J. Austral. Math. Soc. (*Series A*) 36 (1984), 30–52

RECURSIVE CAUSAL MODELS

HARRI KIIVERI, T. P. SPEED and J. B. CARLIN

(Received 30 October 1981)

Communicated by R. L. Tweedie

Abstract

The notion of a recursive causal graph is introduced, hopefully capturing the essential aspects of the path diagrams usually associated with recursive causal models. We describe the conditional independence constraints which such graphs are meant to embody and prove a theorem relating the fulfilment of these constraints by a probability distribution to a particular sort of factorisation. The relation of our results to the usual linear structural equations on the one hand, and to log-linear models, on the other, is also explained.

1980 *Mathematics subject classification* (*Amer. Math. Soc.*): 62 F 99, 60 K 35.
Keywords and phrases: causal models, path analysis, conditional independence, log-linear models.

Introduction

In his initial exposition of path analysis, Wright (1921) introduced into statistics the basic idea of directed graphs whose vertices represent continuous random variables and edges some notion of correlation and causation. Apart from simply depicting the general nature of the linear structural equations which define the causal relations under study, these graphs are also used to write down those partial correlations which must vanish when the equations and the associated distributional assumptions take a standard form, see Blalock (1962). Furthermore, the path analysis rules of Wright (1921, 1934) involve tracing paths in the graph as part of an algorithm giving equations relating the variances and covariances of the random variables. More recently, Goodman (1973a, b) has drawn similar graphs whose vertices correspond to discrete random variables and edges to a

notion of interaction in a probability model of the log-linear type. He has pointed out that in certain examples, these models embody conditional independence constraints on the distribution of the random variables.

In a different context we find that Markov fields over finite *undirected* graphs (that is, probability distributions for random variables identified with the vertices of such graphs which satisfy certain independence constraints defined by the graph) have intimate connexions with the theory of log-linear models, see Darroch *et al.* (1980). A fundamental result in the theory relates Markov fields to so-called nearest-neighbour Gibbs states, and this turns out to include a description of a large class of *independence* or *Markov models* for discrete random variables, see also Speed (1978). Can we do likewise with directed graphs, and does this tie up with path analysis?

Up until now there has been little consistency in the use of graphs in path analysis. Some authors include all possible edges between exogenous variables, making them undirected or bidirectional as they think appropriate, whilst others don't; some include undirectional edges associated with errors in the equations, whereas most authors don't do so, and so on. The difference here are partly explained by varying assumptions concerning the correlation structure on the exogenous variables or the errors in the equations, but there still remains a diversity of practices even when—and this is not always easy to determine—different writers' intentions concerning these issues appear to be the same.

If a standard form of causal graph could be agreed upon, the question of exactly which conditional independence constraints it should be regarded as embodying could then be addressed. These would not depend upon whether or not discrete or continuous random variables were associated with the vertices. Given a satisfactory answer to this question, we would then attempt to describe all joint probability distributions which satisfy the appropriate independence constraints. If successful, the resulting unification of discrete and continuous models, together with the standardisation of terminology and fundamental results which would ensue, should prove of value to those interested in defining, fitting, testing and interpreting causal probability models of data. This has been our program.

In Section 2 we define what we call a *recursive causal graph*, hopefully capturing the essence of the path diagrams associated with *recursive* causal models. These graphs permit neither causal cycles nor simultaneity. We describe the separation properties which help define the independence constraints the graph is meant to embody, and our main theorem relates the fulfilment of these constraints by a probability distribution to a particular sort of factorisation. This theorem is analysed in more detail in Section 3 for Gaussian and Section 4 for multinomial distributions. The relationship of our results to the usual linear

structural equations on the one hand, and to log-linear models, on the other, is also explained in these last two sections. A much more extensive discussion of these ideas with reference and illustrations can be found in Kiiveri and Speed (1982).

2. General results

Our aim in this section is to prove a general result characterising the distribution of what we will be calling a *recursive causal system* of random variables (equivalently, a recursive causal (probability) model). Such systems (models) will always be associated with a particular kind of *graph* and we begin by collecting up some preliminaries concerning these graphs.

2.1. *Causal graphs.* A *causal graph* is an ordered pair $\mathfrak{G} = (V(\mathfrak{G}), E(\mathfrak{G}))$ consisting of a finite set $V(\mathfrak{G}) = V_x(\mathfrak{G}) \cup V_n(\mathfrak{G})$ of *vertices* and a finite set $E(\mathfrak{G}) = E_x(\mathfrak{G}) \cup E_n(\mathfrak{G})$ of *edges*, with vertices in $V_x(\mathfrak{G})$ being termed *exogenous* and those in $V_n(\mathfrak{G})$ *endogenous*; edges in $E_x(\mathfrak{G})$ are *undirected* ones, that is, unordered pairs of distinct exogenous vertices, whilst edges in $E_n(\mathfrak{G})$ are *directed* ones, that is, ordered pairs of distinct vertices, the second element of which is an endogenous vertex. In what follows we denote vertices by natural numbers: $1, 2, 3, \ldots$ or h, i, j; edges are unordered or ordered pairs of vertices and depicted in the usual way, namely $1 - 2$ (undirected) and $i \to j$ (directed) respectively.

EXAMPLE 1. If $V_x(\mathfrak{G}_1) = \{1,2\}$, $V_n(\mathfrak{G}_1) = \{3,4\}$, $E_x(\mathfrak{G}_1) = \{1 - 2\}$, and $E_n(\mathfrak{G}_1) = \{2 \to 3, 3 \to 4, 1 \to 4\}$, then $\mathfrak{G}_1$ may be depicted as in Figure 1.

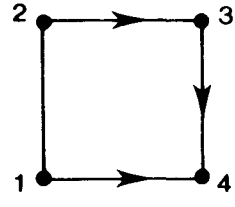

FIGURE 1

We will be adapting standard graph-theoretic notions to our context in which directed and undirected edges coexist, and it is hoped that no confusion will result from so doing. A directed [undirected] *chain* in a causal graph $\mathfrak{G}$ is a sequence $i_0, i_1, \ldots, i_m$ of vertices such that $i_{l-1} \to i_l$ [$i_{l-1} - i_l$] for $l = 1, 2, \ldots, m$, and such a chain is called a *cycle* if $i_0 = i_m$. All of the causal graphs which we consider in

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

this paper are *recursive*, where this term means that the graph in question has *no directed cycles*.

For each $j \in V_n(\mathfrak{G})$ we write $D_j = \{h \in V(\mathfrak{G}): h \to j\}$, and refer to the elements of D_j as *direct causes* of j; $D_j \cup \{j\}$ is denoted by $\overline{D}_j$. Similarly if we write B_j for j and the set of vertices k connected to j via a chain $j \to j_1 \to \cdots \to k$, then $A_j = V(\mathfrak{G}) \setminus B_j$ is termed the set of vertices *anterior* to $j \in V(\mathfrak{G})$; we also write $\overline{A}_j = A_j \cup \{j\}$. The undirected graph with vertices $V_x(\mathfrak{G})$ and edges $E_x(\mathfrak{G})$ will be denoted by $\mathfrak{G}_x$, the subgraph on the exogenous vertices. More generally, the *subgraph* of $\mathfrak{G}$ defined by any subset $B \subseteq V$ of vertices will be denoted by $\langle B \rangle_{\mathfrak{G}}$; its vertices are the elements of B and its edges those in $\mathfrak{G}$ both of whose elements belong to B.

An important object associated with any causal graph $\mathfrak{G}$ is what we call the *underlying undirected graph* $\mathfrak{G}^u$ which has the same set of vertices $V(\mathfrak{G}^u) = V(\mathfrak{G})$, whilst its edges $E(\mathfrak{G}^u)$ are the undirected ones $E_x(\mathfrak{G})$ of $\mathfrak{G}$ together with the additional undirected ones connecting pairs of vertices between which a directed edge exists in $\mathfrak{G}$, that is, $E(\mathfrak{G}^u) = E_x(\mathfrak{G}) \cup \tilde{E}_n(\mathfrak{G})$, where $\tilde{E}_n(\mathfrak{G})$ denotes the directed edges of $\mathfrak{G}$ with their direction omitted.

A triple i, j and k of distinct vertices in $\mathfrak{G}$ is said to be *in configuration* $[>]$ if $i \to k$, $j \to k$ but i and j are not connected by any edge, directed or undirected. This notion, which first appeared in Wermuth (1980), plays a key role in determining the admissible independence statements associated with a causal graph.

If a, b and d are disjoint sets of vertices of $\mathfrak{G}$ we say that a and b are *separated* by d in $\mathfrak{G}^u$ if any chain $i = i_0, i_1, \ldots, i_m = j$ connecting a vertex $i \in a$ with a vertex $j \in b$ necessarily intersects d. Further, we say that a, b and d are in configuration $[>]$ if there is a chain in $\mathfrak{G}^u$ from an element $i \in a$ to an element $j \in b$ which includes a triple $i, j \notin d$ and $k \in d$ in configuration $[>]$ in $\mathfrak{G}$.

Some of our induction arguments will make use of what we will call an extreme endogenous vertex in a causal graph $\mathfrak{G}$, where $i^* \in V(\mathfrak{G})$ is *extreme* if no directed edge $i^* \to j$ exists in $E_n(\mathfrak{G})$. Clearly $\overline{A}_{i^*} = V(\mathfrak{G})$ for such vertices. An easy induction argument proves the validity of the following

LEMMA 1. *Every causal graph has at least one extreme endogenous vertex.*

2.2. *Factorisation of joint densities.* Our main result below concerns factorisations of the joint density $p(\mathbf{x}, \mathbf{y})$ of an array $(\mathbf{X}; \mathbf{Y}) = (X_h: h \in V_x(\mathfrak{G}); Y_j: j \in V_n(\mathfrak{G}))$ indexed by the vertices of a causal graph $\mathfrak{G}$, and it will be convenient to use certain suggestive abbreviations for joint, marginal and conditional densities. (All joint distributions will be given via strictly positive densities with respect to a product measure. In fact all the examples we discuss below are either

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

Gaussian or discrete (multinomial), and these conditions are then satisfied.) In order to illustrate our abbreviations we return to Example 1.

EXAMPLE 1 (continued). Associated with the (recursive) causal graph $\mathcal{G}_1$ we will have four random variables $(X_1, X_2; Y_3, Y_4)$, the X's being termed *exogenous* and the Y's *endogenous variables*. Our later discussion will involve the assumption of independence of Y_3 and X_1 given X_2 and also of Y_4 and X_2 given (X_1, Y_3); we abbreviate the conditions to $1 \perp 3/2$ and $2 \perp 4/1, 3$. Similarly the factorisations of the joint density p of the variables which are equivalent to these independence assertions:

$$p_{\{1,2,3\}}(x_1, x_2, y_3) = \frac{p_{\{1,2\}}(x_1, x_2)p_{\{2,3\}}(x_2, y_3)}{p_{\{2\}}(x_2)}$$

and

$$p(x_1, x_2, y_3, y_4) = \frac{p_{\{1,2,3\}}(x_1, x_2, y_3)p_{\{1,3,4\}}(x_1, y_3, y_4)}{p_{\{1,3\}}(x_1, y_3)}$$

are abbreviated to

$$(123) = \frac{(12)(23)}{(2)} \quad \text{and} \quad (1234) = \frac{(123)(134)}{(13)}.$$

Finally, the factorisation which embodies both of these conditions:

$$p(x_1, x_2, y_3, y_4) = p_{\{1,2\}}(x_1, x_2)p_{3|2}(y_3 | x_2)p_{4|\{1,3\}}(y_4 | x_1, y_3)$$

is abbreviated to

$$(1234) = (12)(3|2)(4|13).$$

This illustration should explain how our abbreviations are intended to be read.

We will be making considerable use of the notions and results concerning *Markov random fields* over finite undirected graphs which can be found in Darroch *et al.* (1980) and Speed (1978, 1979). A distribution (V_x) for a set of random variables indexed by an *undirected* graph $\mathcal{G}_x = (V_x, E_x)$ is said to be *Markov* over $\mathcal{G}_x$ if it satisfies either of the conditions:

Local Markov Condition: For each $h \in V_x$ the conditional distribution $(h | V_x \backslash \{h\})$ of X_h given all the X_g, $g \neq h$, coincides with $(h | \partial h)$, the conditional distribution of X_h given all X_g with $g \in \partial h = \{i: \{h, i\} \in E_x\}$.

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

Global Markov Condition: For disjoint subsets a, b and d of V_x such that d separates a from b in V_x, we have

$$(a \cup b \cup d) = \frac{(a \cup d)(b \cup d)}{(d)}.$$

An extension of the global Markov condition to disjoint subsets $a_1, \ldots, a_m$ and d with a_k and a_l separated by d $(1 \leqslant k < l \leqslant m)$ is readily found to be equivalent to the condition stated here. The general equivalence of the local and global Markov conditions over an arbitrary finite graph does not seem to be explicit in the published literature. It is well known for discrete random variables, where it follows from a characterisation of all corresponding probability distributions, see Speed (1979) for this result (and many references to equivalent ones), while the remarks on page 194 of that paper show how to get the general result.

2.3. *The main theorem*. This subsection is devoted to the statement and proof of the main result of the paper. It is a fairly natural extension of the corresponding result for purely undirected graphs, although it cannot go too far without some restrictions on the type of probability densities under consideration. Each of the important cases—the Gaussian and the multinomial—is discussed later in the paper, and it turns out that statement (1) of the theorem is the lead-in to a reasonable parametrisation, that is, a complete description, of all such probability densities in these two cases. In a sense the theorem together with Proposition 4 below provides a *directed* analogue of the Hammersley-Clifford or $NNG = M$ theorem, so-called in Speed (1979).

THEOREM. *Let* $\mathfrak{G}$ *be a recursive causal graph and* $(\mathbf{X}; \mathbf{Y})$ *a system of random variables indexed by the vertices of* $\mathfrak{G}$:

$$(\mathbf{X}; \mathbf{Y}) = \big(X_h : h \in V_x(\mathfrak{G});\ Y_j : j \in V_n(\mathfrak{G})\big).$$

The following are equivalent for a strictly positive joint density (V):

(1) *The recursive causal factorisation*:

(i) (V_x) *is Markov over the undirected graph* $\mathfrak{G}_x$; *and*

(ii)

$$(\text{RCF}) \qquad (V) = (V_x) \prod_{j \in V_n} (j \mid D_j).$$

(2) *The Global Markov property for causal graphs*:

For all families $a_1, a_2, \ldots, a_m, d$ *of* $m + 1 \geqslant 3$ *pairwise disjoint subsets of* V *satisfying*

(i) $\bigcup_1^m a_l \cup d = V_x$ *or, for some* $j \in V_n$, $\bigcup_1^m a_l \cup d = \bar{A}_j$;

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

and

(ii) *for* $1 \leqslant k < l \leqslant m$ *the sets* a_k *and* a_l *are separated by* d *in* $\mathfrak{G}_x$ *or* $\langle \bar{A}_j \rangle_{\mathfrak{G}}$, *as the case may be, and are not in configuration* $[>]$ *in* $\langle \bar{A}_j \rangle_{\mathfrak{G}^u}$,

we have

$$\text{(GM)} \qquad \left(\bigcup_1^m a_l \cup d \right) = \frac{\prod_1^m (a_l \cup d)}{(d)^{m-1}};$$

(3) *As in* (2) *above but with the* $=$ *in* (i) *replaced by* $\subseteq$;

(4) *The Local Markov property for causal graphs*:

(i) (V_x) *is locally Markov over the undirected graph* $\mathfrak{G}_x$; *and*

(ii) *For all* $j \in V_n$:

$$\text{(LM)} \qquad \left(\bar{A}_j\right) = \frac{\left(\bar{D}_j\right)\left(A_j\right)}{\left(D_j\right)}.$$

As an illustration of the theorem, we return to our example.

EXAMPLE 1 (continued). We have already asserted the equivalence of the factorisation $(1234) = (12)(3\,|\,2)(4\,|\,13)$ with the pair of factorisations $(123) = (12)(23)/(2)$ and $(1234) = (123)(134)/(13)$. These assertions—which are easily checked directly—can now be regarded as an instance of the theorem just stated; for example, $\bar{A}_4 = \{1, 2, 3, 4\}$, $D_4 = \{1, 3\}$, whilst $\bar{D}_4 = \{1, 3, 4\}$.

REMARK. Each assertion in the statement of the theorem has essentially two parts: one concerning (V_x) relative to $\mathfrak{G}_x$, and one concerning other aspects of (V) in relation to $\mathfrak{G}$. The assertions concerning (V_x) and $\mathfrak{G}_x$ are either the same or equivalent by the basic theorem concerning Markov probabilities over *undirected* graphs referred to in Section 2.2, and will not be referred to any further in the proof which follows.

PROOF. (1) *implies* (2). We do this by induction on the cardinality $|\,V_n\,|$ of V_n assuming that $|\,V_x\,| = p \geqslant 1$. Let us suppose that $|\,V_n\,| = q = 1$, that is, assume

$$(1) \qquad (V) = (V_x)(\,p + 1\,|\,D_{p+1}),$$

and suppose that $a_1, a_2, \ldots, a_m$ and d satisfy (i) and (ii) of (2) with union $\bar{A}_{p+1}$. To begin the proof we show that $\bar{D}_{p+1} \subseteq a_{l^*} \cup d$ for some $l^* \in \{1, \ldots, m\}$. If $p + 1 \in d$ and $D_{p+1} \subseteq d$ the result is obvious, so we consider the case when there exists an $i \in D_{p+1}$ and $i \notin d$. For this i there is a (unique) l^* such that $i \in a_{l^*}$. Now suppose that we have a $j \in D_{p+1}$ and $j \in a_l$ for $l \neq l^*$. Then a_l, a_{l^*} and d are in configuration $[>]$, contradicting our assumption. Hence $\bar{D}_{p+1} \subseteq a_{l^*} \cup d$ in this

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

case. On the other hand if $p+1 \in a_{l^*}$ for some l^* and $i \in D_{p+1}$ satisfies $i \in a_l$ for $l \neq l^*$, we contradict the separation assumption. Hence $D_{p+1} \subseteq a_{l^*} \cup d$ in both cases and the assertion follows.

Our proof (of the case $q = 1$) is now completed separately for each of the cases $p+1 \in d$ and $p+1 \notin d$. Let us start with the latter, observing that in this case $\{a_l\colon l \neq l^*\} \cup \{a_{l^*} \backslash \{p+1\}\}$ is a family of m pairwise disjoint subsets of (V_x) separated by $d \subseteq V_x$. By the undirected global Markov property

$$(2) \quad \Big(\bigcup_{l \neq l^*} a_l \cup d \cup a_{l^*} \backslash \{p+1\} \Big) = \frac{\prod_{l \neq l^*} (a_l \cup d) \cdot (a_{l^*} \cup d \backslash \{p+1\})}{(d)^{m-1}}$$

and this part of the proof would be completed if we could include the singleton $\{p+1\}$ in those parts above where it is excluded. Integrate out all variables from both sides of equation (1) except those in $(a_{l^*} \cup d)$; we obtain

$$(3) \qquad (a_{l^*} \cup d) = (a_{l^*} \cup d \backslash \{p+1\})(p+1 \mid D_{p+1}).$$

In a similar way we can integrate our variables† from both sides of (1) until its left-hand side coincides with that of equation (2) except that $p+1$ remains, and we get

$$(4) \qquad \Big(\bigcup_{l} a_l \cup d \Big) = \Big(\bigcup_{l \neq l^*} a_l \cup d \cup a_{l^*} \backslash \{p+1\} \Big)(p+1 \mid D_{p+1}).$$

The desired result now follows from equations (2), (3) and (4).

The remaining case is when $p+1 \in d$. Here we put $d^* = d \backslash \{p+1\}$ and observe that d^* separates $a_1, \ldots, a_m$ in V_x and so we have again by the undirected global Markov property:

$$(5) \qquad (V_x) = \frac{\prod_l (a_l \cup d^*)}{(d^*)^{m-1}}.$$

Now we can integrate out variables from both sides of equation (1) to obtain

$$(a_{l^*} \cup d) = (a_{l^*} \cup d^*)(p+1 \mid D_{p+1})$$

and this combines with equations (5) and (1) to give

$$(6) \qquad (V) = \frac{\prod_{l \neq l^*} (a_l \cup d^*)(a_{l^*} \cup d)}{(d^*)^{m-1}}.$$

Finally, we integrate all variables except those in $a_l \cup d$, $l \neq l^*$, out of equation (6) and get down to

$$(a_l \cup d) = \frac{(a_l \cup d^*)}{(d^*)}(d), \qquad l \neq l^*.$$

† There are none in this case ($q = 1$) but there will be in the inductive step ($q > 1$).

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

This, together with equation (5), gives us what we want. Thus the induction argument has begun.

Suppose now that the implication has been proved for all recursive causal graphs $\mathcal{G}$ having $|V_n| < q$ vertices, $q > 1$, and let us consider such a graph with $|V_n| = q$. Take an extreme endogenous vertex $j^* \in V(\mathcal{G})$ and consider the smaller graph $\mathcal{G}^*$ with j^* and its incident edges removed from $\mathcal{G}$. This satisfies our induction hypothesis, and we now prove the induction step in much the same way that the induction argument was begun. For this reason we present the argument only in outline.

Given a system $a_1,\ldots,a_m, d$ satisfying (i) and (ii) of (2) in the statement with $\bigcup_l a_l \cup d = \bar{A}_j$, we first note that if $j^* \notin \bar{A}_j$, then the result follows from our inductive hypothesis. Thus we need only consider the case $j^* \in \bar{A}_j$, and here we readily observe that $i \in \bar{A}_j$ also holds for all $i \in D_j$, that is, that $D_{j^*} \subseteq \bar{A}_j$. An earlier argument now proves that $D_{j^*} \subseteq a_{l^*} \cup d$ for a unique l^*, and the first part of this proof is indicated.

The remainder of the proof of the induction step goes as before. If $j^* \notin d$ then the $\{a_l, l \neq l^*\}$, $a_{l^*} \setminus \{j^*\}$ and d satisfy the conditions (i) and (ii) of (2) in $\mathcal{G}^*$ and the induction hypothesis together with the earlier argument completes the proof. On the other hand, if $j^* \in d$, then $\{a_l\colon l = 1,\ldots,m\}$ and $d^* = d \setminus \{j^*\}$ can be used; again the details are the same as in the earlier argument. Thus the induction step and so the whole implication is proved.

(2) *implies* (3). This implication will be proved if we can extend any system $\{a_1,\ldots,a_m\}$ and d satisfying 2(i) and (ii) with only $\subseteq$ in 2(i), to a system $\{a_1^*,\ldots,a_m^*\}$ and d with $a_l^* \supseteq a_l$, $l = 1,\ldots,m$ satisfying 2(i) and (ii) but with $=$ in 2(i). For then the variables in $a_l^* \setminus a_l$, $l = 1,\ldots,m$ may be integrated out to prove that the desired factorisation for the original sets is a consequence of that for the enlarged sets.

The desired extension is a purely graph-theoretic matter. We begin with a system $\{a_1,\ldots,a_m\}$ and d satisfying 2(i) and (ii), but with $\bigcup_l a_l \cup d \subseteq \bar{A}_j$ say. Consider the class of all systems $\{b_1,\ldots,b_m\}$ and d which satisfy all the relevant separation properties of 2(i) and (ii), and further, $b_l \supseteq a_l$, $l = 1,\ldots,m$. This is clearly a finite non-empty class and so must possess elements maximal in the componentwise ordering. Let $\mathcal{L}\colon \{a_1^*,\ldots,a_m^*\}$ and d, be such a maximal system, and suppose that $\bigcup_l a_l^* \cup d \subset \bar{A}_j$. Then there exists j^* belonging to $\bar{A}_j$ but not to $\bigcup_l a_l^* \cup d$, and for each l, $1 \leqslant l \leqslant m$, the system $\mathcal{L}_l\colon \{a^*,\ldots,a_l^* \cup \{j^*\},\ldots,a_m^*\}$ and d, must violate one or the other of the restrictions of 2(ii). Let us consider $\mathcal{L}_1$. Then there exists $k \in \{2,\ldots,m\}$ and a chain $j_0 = j^*,\ldots,j_p \in a_k^*$ which either fails to intersect d, and so violates the separation requirement of 2(ii), or meets d in configuration [>], thereby violating the other requirement of 2(ii). In a similar way we may consider $\mathcal{L}_k$; there will exist $l \in \{1,\ldots,m\} \setminus \{k\}$ and a chain

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

$j_0' = j^*, \ldots, j_q' \in a_l^*$ which violates either the separation or the configuration [>] requirement of 2(ii). This gives us four cases, each of which leads to a contradiction, and so we conclude that no such j^* exists.

To see this, suppose that the separation requirement is violated in both cases. Then we will have a chain from $j_p \in a_k^*$ to $j_q' \in a_l^*$ (via j^*) which does not meet d, contrary to our hypothesis about $\mathfrak{L}^*$. The other three possibilities are dealt with in a similar manner and our conclusion follows.

Thus any maximal system $\mathfrak{L}$ has union the whole of $\bar{A}_j$ and the remainder of the proof that (2) implies (3) is as outlined at the beginning.

(3) *implies* (4). This is immediate: simply take $m = 2$, $a_1 = \{j\}$, $a_2 = A_j \setminus D_j$ and $d = D_j$ in (GM) and (LM) follows.

(4) *implies* (1). Once more we use induction on $|V_n(\mathfrak{G})|$. When $|V_n| = 1$, that is, when $V_n = \{p+1\}$, the factorisation (LM) with $j = p+1$ is just (RCF). Thus our induction argument can begin.

Suppose now that the implication is true for all $\mathfrak{G}$ with $|V_n(\mathfrak{G})| < q$, $q > 1$, and that we have a $\mathfrak{G}$ with $|V_n(\mathfrak{G})| = q$. Take an extreme endogenous vertex, j^* say, and notice that $A_{j^*} = V \setminus \{j^*\}$. Thus (LM) with $j = j^*$ gives us

$$(V) = \frac{(V \setminus \{j^*\})(\bar{D}_{j^*})}{(D_{j^*})} = (V \setminus \{j^*\})(j^* \mid D_{j^*})$$

whilst our inductive hypothesis gives us

$$(V \setminus \{j^*\}) = (V_x) \prod_{j \in V_n \setminus \{j^*\}} (j \mid D_j).$$

These last two equations combine to give (RCF) for the whole of V.

Our next result incorporates the work of Wermuth (1980) into the present framework. Decomposable graphs are defined and discussed in Darroch *et al.* (1980); they are simple graphs possessing no cycles of length $n \geqslant 4$ without chords.

COROLLARY. *Suppose that the recursive causal graph $\mathfrak{G}$ of the theorem has no configuration* [>]. *Then each of the conditions* (RCF), (GM) *and* (LM) *is equivalent to*:

(UM) *The joint distribution* (V) *is Markov over the underlying undirected graph* $\mathfrak{G}^u$.

Moreover, if $\mathfrak{G}_x$ *is decomposable, then* $\mathfrak{G}^u$ *is also decomposable.*

PROOF. Let us suppose that a joint distribution (V) over such a $\mathfrak{G}$ satisfies the equivalent conditions of the theorem. Choose an extreme exogenous vertex j^*,

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

noting once more that $\bar{A}_{j^*} = V$. Then for *any* system $a_1, \ldots, a_m$ and d for which a_k and a_l are separated by d in $\mathcal{G}^u$, $1 \leqslant k < l \leqslant m$, we conclude that $a_1, \ldots, a_m$ are mutually conditionally independent given d under (V). But this is just the Markov property of (V) over $\mathcal{G}^u$.

To prove the converse we need to check that there are no *additional independencies* arising from a system $a_1, \ldots, a_m$ and d whose union is contained in $\bar{A}_j$, j not extreme. Suppose that d separates these (pairwise) in $\langle \bar{A}_j \rangle_{\mathcal{G}^u}$ but not in $\mathcal{G}^u$. Then there is a chain $a_k \ni j_0, \ldots, j_{p-1}, j_p, j_{p+1}, \ldots, j_q \in a_l$ connecting some pair a_k and a_l from the system which involves a $j_p \notin \bar{A}_j$, that is, $j_p \in E_j$. Supposing, as we may, that the chain under discussion is a minimal length one having this property, we will derive a contradiction.

Since $\mathcal{G}$ has no instance of configuration [>] we cannot have $j_{p-1} \to j_p \leftarrow j_{p+1}$, and so $j_p \to j_{p-1}$, say, holds. Then $j_{p-1} \to j_{p-2}$ must also hold, for if $j_{p-2} \to j_{p-1}$ then we would need to have $j_{p-2} \to j_p$ or $j_p \to j_{p-2}$ to avoid a configuration [>], but this would contradict minimality of the length of the path. This argument continues down to $j_1 \to j_0$. At no stage can j_r, $0 \leqslant r \leqslant p$, belong to V_x, for every one of them belongs to E_j by construction. But this is just our contradiction, for $j_0 \in a_k \subseteq \bar{A}_j$ was part of our assumptions. Thus separation in $\langle \bar{A}_j \rangle$ coincides with separation in $\mathcal{G}^u$ and the first part of the corollary is proved.

The decomposability of $\mathcal{G}^u$ is proved by induction on $|V_n(\mathcal{G})|$. Suppose that $|V_n| = 1$. By assumption the graph $\mathcal{G}^u$ without $p+1$ and its incident edges contains no r-cycles, $r \geqslant 4$. This must continue to be the case when $p+1$ and its incident edges are included, for an r-cycle, $r \geqslant 4$, involving $p+1$ must include a configuration [>] with $p+1$ at its apex. The inductive step is proved in a similar way with the role of $p+1$ in the foregoing taken by an extreme endogenous vertex. This completes the proof of the corollary.

EXAMPLE 2. Let $V_x(\mathcal{G}_2) = \{1, 2, 3\}$ and $V_n(\mathcal{G}_2) = \{4, 5\}$, with $\mathcal{G}_2$ being as depicted in Figure 2(i) below.

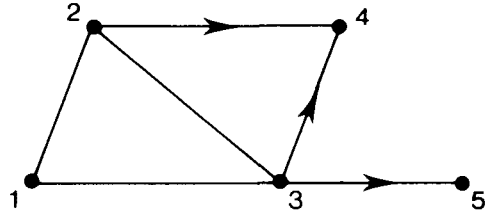

FIGURE 2(i)

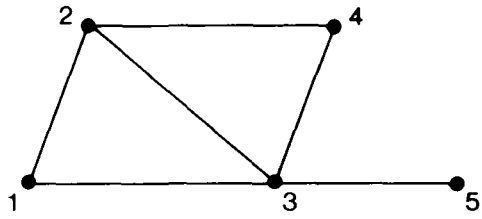

FIGURE 2(ii)

Then any joint distribution (12345) satisfying the causal Markov constraints of $\mathfrak{G}_2$ also satisfies those of $\mathfrak{G}''$, and conversely.

EXAMPLE 3. The graph Figure 3(i) below arises as part of the causal system described as two-wave two-variables, see Kiiveri and Speed (1982).

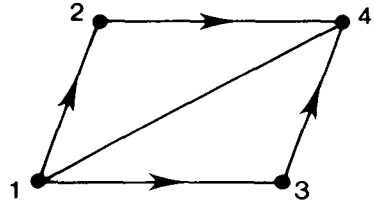

FIGURE 3(i)

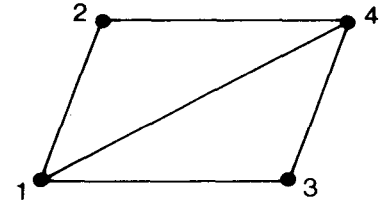

FIGURE 3(ii)

The associated causal factorisation is $(1234) = (1)(2 \mid 1)(3 \mid 1)(4 \mid 123)$ and this corresponds to the single conditional independence constraint 2 and 3 independent given 1. It is clear that there is one instance of configuration [>], involving 4, and so the Markov constraint of the underlying undirected graph Figure 3(ii), namely 2 and 3 independent given 1 and 4, do not coincide with the causal Markov constraints. To see this directly we note that Gaussian random variables with covariance matrix Σ of the form given below satisfy the causal constraints of Figure 3(i) but not those of Figure 3(ii).

$$\Sigma = \begin{array}{c} \begin{matrix} 4 & 3 & 2 & 1 \end{matrix} \\ \begin{bmatrix} 4 & -2 & -2 & 1 \\ -2 & 2 & 1 & -1 \\ -2 & 1 & 2 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \end{array} \begin{matrix} 4 \\ 3 \\ 2 \\ 1 \end{matrix}, \qquad \Sigma^{-1} = \begin{array}{c} \begin{matrix} 4 & 3 & 2 & 1 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 4 \end{bmatrix} \end{array} \begin{matrix} 4 \\ 3 \\ 2 \\ 1 \end{matrix}$$

3. Gaussian distribution

The most thoroughly studied causal systems or causal models are those in which the underlying distributions are Gaussian, see Jöreskog (1977) and Wermuth (1980), although many people treat the subject as an aspect of regression and

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

correlation analysis, not requiring a complete specification of the joint distribution of the random variables under study, see Kang and Seneta (1980) and references therein. It is not hard to see that all the (conditional) independence statements concerning our variables can be interpreted on terms of *zero (partial) correlations*, if we assume only that the random variables have a finite covariance matrix. The structure of Σ^{-1} in 3.1 also lends itself to deriving *recursive systems of linear equations*, and it is to this topic which we now turn. A byproduct of our analysis is a proof of one form of the familiar *path analysis rules*. General references in this area include Boudon (1965), Duncan (1966, 1975), Goldberger and Duncan (1973), Moran (1963) and Simon (1953, 1954).

Throughout this section $\Sigma = (\sigma_{hi})$ will denote the covariance matrix of the random variables $(\mathbf{X}; \mathbf{Y})$, arranged in some order beginning with the p exogenous (X-) variables followed by the q endogenous (Y-) variables. The matrix Σ will be partitioned in a way compatible with $(\mathbf{X}; \mathbf{Y})$ but we place its elements *in the reverse of the usual* order, that is, with σ_{11} in the bottom right-hand corner. All mean values will be taken to be zero.

3.1. *Factorisation of* Σ^{-1}. Most of the results in Section 3 relate to our particular parametrisation of Σ which is a variant of the Choleski-type factorisation used in Wermuth (1980). No use is made of the graph $\mathcal{G}$ in this first lemma; we are simply dealing with $p + q$ random variables labelled as above.

LEMMA 2. *The inverse covariance matrix* Σ^{-1} *of the Gaussian system* $(\mathbf{X}; \mathbf{Y})$ *of random variables has a unique representation* $\Sigma^{-1} = LAL^T$ *where* L *and* A *have the form*

$$L = \begin{matrix} \\ q \\ p \end{matrix} \begin{matrix} \begin{matrix} q & p \end{matrix} \\ \begin{bmatrix} C & 0 \\ B & I \end{bmatrix} \end{matrix} \qquad A = \begin{matrix} \\ q \\ p \end{matrix} \begin{matrix} \begin{matrix} q & p \end{matrix} \\ \begin{bmatrix} \Psi^{-1} & 0 \\ 0 & \Phi^{-1} \end{bmatrix} \end{matrix}$$

with C *lower-triangular and having* $+1$*s downs the diagonal,* Ψ *diagonal, with positive elements, and* Φ *positive definite,* I *denoting the* $p \times p$ *identity matrix.*

PROOF. The easiest way to get this result—which is just a modification of the familiar Choleski decomposition of Σ^{-1}, involving the treatment of the first p variables *en bloc* is to define the matrices L and A and check that $H^{-1} = LAL^T$ actually coincides with Σ^{-1}. We do this as follows:

$$\begin{aligned} & \text{For } j > p,\ i < j, && l_{ij} := -\beta_{ji \cdot \{1,\ldots,i-1,i+1,\ldots,j-1\}}; \\ & \text{for } p < j \leqslant p+q, && \psi_j := \sigma_{jj \cdot \{1,\ldots,j-1\}}; \\ \text{and} \quad & \text{for } 1 \leqslant g, h \leqslant p, && \phi_{gh} := \sigma_{gh}. \end{aligned}$$

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

Here $\beta_{ji \cdot a}$ is the partial regression coefficient of the jth variable on the ith, eliminating the variables with indices in the set a, and $\sigma_{jj \cdot a}$ is the residual variance of the jth variable after eliminating those with indices in a.

Writing $L = (l_{ij})$, $\Phi = (\phi_{gh})$ and $\Psi = \text{diag}(\psi_j)$ (where we have added in 0s and 1s to the definition of L) we readily see that if $H^{-1} = LAL^T$, then $HL = L^{-T}A^{-1}$. Beginning with the bottom right-hand $p \times p$-block and continuing recursively we can check easily from this equation that $\Sigma = H$. We omit the details.

Uniqueness is easily proved and again the details are omitted.

It is worthwhile gathering up some formulae associated with this decomposition of Σ^{-1}; they are all easily checked.

$$\Sigma^{-1} = \begin{bmatrix} C & 0 \\ B & I \end{bmatrix} \begin{bmatrix} \Psi^{-1} & 0 \\ 0 & \Phi^{-1} \end{bmatrix} \begin{bmatrix} C^T & B^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} C\Psi^{-1}C^T & C\Psi^{-1}B^T \\ B\Psi^{-1}C^T & B\Psi^{-1}B^T + \Phi^{-1} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} C^{-T} & -C^{-T}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} \Psi & 0 \\ 0 & \Phi \end{bmatrix} \begin{bmatrix} C^{-1} & 0 \\ -BC^{-1} & I \end{bmatrix}$$

$$= \begin{bmatrix} C^{-T}\Psi C^{-1} + C^{-T}B^T\Phi BC^{-1} & -C^{-T}B^T\Phi \\ -\Phi BC^{-1} & \Phi \end{bmatrix}.$$

For $j > p$,

$$\sum_{i=1}^{j} \sigma_{ik} l_{ik} = \begin{cases} \psi_j & \text{if } k = j, \\ 0 & \text{if } k < j. \end{cases}$$

The factorisation described in the preceding lemma will be called the $(L; A)$ or $(L; \Psi, \Phi)$ or $(C, B; \Psi, \Phi)$ decomposition of Σ^{-1} in what follows. Notice that it does depend on the ordering of the random variables.

For our main result in this section we need the notion of a *strict ordering* of the vertices of a recursive causal graph $\mathfrak{G}$ *compatible with the graph structure*, which is an ordering φ: $V \to \{1, 2, \ldots, |V|\}$ such that

(i) $\varphi(V_x) = \{1, 2, \ldots, |V_x|\}$, (ii) $\varphi(i) < \varphi(j)$ whenever $j \in V_n$ and $i \to j$.

It is not hard to see that for any recursive causal graph $\mathfrak{G}$ there is always at least one compatible strict ordering of $V(\mathfrak{G})$.

The following result concerns random variables $(\mathbf{X}; \mathbf{Y})$ indexed by the vertices of a recursive causal graph $\mathfrak{G}$ and ordered in the same way as these vertices. Their joint Gaussian distribution has density p_Σ, corresponding to mean 0 and covariance matrix Σ.

PROPOSITION 1. *The distribution p_Σ satisfies the equivalent conditions of the theorem if and only if for all strict orderings of $V(\mathfrak{G})$ compatible with $\mathfrak{G}$, the*

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

elements of the associated $(L; \Psi, \Phi)$ *factorisation of* Σ^{-1} *satisfy the zero constraints for all* $g, h \in V_x, j \in V_n$ *and* $i \in V$

$$\text{(i)} \quad \phi^{gh} = 0 \quad \textit{whenever } \{g, h\} \notin E_x(\mathfrak{G});$$

(ZC)

$$\text{(ii)} \quad l_{ij} = 0 \quad \textit{whenever } (i, j) \notin E_n(\mathfrak{G}).$$

REMARK. It is clear from this result that the $(L; \Psi, \Phi)$ parametrisation is a natural one for describing the causal Markov property of a Gaussian distribution. Statistical matters such as the fitting and testing of such models with this parametrisation are discussed in Kiiveri (1982).

PROOF. We will compare the density p_Σ, where Σ has the $(C, B; \Psi, \Phi)$ factorisation, with (1) of the main theorem. Suppose that L and Φ satisfy the zero constraints (i) and (ii) when some strict ordering is used for labelling the Xs and Ys, and hence the elements of Σ. Then a little simplification shows that $-2 \log p_\Sigma$ involves the log of two determinants plus

$$\sum_{g,h} \phi^{gh} x_g x_h + \sum_j \psi_j^{-1} \left(y_j + \sum_{i \in D_j} c_{ji} y_i + \sum_{h \in D_j} b_{jh} x_h \right)^2.$$

But as soon as we recall the interpretations of ψ_j, c_{ji} and b_{jh} given in Lemma 2 this is seen to be just the $-2 \log(V)$ in the form (RCF).

The converse is proved by reversing the above argument.

EXAMPLE 1 (continued). The $(L; A)$ factorisation of the inverse covariance matrix Σ^{-1} of four random variables $(X_1, X_2; Y_3, Y_4)$ whose Gaussian distribution satisfies the causal Markov constraints of $\mathfrak{G}_1$ has the form

$$L = \begin{array}{c} \left[\begin{array}{cc:cc} 1 & 0 & 0 & 0 \\ * & 1 & 0 & 0 \\ \hdashline 0 & * & 1 & 0 \\ * & 0 & 0 & 1 \end{array}\right] \begin{array}{c} 4 \\ 3 \\ 2 \\ 1 \end{array} \\ \begin{array}{cccc} 4 & 3 & 2 & 1 \end{array} \end{array}, \qquad A = \begin{array}{c} \left[\begin{array}{cc:cc} + & 0 & 0 & 0 \\ 0 & + & 0 & 0 \\ \hdashline 0 & 0 & + & * \\ 0 & 0 & * & + \end{array}\right] \begin{array}{c} 4 \\ 3 \\ 2 \\ 1 \end{array} \\ \begin{array}{cccc} 4 & 3 & 2 & 1 \end{array} \end{array}$$

where $*$ (resp. $+$) denote freely-varying real (resp. positive) numbers, and the lower right-hand 2×2 submatrix of A must be positive definite. For example, the element l_{34} of L is in fact $-\beta_{43 \cdot 12}$, whilst $l_{24} = -\beta_{42 \cdot 13} = 0$. Similarly $l_{13} = -\beta_{31 \cdot 2} = 0$, whilst $a_{44} = \sigma_{44 \cdot 123}$.

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

EXAMPLE 2 (continued). The $(L; A)$ factors here have the form

$$L = \left[\begin{array}{ccc:ccc|c} 1 & 0 & 0 & 0 & 0 & 5 \\ 0 & 1 & 0 & 0 & 0 & 4 \\ \hdashline * & * & 1 & 0 & 0 & 3 \\ 0 & * & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 5 & 4 & 3 & 2 & 1 & \end{array}\right], \qquad A = \left[\begin{array}{ccc:ccc|c} + & 0 & 0 & 0 & 0 & 5 \\ 0 & + & 0 & 0 & 0 & 4 \\ \hdashline 0 & 0 & + & * & * & 3 \\ 0 & 0 & * & + & * & 2 \\ 0 & 0 & * & * & + & 1 \\ \hline 5 & 4 & 3 & 2 & 1 & \end{array}\right]$$

but we note that it is not necessary to use these matrices with Example 2 since $\mathfrak{G}_2$ is decomposable. In such cases the non-causal Markov constraints coincide with the causal ones, and positive definite matrices Σ^{-1} having zeros in the positions corresponding to those $\{g, h\} \notin E_x(\mathfrak{G})$, and those $(i, j) \in E_n(\mathfrak{G})$ with $i \notin D_j$, give a more compact description of the associated distribution p_Σ. This coincides with the approach of N. Wermuth (1980), see the expression for Σ^{-1} associated with $\mathfrak{G}_2$ (or $\mathfrak{G}_2^u$) below.

$$\Sigma^{-1} = \left[\begin{array}{ccccc|c} + & & & & & 5 \\ 0 & + & & & & 4 \\ * & * & + & & & 3 \\ 0 & * & * & + & & 2 \\ 0 & 0 & * & * & + & 1 \\ \hline 5 & 4 & 3 & 2 & 1 & \end{array}\right]$$

3.2. *Structural equations.* We can now describe the connexion between our approach to recursive causal systems of random variables and the much more familiar one used in econometrics and elsewhere involving linear equations. Let us begin with a system $(\mathbf{X}; \mathbf{Y})$ of $p + q$ Gaussian random variables having covariance matrix Σ as in 3.1. The following is an easy consequence of Lemma 2.

LEMMA 3. *If Σ^{-1} is decomposed into $(C, B; \Psi, \Phi)$ as in Lemma 2, then* $\mathbf{X}$ *and* $\mathbf{Y}$ *satisfy the linear structural equations*

$$C^T\mathbf{Y} + B^T\mathbf{X} = U, \tag{SE}$$

where U and X are independent Gaussian vectors with covariance matrices Ψ and Φ. Conversely, if $\mathbf{X}$ *and* $\mathbf{Y}$ *satisfy a system such as* (SE), *if* $\mathbf{U}$ *and* $\mathbf{X}$ *are independent with covariance matrices Ψ and Φ, and further, if C is lower triangular with $+1$s down its diagonal and Ψ is diagonal, then the matrices B, C, Ψ and Φ combine as in Lemma 2 to give Σ^{-1}.*

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

PROOF. This result is an immediate consequence of Lemma 2 and the formulae which follow it.

REMARKS. (i) It is perhaps more usual with structural equations to specify that (SE) hold with C, Ψ having the properties stated, and that only the conditional distribution of $\mathbf{U}$ given $\mathbf{X}$ be Gaussian (with mean zero and covariance matrix Ψ). In other words, either $\mathbf{X}$ is not regarded as a random vector, or it is, but no assumptions are made about its distribution. In the latter case such a specification still corresponds to Σ^{-1} having the form LAL^T with L and A having their usual structure. For if $C^T\mathbf{Y} + B^T\mathbf{X}$ is normal with zero mean and covariance matrix Ψ, given $\mathbf{X}$, then $\mathbf{Y}$ has mean $-C^{-T}B^T\mathbf{X}$ and covariance matrix $C^{-T}\Psi C^{-1}$ given $\mathbf{X}$, whence $\mathrm{Var}(Y) = C^{-T}\Psi C^{-1} + C^{-T}B^T\Phi BC^{-1}$ and $\mathrm{Cov}(Y, X) = -C^{-T}B^T\Phi$, where $\Phi = E(\mathbf{X}\mathbf{X}^T)$ is assumed to be finite. These formulas may be compared with those following Lemma 2 and the assertion will then be evident.

A consequence of the remarks just made is the following: any conditional independence statements concerning $(\mathbf{X}; \mathbf{Y})$ involving X-variables only in the conditioning which are valid when the whole system is jointly Gaussian are also valid if we assume only that $\mathbf{Y}$ given $\mathbf{X}$ (equivalently, $\mathbf{U}$ given $\mathbf{X}$ in the above) Gaussian.

(ii) All of the foregoing extends to the case in which only second-order assumptions concerning $\mathbf{U}$ given $\mathbf{X}$ are made; simply replace conditional independence statements by the corresponding zero partial correlation ones.

Turning now to the Markov properties enjoyed by $(\mathbf{X}; \mathbf{Y})$ when they satisfy a set of equations such as (SE) under the further assumptions stated in Lemma 3, we have the following immediate consequence of Proposition 1.

PROPOSITION 2. *A Gaussian system* $(\mathbf{X}; \mathbf{Y})$ *satisfying the equations* (SE) *with* $\mathbf{U}$ *independent of* $\mathbf{X}$ *and having covariance matrices* Ψ, Φ *respectively*, C *lower-triangular with* $+1$*s down the diagonal and* Ψ *diagonal, also satisfies the equivalent conditions of the theorem if and only if* $L = \begin{pmatrix} C & 0 \\ B & I \end{pmatrix}$ *and* Φ *satisfy the zero constraints* (ZC) *of Proposition* 1.

In other words, we can use the theorem to draw causal graph associated with any system of structural equations such as (SE) having zeros in the lower-triangular matrix C (and also in the inverse of the covariance matrix of the exogenous variables), and then make direct conditional independence statements concerning the endogenous variables (and also the exogenous variables) valid under the further assumption that $\mathbf{Y} \mid \mathbf{X}\ [(\mathbf{X}; \mathbf{Y})]$ is Gaussian. Once more we remark that the same argument yield zero partial correlation statements which are generally valid.

3.3. *Path analysis.* There is now quite a large literature on path analysis but few precise formulations or proofs of the so-called *path analysis rule*, see Kang and Seneta (1980) for references. Suppose that $(\mathbf{X}; \mathbf{Y})$ is a Gaussian system of random variables indexed by the vertices of a recursive causal graph $\mathfrak{G}$ in the manner of our earlier discussion. The *path regression coefficient* associated with an edge $i \to j$ of $E_n(\mathfrak{G})$ is simply the coefficient $-l_{ij}$, that is, $\beta_{ji \cdot D_j \setminus \{i\}} = \beta_{ji \cdot A_j \setminus \{i\}}$ hereafter abbreviated to β_{ji}, and one form of the basic rule expresses the covariance σ_{ij} between any of the variables at vertex $i \in V$ and $j \in V_n$ in terms of path regression coefficients and covariances σ_{gh}, $g, h \in V_x$, of pairs of exogenous variables. A more refined rule, which will not be given here, applies when the graph structure assumed involves a decomposable graph $\mathfrak{G}_x$ on the exogenous vertices. For in that case we can further decompose the covariance σ_{gh}, $g, h \in V_x(\mathfrak{G})$ into sums of products of covariances which are associated with edges $\{g, h\} \in E_x(\mathfrak{G})$.

PROPOSITION 3. *In the notation introduced above*

$$\sigma_{ij} = \sum_{1} \beta_{h_1 h_2} \cdots \beta_{h_{r-1} h_r} \sigma_{h_r h_{r+1}} \beta_{h_{r+2} h_{r+1}} \cdots \beta_{h_u h_{u-1}}$$
$$+ \sum_{2} \beta_{i_1 i_2} \cdots \beta_{i_{s-1} i_s} \sigma_{i_s i_s} \beta_{i_{s+1} i_s} \cdots \beta_{i_v i_{v-1}}$$

where $\sum_{1}$ *is the sum over all non-self-intersecting paths*

$$i = h_1 \leftarrow \cdots \leftarrow h_{r-1} \leftarrow h_r - h_{r+1} \rightarrow h_{r+2} \rightarrow \cdots \rightarrow h_u = j, \qquad r \geqslant 1, u \geqslant r + 2,$$

and $\sum_{2}$ *is the sum over all non-self-intersecting paths*

$$i = i_1 \leftarrow \cdots \leftarrow i_{s-1} \leftarrow i_s \rightarrow i_{s+1} \rightarrow \cdots \rightarrow i_v = j, \qquad s \geqslant 1, v \geqslant s + 2.$$

REMARK. In terms of *path-tracing* we are in effect supposing that *every* pair g, h of exogenous vertices is connected by an edge (unless $\sigma_{gh} = 0$). This will in general be inconsistent with the edge structure $E_x(\mathfrak{G})$ assumed over $V_x(\mathfrak{G})$, but as we have already indicated a completely satisfactory but rather more complicated reformulation of the rules exist when $\mathfrak{G}_x$ is decomposable. In practice it is common to have one or the other of the extreme cases: all exogenous variables arbitrarily correlated, or all mutually independent, and in both of these our reformulation is unnecessary.

PROOF. By induction on $|V_n(\mathfrak{G})|$. If $|V_n(\mathfrak{G})| = 1$ then we need only consider $\sigma_{h,p+1}$ where $p = |V_x(\mathfrak{G})|$. By the formula following Lemma 2

$$\sigma_{h,p+1} = \sum_{g \in V_x} \beta_{p+1,g} \sigma_{gh} = \beta_{p+1,h} \sigma_{hh} + \sum_{g \neq h} \beta_{p+1,g} \sigma_{gh}.$$

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

The first term is seen to correspond to Σ_2 if $h \to p + 1 \in E_n(\mathcal{G})$ whilst the second sum corresponds to Σ_1, being over all paths of the form $h - g, g \to p + 1$.

Now assume that the result holds for all causal graphs with fewer than $q = |V_n(\mathcal{G})|$, $q > 1$, endogenous vertices, and let us consider an extreme exogenous vertex j^* of $\mathcal{G}$. We need only check that σ_{ij^*} takes the form of our statement, for all other covariances have that form by the inductive hypothesis. Once more we use the formula following Lemma 2, and this time it reads

$$\sigma_{ij^*} = \sum_{k \in D_{j^*}} \beta_{j^*k}\sigma_{ik}.$$

But for $k \in D_{j^*}$, $i \in V \setminus \{j^*\}$ our inductive hypothesis tells us that (in an obvious notation)

$$\sigma_{ik} = \sum_{1}^{(k)} + \sum_{2}^{(k)}$$

whence

$$\sigma_{ij^*} = \sum_{k \in D_{j^*}} \left\{ \beta_{j^*k} \sum_{1}^{(k)} + \beta_{j^*k} \sum_{2}^{(k)} \right\} = \sum_{1} + \sum_{2}$$

completing the proof of the inductive step and so the proposition.

EXAMPLE 2 (continued). Applying the rule just given to calculate σ_{45} we find that

$$\sigma_{45} = \beta_{42}\sigma_{23}\beta_{53} + \beta_{43}\sigma_{33}\beta_{53}$$

these being the sums over the paths $4 \leftarrow 2 - 3 \to 5$ and $4 \leftarrow 3 \to 5$ respectively.

We close this section with some remarks on the relation between the above and the work of others. Moran (1961) operates within a framework similar to ours, making Markov-type conditional independence assumptions concerning his system of random variables. These (Assumption II) suffice to give him a form of our Proposition 3, but do not characterise the systems. More recently Kang and Seneta (1980) prove results which relate closely to the material concerning Gaussian arrays. Specifically, their Lemma 1 is a second-order version of part of the main theorem and their Lemma 3 is a more general version of our Proposition 3. Finally, Wermuth (1980) considers the relationship between the pattern of zeros of Σ^{-1} and that of L in $\Sigma^{-1} = LA^{-1}L^T$, proving that they are essentially the same iff the corresponding Gaussian distribution (or graph) is decomposable. It is in this paper that the condition we term [>] (there called reducible zero pattern) is introduced.

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

4. Discrete distributions

One of the main reasons why an independence formulation of the basic results of recursive causal models is desirable is their immediate applicablity to discrete data. In this section we examine the problems of parametrising discrete recursive causal models, and relate such models to the more familiar *log-linear models* for discrete data, see Goodman (1972, 1973a, b) and Fienberg (1977). We begin with some extra notation and terminology.

Let us suppose that the exogenous variable X_h takes value x_h from a finite set $\mathcal{X}_h$, $h \in V_x(\mathcal{G})$, and similarly that Y_j takes values y_j from $\mathcal{Y}_j$, $j \in V_n(\mathcal{G})$. Then the full array $(\mathbf{X}; \mathbf{Y})$ takes values $(\mathbf{x}; \mathbf{y})$ from $\prod_{h \in V_x} \mathcal{X}_h \times \prod_{j \in V_n} \mathcal{Y}_j = \mathcal{X} \times \mathcal{Y}$ and throughout this section we will suppose that for all $(\mathbf{x}; \mathbf{y})$ we have the positivity constraint:

$$p(\mathbf{x}; \mathbf{y}) = \mathbf{P}(X_h = x_h, h \in V_x; Y_j = y_j, j \in V_n) > 0.$$

If $A \subseteq V$ we write $\mathbf{x}_A$ [resp. $\mathbf{y}_A$] for $(x_h: h \in V_x \cap A)$ [resp. $(y_j: j \in V_n \cap A)$]. In order to relate our main theorem to log-linear models, we need to refer to the vector space S of all real-valued functions on $\mathcal{X} \times \mathcal{Y}$, and to the subspaces $S(A)$, $A \subseteq V$, of functions depending only on $(\mathbf{x}_A; \mathbf{y}_A)$. They have been discussed in Speed (1979). (There, however, the subspace $S(A)$ is denoted by E_A; we have changed notation to avoid confusion with edge sets.)

For a probability distribution p over $\mathcal{X} \times \mathcal{Y}$ and for $j \in V_n$ let us write p_j for the *marginal distribution* of the variables indexed by $\bar{A}_j$, and θ_j for the *conditional distribution* of Y_j given $(\mathbf{X}_{A_j}; \mathbf{Y}_{A_j})$. Note that θ_j depends only on $\bar{A}_j$; indeed

$$\theta_j = p_j / \sum_j p_j$$

where Σ_j denotes a summation over all $y_j \in \mathcal{Y}_j$. Furthermore, write p_x for the marginal distribution over V_x.

The following reformulation of the main theorem shows that a recursive causal model for discrete data is, in general, the *conjunction of a set of log-linear models* for the full array and certain of its marginals. Recall, see Speed (1978), that a *maximal clique* in an undirected graph is a set of vertices each pair of which is connected by an edge, and is maximal with respect to this property. The set of all maximal cliques of $\mathcal{G}_x = (V_x(\mathcal{G}), E_x(\mathcal{G}))$ is denoted by $\mathcal{C}_x$.

PROPOSITION 4. *A probability distribution p satisfies the equivalent conditions of the main theorem if and only if*

(i) $\log p_x \in S(\mathcal{C}_x) = \Sigma\{S(a): a \in \mathcal{C}_x\}$; *and for all* $j \in V_n$,

(ii) $\log p_j \in S(A_j) + S(\bar{D}_j)$.

https://doi.org/10.1017/S1446788700027312 Published online by Cambridge University Press

PROOF. These conditions are just a reformulation of (4) from the main theorem using the Hammersley-Clifford theorem, see for example the main result of Speed (1979) for a proof in the present spirit.

Many instances of this result, with a different parametrisation, can be found in the papers of Goodman (1972, 1973a, b; 1974a, b).

The subspace sum $S(A_j) + S(\bar{D}_j)$ may be written as

$$S(A_j) + \left[S(\bar{D}_j) \ominus S(A_j)\right] = S(A_j) + \left[S(\bar{D}_j) \ominus S(D_j)\right],$$

where $\ominus$ denotes orthogonal complement in the usual inner product. This fact is a consequence of the fact that the projections onto the various subspaces $S(A) \subseteq S$ all commute, and that $A_j \cap \bar{D}_j = D_j$. Thus we see that if $p_j = \exp(\xi_j + \eta_j)$, $\xi_j \in S(A_j)$, $\eta_j \in S(\bar{D}_j) \ominus S(D_j)$, then θ_j may be represented as

$$\theta_j = \exp \eta_j / \sum_j \exp \eta_j,$$

and furthermore, the $\eta_j \in S(\bar{D}_j) \ominus S(D_j)$ is *unique*. Putting this into (1) of the main theorem we see that a probability p over $\mathcal{X} \times \mathcal{Y}$ which satisfies the causal Markov constraints has a unique representation

$$p = p_x \prod_{j \in V_n} \frac{\exp \eta_j}{\Sigma_j \exp \eta_j}, \quad \text{where } \eta_j \in S(\bar{D}_j) \ominus S(D_j), j \in V_n.$$

Further, one can easily prove that the $\{\eta_j\}$ are pairwise orthogonal. With p represented in this form we see that it is possible to restrict even further the higher-order interactions between an endogenous variable and its direct causes without disturbing the causal Markov constraints. Thus causal modelling with discrete data has two aspects: the underlying causal model, and the higher-order interactions just mentioned.

In closing this section we remark that when $\mathfrak{G}$ contains no configurations $[>]$ the causal Markov constraints collapse into a single set of log-linear constraints, those associated with what Darroch *et al.* (1980) call a *graphical log-linear model.*

5. Acknowledgements

This paper began as an honours dissertation, Carlin (1977), by one of the authors. Its further development was greatly assisted by access to then unpublished work of Kang and Seneta (1980) and Wermuth (1980), and many thanks are due to these authors for their kindness. The ideas were discussed in a seminar Speed (1978a) at the University of Copenhagen, and further thanks are due to all of the active participants in that seminar, especially Steffen Lauritzen.

Its final form is the basis for the approach adopted in a thesis, Kiiveri (1982), on causal models, and the author gratefully acknowledges the Commonwealth Postgraduate Research Scholarship held during the preparation of that thesis.

References

H. M. Blalock, Jr. (1962), 'Four-variable causal models and partial correlations', *Amer. J. Sociology* **68**, 182–194. Reprinted as Chapter 2 in Blalock (1971).

H. M. Blalock, Jr. (1971), *Causal models in the social sciences* (Macmillan Press Ltd., London).

R. Boudon (1965), 'A method of linear causal analysis: dependence analysis', *Amer. Sociological Rev.* **30**, 365–374.

J. B. Carlin (1977), *Causal models: an attempt at a unified approach* (Honours thesis, Department of Mathematics, University of Western Australia).

J. N. Darroch, S. L. Lauritzen and T. P. Speed (1980), 'Log-linear models for contingency tables and Markov fields over graphs', *Ann. Statist.* **8**, 522–539.

O. D. Duncan (1966), 'Path analysis: sociological examples', *Amer. J. Sociology* **72**, 1–16. Reprinted as Chapter 7 in Blalock (1971).

O. D. Duncan (1975), *Introduction to structural equation models* (Academic Press, New York).

S. E. Fienberg (1977), *The analysis of cross classified categorical data* (MIT Press, Cambridge, Massachusetts).

A. S. Goldberger and O. D. Duncan (1973), *Structural equation models in the social sciences* (Seminar Press, New York).

L. A. Goodman (1972), 'A general model for the analysis of surveys', *Amer. J. Sociology* **77**, 1035–1086.

L. A. Goodman (1973a), 'The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach', *Biometrika* **60**, 179–192.

L. A. Goodman (1973b), 'Causal analysis of data from panel studies and other kinds of surveys', *Amer. J. Sociology* **78**, 1135–1191.

L. A. Goodman (1974a), 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika* **61**, 215–231.

L. A. Goodman (1974b), 'The analysis of systems of qualitative variables when some of the variables are unobservable, Part I—a modified latent structure approach', *Amer. J. Sociology* **79**, 1179–1259.

D. R. Heise (1975), *Causal analysis* (John Wiley & Sons, New York).

K. G. Jöreskog (1977), 'Structural equation models in the social sciences: Specification, estimation and testing', *Applications of statistics*, pp. 265–287, edited by P. R. Krishnaiah (North-Holland Publishing Co.).

K. M. Kang and E. Seneta (1980), 'Path analysis: an exposition', *Developments in statistics* 3, pp. 217–246, edited by P. R. Krishnaiah (Academic Press).

H. Kiiveri (1982), A unified theory of causal models (Ph.D. thesis, in preparation).

H. Kiiveri and T. P. Speed (1982), 'The structural analysis of multivariate data: a review', *Sociological Methodology*, to appear.

P. A. P. Moran (1961), 'Path coefficients reconsidered', *Austral. J. Statist.* **3**, 87–93.

H. A. Simon (1953), 'Causal ordering and identifiability', *Studies in econometric method*, Chapter 3, edited by William C. Hood and Tjalling C. Koupmans. Cowles Commission for Research in Economics (John Wiley & Sons, New York). Reprinted in Simon (1957).

H. A. Simon (1954), 'Spurious correlation: a causal interpretation', *J. Amer. Statist. Assoc.* **49**, 467–479. Reprinted in Simon (1957).

T. P. Speed (1978), 'Relations between models for spatial data, contingency tables and Markov fields on graphs', *Proceedings of the conference on spatial patterns and processes*, edited by R. L. Tweedie, *Supplement to Advances in Applied Probability*.

T. P. Speed (1978a), *Graphical methods in the analysis of data* (Lecture notes issued at the University of Copenhagen Institute of Mathematical Statistics, 111 pp.).

T. P. Speed (1979), 'A note on nearest-neighbour Gibbs and Markov probability', *Sankhyā Ser. A* **41**, 184–197.

N. Wermuth (1980), 'Linear recursive equations, covariance selection and path analysis', *J. Amer. Statist. Assoc.* **75**, 963–972.

S. Wright (1921), 'Correlation and causation', *J. Agric. Research* **20**, 557–585.

S. Wright (1934), 'The method of path coefficients', *Ann. Math. Statist.* **5**, 161–215.

Department of Mathematics
University of Western Australia
Nedlands, W.A. 6009
Australia

CSIRO
Division of Mathematics and Statistics
P.O. Box 1965, Canberra City
ACT 2601
Australia

Department of Statistics
1 Oxford Street
Harvard University
Cambridge, Massachusetts 02138
U.S.A.