confounder. The direction of the arrow we have drawn between L and U is of no consequence, but merely indicates the possibility of dependence between the observed and the unobserved covariates (a better representation would be in terms of a *chain graph*, rather than a DAG: Dawid, 2002). By redefining U to be (U, L), we could without loss of generality assume that L is in fact a function of U.

8.3 *A Non-Confounder*

If we are willing to assume (35) and (36), under what further conditions are we able to deduce the desired properties (27) and (28)? It is easy to see that (27) follows automatically. As for (28), it may again be verified (*e.g.* by applying the moralisation criterion to the resulting graph) that this will follow if *either* of the arrows marked a and b in Figure 18 is absent. These missing arrows

correspond, respectively, to the following conditional independence properties:

Irrelevance (b missing):

$$Y \perp\!\!\!\perp U \mid (L, T). \tag{37}$$

Stratified randomisation (a missing):

$$T \perp\!\!\!\perp U \mid (I, L, \sigma). \tag{38}$$

Condition (37) requires that the conditional distribution of response Y, as it depends on treatment T and the sufficient covariate (L, U), is in fact completely determined by T and L alone (U is irrelevant). Since that conditional distribution was assumed unaffected by intervention, this immediately implies that L is a sufficient covariate.

Condition (38) requires

$$T \perp\!\!\!\perp U \mid (I = E, L, \sigma). \tag{39}$$

This will typically hold, since L is the full extent of the covariate information that will be available to the experimenter as input to his treatment strategy σ. In addition, and more restrictively, (38) requires:

$$T \perp\!\!\!\perp U \mid (I = N, L). \tag{40}$$

That is to say, Nature's conditional distribution for treatment given the full covariate (L, U) should in fact be completely determined by the observed covariate L: Nature would then be acting as if she were randomising treatment allocation within strata defined by L.

We see once again that, in order to show that we can apply the 'no confounding' conditions (27) and (28), and so answer causal queries from observational data, it is sufficient to identify *just one* potential confounder U (*i.e.* such that (35) and (36) hold), for which we can accept either (37) or (38). We do not need to think about *all* possible potential confounders.

Again we point out that, in more structured versions of this or similar problems, the desired conclusions (*e.g.* (28)) may be derivable from the initial assumptions (*e.g.* (35) and (36)) by the addition of other conditions, more complex than the simple sufficient conditions (37) or (38). And again, such derivations can always be conducted by standard (graphical or non-graphical) manipulations of conditional independence properties.

9 Functional DAGs

We devote the rest of this paper to consideration of *functional models*, another representation for probabilistic and causal structures, based on functional rather than probabilistic dependence relations. These form the basis of Pearl's more recent approach to causal modelling (misleadingly, he calls them 'probabilistic causal models'). We shall relate these to our previously considered representations, as well as to models incorporating *counterfactuals*.

Consider the *functional DAG* of in Figure 19.

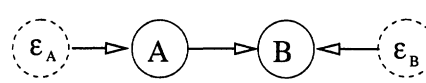

Figure 19. *Functional DAG, A causes B.*

Here the observable 'domain variables' are A and B; ε_A and ε_B represent independent unobserved

'exogenous random errors', indicated by dashed circles. The hollow arrows are used to indicate externally specified *deterministic* dependencies, given by functional relations of the form:

$$A = g_A(\varepsilon_A), \tag{41}$$

$$B = g_B(A; \varepsilon_B). \tag{42}$$

These can also be considered as special (degenerate) forms for the conditional probability specifications at nodes A and B, so that Figure 19 is indeed a DAG representation of the joint probabilistic structure of $(\varepsilon_A, \varepsilon_B, A, B)$.

In order to complete the description of the model, we need to specify the dependence of each node in Figure 19 on its parents: that is, the marginal distributions of ε_A and ε_B, and the functional relationships g_A and g_B in (41) and (42). Having done this we can obtain the joint distribution of $(\varepsilon_A, \varepsilon_B, A, B)$, and thus the joint distribution for (A, B) by marginalising over $(\varepsilon_A, \varepsilon_B)$. It is not hard to see that, by suitable choice of the functions g_A, g_B and the distributions of ε_A and ε_B, we can obtain any desired joint distribution for (A, B) in this way. Thus suppose that this is represented by the DAG model of Figure 3 (which imposes no restriction on the joint distribution for (A, B)), together with external specification of G_A, the distribution function of A, and (for each a) of G_B^a, the conditional distribution function of B, given $A = a$ (extensions to the multivariate case are straightforward). Define ε_A and ε_B to be independently uniform on $[0, 1]$, $g_A(\cdot) \equiv G_A^{-1}(\cdot)$, and $g_B(a; \cdot) \equiv (G_B^a)^{-1}(\cdot)$. It is easily verified that the implied marginal distribution of A, and the conditional distributions for B given A, are then exactly as desired. It is also easy to see that this is just one of many distinct ways in which we could specify a functional DAG that induces the desired distributions for A, and for B given A.

We can similarly represent any probabilistic DAG by means of a functional DAG, constructed by adding a new exogenous 'error node' ε_v (again represented by a dashed circle) feeding into each domain node v, different error variables being taken as mutually independent. Thus a functional DAG representing the probabilistic DAG of Figure 1 would look like Figure 20, with, again, the functional

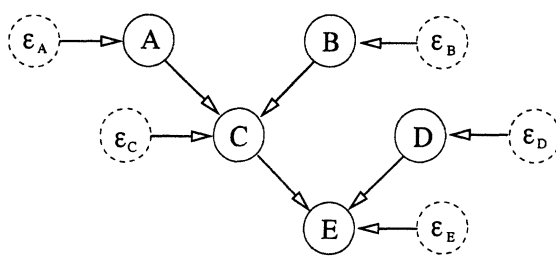

Figure 20. *Functional DAG.*

relationships and the distributions of the error variables externally specified. It is readily seen, using the moralisation criterion, that the conditional independence properties between the domain variables embodied in Figure 20 are identical with those embodied in Figure 1. And once again, for each node v, we can reproduce any desired specification of $p(v|\, \mathrm{pa}^0(v))$ by suitable choice of the distribution of the corresponding error variable ε_v and the function g_v in $v = g_v(\mathrm{pa}^0(v), \varepsilon_v)$. Consequently we can recreate an arbitrary specification of a probabilistic DAG structure by starting from some functional model, in which all the randomness is confined to the error variables. We reiterate, however, that the details of such functional representations are far from unique. The following example makes this clear.

Example 3. For the joint distribution $p(a, b) = p(a)\, p(b|\, a)$ of Example 1, associated with the DAG of Figure 3, we introduce $\varepsilon_B = (\varepsilon_{B,1}, \varepsilon_{B,2})$, having a bivariate normal distribution with

each margin $\mathcal{N}(0, 1)$, and correlation ρ. Define g_B by: $g_B(a; \varepsilon_B) = \mu_a + \varepsilon_{B,a}$. We can complete the functional description, introducing an additional variable ε_A and equation $A = g_A(\varepsilon_A)$, in a variety of ways, to recreate the marginal distribution of A: the specific details of this are not relevant to our current purpose. We then obtain a functional representation, as described by Figure 19, of our original problem. Note particularly that the desired conditional distribution of B given $A = a$, *viz.* $\mathcal{N}(\mu_a, 1)$ $(a = 1, 2)$, will follow from (42), *no matter what may be the value assigned to* ρ. □

We thus see that there can be quite distinct functional models which represent the same probabilistic DAG. In the above example, the differences are confined to the form of the distribution for ε_B. In general we could also vary the functional relationships. (These arbitrarinesses are over and above the variety of ways by which the same joint distribution may be represented by different probabilistic DAG diagrams, as in the case of Figures 3 and 4). Some of the consequences and difficulties that flow from this arbitrariness are considered in detail in Dawid (2000), the above example being essentially the same as Example 1 of that paper—see also Section 10 below.

9.1 Latent Variable Models

A functional DAG is a special case of a *DAG with unobserved (latent) variables*—in this case, the ε's. A more general case arises when we do not insist that the arrows in a diagram such as Figure 20 be hollow, i.e. the dependence of C on (A, B, ε_C), *etc.*, is allowed to be probabilistic rather than deterministic. In this case too, the implied joint distribution for the original domain variables (A, B, C, D, E) will have the independence properties expressed in Figure 1 (in general, however, if we marginalise out over certain variables in a DAG model, the joint distribution over the remainder may not itself be describable in terms of any DAG). This gives yet more ways in which we can represent that original structure—by introducing, and then ignoring, additional variables. And once again, there is a very wide variety of ways in which this could be done. In certain problems, the additional variables introduced might themselves be unobserved domain variables, with clear external meaning, and in such cases it could indeed be helpful to expand the DAG in this way—an example of this is analysed in Dawid (2002). However, when the additional variables are pure mathematical fictions, introduced merely so as to reproduce the desired probabilistic structure of the domain variables, there seems absolutely no good reason to include them in the model. Moreover there is a danger that, if we incorporate into our model terms which do not have any clear external referent, we may all too easily lose sight of this fact, and attempt to make inference about them, or impose additional conditions on them, so leading us to conclusions that might appear to be meaningful (because they can be expressed mathematically in terms of the ingredients we have chosen to include in our model), but which in fact have no scientific basis (since those ingredients are themselves largely arbitrary, and this arbitrariness may feed through to our 'conclusions').

9.2 Functional Intervention Models

Just as we were able to extend the semantics of a probabilistic DAG to model certain assumption about the effects of interventions, so we can introduce an intervention semantics into a functional model. In the probabilistic case we needed to know the various conditional probability distributions associated with each node, and describe the effect of interventions on these. Now we need to know the functional relationships and error distributions, and describe how interventions are supposed to affect these. Thus consider the effect of 'setting' C to c_0 in Figure 20. One way of modelling this is as follows (though we again emphasise that this is by no means the only possibility, and whether or not it provides a good model must remain an empirical question). We suppose that, on setting C to c_0, the original functional relation $C = g_C(A, B; \varepsilon_C)$ is now deleted, to be replaced by: $C = c_0$; while *all other functional relations remain unchanged.*

To complete the specification, we still need to describe the effect of the intervention on the error variables. The assumption we shall make is that there is *no effect* of intervention on the (joint) *distribution* of the errors. We may describe this by saying that the errors are supposed *insensitive* to intervention. Note that this is different from the assumption usually made, that the actual *values* of the errors are unchanged by intervention, a situation that may be described as *unresponsiveness*. In full generality neither condition need imply the other, although in a functional intervention model unresponsiveness implies insensitivity (Heckerman & Shachter, 1995).

Just as for a probabilistic intervention DAG, the assumptions implicit in a functional intervention DAG can be represented more explicitly and usefully by its augmented form, incorporating an additional decision node F_v for each domain node v. The augmented functional intervention DAG corresponding to Figure 19 is as in Figure 21.

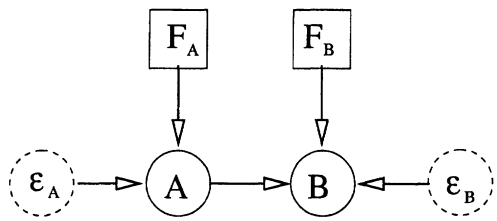

Figure 21. *Augmented functional DAG, A causes B.*

Equation (42) is now replaced by:

$$B = h_B(A; \varepsilon_B; F_B), \tag{43}$$

where $h_B(a; u; \emptyset) \equiv g_B(a; u)$, while, for $x \neq \emptyset$, $h_B(a; u; x) \equiv x$. Equation (41) is adjusted similarly. Apart from the fact that we never associate interventions with the error nodes, the augmented functional DAG is related to the functional DAG (which is after all just a special probabilistic DAG) in exactly the way we have described in Section 6. Note that the assumption that the ε's be insensitive to (*i.e.* independent of) the F's is encoded explicitly, by the standard DAG semantics, in Figure 21.

Again, we can readily represent variations in the nature and scope of interventions. For example, the 'functional influence diagram' of Figure 22, together with Equation (42), represents a situation

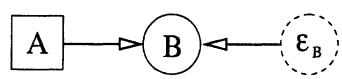

Figure 22. *Functional influence diagram, A causes B.*

in which A is a pure decision node, entirely under experimental control, while no direct intervention on B is possible. (Indeed, this more restricted structure is all that is really needed for our description of Example 3).

It is easy to see that, if we start with a functional DAG representation of a probabilistic DAG, then the above defined functional intervention DAG is likewise equivalent to the associated probabilistic intervention DAG, in the sense that the probabilistic consequences of any intervention will be the same in both descriptions. Similarly, any probabilistic influence diagram, such as that of Figure 5, can be represented by means of a *functional influence diagram*, having, for each random domain node, a new exogenous error node as an additional parent. The distributions and functions can be chosen (though, we again note, non-uniquely) to represent any desired partial distribution for random nodes given decision nodes.

It would seem that we could use either the probabilistic or the functional representation (and corresponding semantics), interchangeably, with the same consequences. However, care is required. This is indeed true to the extent that the causal queries we address to the graph relate only to 'effects of causes', such as "What is the conditional distribution of B given $C = c$, after having intervened to set $A = a$?", as in Section 4.2 above. But the very increase in the richness of the structure in moving to a functional representation makes it possible to ask, and seemingly to answer, other questions, which cannot be addressed through regular intervention DAGs. Typically the answers we obtain will then depend on the specific additional arbitrary choices we have needed to make to construct a functional DAG representation of our intervention DAG. Since we will never be able to distinguish between such alternative functional representations when (as we assume) we cannot observe the ε's, we shall not be able to reach a non-ambiguous answer to such a question—unless we arbitrarily, and without any empirical support, rule out certain choices for the functional model in favour of others.

10 Counterfactuals

Suppose that, for the situation modelled by Example 2, with Figure 3 as the intervention DAG, we have in fact intervened to set $A = 1$, and have observed $B = b$. Consider the question:

> "What would have been the value of B had we set $A = 2$?"

This is a question which seems, at least on the face of it, to make sense. It is what is termed a *counterfactual query*, since it relates to a circumstance ($A = 2$) which runs directly counter to known facts (explicitly, that $A = 1$).

A popular approach to such questions (Rubin, 1974, 1978) has been to introduce 'potential response' variables B_1, B_2, where B_a ($a = 1, 2$) is intended to represent "the value B would take if A were set to a". Since we cannot set A to two distinct values at the same time, it is never possible to observe both B_1 and B_2 simultaneously: they are *complementary variables*. When we have in fact set A to 1, B_2 is, by definition, unobservable: at this point it becomes a *counterfactual variable*. Counterfactual queries and quantities have seemed to many investigators to be of fundamental importance for understanding and implementing causal explanation. A critical account of counterfactuals may be found in Dawid (2000).

If we start with a functional intervention model, we can represent potential responses explicitly in terms of error variables, and thus produce answers to counterfactual queries. Thus suppose we use the functional model described by Figure 19 and Equations (41) and (42). We then have

$$B_1 = g_B(1; \varepsilon_B), \tag{44}$$

$$B_2 = g_B(2; \varepsilon_B). \tag{45}$$

Knowing the function g_B and the distribution of ε_B, we can derive the joint distribution of (B_1, B_2), and hence the conditional distribution of the counterfactual response B_2, given the observed response $B_1 = b$. Since ε_B is supposed insensitive to the setting of A, this distribution remains appropriate even after an intervention on A, and hence supplies a probabilistic answer to the above counterfactual query. (Similar analyses can be conducted in non-functional models with latent variables, although additional conditional independence assumptions are then required (Dawid, 2000; §12)).

To continue, suppose that we specialise by imposing the assumptions in Example 3, for some specified value of ρ. Then (B_1, B_2) will have a bivariate normal distribution with mean vector (μ_1, μ_2), unit variances, and correlation ρ. Thus the desired distribution for B_2, given $B_1 = b$, will be normal with mean $\mu_2 + \rho(b - \mu_1)$ and variance $(1 - \rho^2)$. A disturbing feature of this solution is its dependence on the value of ρ, even though all values for $\rho \in [-1, 1]$ correspond to the same probabilistic intervention DAG. That is to say, from observations on the system, including cases in which interventions are made, it is completely impossible to learn anything about the value of

ρ in an assumed model such as that of Example 3; nevertheless, we need to know this value in order to apply the above reasoning and extract a definite answer to our counterfactual query. It is for reasons such as this that Dawid (2000) regards the introduction of counterfactual quantities into statistical models as full of danger, and best avoided. And indeed, although counterfactual models have been used to develop important results in causal inference, such as the G-computation formula (Robins, 1986, 1987) or bounds on causal effects under partial compliance (Robins, 1989; Pearl, 2000, Chapter 8), these same results can be expressed more clearly, and derived more easily, in the language of influence diagrams, with no reference to functional relations or counterfactuals (Dawid, 2002; Dawid *et al.*, 2002). The only analyses that cannot be re-expressed in this way are those where the 'answer' depends on the arbitrary way in which we choose between observationally indistinguishable functional models (as in the above analysis of Example 3).

11 Canonical Functional Models

We have seen how, given a functional intervention DAG, we can construct complementary quantities, such as (B_1, B_2), with a fully specified joint distribution. We have also seen how problematic it can be to learn from data about the functions, or the joint distribution of complementaries, in such a model.

In the alternative approach of Balke & Pearl (1994), one starts by boldly assuming the simultaneous existence, and joint distribution, of the collection of complementary quantities. This constitutes what we may call a *counterfactual model*. We can then associate, with any counterfactual model, a functional model, in a natural way. Thus for the case of a binary decision variable A, and response variable B, we would consider, together, the complementary versions (B_1, B_2) of B, under the two settings for A, with a supposed joint distribution. In that case we can *deduce* the appropriateness of a functional representation, as in Figure 23, with the specific functional relationship

$$g_B(a; (b_1, b_2)) := b_a. \tag{46}$$

This is seen to be of exactly the same form as Figure 22, with the identification of the 'error' ε_B

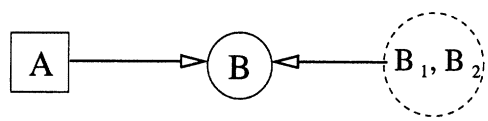

Figure 23. *Canonical functional model.*

as the collection (B_1, B_2) of complementary responses. Such a functional model, constructed from and representing supposedly pre-existing complementary variables, may be termed *canonical*, and Equation (46) a *canonical functional relationship*.

We see then that functional and counterfactual models are essentially equivalent. In particular, any functional model may be replaced by its canonical version, with the original ε_B replaced by (B_1, B_2), as defined by Equations (44) and (45), having the appropriate induced joint distribution, and with the functional relationship now given by the canonical function (46). This construction is illustrated in Figure 24.

For all purposes of addressing counterfactual queries, the canonical model contains all the relevant information—specifically, this is coded into the joint distribution of (B_1, B_2). But once again we emphasise that this joint distribution is to a large extent arbitrary, since from scientific investigations (involving setting A either to 1 or to 2) we can only learn the separate marginal distributions of B_1 and of B_2, but nothing whatsoever about their dependence—which affects the answers to counterfactual

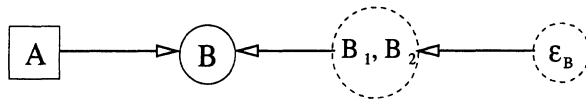

Figure 24. *Construction of canonical functional model.*

queries, as illustrated in Example 3. From this point of view the canonical model is not really canonical, since it is not completely specified by the properties of the associated intervention DAG.

12 Conclusions

We have demonstrated the expressive power and fruitfulness of various forms of probabilistic influence diagram as representations of complex causal structures. Once the appropriate diagram is drawn, no new concepts, tools or techniques are required, beyond those already developed for the analysis of probabilistic DAG models. Using a probabilistic influence diagram to represent and manipulate causal models is adequate to address a very wide range of problems relating to inference about the effects of causes, and in no case is any advantage to be gained by elaborating the model with functional or counterfactual elements. (Inference about 'causes of effects' may require some form of counterfactual reasoning, but cannot be satisfactorily addressed without deeper critical analysis (Dawid, 2000, Part 3)).

There are certainly limitations to such representations of causal problems by influence diagrams, but these are no different in principle from those that limit the use of graphical representations in purely probabilistic problems. For example, it is well known that the number of probabilistic conditional independence structures over a set of variables exceeds by far those that can be represented by means of a graph (Studený, 2001). Moreover, graphs are poor at representing restricted conditional independencies, that hold for some values of certain conditioning variables but not others: for example, we might wish to impose additional independencies when a certain variable is 'set', rather than arising naturally. A more general representation of a decision problem would be by means of a *decision tree* (Raiffa, 1968). However, this does not allow for explicit encoding of conditional independence properties, which remain hidden in numerical probability assignments. But even in the absence of a suitable graphical representation, the algebraic skeleton of our general approach, which proceeds by introducing decision variables, strategy variables and parameter variables in addition to domain variables, and manipulating conditional independence properties between all of these according to standard rules (Dawid, 1979), can still be used to express and conduct causal analyses, in an explicit and straightforward fashion. Dawid *et al.* (2002) describes and analyses, in exactly such terms, the fundamental conditions needed to justify the use of the general G-computation formula. Other approaches, by leaving many of the required assumptions implicit, are simply unable to express such conditions clearly.

Acknowledgments

I am grateful to Jamie Robins and Vanessa Didelez for helpful comments.

References

Balke, A.A. & Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, Eds. R.L. de Mantaras and D. Poole, pp. 46–54.

Cowell, R.G., Dawid, A.P., Lauritzen, S. & Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*. New York: Springer.

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

Dawid, A.P. (1980). Conditional independence for statistical operations. *Annals of Statistics*, **8**, 598–617.
Dawid, A.P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, **95**, 407–448.
Dawid, A.P. (2001). Some variations on variation independence. In *Artificial Intelligence and Statistics 2001*, Eds. T. Jaakkola and T. Richardson, pp. 187–191. San Francisco, CA: Morgan Kaufmann Publishers.
Dawid, A.P. (2002a). Causal inference using influence diagrams: The problem of partial compliance (with discussion). In *Highly Structured Stochastic Systems*, Eds. P. Green, N. Hjort and S. Richardson. Oxford University Press. To appear.
Dawid, A.P. (2002b). Discussion of Lauritzen & Richardson (2002). *Journal of the Royal Statistical Society, Series B*, **64**. To appear.
Dawid, A.P., Didelez, V. & Murphy, S. (2001). On the conditions underlying the estimability of causal effects from observational data. Manuscript in preparation.
Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, **17**, 333–353.
Galles, D. & Pearl, J. (1995). Testing identifiability of causal effects. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Eds. P. Besnard and S. Hanks, pp. 185–195. San Francisco, CA: Morgan Kaufmann Publishers.
Glymour, C., Spirtes, P. & Richardson, T. (1999). On the possibility of inferring causation from association without background knowledge. In *Computation, Causation, and Discovery*, Eds. C. Glymour and G.F. Cooper, pp. 323–331. Cambridge, MA: MIT Press.
Heckerman, D. & Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, **3**, 405–430.
Howard, R.A. & Matheson, J.E. (1984). Influence diagrams. In *Readings in the Principles and Applications of Decision Analysis*, Eds. R.A. Howard and J.E. Matheson. Menlo Park, CA: Strategic Decisions Group.
Lauritzen, S.L. (2000). Causal inference from graphical models. In *Complex Stochastic Systems*, Eds. O.E. Barndorff-Nielsen, D.R. Cox and C.Klüppelberg, chapter 2, pp. 63–107. London: CRC Press.
Lauritzen, S.L., Dawid, A.P., Larsen, B.N. & Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.
Lauritzen, S.L. & Nilsson, D. (2001). Representing and solving decision problems with limited information. *Management Science*, **47**, 1235–1251.
Lauritzen, S.L. & Richardson, T.S. (2002). Chain graph models and their causal interpretation (with Discussion). *Journal of the Royal Statistical Society, Series B*, **64**. To appear.
Nilsson, D. & Lauritzen, S.L. (2000). Evaluating influence diagrams using LIMIDs. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, Eds. C. Boutilier and M. Goldszmidt, pp. 436–445. San Francisco, CA: Morgan Kaufmann Publishers.
Oliver, R.M. & Smith, J.Q. (1990). *Influence Diagrams, Belief Nets and Decision Analysis*. Chichester, UK: John Wiley and Sons.
Pearl, J. (1993a). Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, pp. 391–401.
Pearl, J. (1993b). Comment: Graphical models, causality and intervention. *Statistical Science*, **8**, 266–269.
Pearl, J. (1995). Causal diagrams for empirical research (with Discussion). *Biometrika*, **82**, 669–710.
Pearl, J. (2000). *Causality*. Cambridge University Press.
Pearl, J. & Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Eds. P. Besnard and S. Hanks, pp. 444–453. San Francisco, CA: Morgan Kaufmann Publishers.
Raiffa, H. (1968). *Decision Analysis*. Reading, MA: Addison-Wesley.
Robins, J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–1512.
Robins, J.M. (1987). Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect". *Computers and Mathematics with Applications*, **14**, 923–945.
Robins, J.M. (1989). The analysis of randomized and nonrandomized aids treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on Aids*, Eds. L. Sechrest, H. Freeman and A. Mulley, pp. 113–159. NCSHR, U.S. Public Health Service.
Robins, J.M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics, Vol. **120**, Ed. M. Berkane, pp. 69–117. New York: Springer-Verlag.
Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.
Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, **6**, 34–68.
Savage, L.J. (1954). *The Foundations of Statistics*. New York: John Wiley and Sons.
Shachter, R.D. (1986). Evaluating influence diagrams. *Operations Research*, **34**, 871–882.
Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction and Search*. New York: Springer-Verlag.
Studený, M. (2001). *On non-graphical description of models of conditional independence structure.* Paper presented at HSSS Workshop on Stochastic Systems for Individual Behaviours, 22–23 January 2001, Louvain la Neuve, Belgium.
Verma, T. & Pearl, J. (1990). Causal networks: Semantics and expressiveness. In *Uncertainty in Artificial Intelligence 4*, Eds. R.D. Shachter, T.S. Levitt, L.N. Kanal and J.F. Lemmer, pp. 69–76. Amsterdam: North-Holland.
Verma, T. & Pearl, J. (1991). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence 6*, Eds. P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer, pp. 255–268. Amsterdam: North-Holland.

Résumé

On considère une variété de moyens par lesquelles on peut réprésenter les modèles probabilistes et causals en forme graphique. En ajoutant à nos graphiques des nœuds pour réprésenter des paramètres, des décisions, etc., on obtient une généralisation des diagrammes d'influence qui supporte la modélisation et l'inference significatives, et qui n'exige que des concepts et des méthodes déja connus dans le cas purement probabiliste. On rapporte ces réprésentations à des autres, en particulier les modèles fonctionels, et on présente des arguments et des exemples pour démontrer leur supériorité.

[*Received June 2001, accepted October 2001*]