Likelihood-based Causal Inference*

Qing Yao†
Harvard University
qyao@hsph.harvard.edu

David Tritchler‡
Ontario Cancer Institute
and
University of Toronto
tritchle@oci.utoronto.ca

SUMMARY

A method is given which uses subject matter assumptions to discriminate recursive models and thus point toward possible causal explanations. The assumptions alone do not specify any order among the variables — rather just a theoretical absence of direct association. We show how these assumptions, while not specifying any ordering, can when combined with the data through the likelihood function yield information about an underlying recursive order. We derive details of the method for multinormal random variables and apply the procedure to a simulated example.

Key words : DIRECTED ACYCLIC GRAPHS; LINEAR RECURSIVE REGRESSION; MAXIMUM LIKELIHOOD; MONTE CARLO; CAUSAL ORDERING.

1 INTRODUCTION

Starting from Sewall Wright (1934), directed graphs have been used to represent structures in which variables 'cause' or 'influence' other variables. Nodes of the graph are used to represent variables and an arrow from one variable to another indicates that the first has a direct causal influence on the second, an influence not blocked by holding constant others considered.

If the graphs are restricted to directed acyclic graphs (DAGs) by prohibiting directed cycles, then there exists an ordering of the vertices in the DAG consistent with the direction of the edges, in that all variables are ordered after their causes in a causal or temporal sense. Conversely, an ordering of the variables can specify a recursive model for which statistical analysis is routine (see section 3), and which will define a DAG.

Thus there is a correspondence between causal descriptions, DAGs, and recursive statistical models. A search for likely causal explanations can be implemented as a search for good-fitting recursive

*The authors thank David Andrews, Paul Corey, Michael Escobar, Wayne Oldford, Judea Pearl and Robert Tibshirani for helpful comments and discussion.

†Supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada.

‡Supported by a research grant from the Natural Sciences and Engineering Research Council of Canada.

models. The interpretation of such a model is controversial. Cox and Wermuth (1993) prefer to restrict the term causal to "situations when there is some understanding of an underlying process" and indicate that recursive models "could be consistent with a causal explanation". Pearl and Verma (1991) and Spirtes et al (1993) regard directed association in a good-fitting DAG as a definition of causality. Our view is similar to Cox and Wermuth's, but the causal interpretation of recursive models is not central to this paper; our focus is the identification of recursive models. Throughout, we use the term causal to indicate a directed edge in a DAG, as this corresponds with the intuition for such models; we make no claims about an underlying mechanism. We use the term 'causal ordering' synonymously with 'recursive ordering', the sequence of variables in a recursive model, again in accordance with intuition.

It is standard practice to rely on purely subject matter considerations to specify the causal ordering of variables. An important question is: can we obtain any information on the ordering from the data? We argue in section 3 that, in the absence of external assumptions, we cannot. Some researchers (Pearl and Verma, 1991; Spirtes et al, 1993) have emphasized stability assumptions which a priori rule out certain probability models, in a way that allows causal inferences to be made. We discuss these assumptions in section 8 and state reservations about relying on them.

We agree with Cox and Wermuth (1993) that "it is unrealistic to think that causality could be established from a single empirical study or even from a number of studies of similar form without injection of external information". In this paper we give a method which uses subject matter assumptions to discriminate recursive models and thus point toward possible causal explanations. The assumptions alone do not specify any order among the variables — rather just a theoretical absence of direct association. We show how these assumptions, while not specifying any ordering, can when combined with the data through the likelihood function yield information about the underlying recursive order.

The next section introduces DAGs, linear orders, and recursive models. Section 3 describes the recursive analysis of normally distributed data and gives a method for extracting and interpreting information about causal order, deriving the details for the Gaussian linear model. Section 4 describes the information yielded by the method, which can take the form of a partial order which indicates the relative order of pairs of variables. The method is applied to a simulated data in section 5. Section 6 gives a Bayesian Monte Carlo computational method based on importance sampling. Section 7 discusses causation in the context of our method. The last section discusses some limitations of our method and other methods, and makes some concluding comments.

2 DAGS AND RECURSIVE MODELS

A directed graph is a graph representing random variables and the statistical dependencies among them. Random variables are represented by nodes of the graph, and the direct influence of a variable U upon V is depicted by an arrow (synonymously, directed edge) from node U to node V. If there is a directed edge $U \to V$, we call U a parent and V a child of U, and U is considered a cause of V. A directed path is a sequence $Z_{i_1} \to Z_{i_2} \to \cdots \to Z_{i_n}$ of distinct nodes; here we say that Z_{i_n} is a descendent of Z_{i_1} and Z_{i_1} is an ancestor of Z_{i_n}. A cycle is a directed path from a node to itself. A directed acyclic graph is a directed graph with no cycles.

2.1 Markov properties of DAGs

Kiiveri and Speed (1982) related causal influences represented by a directed graph without cycles to conditional independence constraints. The connection, by what are called Markov conditions assures that any direct dependencies between random variables X and Y can not be explained away by holding constant others considered.

The (local) Markov condition specifies that every vertex is statistically independent of its non-descendants given its parents. This is equivalent to the recursive factorization of the joint distribution of the vertices

$$f(\mathbf{V}) = \prod_{v \in \mathbf{V}} f(v|parents(v)), \tag{1}$$

where $f(v) \neq 0$. This factorization uniquely determines the joint distribution. The Markov condition is intuitively desirable in a model for causal influence in that immediate causes should shield v from every variable that is not a consequence of v. Recent work allows us to deduce from the DAG all conditional and marginal independencies implied by the factorization (1) (Pearl, 1988; Lauritzen et al, 1990).

2.2 Orderings and recursive models

Suppose that we label the nodes by numbering them so that $i < j \Rightarrow i \in nd(j)$ in any feasible DAG model, where $nd(j)$ is the set of non-descendents of node j. That is, arrows always point from low numbers to higher numbers in accordance with causal or temporal order. Thus if i is a potential cause of j, $i < j$. This labeling maps a variable to its temporal or causal order i, while the inverse maps i to $X(i)$. Thus i and $X(i)$ are synonymous for a variable ordered *ith*. We write $X(i) \prec X(j)$ whenever $i < j$. For sets $A = \{a_k\}$ and $B = \{b_l\}$, we write $A \prec B$ whenever $a_k \prec b_l$ for all such pairs. There are several notions implicit in the numbering process. We cannot have $i < i$, so causation is irreflexive. For any two variables we must have either $i < j$ or $j < i$, but not both. Thus causation is asymmetric. Also, there is an underlying temporal or causal order determining for every pair of variables which one influences the other assuming that a direct association exists, so we are defining a complete relation. This is in tune with two ideas: i) every direct statistical association not due to latent factors is causal (Reichenbach, 1956) and thus directed and ii) our set of variables is comprehensive and no associations are due to latent variables which are common causes of pairs of observed variables, since such an association would be symmetric. Finally, properties of the integers express the requirement that causation or temporal precedence is transitive.

The numbering of variables defines a complete, irreflexive, assymetric, transitive relation, which is formally referred to as a linear order or ordering. Viewed as a relation, it is the set of all ordered pairs of variables $(X(i), X(j)), i < j$. We may also write such an ordered pair as $X(i) \prec X(j)$ for emphasis. An ordering can be written as $X(1)X(2)\cdots X(n)$ or, alternatively, as $X(1) \prec X(2) \cdots \prec X(n)$. It can also be viewed as a complete directed graph with edges $X(i) \rightarrow X(j)$ for $i < j$. The transitivity of the relation implies that the graph is acyclic, so the ordering is equivalent to a complete DAG. We can obtain DAG models for the data by deleting edges from the complete DAG

specified by an ordering. Pearl (1988) and Lauritzen et al (1990) show that the following property, the local well numbering property, is equivalent to the Markov property.

$$X(i) \perp \{X(j); j < i\} \mid parents(X(i)) \tag{2}$$

The set $\{X(j); j < i\}$, will be referred to as $pred(X(i))$, the predecessors of $X(i)$. Given an ordering we can determine the parents of every variable by a recursive series of analyses of the ordered sets $\{X(2), X(1)\}, \{X(3), X(2), X(1)\}, \cdots, \{X(n), X(n-1), \cdots, X(1)\}$, where each variable is regressed on its predecessors. The resulting graph will possess the Markov property due to the equivalence of the Markov property and (2).

3 INFERENCE FOR LINEAR RECURSIVE MODELS

3.1 Linear recursive regression

Recursive equations have been related to path analysis and the statistical theory of covariance selection (Wermuth, 1980), and later to graphical models (Wermuth and Lauritzen, 1983). Wermuth and Lauritzen (1990) and Cox and Wermuth (1993) extend the system to a broader context, introducing systems of block-recursive regression for chain graphs.

Throughout, we assume a linear relationship between the variables in a p-dimensional random vector $\mathbf{X} = (X_1, X_2, ..., X_p)^T$. For simplicity, we also assume that X_i follows an origin-centered, non-degenerate continuous distribution with covariance matrix $\mathbf{\Sigma}$. Therefore for a linear order ϕ, if the variables are indexed, i.e. $X(i) = X_i$ we may write the above recursive analysis in the form of linear recursive equations as

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \beta_{21} X_1 + \epsilon_2 \\ X_3 &= \beta_{31\cdot 2} X_1 + \beta_{32\cdot 1} X_2 + \epsilon_3 \\ &\vdots \\ X_p &= \beta_{p1\cdot 23\cdots p-1} X_1 + \beta_{p2\cdot 13\cdots p-1} X_2 + \cdots + \beta_{p,p-1\cdot 12\cdots p-2} X_{p-1} + \epsilon_p, \end{aligned} \tag{3}$$

where ϵ_i's are errors assumed to be independently distributed with mean zero and variance d_{ii}. Or we can write the above equations as

$$\mathbf{BX} = \epsilon, \tag{4}$$

where $\mathbf{B}$ is a lower triangle matrix with diagonal elements 1, lower off-diagonal elements coefficients of the above equations, and $\epsilon \sim f(0, \mathbf{D})$ with $\mathbf{D}$ a diagonal matrix (d_{ii}). In the equation corresponding to X_i, only predecessors of X_i appear on the right hand side, so the non-zero coefficients correspond to parents of X_i. Taking variances of both sides of equation (4) gives

$$\mathbf{\Sigma} = \mathbf{B^{-1}DB^{-T}}. \tag{5}$$

For a particular order, $(\mathbf{B}, \mathbf{D})$ is a one-to-one transformation of $\mathbf{\Sigma}$. Hence the linear recursive equations completely capture the covariance structure of the data. We further impose the condition

that ϵ (and thus $\mathbf{X}$) is normally distributed, so that

$$\beta_{ij \cdot 1 \cdots j-1, j+1, \cdots i-1} = 0 \qquad \textbf{iff} \qquad X_i \perp X_j | (X_1 \cdots X_{j-1}, X_{j+1} \cdots X_{i-1}). \tag{6}$$

We may express this last conditional independence as $X_i \perp X_j \mid pred(X_i, X_j)$.

If the ordering of variables is given, then we use maximum likelihood to estimate the regression parameters of the linear recursive equations, to learn which independencies are implied by the data. Because of the separability of parameters (Whittaker, 1994) implied by (3), the MLEs of the recursive system can be estimated by each regression separately; the MLEs for each equation are the LSEs of each regression.

3.2 Discriminating orderings

The order of variables is often unknown, and it would be useful to be able to obtain some information about the causal ordering of the variables from the data. Unfortunately, without restrictions, all orderings give the same maximum likelihood when linear recursive equations are fit. This is because for any ordering, with no restrictions, the decomposition (5) holds and $(\mathbf{B}, \mathbf{D})$ is a unique one to one transformation of $\mathbf{\Sigma}$. Thus by the composition of functions there exists a unique one to one transformation between the parameters $(\mathbf{B_1}, \mathbf{D_1})$ and $(\mathbf{B_2}, \mathbf{D_2})$ for any two orderings. Thus the likelihood of these parameterizations will agree, and the observed data by itself will not suggest any ordering information.

However, we can discriminate the fit of different orderings by injecting external information; if we fix different parameters, and maximize the likelihood over the remaining parameters, the maximum likelihoods will vary. In a relatively well-understood domain, we may be able to specify that some variables can not directly influence each other. That is, we may state that for a pair of variables, neither variable can be a parent of the other. By Verma and Pearl (1990), then they are independent given their ancestors; we write this as $X_i \perp X_j \mid An(X_i, X_j)$. Although this information does not tell us anything directly about ordering of variables, we will show that it will give us information about the underlying structure when combined with the data through the likelihood function. Making some assumptions about infeasible associations among the variables, we will draw statistical inferences about the ordering of variables using maximum likelihood.

To assume X_i and X_j are not directly related means X_i and X_j are not adjacent in the underlying DAG, which is modelled in the recursive system as $\beta_{X_i X_j \cdot pred(X_i, X_j)} = 0$ according to (6). By applying such assumptions to the recursive system, i.e. restricting some regression coefficients to zero, for different orderings the likelihood will vary. The assumptions have injected some ordering information into the system by implying different parameterizations and restrictions for different orderings, in that the set $pred(X_i, X_j)$ varies for different orderings. Therefore given such assumptions or restrictions, the likelihood can tell us which orderings are consistent with the data and assumptions. Note that these assumptions are symmetric with respect to X_i and X_j; that is, they do not directly express any information about order.

To determine which orders are consistent with the data under the assumptions we use the generalized log-likelihood ratio test. For a particular ordering ϕ, define $l(\phi)$ to be the log-likelihood for

the system (4) when the vector $\mathbf{X}$ is ordered as ϕ. The log-likelihood $l(\phi)$ is subject to restrictions on regression coefficients based on subject matter knowledge and maximized over the remaining parameters, for ordering ϕ. For large samples, goodness of fit is tested using the statistic

$$R(\phi) = -2[Max\{l(\phi)\} - Max\{l_0\}] \sim \chi^2_{d.f.}, \tag{7}$$

where ϕ is the linear order determining the recursive model, $Max\{l(\phi)\}$ is the maximum log-likelihood for the restricted model, $Max\{l_0\}$ is for the unrestricted model and does not depend on the ordering, and $d.f.$ = Number of restrictions (Whittaker, 1990). For different orderings the restrictions are on different parameters, i.e. $\beta_{ij \cdot pred(i,j)} = 0$ changes meaning since the set $pred(i,j)$ depends on the ordering. If the assumptions are true, in the limit the true ordering fits the data perfectly, and gives the same likelihood as the unrestricted case (Yao, 1994). Generally, if assumptions are true, for sufficiently large samples the true ordering should fit the data well and give a relatively high maximum likelihood. Theory does not, however, indicate that an incorrect order must be rejected. Conceivably a wrong ordering could also fit the data well; other external information or subject matter knowledge is the only way to choose among the orderings which yield a high maximized likelihood. If no ordering gives good fit, the assumptions should be re-investigated, or there may be important variables omitted from the analysis.

4 INTERPRETING SETS OF ORDERINGS

4.1 Equivalent sets

The general idea of the algorithm is to permute orders of variables, then for each ordering ϕ to calculate the maximum likelihood with restrictions, and finally select the orderings with low values of $R(\phi)$ of (7). However, there is no need to calculate the restricted maximum likelihood for all linear orders, as some orderings are equivalent with respect to the restrictions. Equivalent orderings are determined by the following theorems.

Theorem 1 *If we have assumptions* $i_1 \perp j_1 | pred(i_1, j_1)$, $\cdots$, $i_m \perp j_m | pred(i_m, j_m)$, *and all* $i_k < j_k$, *then for an order* $\mathbf{A_0} \prec j_1 \prec \mathbf{A_1} \prec \cdots \prec j_m \prec \mathbf{A_m}$, *where* $\mathbf{A_k}$*'s are the sets between* j_{k-1} *and* j_k, *permutation within any* $\mathbf{A_k}$*'s,* $k = 1, 2, \cdots, m$, *does not affect the maximum likelihood.*

Theorem 2 *Under the conditions of Theorem 1, if* i_k *and* j_k *are adjacent in the ordering, they may be permuted without affecting the maximum likelihood.*

Corollary 1 *If the only assumption is* $i \perp j | pred(i,j)$, *then for an order* $\mathbf{A} \prec j \prec \mathbf{B}$, *where* $i \in A$, $\mathbf{A}$ *is the set preceding* j *and* $\mathbf{B}$ *is the set following* j, *permutation within* $\mathbf{A}$ *and* $\mathbf{B}$ *does not affect the maximum likelihood.*

Corollary 2 *If the only assumption is* $i \perp j | pred(i) \cup pred(j)$, *then for an order* $\mathbf{A} \prec i \prec j \prec \mathbf{B}$, *switching* i *and* j *in the ordering does not affect the maximum likelihood.*

Corollary 3 *If the only assumption is* $i \perp j | pred(i) \cup pred(j)$, *then for an order* $\mathbf{A} \prec i \prec \mathbf{B} \prec j \prec \mathbf{C}$, *switching* i *and* j *in the ordering does not affect the maximum likelihood.*

Thus an assumption about i and j does not yield information about the relative order of i and j.

Proofs of these Theorems and further details can be found in Yao (1994). The theorems indicate the resolution of the method. When no assumptions are made, there is a single equivalent set composed of all orderings. As more assumptions are introduced, the ability to discriminate between orderings improves, in that the equivalent sets become smaller.

We call the set of orderings with the same maximized likelihood by Theorem 1 and 2, the equivalent set. The first step in interpreting the outcome of the above analysis is to eliminate orderings which conflict with subject-matter knowledge about order. We will assume that all infeasible orderings have been eliminated from the equivalent sets.

4.2 Partial orderings

Since directed graphs are so useful for representing causal relations, a graphical interpretation of equivalent sets is desirable. We do this by defining the partial order corresponding to an equivalent set. Any set Θ of orderings, and in particular an equivalent set, defines a partial order Γ as follows: for an ordered pair (x, y), $(x, y) \in \Gamma \Leftrightarrow (x, y) \in \phi$ for every ordering $\phi \in \Theta$ (Dushnik and Miller, 1941). We can thus interpret the equivalent set with the lowest value of $R(\cdot)$ given by (7) as defining the partial order of the variables which is the most consistent with the data and assumptions. The ordered pairs in the partial order derived from an equivalent set are directed edges in the underlying causal graph.

There may be multiple equivalent sets which are accepted by the test (7). These may be entertained as competing explanations of the data and judged individually on their scientific merit. Alternatively, we may pool a number of equivalent sets strongly supported by the data. This expanded set of orderings will determine a partial order in the same fashion as a single equivalent set, by intersecting ordered pairs according to Dushnik and Miller's definition. To guide the pooling of equivalent sets, we may use a hypothesis test based on equation (7) to form a test based confidence set which is a union of equivalent sets. That is for each order ϕ, use the null distribution (7) to accept or reject ϕ at level α. Then the set composed of all accepted equivalent sets will have confidence level α. i.e. P(true order in the confidence set) = P(true order is not rejected by (7)) = 1 - P(order rejected | order is true) = 1 - α or P(order is not in the confidence set | order is true) = α. When the true ordering is in the confidence set, if every member of that set shares an ordered pair (x, y), then the true order must possess that property. As a result, every pair in the resulting partial order agrees with that implied by the underlying true ordering at the significance level α.

4.3 Non-graphical information

Section 4.2 described the relationship, expressible graphically, implied by selecting a set of orderings. However, a set of orderings also implies non-graphical relationships. For example, consider

variables A, B, and C, along with the condition $A \perp B \mid C$. An equivalent set for that condition is $\{CAB, CBA, BCA, ACB\}$. The partial order that is the union of these four orderings is null - there are no ordered pairs in the relation. However, selecting that equivalent set does rule out the orderings ABC and BAC. This tells us that we should not consider models for which C is preceded by *both* A and B. However, this information is not a binary relation, and thus is not expressible graphically.

5 AN EXAMPLE

For example, consider four variables U, V, W, Z from a multivariate normal distribution. Among these variables, we assume that V is not feasibly directly associated with W, nor is U with Z. Expressed statistically, we assume $U \perp Z|pred(U, Z)$ and $V \perp W|pred(V, W)$. With these two restrictions on the linear recursive coefficients, we calculate the maximum likelihood for all equivalent sets. The equivalent sets and their goodness-of-fit statistics (on 2 degrees of freedom) for a simulated data set are given below.

1) ZWUV, WZUV, WUZV, UWZV	R = 0.75;
2) UZVW, ZUVW, ZUWV, UZWV	R = 55.04;
3) ZVUW, VZUW, VUZW, UVZW	R = 225.18;
4) UVWZ, UWVZ, WUVZ, VUWZ	R = 355.54;
5) ZWVU, WZVU, VZWU, ZVWU	R = 381.27;
6) WVZU, WVUZ, VWUZ, VWZU	R = 565.92.

A 95 percent confidence set of orderings consists only of equivalent set 1. This set corresponds to the partial ordering $\{U \prec V, W \prec V, Z \prec V\}$. The relative order of U, W and Z is unknown, but we have been able to conclude that they all precede V with 95 percent confidence, which agrees with the model used to generate the data. If equivalent set 2 had been included in the confidence set we would have inferred only $U \prec V$ and $Z \prec V$.

6 AN ALTERNATIVE STRATEGY

Our method involves permutation and hence exponential computational complexity. Thus, to evaluate posterior probabilities of arbitrary properties of the underlying linear order, we derive an alternative Bayesian importance sampling method.

Let $f(\mathbf{x}; \beta|\phi_i)$ be the restricted likelihood function for a given linear order ϕ_i where restrictions express assumptions as described in section 3. We denote $\hat{f}(\mathbf{x}; \beta|\phi_i)$ as the maximized estimate of the function. If we sample linear orders $\phi_i, i = 1, 2, \cdots, M$ from a prior distribution $p(\phi)$, then for any function $g(\phi)$, we prove the following convergence property (Yao, 1994).

Theorem 3 *Under regularity conditions such that the MLE of β is strongly consistent, as the*

number of observations n and M go to infinity,

$$[\sum_{i=1}^{M} \hat{f}(\mathbf{x};\beta|\phi_i)]^{-1} \sum_{i=1}^{M} g(\phi_i)\hat{f}(\mathbf{x};\beta|\phi_i) \xrightarrow{a.s.} E_{\phi|\mathbf{x}}[g(\phi)], \tag{8}$$

where $E_{\phi|\mathbf{x}}[\cdot]$ is an expectation over the posterior distribution of the orderings.

Let $g(\phi)$ be an indicator function for any given pair of variables, eg. for the pair X and Y, we define

$$g_{xy}(\phi) = \begin{cases} 1, & if \quad X \prec Y \in \phi \\ 0, & o.w. \end{cases} \tag{9}$$

Then $E_{\phi|\mathbf{x}}[g_{xy}(\phi)]$ measures the posterior probability of "$X \prec Y$ implied in the ordering ϕ" under the restrictions. Geweke (1989) gives the importance sampling error of the estimated posterior probability.

$$\sqrt{\sum_{i=1}^{m}\{g_{xy}(\phi_i) - E_{\phi|\mathbf{x}}[g(\phi)]\}^2 w_i^2 / \sum_{i=1}^{m} w_i^2}, \tag{10}$$

where $w_i = \hat{f}(\mathbf{x};\beta,\phi_i)$.

To illustrate, we apply the procedure to the example of the previous section. For this small example we exhaustively sample all orderings. Defining $g_{uv}(\phi) = 1$ if $U \prec V$ in ϕ, and zero otherwise, we compute $Prob(U \prec V) = E[g_{uv}(\phi)] \approx 1.0$. Similar calculations give $Prob(W \prec V) \approx 1.0$, $Prob(Z \prec V) \approx 1.0$ and $Prob(W \prec Z) = 0.5$.

7 CAUSATION

A key assumption attached to the system of equations (4) is that

$$\epsilon \sim f(0, \mathbf{D}), \tag{11}$$

with $\mathbf{D}$ a diagonal matrix (d_{ii}). We now show how this assumption relates to causation, and to the validity of our method.

Stone (1993) shows that if

$$\epsilon_i \perp X_j \mid \{X_1, \ldots, X_{i-1}\}/X_j, 1 \leq j \leq i-1 \tag{12}$$

in the Gaussian linear model, then $\beta_j = 0$ implies that X_j causes X_i, assuming that the temporal or causal order is consistent with the indexing of the variables. For any positive distribution, (12) holds for all $j = 1, \ldots, i-1$ if and only if

$$\epsilon_i \perp X_1, \ldots, X_{i-1}. \tag{13}$$

This is the usual requirement for the consistency of the least square estimate (Bowden and Turkington, 1984). It is easy to show that in the recursive system (4), the assumption (11) is sufficient

to establish (12) and (13) for $i = 1, \ldots, p$. Thus the usual assumption about error structure implies that causation can be tested by examining the regression coefficients when the linear model is correct and the variables are ordered correctly.

The validity of our proposed method rests on the assumption that if subject-matter knowledge specifies that $X_i \perp X_j \mid An(X_i, X_j)$, i.e. they are not directly related in the underlying DAG, then $X_i \perp X_j \mid pred(X_i, X_j)$ in the correctly ordered recursive model. Note that $pred(X_i, X_j)$ consists of variables actually observed and may be a proper subset of $An(X_i, X_j)$ since the underlying system could include unobserved ancestors. It can be shown that the diagonal covariance structure (11) insures that $X_i \perp X_j \mid An(X_i, X_j)$ implies $X_i \perp X_j \mid pred(X_i, X_j)$.

We have thus shown that the usual assumptions about the error structure of the linear recursive system validate our method, and also allow zero regression coefficients to indicate causation when the ordering is correct. This is not to say that causal inferences may be made routinely. Rather, it emphasizes what an extremely strong assumption (11) is. It amounts to assuming that all relevant variables have been observed, i.e. there is no confounding. For observational data the domain being studied must be very well-understood before we can confidently assume that all relevant factors are being modelled.

8 DISCUSSION

Pearl and Verma (1991) and Spirtes et al. (1993) give valuable algorithms for generating causal models from data. Verma and Pearl (1992) give an algorithm which decides if there is a DAG which explains a list of conditional independence statements. These methods assume, besides the Markov condition, two more conditions, minimality and stability, which enable causal inference from observed data (cross-sectional) without prior assumptions about the nature of variables. Informally, the minimality condition says that each edge in the DAG prevents some conditional independence that would otherwise obtain, and the stability condition says that the only independencies in data are a consequence of the Markov assumptions and can not be caused by a mere cancellation of the effects of other dependencies. Though theoretically cancellation occurs with measure zero (Spirtes et al., 1993), judging independencies in practice relies on statistical tests which will support null association with appreciable probability for a range of parameter values. For limited sample size, the range may be great. Hence in practice we can not assume that approximate cancellation is impossible, and prefer to rely on assumptions that have subject matter justification, as opposed to stability which demands that probabilities in general behave in a certain way.

As in any observational study, omitted variables can lead to false inferences with our method, and results should be interpreted with the appropriate caution. We are proposing this as a method for the generation of possible causal explanations, not a substitute for randomized experiments.

If we are investigating a domain where there is confidence that all the relevant variables are known, and there is sufficient knowledge to make solid assumptions of the type required here, the inferences yielded by the method of this paper will have commensurate validity. Essentially, the more knowledge we start with, the more we can confidently infer.

The method is more general than the multivariate normal case. It pertains to any recursive model for which the parameters are separable and can express pairwise independence.

REFERENCES

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

Cox, D. R. and Wermuth, N. (1993) Linear Dependencies Represented by Chain Graphs (with discussion). *Statistical Science*, **8**, 204-283.

Dawid, A. (1979) Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc.* B, **41**, 1-31.

Dushnik, B. and Miller, E.W. (1941) Partially ordered sets. *American Journal of Mathematics*, **63**, 600-610.

Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317-1339.

Kiiveri, H. and Speed, T. (1982) Structural analysis of multivariate data: A review. *Social Methodology* (ed. Leinhardt, S.). Jossey-Bass, San Francisco.

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

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

Pearl, J. and Verma, T. (1991) A theory of inferred causation. Principles of Knowledge Representation and reasoning: *Proceedings of the Second International Conference*, Morgan Kaufmann, San Mateo, CA.

Reichenbach, H. (1956) *The direction of time.* University of California Press, Berkeley, CA.

Spirtes, P , Glymour, C, and Scheines, R. (1993) Causation, prediction, and search. *Lecture notes in statistics*, **81**. Springer-Verlag, New York.

Stone, R. (1993) The assumptions on which causal inferences rest. *J. Roy. Statist. Soc.* B, **55**, 455-466.

Verma, T.S. and Pearl, J. (1990) Equivalence and synthesis of causal models. In *Proceedings of the Conference on Uncertainty in AI*, Cambridge, MA., July, 1990.

Verma, T.S. and Pearl, J. (1992) An algorithm for deciding if a set of observed independencies has a causal explanation: *Uncertainty in Artificial Intelligence* (D. Dubois, M.P. Wellman, B. D'Ambrosio and P. Smets, eds), **8**, 323-330. Morgan Kaufmann, San Mateo, CA.

Wermuth, N. (1980) Linear recursive equations, covariance selection and path analysis. *J. Am. Statist. Ass.*, **75**, 963-972.

Wermuth, N. and Lauritzen, S. (1983) Graphical and recursive models for contingency tables. *Biometrika* **72**, 537-552.

Wermuth, N. and Lauritzen, S. L. (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. R. Statist. Soc.* B, **52**, 21-72.

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics.* Wiley, New York.

Wright, S. (1934) The method of path coefficients. *Ann. Math. Stat*, **5**, 161-215.

Yao, Q. (1994) Ph.D dissertation. University of Toronto.