Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press: Menlo Park, CA, 574-579, August 2002.

TECHNICAL REPORT
R-298
January 2002

A New Characterization of the Experimental Implications of Causal Bayesian Networks

Jin Tian and Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
{jtian, judea }@cs.ucla.edu

Abstract

We offer a complete characterization of the set of distributions that could be induced by local interventions on variables governed by a causal Bayesian network. We show that such distributions must adhere to three norms of coherence, and we demonstrate the use of these norms as inferential tools in tasks of learning and identification. Testable coherence norms are subsequently derived for networks containing unmeasured variables.

Introduction

The use of graphical models for encoding distributional and causal information is now fairly standard (Pearl 1988; Spirtes, Glymour, & Scheines 1993; Heckerman & Shachter 1995; Lauritzen 2000; Pearl 2000; Dawid 2001). The most common such representation involves a *causal Bayesian network*, namely, a directed acyclic graph (DAG) G which, in addition to the usual conditional independence interpretation, is also given a causal interpretation. This additional feature permits one to infer the effects of interventions, such as policy decisions and ordinary actions. Specifically, if an external intervention fixes any set T of variables to some constants t, the DAG permits us to infer the resulting post-intervention distribution, denoted by $P_t(v)$,[1] from the pre-intervention distribution $P(v)$.

In this paper, we seek a characterization for the set of interventional distributions, $P_t(v)$, that could be induced by some causal Bayesian network. Whereas (Pearl 2000, pp.23-4) has given such characterization relative to a given network, we assume that the underlying network, if such exists, is unknown. Given a collection of distribution functions, $P_t(s)$, each obtained by observing a set S of variables under experimental conditions $T = t$, we ask whether the collection is compatible with the predictions of some underlying causal Bayesian network, and we identify three properties (of the collection) that are both necessary and sufficient for the existence of such an underlying network. We subsequently identify necessary properties of distributions induced by causal Bayesian networks in which some of the variables are unmeasured. We further show how these properties can be used as symbolic inferential tools for predicting the effects of actions from nonexperimental data in the presence of unmeasured variables. The Conclusion Section outlines the use of these properties in learning tasks which aim at uncovering the structure of the network.

[1](Pearl 1995; 2000) used the notation $P(v|set(t))$, $P(v|do(t))$, or $P(v|\hat{t})$ for the post-intervention distribution, while (Lauritzen 2000) used $P(v||t)$.

Causal Bayesian Networks and Interventions

A causal Bayesian network (also known as a *Markovian model*) consists of two mathematical objects: (i) a DAG G, called a *causal graph*, over a set $V = \{V_1, \ldots, V_n\}$ of vertices, and (ii) a probability distribution $P(v)$, over the set V of discrete variables that correspond to the vertices in G. The interpretation of such a graph has two components, probabilistic and causal.[2] The probabilistic interpretation views G as representing conditional independence restrictions on P: Each variable is independent of all its non-descendants given its direct parents in the graph. These restrictions imply that the joint probability function $P(v) = P(v_1, \ldots, v_n)$ factorizes according to the product

$$P(v) = \prod_i P(v_i|pa_i) \tag{1}$$

where pa_i are (values of) the parents of variable V_i in G.

The causal interpretation views the arrows in G as representing causal influences between the corresponding variables. In this interpretation, the factorization of (1) still holds, but the factors are further assumed to represent autonomous data-generation processes, that is, each conditional probability $P(v_i|pa_i)$ represents a stochastic process by which the values of V_i are assigned[3] in response to

[2]A more refined interpretation, called *functional*, is also common (Pearl 2000, chapter 1), which, in addition to interventions, supports counterfactual readings. The functional interpretation assumes strictly deterministic, functional relationships between variables in the model, some of which may be unobserved. Complete axiomatizations of deterministic counterfactual relations are given in (Galles & Pearl 1998; Halpern 1998). We are not aware of an axiomatization of the probabilistic predictions of functional models.

[3]In contrast with functional models, here the probability of each V_i, not its precise value, is determined by the other variables in the model.

Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press: Menlo Park, CA, 574-579, August 2002.

TECHNICAL REPORT
R-298
January 2002

the values pa_i (previously chosen for V_i's parents), and the stochastic variation of this assignment is assumed independent of the variations in all other assignments in the model. Moreover, each assignment process remains invariant to possible changes in the assignment processes that govern other variables in the system. This modularity assumption enables us to predict the effects of interventions, whenever interventions are described as specific modifications of some factors in the product of (1). The simplest such intervention, called *atomic*, involves fixing a set T of variables to some constants $T = t$, which yields the post-intervention distribution

$$P_t(v) = \begin{cases} \prod_{\{i|V_i \notin T\}} P(v_i|pa_i) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \tag{2}$$

Eq. (2) represents a truncated factorization of (1), with factors corresponding to the manipulated variables removed. This truncation follows immediately from (1) since, assuming modularity, the post-intervention probabilities $P(v_i|pa_i)$ corresponding to variables in T are either 1 or 0, while those corresponding to unmanipulated variables remain unaltered.[4] If T stands for a set of treatment variables and Y for an outcome variable in $V \setminus T$, then Eq. (2) permits us to calculate the probability $P_t(y)$ that event $Y = y$ would occur if treatment condition $T = t$ were enforced uniformly over the population. This quantity, often called the "causal effect" of T on Y, is what we normally assess in a controlled experiment with T randomized, in which the distribution of Y is estimated for each level t of T.

Let $\mathbf{P}_*$ denote the set of all interventional distributions

$$\mathbf{P}_* = \{P_t(v) | T \subseteq V, t \in Dm(T)\} \tag{3}$$

where $Dm(T)$ represents the domain of T. In the next section, we will give a set of properties that fully characterize the $\mathbf{P}_*$ set.

Interventional Distributions in Markovian Models

The $\mathbf{P}_*$ set induced from a Markovian model must satisfy three properties: effectiveness, Markov, and recursiveness.

Property 1 (Effectiveness) *For any set of variables T,*

$$P_t(t) = 1. \tag{4}$$

Effectiveness states that, if we force a set of variables T to have the value t, then the probability of T taking that value t is one.

For any set of variables S disjoint with T, an immediate corollary of effectiveness reads:

$$P_{t,s}(t) = 1, \tag{5}$$

which follows from

$$P_{t,s}(t) \geq P_{t,s}(t, s) = 1. \tag{6}$$

[4]Eq. (2) was named "Manipulation Theorem" in (Spirtes, Glymour, & Scheines 1993), and is also implicit in Robins' (1987) G-computation formula.

Equivalently, if $T_1 \subseteq T$, then

$$P_t(t_1) = \begin{cases} 1 & \text{if } t_1 \text{ is consistent with } t. \\ 0 & \text{if } t_1 \text{ is inconsistent with } t. \end{cases} \tag{7}$$

We further have that, for $T_1 \subseteq T$ and S disjoint of T,

$$P_t(s, t_1) = \begin{cases} P_t(s) & \text{if } t_1 \text{ is consistent with } t. \\ 0 & \text{if } t_1 \text{ is inconsistent with } t. \end{cases} \tag{8}$$

Property 2 (Markov) *For any two disjoint sets of variables S_1 and S_2,*

$$P_{v \setminus (s_1 \cup s_2)}(s_1, s_2) = P_{v \setminus s_1}(s_1) P_{v \setminus s_2}(s_2). \tag{9}$$

An equivalent form of the Markov property is: For any set of variables $T \subseteq V$,

$$P_t(v \setminus t) = \prod_{\{i|V_i \in V \setminus T\}} P_{v \setminus \{v_i\}}(v_i). \tag{10}$$

Eq. (10) can be obtained by repeatedly applying Eq. (9), and Eq. (9) follows from Eq. (10) as follows:

$$\begin{aligned} P_{v \setminus (s_1 \cup s_2)}(s_1, s_2) &= \prod_{V_i \in S_1 \cup S_2} P_{v \setminus \{v_i\}}(v_i) \\ &= \prod_{V_i \in S_1} P_{v \setminus \{v_i\}}(v_i) \prod_{V_i \in S_2} P_{v \setminus \{v_i\}}(v_i) \\ &= P_{v \setminus s_1}(s_1) P_{v \setminus s_2}(s_2). \end{aligned} \tag{11}$$

Definition 1 *For two single variables X and Y, define "X affects Y", denoted by $X \rightsquigarrow Y$, as $\exists W \subset V, w, x, y$, such that $P_{x,w}(y) \neq P_w(y)$. That is, X affects Y if, under some setting w, intervening on X changes the distribution of Y.*

Property 3 (Recursiveness) *For any set of variables $\{X_0, \ldots, X_k\} \subseteq V$,*

$$(X_0 \rightsquigarrow X_1) \wedge \ldots \wedge (X_{k-1} \rightsquigarrow X_k) \Rightarrow \neg(X_k \rightsquigarrow X_0). \tag{12}$$

Property 3 is a stochastic version of the (deterministic) recursiveness axiom given in (Halpern 1998). It comes from restricting the causal models under study to those having acyclic causal graphs. For $k = 1$, for example, we have $X \rightsquigarrow Y \Rightarrow \neg(Y \rightsquigarrow X)$, saying that for any two variables X and Y, either X does not affect Y or Y does not affect X. (Halpern 1998) pointed out that, recursiveness can be viewed as a collection of axioms, one for each k, and that the case of $k = 1$ alone is not enough to characterize a recursive model.

Theorem 1 (Soundness) *Effectiveness, Markov, and recursiveness hold in all Markovian models.*

Proof: All three properties follow from the factorization of Eq. (2).

Effectiveness From Eq. (2), we have

$$P_t(T = t') = 0 \quad \text{for } t' \neq t, \tag{13}$$

and since

$$\sum_{t' \in Dm(T)} P_t(t') = 1, \tag{14}$$

we obtain the effectiveness property of Eq. (4).

Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press: Menlo Park, CA, 574-579, August 2002.

TECHNICAL REPORT
R-298
January 2002

Markov From Eq. (2), we have

$$P_t(v \setminus t) = P_t(t, v \setminus t) = \prod_{V_i \in V \setminus T} P(v_i|pa_i). \quad (15)$$

Letting $T = V \setminus \{V_i\}$ in Eq. (15) yields

$$P_{v \setminus \{v_i\}}(v_i) = P(v_i|pa_i). \quad (16)$$

Substituting Eq. (16) back into Eq. (15), we get the Markov property (10), which is equivalent to (9).

Recursiveness Assume that a total order over V that is consistent with the causal graph is $V_1 < \cdots < V_n$, such that V_i is a nondescendant of V_j if $V_i < V_j$. Consider a variable V_j and a set of variables $S \subseteq V$ which does not contain V_j. Let $B_j = \{V_i | V_i < V_j, V_i \in V \setminus S\}$ be the set of variables not in S and ordered before V_j, and let $A_j = \{V_i | V_j < V_i, V_i \in V \setminus S\}$ be the set of variables not in S and ordered after V_j. First we show that

$$P_{v_j,s}(b_j) = P_s(b_j). \quad (17)$$

We have

$$\begin{aligned} P_{v_j,s}(b_j) &= \sum_{a_j} P_{v_j,s}(a_j, b_j) \\ &= \sum_{a_j} P_{v_j,s,a_j}(b_j) P_{v_j,s,b_j}(a_j), \quad \text{(by Eq. (9))} \quad (18) \end{aligned}$$

where $P_{v_j,s,a_j}(b_j) = \prod_{\{i|V_i \in B_j\}} P(v_i|pa_i)$ is a function of b_j and its parents. Since all variables in A_j are ordered after the variables in B_j, $P_{v_j,s,a_j}(b_j)$ is not a function of a_j. Hence Eq. (18) becomes

$$\begin{aligned} P_{v_j,s}(b_j) &= P_{v_j,s,a_j}(b_j) \sum_{a_j} P_{v_j,s,b_j}(a_j) \\ &= P_{v_j,s,a_j}(b_j) \quad (19) \end{aligned}$$

Similarly,

$$\begin{aligned} P_s(b_j) &= \sum_{v_j,a_j} P_s(v_j, a_j, b_j) \\ &= \sum_{v_j,a_j} P_{v_j,s,a_j}(b_j) P_{s,b_j}(v_j, a_j) \\ &= P_{v_j,s,a_j}(b_j) \sum_{v_j,a_j} P_{s,b_j}(v_j, a_j) = P_{v_j,s,a_j}(b_j) \quad (20) \end{aligned}$$

Eq. (17) follows from (19) and (20).

From Eq. (17), we have that, for any two variables $V_i < V_j$ and any set of variables S,

$$P_{v_j,s}(v_i) = P_s(v_i), \quad (21)$$

which states that if X is ordered before Y then Y does not affect X, based on our definition of "X affects Y". Therefore, we have that if X affects Y then X is ordered before Y, or

$$X \rightsquigarrow Y \Rightarrow X < Y. \quad (22)$$

Recursive property (12) then follows from (22) because the relation "$<$" is a total order.

□

To facilitate the proof of the completeness theorem, we give the following lemma.

Lemma 1 *(Pearl 1988, p.124) Given a DAG over V, if a set of functions $f_i(v_i, pa_i)$ satisfy*

$$\sum_{v_i \in Dm(V_i)} f_i(v_i, pa_i) = 1, \text{ and } 0 \le f_i(v_i, pa_i) \le 1, \quad (23)$$

and $P(v)$ can be decomposed as

$$P(v) = \prod_i f_i(v_i, pa_i), \quad (24)$$

then we have

$$f_i(v_i, pa_i) = P(v_i|pa_i), \quad i = 1, \ldots, n. \quad (25)$$

Theorem 2 (Completeness) *If a $\mathbf{P}_*$ set satisfies effectiveness, Markov, and recursiveness, then there exists a Markovian model with a unique causal graph that can generate this $\mathbf{P}_*$ set.*

Proof: Define a relation "$\prec$" as: $X \prec Y$ if $X \rightsquigarrow Y$. Then the transitive closure of $\prec$, $\prec^*$, is a partial order over the set of variables V from the recursiveness property as shown in (Halpern 1998). Let "$<$" be a total order on V consistent with $\prec^*$. We have that

$$\text{if } X < Y \text{ then } P_{y,s}(x) = P_s(x) \quad (26)$$

for any set of variables S. This is because if $P_{y,s}(x) \neq P_s(x)$, then $Y \rightsquigarrow X$, and therefore $Y \prec X$, which contradicts the fact that $X < Y$ is consistent with $\prec^*$.

Define a set PA_i as a minimal set of variables that satisfies

$$P_{pa_i}(v_i) = P_{v \setminus \{v_i\}}(v_i). \quad (27)$$

We have that

$$\text{if } V_i < V_j, \text{ then } V_j \notin PA_i. \quad (28)$$

Otherwise, assuming $V_j \in PA_i$ and letting $PA'_i = PA_i \setminus \{V_j\}$, from Eqs. (26) and (27) we have

$$P_{pa'_i}(v_i) = P_{pa'_i, v_j}(v_i) = P_{v \setminus \{v_i\}}(v_i), \quad (29)$$

which contradicts the fact that PA_i is minimal. From Eq. (28), drawing an arrow from each member of PA_i toward V_i, the resulting graph G is a DAG.

Substituting Eq. (27) into the Markov property (10), we obtain, for any set of variables T,

$$P_t(v \setminus t) = \prod_{\{i|V_i \notin T\}} P_{pa_i}(v_i). \quad (30)$$

By Lemma 1, we get

$$P_{pa_i}(v_i) = P(v_i|pa_i). \quad (31)$$

From Eqs. (30), (31), and the effectiveness property (8), Eq. (2) follows. Therefore, a Markovian model with a causal graph G can generate this $\mathbf{P}_*$ set.

Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press: Menlo Park, CA, 574-579, August 2002.

TECHNICAL REPORT
R-298
January 2002

Next, we show that the set PA_i is unique. Assuming that there are two minimal sets PA_i and PA'_i both satisfying Eq. (27), we will show that their intersection also satisfies Eq. (27). Let $A = PA_i \cap PA'_i$, $B = PA_i \setminus A$, $B' = PA'_i \setminus A$, and $S = V \setminus (PA_i \cup PA'_i \cup \{V_i\})$. From the Markov property Eq. (9), we have

$$\begin{aligned} P_a(b, b', s, v_i) &= P_{a,v_i}(b, b', s) P_{v\setminus\{v_i\}}(v_i) \\ &= P_{a,v_i}(b, b', s) P_{a,b}(v_i) \end{aligned} \quad (32)$$

Summing both sides of (32) over B' and S, we get

$$P_a(b, v_i) = P_{a,v_i}(b) P_{a,b}(v_i). \quad (33)$$

Substituting $P_{pa_i}(v_i)$ with $P_{pa'_i}(v_i)$ in (33), we get

$$P_a(b, v_i) = P_{a,v_i}(b) P_{a,b'}(v_i). \quad (34)$$

Summing both sides of (34) over B, we obtain

$$P_a(v_i) = P_{a,b'}(v_i) = P_{pa'_i}(v_i), \quad (35)$$

which says that the set $A = PA_i \cap PA'_i$ also satisfies Eq. (27). This contradicts the assumption that both PA_i and PA'_i are minimal. Thus PA_i is unique. □

A Markovian model also satisfies the following properties.

Property 4 *If a set B is composed of nondescendants of a variable V_j, then for any set of variables S,*

$$P_{v_j,s}(b) = P_s(b). \quad (36)$$

Proof: If B is disjoint of S, Eq. (36) follows from Eq. (17) since $B \subseteq B_j$. If B is not disjoint of S, Eq. (36) follows from the Effectiveness property and Eq. (17). □

Property 5 *For any set of variables $S \subseteq V \setminus (PA_i \cup \{V_i\})$,*

$$P_{pa_i,s}(v_i) = P_{pa_i}(v_i). \quad (37)$$

Proof: Let $S' = V \setminus (PA_i \cup \{V_i\} \cup S)$.

$$\begin{aligned} P_{pa_i,s}(v_i) &= \sum_{s'} P_{pa_i,s}(s', v_i) \\ &= \sum_{s'} P_{v\setminus\{v_i\}}(v_i) P_{pa_i,s,v_i}(s') \quad \text{(by Eq. (9))} \\ &= P_{pa_i}(v_i) \sum_{s'} P_{pa_i,s,v_i}(s') \quad \text{(by Eq. (27))} \\ &= P_{pa_i}(v_i) \end{aligned} \quad (38)$$

□

Property 6

$$P_{pa_i}(v_i) = P(v_i|pa_i). \quad (39)$$

Property 6 has been given in Eq. (31).

Property 7 *For any set of variables $S \subseteq V$, and $V_i \notin S$,*

$$P_s(v_i|pa_i) = P(v_i|pa_i), \quad \textit{for } pa_i \textit{ consistent with } s. \quad (40)$$

Proof: Let $S' = V \setminus (PA_i \cup \{V_i\} \cup S)$. Assuming that pa_i is consistent with s, we have

$$\begin{aligned} P_s(v_i, pa_i) &= \sum_{s'} P_s(v_i, pa_i, s') \\ &= \sum_{s'} P_{v\setminus\{v_i\}}(v_i) P_{s,v_i}(pa_i, s') \quad \text{(by Eq. (9))} \\ &= P(v_i|pa_i) \sum_{s'} P_{s,v_i}(pa_i, s') \quad \text{(by Eq. (16))} \\ &= P(v_i|pa_i) P_{s,v_i}(pa_i) \\ &= P(v_i|pa_i) P_s(pa_i) \quad \text{(by Property 4)} \end{aligned} \quad (41)$$

which leads to Eq. (40). □

Interventional Distributions in Semi-Markovian Models

When some variables in a Markovian model are unobserved, the probability distribution over the observed variables may no longer be decomposed as in Eq. (1). Let $V = \{V_1, \ldots, V_n\}$ and $U = \{U_1, \ldots, U_{n'}\}$ stand for the sets of observed and unobserved variables respectively. If no U variable is a descendant of any V variable, then the corresponding model is called a *semi-Markovian* model. In a semi-Markovian model, the observed probability distribution, $P(v)$, becomes a mixture of products:

$$P(v) = \sum_u \prod_i P(v_i|pa_i, u^i) P(u) \quad (42)$$

where PA_i and U^i stand for the sets of the observed and unobserved parents of V_i, and the summation ranges over all the U variables. The post-intervention distribution, likewise, will be given as a mixture of truncated products

$$P_t(v) = \begin{cases} \sum_u \prod_{\{i|V_i \notin T\}} P(v_i|pa_i, u^i) P(u) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \quad (43)$$

If, in a semi-Markovian model, no U variable is an ancestor of more than one V variable, then $P_t(v)$ in Eq. (43) factorizes into a product as in Eq. (2), regardless of the parameters $\{P(v_i|pa_i, u^i)\}$ and $\{P(u)\}$. Therefore, for such a model, the causal Markov condition holds relative to G_V (the subgraph of G composed only of V variables), that is, each variable V_i is independent on all its non-descendants given its parents PA_i in G_V. And by convention, the U variables are usually not shown explicitly, and G_V is called the causal graph of the model.

The causal Markov condition is often assumed as an inherent feature of causal models (see e.g. (Kiiveri, Speed, & Carlin 1984; Spirtes, Glymour, & Scheines 1993)). It reflects our two basic causal assumptions: (i) include in the model every variable that is a cause of two or more other variables in the model; and (ii) Reichenbach's (1956) common-cause assumption, also known as "no correlation

Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press: Menlo Park, CA, 574-579, August 2002.

TECHNICAL REPORT
R-298
January 2002

without causation," stating that, if any two variables are dependent, then one is a cause of the other *or* there is a third variable causing both.

If two or more variables in V are affected by unobserved confounders, the presence of such confounders would not permit the decomposition in Eq. (1), and, in general, $P(v)$ generated by a semi-Markovian model is a mixture of products given in (42). However, the conditional distribution $P(v|u)$ factorizes into a product

$$P(v|u) = \prod_i P(v_i|pa_i, u^i), \tag{44}$$

and we also have

$$P_t(v|u) = \begin{cases} \prod_{\{i|V_i \notin T\}} P(v_i|pa_i, u^i) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \tag{45}$$

Therefore all Properties 1–7 hold when we condition on u. For example, the Markov property can be written as

$$P_{v\setminus(s_1 \cup s_2)}(s_1, s_2|u) = P_{v\setminus s_1}(s_1|u) P_{v\setminus s_2}(s_2|u). \tag{46}$$

Let $\mathbf{P}_*(u)$ denote the set of all conditional interventional distributions

$$\mathbf{P}_*(u) = \{P_t(v|u) | T \subseteq V, t \in Dm(T)\} \tag{47}$$

Then $\mathbf{P}_*(u)$ is fully characterized by the three properties effectiveness, Markov, and recursiveness, conditioning on u.

Let $\mathbf{P}_*$ denote the set of all interventional distributions over observed variables V as in (3). From the properties of the $\mathbf{P}_*(u)$ set, we can immediately conclude that the $\mathbf{P}_*$ set satisfies the following properties: effectiveness (Property 1), recursiveness (Property 3), Property 4, and Property 5, while Markov (Property 2), Property 6, and Property 7 do not hold. For example, Property 5 can be proved from its conditional version,

$$P_{pa_i,s}(v_i|u) = P_{pa_i}(v_i|u), \tag{48}$$

as follows

$$\begin{aligned} P_{pa_i,s}(v_i) &= \sum_u P_{pa_i,s}(v_i|u) P(u) \\ &= \sum_u P_{pa_i}(v_i|u) P(u) = P_{pa_i}(v_i). \end{aligned} \tag{49}$$

Significantly, the $\mathbf{P}_*$ set must satisfy inequalities that are unique to semi-Markovian models, as opposed, for example, to models containing feedback loops. For example, from Eq. (43), and using

$$P(v_i|pa_i, u^i) \leq 1, \tag{50}$$

we obtain the following property.

Property 8 *For any three sets of variables, T, S, and R, we have*

$$P_{tr}(s) \geq P_t(r, s) + P_r(t, s) - P(t, r, s) \tag{51}$$

Additional inequalities, involving four or more subsets, can likewise be derived by this method. However, finding a set of properties that can completely characterize the $\mathbf{P}_*$ set of a semi-Markovian causal model remains an open challenge.

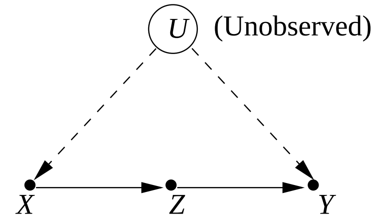

Figure 1:

Applications in the Identification of Causal Effects

Given two disjoint sets T and S, the quantity $P_t(s)$ is called the *causal effect* of T on S. $P_t(s)$ is said to be *identifiable* if, given a causal graph, it can be determined uniquely from the distribution $P(v)$ of the observed variables, and is thus independent of the unknown quantities, $P(u)$ and $P(v_i|pa_i, u^i)$, that involve elements of U. Identification means that we can learn the effect of the action $T = t$ (on the variables in S) from sampled data taken prior to actually performing that action. In Markovian models, all causal effects are identifiable and are given in Eq. (2). When some confounders are unobserved, the question of identifiability arises. Sufficient graphical conditions for ensuring the identification of $P_t(s)$ in semi-Markovian models were established by several authors (Spirtes, Glymour, & Scheines 1993; Pearl 1993; 1995) and are summarized in (Pearl 2000, Chapters 3 and 4). Since

$$P_t(s) = \sum_u P_t(s|u) P(u), \tag{52}$$

and since we have a complete characterization over the set of conditional interventional distributions ($\mathbf{P}_*(u)$), we can use Properties 1–3 (conditioning on u) for identifying causal effects in semi-Markovian models.

The assumptions embodied in the causal graph can be translated into the language of conditional interventional distributions as follows:

For each variable V_i,

$$P_{v\setminus\{v_i\}}(v_i|u) = P_{pa_i}(v_i|u^i). \tag{53}$$

The Markov property (10) conditioning on u then becomes

$$P_t(v \setminus t|u) = \prod_{\{i|V_i \in V\setminus T\}} P_{pa_i}(v_i|u^i). \tag{54}$$

The significance of Eq. (54) rests in simplifying the derivation of elaborate causal effects in semi-Markov models. To illustrate this derivation, consider the model in Figure 1, and assume we need to derive the causal effect of X on $\{Z, Y\}$, a task analyzed in (Pearl 2000, pp.86-8) using do-calculus. Applying (54) to $P_x(y, z|u)$, (with x replacing t), we obtain:

$$\begin{aligned} P_x(y, z) &= \sum_u P_x(y, z|u) P(u) \\ &= \sum_u P_z(y|u) P_x(z) P(u) \\ &= P_x(z) P_z(y) \end{aligned} \tag{55}$$

Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press: Menlo Park, CA, 574-579, August 2002.

TECHNICAL REPORT
R-298
January 2002

Each of these two factors can be derived by simple means; $P_x(z) = P(z|x)$ because Z has no unobserved parent, and $P_z(y) = \sum_{x'} P(y|x', z)P(x')$ because X blocks all back-door paths from Z to Y (they can also be derived by applying (54) to $P(x, y, z|u)$). As a result, we immediately obtain the desired quantity:

$$P_x(y, z) = P(z|x) \sum_{x'} P(y|x', z)P(x'), \qquad (56)$$

a result that required many steps in do-calculus.

In general, from (54), we have

$$P_t(v \setminus t) = \sum_u \prod_{\{i|V_i \in V \setminus T\}} P_{pa_i}(v_i|u^i)P(u). \qquad (57)$$

Depending on the causal graph, the right hand side of (57) may sometimes be decomposed into a product of summations as

$$P_t(v \setminus t) = \prod_j \sum_{n_j} \prod_{V_i \in S_j} P_{pa_i}(v_i|u^i)P(n_j)$$
$$= \prod_j P_{v \setminus s_j}(s_j), \qquad (58)$$

where N_j's form a partition of U and S_j's form a partition of $V \setminus T$. Eq. (55) is an example of such a decomposition. Therefore the problem of identifying $P_t(v \setminus t)$ is reduced to identifying some $P_{v \setminus s_j}(s_j)$'s. Based on this decomposition, a method for systematically identifying causal effects is given in (Tian & Pearl 2002). This method provides an alternative inference tool for identifying causal effects. At this stage of research it is not clear how the power of this method compares to that of do-calculus.

Conclusion

We have shown that all experimental results obtained from an underlying Markovian causal model are fully characterized by three norms of coherence: Effectiveness, Markov, and Recursiveness. We have further demonstrated the use of these norms as inferential tools for identifying causal effects in semi-Markovian models. This permits one to predict the effects of actions and policies, in the presence of unmeasured variables, from data obtained prior to performing those actions and policies.

The key element in our characterization of experimental distributions is the generic formulation of the Markov property (9) as a relationship among three experimental distributions, instead of the usual formulation as a relationship between a distribution and a graph (as in (1)). The practical implication of this formulation is that violations of the Markov property can be detected without knowledge of the underlying causal graph; comparing distributions from just three experiments, $P_{v \setminus (s_1 \cup s_2)}(s_1, s_2)$, $P_{v \setminus s_1}(s_1)$, and $P_{v \setminus s_2}(s_2)$, may reveal such violations, and should allow us to conclude, prior to knowing the structure of G, that the underlying data-generation process is non-Markovian. Alternatively, if our confidence in the Markovian nature of the data-generation process is unassailable, such a violation would imply that the three experiments were not conducted on the same population, under the same conditions, or that the experimental interventions involved had side effects and were not properly confined to the specified sets S_1, S_2, and $S_1 \cup S_2$.

This feature is useful in efforts designed to infer the structure of G from a combination of observational and experimental data; a single violation of (9) suffices to reveal that unmeasured confounders exist between variables in S_1 and those in S_2. Likewise, a violation of any inequality in (51) would imply that the underlying model is not semi-Markovian; this means that feedback loops may operate in data generating process, or that the interventions in the experiments are not “atomic”.

Acknowledgements

This research was supported in parts by grants from NSF, ONR, AFOSR, and DoD MURI program.

References

Dawid, A. 2001. Influence diagrams for causal modelling and inference. *International Statistical Review*. Forthcoming.

Galles, D., and Pearl, J. 1998. An axiomatic characterization of causal counterfactuals. *Foundations of Science* 3(1):151–182.

Halpern, J. 1998. Axiomatizing causal reasoning. In Cooper, G., and Moral, S., eds., *Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann. 202–210.

Heckerman, D., and Shachter, R. 1995. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research* 3:405–430.

Kiiveri, H.; Speed, T.; and Carlin, J. 1984. Recursive causal models. *Journal of Australian Math Society* 36:30–52.

Lauritzen, S. 2000. Graphical models for causal inference. In Barndorff-Nielsen, O.; Cox, D.; and Kluppelberg, C., eds., *Complex Stochastic Systems*. London/Boca Raton: Chapman and Hall/CRC Press. chapter 2, 67–112.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligence Systems*. San Mateo, CA: Morgan Kaufmann.

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

Pearl, J. 1995. Causal diagrams for experimental research. *Biometrika* 82:669–710.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. NY: Cambridge University Press.

Reichenbach, H. 1956. *The Direction of Time*. Berkeley: University of California Press.

Robins, J. 1987. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases* 40(Suppl 2):139S–161S.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag.

Tian, J., and Pearl, J. 2002. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles.