Scand J Statist 16: 273–306, 1989

Mixed Graphical Association Models

STEFFEN L. LAURITZEN

University of Aalborg

ABSTRACT. The paper surveys the mathematical and statistical theory of mixed graphical association models as introduced by Lauritzen & Wermuth (1984), concerned with description of associations between variables, some of which are allowed to be quantitative and some qualitative. The models originate partly in statistical physics with early work of Gibbs and partly in genetics with work of Wright on so-called path analysis. All models are determined by restrictions of distributional type supplemented by conditional independence restrictions. The models extend and unify a range of statistical techniques that are well established, primarily in the social sciences.

Key words: block-recursive models, covariance selection, contingency tables, decomposable graphs, log-linear models, Markov properties, multivariate analysis, path analysis, recursive models

1. Introduction

Mixed graphical association models were introduced by Lauritzen & Wermuth (1984). The models are concerned with description of associations between variables, some of which can be quantitative and some qualitative. Each model is represented by a graph where qualitative (discrete) variables are represented by dots and quantitative (continuous) variables by circles. Connections between variables are either lines, representing symmetric associations between variables, or arrows, representing directional influences. Each missing connection represents a conditional independence statement. The basic theory and description of the models and distributions involved can be found in Lauritzen & Wermuth (1989), the graph theory needed to understand the structure of the models has been developed in Leimer (1985, 1989a), further developments of the theory has been made in Lauritzen (1985), Frydenberg (1986, 1990) and Frydenberg & Lauritzen (1989), development of algorithms and software in Edwards (1987, 1989a) and Frydenberg & Edwards (1989) and case studies and discussions of the interpretation and practical use of the models are in Edwards (1989a), Wermuth & Lauritzen (1989) and Wermuth (1989).

The models have a long pre-history. They originate partly in statistical physics with early work of Gibbs (1902), and partly in genetics with work of Wright (1921, 1923, 1934) on path analysis.

In the case of continuous variables models of this type have primarily played a role in the social sciences (see, for example, Wold, 1954, 1960; Blalock, 1971; Goldberger & Duncan, 1973; Jöreskog, 1981; Kiiveri & Speed, 1982; Holland, 1986; and references therein for more details of the developments). Important precursors in the discrete case are Birch (1963), Goodman (1970, 1973), Bishop, Fienberg & Holland (1975), Haberman (1974) and Andersen (1974).

In the last decades much work has been directly involving graphical models in the case of only one kind of variables. References are, for example, Dempster (1972), Wermuth (1976a, b, 1980) and Porteous (1985b) in the continuous case and Darroch, Lauritzen & Speed (1980), Lauritzen (1982), Edwards & Kreiner (1983), Edwards & Havránek (1985, 1987) and Whittaker (1982, 1984) in the discrete case.

Recently attention has been focusing on connections between models of this type and the handling of expert systems and other types of decision support systems (see Lauritzen & Spiegelhalter, 1988, for a range of references to this part of the literature, and in particular the recent book by Pearl, 1988).

The paper falls in two parts. First we review the mathematical basis for the models in such a way that most proofs are omitted and appropriate references given instead. Then we discuss the statistical theory, based on selected examples.

2. Graph theory

2.1. *Notation and terminology*

A *graph*, as we use it throughout this paper, is a pair $\mathcal{G}=(V, E)$, where V is a finite set of *vertices* and the set of *edges* E is a subset of the set $V\times V$ of ordered pairs of distinct vertices. Thus our graphs are *simple*, i.e. there are no multiple edges and they have no loops. Edges $(\alpha,\beta)\in E$ with both (α,β) and (β,α) in E are called *undirected*, whereas an edge (α,β) with its *opposite* (β,α) not being contained in E are called *directed*.

We are interested in graphs where the vertices are *marked* in the sense that they are partitioned into two groups, such that the vertex set has the structure

$$V=\Delta\dot{\cup}\Gamma.$$

We then use the term *marked graph*. The vertices in the set Δ represent qualitative variables and those in Γ quantitative variables. Therefore we say that the vertices in Δ are *discrete* and those in Γ are *continuous*. A graph is *pure* if it has only one kind of vertex.

A graph is a visual object. It is conveniently represented by a picture, where we use a *dot* for a discrete vertex and a circle for a continuous. Further a *line* joining α to β represents an undirected edge, whereas an *arrow* from α, pointing towards β is used for a directed edge (α,β) with $(\beta,\alpha)\notin E$. Fig. 1 contains an illustration of these conventions. Correspondingly we sometimes use the notation

$$\alpha\to\beta \quad \alpha\sim\beta$$

$$\alpha\not\to\beta \quad \alpha\not\sim\beta$$

to signify that

$$(\alpha,\beta)\in E \qquad (\alpha,\beta)\in E\wedge(\beta,\alpha)\in E$$

$$(\alpha,\beta)\notin E \qquad (\alpha,\beta)\notin E\wedge(\beta,\alpha)\notin E.$$

Note that $\alpha\not\sim\beta$ then means that there is no arrow between α and β and that, for example, $\alpha\to\beta\wedge\beta\to\alpha\Leftrightarrow\alpha\sim\beta$.

If the graph has only undirected edges (lines) it is an *undirected* graph and if all edges are directed (arrows), the graph is said to be *directed*.

The *symmetrization* $\mathcal{G}^{\sim}$ of a graph $\mathcal{G}$ is the undirected graph obtained from $\mathcal{G}$ by substituting lines for arrows. We shall also use the expression that $\mathcal{G}^{\sim}$ is *the undirected graph corresponding to* $\mathcal{G}$. See also Fig. 1 for an example of this construction.

If $A\subseteq V$ is a subset of the vertex set, it induces a subgraph $\mathcal{G}_A=(A, E_A)$, where the edge set $E_A=E\cap A\times A$ is obtained from $\mathcal{G}$ by keeping edges with both endpoints in A.

A graph is *complete* if all vertices are joined by an arrow or a line. A subset is *complete* if it induces a complete subgraph. A complete subset that is maximal (w.r.t. $\subseteq$) is called a *clique*.

If there is an arrow from α pointing towards β, α is said to be a *parent* of β and β a *child* of α. The set of parents of β is denoted as $\mathrm{pa}(\beta)$ and the set of children of α as $\mathrm{ch}(\alpha)$.

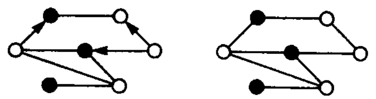

Fig. 1. A marked graph, together with its corresponding undirected version. Discrete vertices are represented by dots and continuous vertices by circles. Directed edges are represented by arrows and undirected edges by lines.

If there is a line between α and β, α and β are said to be *adjacent* or *neighbours*. The neighbours of a vertex α is denoted as $\mathrm{ne}(\alpha)$.

The expressions $\mathrm{pa}(A)$, $\mathrm{ch}(A)$ and $\mathrm{ne}(A)$ denote the collection of parents, resp. children and neighbours, of vertices in A that are not themselves elements of A:

$$\mathrm{pa}(A)=\cup_{\alpha\in A}\mathrm{pa}(\alpha)\setminus A$$

$$\mathrm{ch}(A)=\cup_{\alpha\in A}\mathrm{ch}(\alpha)\setminus A$$

$$\mathrm{ne}(A)=\cup_{\alpha\in A}\mathrm{ne}(\alpha)\setminus A.$$

The *boundary* $\mathrm{bd}(A)$ of a subset A of vertices is the set of vertices in $V\setminus A$ that are parents or neighbours to vertices in A, i.e. $\mathrm{bd}(A)=\mathrm{pa}(A)\cup\mathrm{ne}(A)$. The *closure* of A is $\mathrm{cl}(A)=A\cup\mathrm{bd}(A)$. A subset is *closed* if $A=\mathrm{cl}(A)$. See Fig. 2 for further illustration.

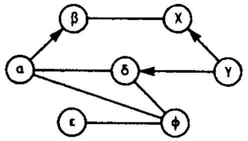

Fig. 2. Illustration of graph-theoretic concepts. We have $\alpha\to\beta$ and also $\alpha\to\delta$ but $\delta\not\to\chi$, $\alpha\not\sim\chi$ whereas, for example, $\varepsilon\sim\phi$. Also $\mathrm{pa}(\chi)=\{\gamma\}$ and $\mathrm{ch}(\gamma)=\{\delta,\chi\}$ as well as $\mathrm{bd}(\delta)=\{\alpha,\phi,\gamma\}$. The connectivity component of β is $\mathrm{co}(\beta)=\{\beta,\chi\}$.

A *path* of length n from α to β is a sequence $\alpha=\alpha_0,\ldots,\alpha_n=\beta$ of distinct vertices such that $(\alpha_{i-1},\alpha_i)\in E$ for all $i=1,\ldots,n$. If there is a path from α to β we say that α *leads to* β and write $\alpha\mapsto\beta$. The set of vertices α such that $\alpha\mapsto\beta$ are the *ancestors* $\mathrm{an}(\beta)$ of β and the *descendants* $\mathrm{de}(\alpha)$ of α are the vertices β such that $\alpha\mapsto\beta$. The *non-descendants* are $\mathrm{nd}(\alpha)=V\setminus[\mathrm{de}(\alpha)\cup\{\alpha\}]$. A set of vertices A is *ancestral* if $\mathrm{an}(A)\subseteq A$. The intersection of two ancestral sets is again ancestral such that the smallest closed ancestral set containing a given set B is well-defined. We denote this by $\mathrm{An}(B)$. Thus $\mathrm{An}(B)$ contains all parents and ancestors of B, all neighbours of B and their ancestors, etc.

If both $\alpha\mapsto\beta$ and $\beta\mapsto\alpha$ we say that α and β *connect* and write $\alpha\rightleftharpoons\beta$. Clearly $\rightleftharpoons$ is an equivalence relation and the corresponding equivalence classes $\mathrm{co}(\alpha)$ where then

$$\beta\in\mathrm{co}(\alpha)\Leftrightarrow\alpha\rightleftharpoons\beta,$$

are the *connectivity components* of $\mathcal{G}$. If $\alpha\in A\subseteq V$, the symbol $\mathrm{co}(\alpha)_A$ denotes the connectivity component of α in $\mathcal{G}_A$.

A *walk* of length n from α to β is a sequence $\alpha=\alpha_0,\ldots,\alpha_n=\beta$ of distinct vertices such that $\alpha_{i-1}\to\alpha_i$ or $\alpha_i\to\alpha_{i-1}$ for all $i=1,\ldots,n$.

A subset S is said to *separate A from B* if all walks from vertices $\alpha \in A$ to $\beta \in B$ intersect S.

An *n-cycle* is a path of length n with the modification that $\alpha = \beta$, i.e. it begins and ends in the same point. The cycle is said to be *directed* if it contains an arrow.

A class of graphs of special interest to us is the class of *chain graphs*, so denoted in Lauritzen & Wermuth (1989). These are the graphs where the vertex set V can be partitioned into numbered subsets, forming a so-called *dependence chain* $V = V(1) \cup \ldots \cup V(T)$ such that all edges between vertices in the same subset are undirected and all edges between different subsets are directed, pointing from the set with lower number to the one with higher number. As shown in Frydenberg (1989), these are characterized by *not having any directed cycles* and the connectivity components form a finest possible partitioning of such a graph, the elements of which are the *chain components of* $\mathscr{G}$. The graph in Fig. 2 is a chain graph. Its chain components are $\{\alpha, \delta, \varepsilon, \phi\}$, $\{\gamma\}$, $\{\beta, \chi\}$. The chain components are most easily found by removing all arrows before taking connectivity components.

For a chain graph $\mathscr{G}$ we define as in Frydenberg (1989) (in analogy with Lauritzen & Spiegelhalter, 1988) its *moral graph* $\mathscr{G}^m$ as the undirected graph with the same vertex set but with α and β adjacent in $\mathscr{G}^m$ if and only if either $\alpha \rightarrow \beta$ or $\beta \rightarrow \alpha$ or if there are γ_1, γ_2 in the same chain component such that $\alpha \rightarrow \gamma_1$ and $\beta \rightarrow \gamma_2$. In the graph of Fig. 2, the moral graph is obtained by adding an edge between α and γ that both have children in the same chain component $\{\beta, \chi\}$, and then ignoring directions.

In the special case of a directed, acyclic graph the moral graph is obtained from the original graph by "marrying parents" with a common child and then dropping directions on arrows.

2.2. *Decompositions of marked graphs*

In the present subsection we shall be concerned with decompositions and decomposable graphs. The notions pertain to *undirected* graphs and have deep connections to many areas, other than the statistical analysis of association such as in Andersen (1974) and Haberman (1974). This includes general graph theory (Diestel, 1987), the four-colour problem (Wagner, 1937), measure theory (Kellerer, 1964a, b; Vorob'ev, 1962, 1963), the solution of systems of linear equations (Parter, 1961; Rose, 1970), game theory (Vorob'ev, 1967), relational databases (Beeri *et al.*, 1981, 1983) and expert systems (Lauritzen & Spiegelhalter, 1988). See these as well as Lauritzen, Speed & Vijayan (1984) and Golumbic (1980) for further discussion and references.

In the pure case, decomposable graphs are well-studied objects although they usually appear under other names such as, for example, *rigid circuit* (Dirac, 1961), *chordal* (Gavril, 1972) or *triangulated* (Berge, 1973; Rose, 1970).

In the case of a marked graph, the notion of a decomposable graph was first introduced by Lauritzen & Wermuth (1984) and studied further by Leimer (1985, 1989a). The present developments follow the latter reference rather closely although there are minor differences in terminology. Since the notion is so central, we state formally that

Definition 1

A triple (A, B, C) *of disjoint subsets of the vertex set* V *of an undirected, marked graph* $\mathscr{G}$ *is said to form a* (*strong*) decomposition *of* $\mathscr{G}$ *if* $V = A \cup B \cup C$ *and the three conditions below all hold:*

(*i*) *C separates A from B*
(*ii*) *C is a complete subset of V*
(*iii*) $C \subseteq \Delta \vee B \subseteq \Gamma$.

When this is the case we say that (A, B, C) *decomposes* $\mathscr{G}$ into the *components* $\mathscr{G}_{A \cup C}$ and $\mathscr{G}_{B \cup C}$. Occasionally, to avoid misunderstanding, we shall use the qualifier *strong* to distinguish from a *weak decomposition* which is defined below.

Definition 2

A triple (A, B, C) *of disjoint subsets of the vertex set* V *of an undirected, marked graph* $\mathscr{G}$ *is said to form a* weak decomposition *of* $\mathscr{G}$ *if* $V = A \cup B \cup C$ *and the two conditions below both hold:*

(*i*) *C separates A from B*
(*ii*) *C is a complete subset of V,*

i.e. a triple (A, B, C) that satisfies (i) and (ii) of definition 1, but not necessarily (iii). In the pure cases (iii) holds automatically and all weak decompositions are also decompositions. Note that what we have chosen to call a decomposition (without a qualifier) is what Leimer (1989a) calls a strong decomposition. Note also that we allow some of the sets in (A, B, C) to be empty. If the sets A and B in (A, B, C) are both non-empty, we say that the decomposition is *proper*. Fig. 3 illustrates the notions of (strong) and weak decompositions.

A decomposable graph is one that can be successively decomposed into its cliques. Again we choose to state this formally through a recursive definition as

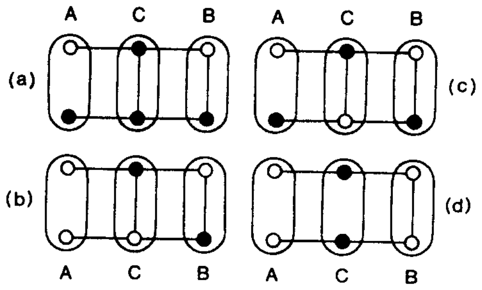

Fig. 3. Illustration of the notions of (strong) and weak decompositions. In (a) we see a decomposition with $C \subseteq \Delta$ and in (b) with $a \subset \Gamma$. In (c) the decomposition is only weak because none of these two conditions is fulfilled. In (d) we do *not* have a decomposition because the separator C is not complete.

Definition 3

An undirected, marked graph is said to be decomposable *if it is complete, or if there exists a proper decomposition* (A, B, C) *into decomposable subgraphs* $\mathscr{G}_{A \cup C}$ *and* $\mathscr{G}_{B \cup C}$.

The definition makes sense because the decomposition is assumed to be proper, such that both subgraphs $\mathscr{G}_{A \cup C}$ and $\mathscr{G}_{B \cup C}$ have fewer vertices than the original graph $\mathscr{G}$.

Analogously, a *weakly decomposable* graph is one that can be decomposed into cliques by weak decompositions. It is then obvious that *any decomposable graph is weakly decomposable*. The converse is in general false. The smallest graph that is not decomposable is given on Fig. 4. The only candidate for a separating set C is the vertex (c) but since this is a circle, they are not allowed dots in both sets A and B (then condition (iii) of definition 1 would be violated). But the graph is weakly decomposable because $(A = \{a\}, B = \{b\}, C = \{c\})$ is a weak decomposition of the graph.

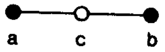

Fig. 4. The smallest, non-decomposable graph. The graph is weakly decomposable.

A *triangulated graph* is an undirected graph with the property that every cycle of length $n\geq 4$ possesses a *chord*, i.e. two non-consecutive vertices that are neighbours. A classical result states that

Proposition 1
The following conditions are equivalent for an undirected, marked graph $\mathcal{G}$:

- *(i) $\mathcal{G}$ is weakly decomposable*
- *(ii) $\mathcal{G}$ is triangulated*
- *(iii) Every minimal (α, β)-separator is complete.*

Proof. See for example Golumbic (1980). □

The smallest graph that is not weakly decomposable is therefore a 4-cycle and shown in Fig. 5.

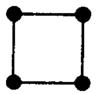

Fig. 5. The smallest graph is not weakly decomposable.

An elegant construction, due to Leimer (1989a), makes it possible to take full advantage of the wealth of knowledge pertaining to triangulated graphs when discussing decomposable, marked graphs. If $\mathcal{G}$ is an undirected, marked graph its *stargraph* is denoted as $\mathcal{G}^*=(V^*, E^*)$ and constructed as follows: the vertex set V is to be extended with an extra vertex, denoted $*$, such that $V^*=V\cup\{*\}$. This extra vertex, the *star*, is then connected to all discrete vertices by a line, such that $E^*=E\cup\{(*, \delta), (\delta, *), \delta\in\Delta\}$. The construction of stargraphs is illustrated on Fig. 6.

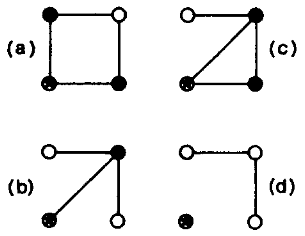

Fig. 6. The construction of stargraphs. In (a), which is the stargraph of the non-decomposable graph from Fig. 4, the star creates a 4-cycle whilst the other three decomposable graphs have triangulated stargraphs.

The astonishing fact which makes the construction so valuable is contained in the following result:

Proposition 2 (Leimer)
A marked graph $\mathcal{G}$ is decomposable if and only if $\mathcal{G}^*$ is triangulated.

Proof. See Leimer (1989a).

It follows from proposition 2 that a forbidden path characterization exists for decomposable graphs. The following result is also due to Leimer (1989a):

Corollary 1
An undirected, marked graph is decomposable if and only if it is triangulated and does not contain any path $(\delta_1=\alpha_0, \ldots, \alpha_n=\delta_2)$ between two non-adjacent discrete vertices passing through only continuous vertices, i.e. with $\delta_1 \not\sim \delta_2$ and $\alpha_i\in\Gamma$ for $0<i<n$.

Proof. The forbidden paths are exactly those that give rise to chordless cycles in $\mathcal{G}^*$. □

We have illustrated the typical forbidden path in Fig. 7. See also Figs 4 and 6, the latter indicating clearly how the forbidden path creates a 4-cycle in the stargraph.

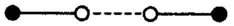

Fig. 7. Path which is forbidden in a decomposable graph.

2.3. *Simplicial subsets and perfect sequences*

Closely related to the notion of a decomposition is the notion of a *simplicial subset*, which is a subset A that satisfies the following two conditions

(i) bd(A) is complete
(ii) $A\subseteq\Gamma\wedge \mathrm{bd}(A)\subseteq\Delta$.

A subset is said to be *weakly* simplicial, if the first of these conditions is satisfied. Clearly, when a subset is (weakly) simplicial the triple $(V\backslash \mathrm{cl}(A), A, \mathrm{bd}(A))$ is a (weak) decomposition of $\mathcal{G}$. The notion is illustrated on Fig. 8. A vertex α is said to be (weakly) simplicial if the subset $\{\alpha\}$ is.

A sequence $(C_1, \ldots, C_k)$ of complete sets in $\mathcal{G}$ such that for all $j>1$, R_j is simplicial in $\mathcal{G}_{H_j}$, where

$$H_j=(C_1\cup\ldots\cup C_j), \quad R_j=C_j\backslash H_{j-1}$$

is said to be *perfect*. H_j are the *histories* and R_j the *residuals* of the sequence. Thus when we let S_j denote the *separators* $S_j=H_{j-1}\cap C_j$, we have that if a sequence of cliques is perfect then

$$(H_{j-1}\backslash C_j, R_j, S_j) \quad \text{decomposes} \quad \mathcal{G}_{H_j}.$$

A *perfect numbering* of the vertices V of $\mathcal{G}$ is a numbering $(\alpha_1, \ldots, \alpha_k)$ such that

$$\mathrm{bd}(\alpha_j)\cap\{\alpha_1, \ldots, \alpha_{j-1}\}, \; j>1$$

is a sequence of complete sets. Similarly we define *weakly perfect* sequences and numberings as those where the relevant sets are only weakly simplicial.

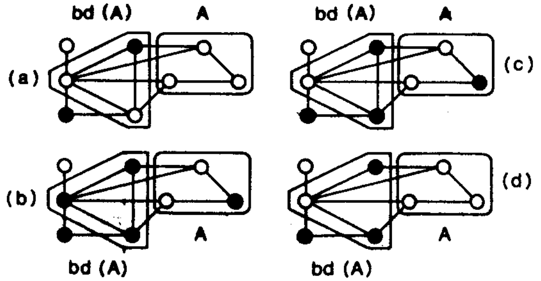

Fig. 8. Simplicial subsets. In (a), A is simplicial because $A\subseteq\Gamma$ and bd(A) is complete. In (b), A is simplicial because bd(A)$\subseteq\Delta$ and bd(A) is complete. In (c), A is only weakly simplicial because $A\cap\Delta\neq\emptyset$ and bd(A)$\cap\Gamma\neq\emptyset$. In (d), A is not even weakly simplicial because bd(A) is not complete.

Perfect sequences play important roles in the understanding and manipulation of decomposable graphs. Partly because their existence characterize decomposable graphs but also because they form the basis for recursive computational procedures. The characterization results are stated below. The proof is in Leimer (1989a), where also fast algorithms for finding such numberings are described

Proposition 3
The following conditions are equivalent for an undirected, marked graph $\mathscr{G}$:

(*i*) *The vertices of $\mathscr{G}$ admit a perfect numbering*
(*ii*) *The cliques of $\mathscr{G}$ admit a perfect numbering*
(*iii*) *The graph $\mathscr{G}$ is decomposable.*

3. Markov properties

3.1. *Markov fields over undirected graphs*

For an undirected graph $\mathscr{G}=(V, E)$ we consider a random field with V as index set, i.e. a collection of random variables (X_v), $v\in V$ taking values in probability spaces $\mathscr{X}_v$, $v\in V$. The probability spaces in the present paper are either the real line or finite sets. For A being a subset of V we let $\mathscr{X}_A=\times_{v\in A}\mathscr{X}_v$ and further $\mathscr{X}=\mathscr{X}_V$. Typical elements of $\mathscr{X}_A$ are denoted as $x_A=(x_v)_{v\in A}$ and so on. Then a probability measure of P on $\mathscr{X}$ is said to *factorize* according to $\mathscr{G}$ if there exists non-negative functions ψ_A defined on $\mathscr{X}_A$ for only complete subsets A, and a product measure $\mu=\otimes_{v\in V}\mu_v$ on $\mathscr{X}$, such that

$$P=f\cdot\mu \quad \text{where} \quad f(x)=\prod_A \psi_A(x_A).$$

The functions ψ_A are referred to as *factor potentials* of P. These are not uniquely determined. There is arbitrariness in the choice of μ, but also groups of functions ψ_A can be multiplied together or split up in different ways. In fact one can without loss of generality assume—although this is not always practical—that only cliques (maximal complete subsets) appear as the sets A, i.e. that

$$f=\prod_{c\in\mathscr{C}}\psi_C, \tag{1}$$

where $\mathscr{C}$ is the set of cliques of $\mathscr{G}$.

Associated with this structure there is a range of Markov properties that in general can be different. A probability measure P on $\mathscr{X}$ is said to obey

(L) *the local Markov property*, relative to $\mathscr{G}$, if for any vertex $\alpha\in V$

$\alpha\perp\!\!\!\perp V\backslash\mathrm{cl}(\alpha)\,|\,\mathrm{bd}(\alpha)$.

(P) *the pairwise Markov property*, relative to $\mathscr{G}$, if for any pair (α,β) of non-adjacent vertices

$\alpha\perp\!\!\!\perp\beta\,|\,V\backslash\{\alpha,\beta\}$,

(G) *the global Markov property*, relative to $\mathscr{G}$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $\mathscr{G}$

$A\perp\!\!\!\perp B\,|\,S$.

Here we have used the notation "$A\perp\!\!\!\perp B|S$", etc., as short for "X_A is conditionally independent of X_B given X_S". See Dawid (1979, 1980) for detailed properties of conditional independence.

It is obvious that we have

$$(\mathrm{G})\Rightarrow(\mathrm{L})\Rightarrow(\mathrm{P}) \tag{2}$$

but in general the three properties are different (see Speed, 1979, for a discussion). In fact, if P admits a density w.r.t. μ which is strictly positive, the properties are equivalent, as shown, for example, by Pearl & Paz (1986) (see also Pearl, 1988). It is easy to see that in general it is true that *if P factorizes according to $\mathscr{G}$, it also obeys the global Markov property.*

The global Markov property (G) is important because it gives a general criterion for deciding when two groups of variables A and B are conditionally independent given a third group of variables S. And this criterion can, apart from degenerate cases, not be improved (see Frydenberg, 1990), in the sense that if A and B are not separated from S then there will be factorizing probabilities P, such that $A\perp\!\!\!\perp B|C$ will not hold.

In our special case where P admits a strictly positive density w.r.t. μ, each of the equivalent Markov properties imply that P factorizes, a result known as the Hammersley–Clifford theorem (see Speed, 1979, for a discussion of the problems involved). It therefore makes sense to use the term *Markov property* for any of the four properties.

3.2. *Markov fields over directed acyclic graphs*

Before we proceed to the case of a general chain graph we consider the same set-up as in the previous subsection just that now the graph $\mathscr{G}$ is assumed to be directed and acyclic. The Markov property on a directed acyclic graph was first studied systematically in Kiiveri, Speed & Carlin (1984), but the understanding has recently developed considerably (see, for example, Pearl, 1986; Verma, 1988; Pearl & Verma, 1987; Smith, 1989; Geiger & Pearl, 1988; Lauritzen *et al.*, 1989).

We say that P admits a *recursive factorization* according to $\mathscr{G}$ if there exist non-negative functions, henceforth referred to as *kernels*, $k^v(\cdot,\cdot)$, $v \in V$ defined on $\mathscr{X}_v \times \mathscr{X}_{\mathrm{pa}(v)}$ such that

$$\int k^v(y_v, x_{\mathrm{pa}(v)})\mu_v(dy_v)=1$$

and

$$P=f\cdot\mu \quad \text{where} \quad f(x)=\prod_{v\in V} k^v(x_v, x_{\mathrm{pa}(v)}).$$

It is an easy induction argument to show that if P admits a recursive factorization as above, then the kernels $k^v(\cdot, x_{\mathrm{pa}(v)})$ are in fact densities for the conditional distribution of X_v, given $X_{\mathrm{pa}(v)}$. Also it is immediate (as was noted in Lauritzen & Spiegelhalter, 1988), that if we form the undirected, moral graph $\mathscr{G}^m$ (marrying parents and dropping directions) such as described towards the end of subsection 2.1, we have

Lemma 1
If P admits a recursive factorization according to the directed, acyclic graph $\mathscr{G}$, it factorizes according to the moral graph $\mathscr{G}^m$ and obeys therefore the global Markov property relative to $\mathscr{G}^m$.

Proof. The factorization follows from the fact that, by construction, the sets $\{v\}\cup\mathrm{pa}(v)$ are complete in $\mathscr{G}_m$ and we can therefore let $\psi_{\{v\}\cup\mathrm{pa}(v)}=k^v$. The remaining part of the statement follows from the remarks in the previous subsection. □

It then follows (see Lauritzen *et al.*, 1989) that we have:

Proposition 4.
Let P factorize recursively according to $\mathscr{G}$. Then

$$A \perp\!\!\!\perp B \mid S$$

wherever A and B are separated by S in $(\mathscr{G}_{\mathrm{An}(A\cup B\cup S)})^m$, the moral graph of the smallest ancestral set containing $A\cup B\cup S$.

The property in proposition 4 shall be refered to as the *directed global Markov property* (DG). Note that our terminology here is different from that used by Kiiveri, Speed and Carlin (1984). One can show that this global directed Markov property has the same role as the global Markov property in the case of an undirected graph, in the sense that it gives the sharpest possible rule for reading conditional independence relations off the directed graph. The procedure is illustrated in the following

Example 1. Consider a directed Markov field on the graph in Fig. 9 and the problem of deciding, whether $a \perp\!\!\!\perp b \mid S$? The moral graph of the smallest ancestral set containing all the variables involved is shown in Fig. 10.
It is immediate that S separates a from b in this graph, implying $a \perp\!\!\!\perp b \mid S$. □

To complete this subsection we say that P obeys the *directed local Markov property* (DL) if any variable is conditionally independent of its non-descendants, given its parents:

$$v \perp\!\!\!\perp \mathrm{nd}(v)\backslash\mathrm{pa}(v) \mid \mathrm{pa}(v).$$

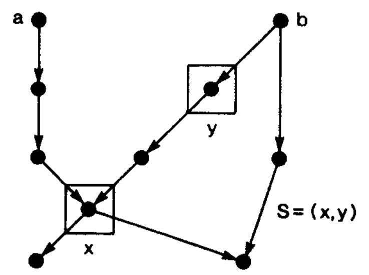

Fig. 9. The directed, global Markov property. Is $a \perp\!\!\!\perp b \mid S$?

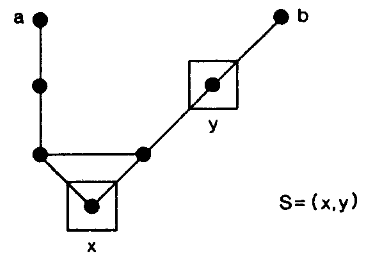

Fig. 10. The moral graph of the smallest ancestral set in the graph of Fig. 9 containing $\{a\}\cup\{b\}\cup S$. Clearly S separates a from b in this graph, implying $a \perp\!\!\!\perp b \mid S$.

The latter coincides with the terminology of Kiiveri, Speed & Carlin (1984). In contrast with the undirected case we have that all the three properties (DF), (DL) and (DG) are equivalent just assuming existence of the density f, stated formally as:

Proposition 5
Let $\mathscr{G}$ be a directed, acyclic graph. For a probability measure P on $\mathscr{X}$ which is absolutely continuous w.r.t a product measure μ, the following conditions are equivalent:

(DF) *P admits a recursive factorization according to $\mathscr{G}$*
(DG) *P obeys the global directed Markov property, relative to $\mathscr{G}$*
(DL) *P obeys the local directed Markov property, relative to $\mathscr{G}$.*

Proof. See Lauritzen *et al.* (1989). □

Since the three conditions in proposition 5 are equivalent it makes sense to speak of a *directed Markov field* as one where any of the conditions is satisfied.

3.3. *Markov fields over chain graphs*

In the present subsection we deal with general chain graphs $\mathcal{G}=(V, E)$ but otherwise the usual set-up and all probability measures are assumed to have positive densities. A complete discussion of the chain graph Markov property has been given by Frydenberg (1989) and some parts of this shall be summarized here. For our purposes it suffices to use the definition in Lauritzen & Wermuth (1989) which is as follows.

Since the graph is a chain graph, the vertex set can be partitioned as $V=V(1)\cup\ldots\cup V(T)$ such that each of the sets $V(t)$ only has lines between vertices and arrows point from vertices in sets with lower number to those with higher number. We then introduce the set of *concurrent* variables $C(t)=V(1)\cup\ldots\cup V(t)$ and say that P satisfies the *pairwise chain Markov property* (PCM) if for any pair (α, β) of non-adjacent vertices we have

$$\alpha \perp\!\!\!\perp \beta \mid C(t^*)\setminus\{\alpha, \beta\},$$

where t^* is the smallest t that has $\alpha, \beta \in C(t)$. It can be shown (Frydenberg, 1989), that—in the case of positive densities—this property does not depend on the actual partitioning, but only depends on the graph $\mathcal{G}$. In the case where $\mathcal{G}$ is undirected, we get the usual Markov property and if the chain graph is directed, we get the directed Markov property described in the previous subsection.

It has been shown by Frydenberg (1989) that the pairwise property above is equivalent to the *global chain Markov property* (GCM):

$$A \perp\!\!\!\perp B \mid S$$

whenever A and B are separated by S in $(\mathcal{G}_{An(A\cup B\cup S)})^m$, the moral graph of the smallest ancestral set containing $A\cup B\cup S$. In contrast to the directed and undirected case, we do at present not know for sure that this criterion cannot be sharpened.

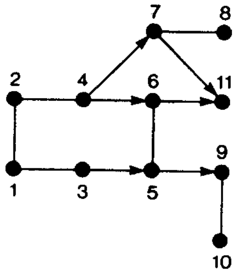

Fig. 11. A chain graph with chain components $\{1, 2, 3, 4\}$, $\{5, 6\}$, $\{7, 8\}$, $\{9, 10\}$, $\{11\}$. Is $3 \perp\!\!\!\perp 8 \mid \{2, 5\}$? Is $3 \perp\!\!\!\perp 8 \mid 2$?

Example 2. As an illustration of this, consider the graph in Fig. 11 which is taken from Frydenberg (1989), and the question of deciding if $3 \perp\!\!\!\perp 8 \mid \{2, 5\}$. The smallest ancestral set containing these variables is the set $\{1, 2, 3, 4, 5, 6, 7, 8\}$. The moral graph of this adds an edge between 3 and 4, because these both have children in the chain component $\{5, 6\}$. Thus the graph in Fig. 12 appears. Since there is a path between 3 and 8 circumventing 2 and 5 in this graph, we cannot conclude that $3 \perp\!\!\!\perp 8 \mid \{2, 5\}$.

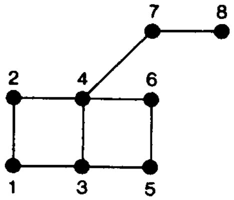

Fig. 12. Moral graph of smallest ancestral set in the graph of Fig. 11 containing $\{2, 3, 5, 8\}$. A connection between 3 and 4 has been introduced since these both have children in the same chain component $\{5, 6\}$. We cannot conclude $3 \perp\!\!\!\perp 8 \mid \{2, 5\}$.

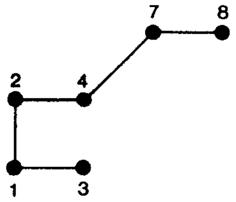

Fig. 13. Moral graph of smallest ancestral set in the graph of Fig. 11 containing $\{2, 3, 8\}$. We conclude that $3 \perp\!\!\!\perp 8 \mid 2$.

If we instead consider the question whether $3 \perp\!\!\!\perp 8 \mid 2$, the smallest ancestral set becomes $\{1, 2, 3, 4, 7, 8\}$, no edge has to be added between 3 and 4 and Fig. 13 reveals that $3 \perp\!\!\!\perp 8 \mid 2$.

4. Distributions and models

4.1. *CG-distributions and -regressions*

We shall briefly review some standard notation from Lauritzen & Wermuth (1989). Recall that the set of variables V is partitioned as $V=\Delta\cup\Gamma$ into variables of *discrete* (Δ) and *continuous* (Γ) type. A typical element of the joint state space is denoted as in one of the possibilities below

$$x=(x_\alpha)_{\alpha\in V}=(i, y)=\{(i_\delta)_{\delta\in\Delta}, (y_\gamma)_{\gamma\in\Gamma}\},$$

where i_δ are qualitative and y_γ are real valued. A particular combination $i=(i_\delta)_{\delta\in\Delta}$ is referred to as a *cell* and the set of cells is denoted by $\mathcal{I}$. The joint distribution of the variables is supposed to have a density f with

$$\log f(x)=\log f(i, y)=g(i)+h(i)'y-\tfrac{1}{2}y'K(i)y,$$

in which case we say that X follows a *CG-distribution* which is equivalent to the statements

$$p(i)=P\{X_\Delta=i\}>0 \quad\text{and}\quad \mathcal{L}(X_\Gamma \mid X_\Delta=i)=\mathcal{N}_{|\Gamma|}\{\xi(i), \Sigma(i)\},$$

where $X_A=(X_a)_{a\in A}$ and so on, $\xi(i)=K(i)^{-1}h(i)$ and $\Sigma(i)=K(i)^{-1}$ is positive definite. The triple

The triple (g, h, K) constitutes the *canonical characteristics* of the distribution and (p, ξ, Σ) are the *moment characteristics*. Note that combinations of the elements of these triples can be used to parametrize the distributions. For example, the *standard mixed characteristics* given as (p, h, K) are often convenient. We say that the distribution is *homogeneous* if the covariance is independent of i, i.e. if $\Sigma(i)=\Sigma$ or, equivalently, if $K(i)=K$. The homogeneous distributions were, for example, studied by Olkin & Tate (1961).

From Lauritzen & Wermuth (1989) we have the following facts pertaining to the behaviour of marginal and conditional distributions of variables that follow CG-distributions. For a subset A of V we denote the marginal density of X_A by f_A and for $B=V\setminus A$ the conditional density of X_B given $X_A=x_A$ as $f_{B|A}$.

Proposition 6
For a CG-density f we have:

- *For all x_A, $f_{B|A}$ is a CG-density.*
- *If $B\subseteq\Gamma$, f_A is a CG-density.*
- *If $B\subseteq\Delta$ and $B\perp\!\!\!\perp\Gamma\,|\,\Delta\setminus B$, f_A is a CG-density.*

Proof. This is straightforward (and in Lauritzen & Wermuth, 1989). □

The densities $f_{B|A}(\cdot\,|x_A)$ are called *CG-regressions* and *homogeneous* CG-regressions if the original density f is homogeneous. They are described in detail in Lauritzen & Wermuth (1989).

In particular we are interested in the interplay between the distributional properties and the various Markov properties. For an undirected graph we introduce the symbols $M(\mathscr{G})$ for the set of CG-distributions that are also Markov relative to $\mathscr{G}$. Similarly $M_H(\mathscr{G})$ is the set of homogeneous Markovian CG-distributions. Further, the symbols for marginal and conditional distributions:

$$M(\mathscr{G})_A=\{f_A\,|\,f\in M(\mathscr{G})\}$$

$$M_H(\mathscr{G})_A=\{f_A\,|\,f\in M_H(\mathscr{G})\}$$

$$M(\mathscr{G})^A=\{f_{B|A}\,|\,f\in M(\mathscr{G})\}$$

$$M_H(\mathscr{G})^A=\{f_{B|A}|f\in M_H(\mathscr{G})\}.$$

The distributions in $M(\mathscr{G})_A$ (nor $M_H(\mathscr{G})_A$) will not for general subsets A be CG-distributions but those in $M(\mathscr{G})^A$ will always be CG-regressions and also Markovian on $\mathscr{G}_B$ (see Lauritzen & Wermuth, 1989). But we have:

Proposition 7
If B is a simplicial subset, then

$$M(\mathscr{G})_A=M(\mathscr{G}_A) \quad and \quad M_H(\mathscr{G})_A=M_H(\mathscr{G}_A).$$

Proof. Lauritzen & Wermuth (1989). □

4.2. *The graphical association models*

We now have the necessary prerequisites to specify the classes of joint distributions that constitute the models of our concern.

Consider first an undirected graph $\mathscr{G}$. Thus the graphical association model determined by $\mathscr{G}$ is given by the family of densities $M(\mathscr{G})$, and the homogeneous graphical association model by $M_H(\mathscr{G})$. In the pure cases there is no difference between the homogeneous and non-homogeneous cases since then $M(\mathscr{G})=M_H(\mathscr{G})$. In the pure discrete case these models are known as the graphical log-linear models studied in Darroch, Lauritzen & Speed (1980) and in the pure continuous case as covariance selection models introduced by Dempster (1972). As shown in Lauritzen & Wermuth (1989) these models are characterized by their canonical characteristics having interaction expansions (cf. Darroch & Speed, 1983), as

$$g(i)=\sum_{d\subseteq\Delta}\lambda_d(i),\quad h(i)=\sum_{d\subseteq\Delta}\eta_d(i),\quad K(i)=\sum_{d\subseteq\Delta}\Psi_d(i),$$

where functions of the type ϕ_d only depend on i through i_d and where

$$\lambda_d(i)=0 \quad \text{unless } d \text{ is complete in } \mathscr{G}$$

$$\eta_d(i)_\gamma=0 \quad \text{unless } d\cup\{\gamma\} \text{ is complete in } \mathscr{G}$$

$$\psi_d(i)_{\gamma,\mu}=0 \quad \text{unless } d\cup\{\gamma,\mu\} \text{ is complete in } \mathscr{G}.$$

In the homogeneous case, the *mixed quadratic interactions* $\Psi_d(i)$, $d\neq\emptyset$ are all equal to zero. The terms λ_d are the *pure discrete interactions*, and $\eta_d(i)$, $d\neq\emptyset$ are the *mixed linear interactions*. Because of this interaction representation of the densities, we use the term *graphical interaction model* in the case of an undirected graph. Edwards (1989b) has extended the class of models to hierarchical mixed interaction models.

In the case of a directed acylic graph $\mathscr{G}$ we proceed as follows. As explained in subsection 3.2, a directed Markov field is specified through the kernels k^v, specifying the conditional distribution of any variable, given the values of its parent variables. In this case we say that $f\in M(\mathscr{G})$ resp. $f\in M_H(\mathscr{G})$ if $f=\Pi_v k^v$ where k^v are the CG-regressions

$$k^v\in M(\mathscr{G}^{\sim}_{\{v\}\cup\overline{\mathrm{pa}(v)}})^{\mathrm{pa}(v)} \quad \text{resp.} \quad k^v\in M_H(\mathscr{G}^{\sim}_{\{v\}\cup\overline{\mathrm{pa}(v)}})^{\mathrm{pa}(v)}.$$

The overlining indicates that all lines between vertices in pa(v) have been added. Recall that $\mathscr{G}^{\sim}$ is obtained from $\mathscr{G}$ by dropping directions on edges. The models $M(\mathscr{G})$ and $M(\mathscr{G}^{\sim})$ are in general different, both concerning distributional type and Markov properties, but for some special graphs they are identical as stated precisely below.

Proposition 8
If all $v\in V$ are simplicial in $\mathscr{G}^{\sim}_{\{v\}\cup\mathrm{pa}(v)}$ then

$$M(\mathscr{G})=M(\mathscr{G}^{\sim}) \quad and \quad M_H(\mathscr{G})=M_H(\mathscr{G}^{\sim}).$$

Proof. Lauritzen & Wermuth (1989). □

Frydenberg (1990) has shown that the converse to the above is true.

In the pure cases there is as usual no difference between the homogeneous and the general case. The pure continuous case was discussed by Wermuth (1980) and the pure discrete case by Wermuth & Lauritzen (1983). See also Birch (1963) for an early example in the case of three variables.

In the general case of a chain graph $\mathscr{G}$ the construction is similar. Instead of single vertices we have to consider the set Ω of chain components ω of $\mathscr{G}$. Then we say that $f\in M(\mathscr{G})$ resp. $f\in M_H(\mathscr{G})$ if $f=\Pi_{\omega\in\Omega}k^\omega$ where k^ω are the multivariate CG-regressions

$$k^\omega\in M(\mathscr{G}^{\sim}_{\omega\cup\overline{\mathrm{pa}(\omega)}})^{\mathrm{pa}(\omega)} \quad \text{resp.} \quad k^\omega\in M_H(\mathscr{G}^{\sim}_{\omega\cup\overline{\mathrm{pa}(\omega)}})^{\mathrm{pa}(\omega)}.$$

This definition is shown by Frydenberg (1990) to coincide with the definition given in Lauritzen & Wermuth (1989).

In the pure cases there is as usual no distinction between the general and the homogeneous case. The pure discrete case has, in a slightly different framework, been studied by Goodman (1973) and Asmussen & Edwards (1983) and the pure continuous case briefly by Porteous (1985a). Again we have found from Lauritzen & Wermuth (1989) the analogous result to proposition 8.

Proposition 9
If all $\omega \in \Omega$ are simplicial in $\mathscr{G}^{\sim}_{\omega \cup \mathrm{pa}(\omega)}$ then

$$M(\mathscr{G})=M(\mathscr{G}^{\sim}) \quad \textit{and} \quad M_H(\mathscr{G})=M_H(\mathscr{G}^{\sim}).$$

And the converse holds as well (cf. Frydenberg, 1989, 1990).

Necessary and sufficient conditions for the Markov properties of two chain models to coincide has been given by Frydenberg (1989). Further results giving conditions for models to be equal when their graphs have the same vertex set and the same connections between vertices, just with possibly different directions, can be found in Lauritzen & Wermuth (1989) and Wermuth & Lauritzen (1989), but the general result waits to be formulated.

As discussed in Wermuth & Lauritzen (1989), these conditions are both helpful when models are to be interpreted and also when models are to be fitted. We shall give examples of this later in the paper.

5. General notation

Throughout the paper we consider the situation with a sample $(x^1, \ldots, x^n)$ of size n from a distribution with density f which is unknown apart from it belonging to $M(\mathscr{G})$ (or $M_H(\mathscr{G})$) for a marked graph $\mathscr{G}$.

It is convenient to introduce the following notation for standard sampling statistics, where A is an arbitrary subset of V, and where we abbreviate to let $i_A=i_{A\cup\Delta}$, $\mathscr{I}_A=\mathscr{I}_{A\cup\Delta}$, $y_A=y_{A\cup\Gamma}$ and so on.

$$d(i_A)=\{v \mid i^v_A=i_A\}$$

$$n(i_A)=|d(i_A)|=\text{the number of observations in cell } i_A$$

$$s(i_A)=\sum_{v\in d(i_A)} y^v=\text{the sum of the } y\text{-values in cell } i_A$$

$$\bar{y}(i_A)=s(i_A)/n(i_A)$$

$$ssd(i_A)=\sum_{v\in d(i_A)} \{y^v-\bar{y}(i_A)\}\{y^v-\bar{y}(i_A)\}'$$

$$=\text{the sum of squares of deviations from the mean in cell } i_A$$

$$ssd(A)=\sum_{i_A\in\mathscr{I}_A} ssd(i_A)$$

$$=\text{the “within” } ssd \text{ in the marginal table } \mathscr{I}_A$$

$$ssd=ssd(V).$$

Strictly speaking the quantities $\bar{y}(i_A)$ and hence also $ssd(i_A)$ are only defined for $n(i_A)>0$, but

they can be assigned any value—0, say—in the case $n(i_A)=0$. In common with standard practice, we shall use capital letters for the random variables corresponding to all the quantities above. An expression for $B\subset V$ such as $ssd_B(i_A)$ denotes the submatrix $\{ssd(i_A)_{\gamma\mu}\}_{\gamma,\mu\in B\cap\Gamma}$, and similarly for the other quantities.

6. Graphical interaction models

6.1. Saturated interaction models

Basic elements in all models are the *saturated interaction models* given by $M(\mathscr{G})$ where $\mathscr{G}$ is an undirected graph with all edges present. Then the restrictions only pertain to the distributional type. From Frydenberg & Lauritzen (1989) we have the sampling distributions of the sufficient statistics as described below, where $\mathscr{N}_k(\xi, \Sigma)$ is the multivariate Gaussian distribution with mean ξ and covariance matrix Σ and $\mathscr{W}_k(\Sigma, f)$ is the corresponding k-dimensional Wishart distribution with f degrees of freedom (cf. Anderson, 1984). We shall here omit a discussion of the homogeneous case.

Proposition 10
The set of statistics $(N(i), \bar{Y}(i), SSD(i))_{i\in\mathscr{I}}$ is minimal sufficient and has a sampling distribution which can be described as follows:

- *All components of $\{\bar{Y}(i), SSD(i)\}_{i\in\mathscr{I}}$ are conditionally independent and independent of $(I^1, \ldots, I^n)$ for a given table of counts $\{N(i)\}_{i\in\mathscr{I}}$.*
- *For all $i\in\mathscr{I}$ we have $\{\bar{Y}(i), SSD(i)\} \perp\!\!\!\perp \{N(j)\}_{j\neq i} \mid N(i)$.*
- *$\{N(i)\}_{i\in\mathscr{I}}$ has a multinomial distribution with parameters n and $p=\{p(i)\}_{i\in\mathscr{I}}$.*
- *$\mathscr{L}\{\bar{Y}(i) \mid N(i)=n(i)\}=\mathscr{N}_{|\Gamma|}\{\xi(i), n(i)^{-1}\Sigma(i)\}$, when $n(i)>0$.*
- *$\mathscr{L}\{SSD(i) \mid N(i)=n(i)\}=\mathscr{W}_{|\Gamma|}\{\Sigma(i), n(i)-1\}$, when $n(i)>1$.*

Proposition 11.
The likelihood function for the saturated model attains its maximum if and only if $ssd(i)$ is positive definite for all $i\in\mathscr{I}$ which for all $f\in M(\mathscr{G})$ is almost surely equal to the event

$$\bigcap_{i\in\mathscr{I}} \{n(i)>|\Gamma|\}.$$

Then the maximum likelihood estimate is given as

$$\hat{p}(i)=n(i)/n, \quad \hat{\xi}(i)=\bar{y}(i), \quad \hat{\Sigma}(i)=ssd(i)/n(i).$$

Note that the number of observations in any given cell is random and that the critical events $\{n(i)\leq|\Gamma|\}$ thus have positive probability. The practical consequence of this is that saturated models with many cells typically require large datasets. For any given set of data one can of course check the condition $n(i)>|\Gamma|$ just as easily as if it had been non-random.

It is of interest to consider the behaviour of other estimates of the concentration matrices $K(i)$ than the maximum likelihood estimator. Usual practice would conform with normalizing $ssd(i)$ by $n(i)-1$ but the desire for unbiased estimates of the interaction parameters suggests that normalization with $N(i)-|\Gamma|-2$ could be more appropriate. Note that the term unbiased has to be interpreted with caution. The distribution of $SSD(i)$ is mixed Wishart and contains a component where $SSD(i)$ is not invertible. Thus unbiasedness of $\hat{K}(i)$ has to be discussed conditionally on $n(i)$, and then only for $n(i)>|\Gamma|+3$ because otherwise expectation in the inverse Wishart distribution is not finite.

6.2. *General interaction models*

In the case where the model is not saturated, the identification of Markov distributions with those where interactions are missing, automatically leads to likelihood functions of exponential families (Barndorff-Nielsen, 1978), and the problem of maximizing the likelihood function can be solved iteratively, using the algorithm described in Frydenberg & Edwards (1989) and implemented in the program MIM (Edwards, 1987), that also has other facilities to be used in the analysis of mixed interaction models. The algorithm used in MIM specializes in the pure discrete case to iterative proportional scaling (Darroch & Ratcliff, 1972), and in the pure continuous case to the analogous algorithm for covariance selection models described by Speed & Kiiveri (1986). It should be noted that the algorithm is known to be convergent in the two pure cases whereas no proof of convergence in the general case is available.

If the graph is decomposable, the estimates can be found explicitly, such as described in Frydenberg & Lauritzen (1989). We shall illustrate the issues involved by three examples depicted in Fig. 14.

Models of this type are primarily appropriate as building blocks for models with response structure in the variables, but can have some independent interest. The variables could for example be students' marks in four subjects, where the marks a, b are given as pass/fail and c, d on a quantitative scale.

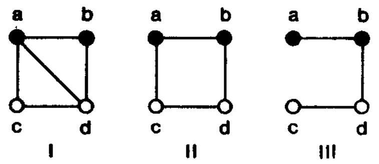

Fig. 14. Three graphical interaction models. Model II is not decomposable and iteration is needed to maximize the likelihood function.

Example 3. Consider the model given by graph I in Fig. 14. The only interactions which are not allowed are such that involve variables b and c. This does not affect the first component g of the canonical characteristics, but only h and K. Thus the log-density can be written as

$$\log f = g(i_{ab}) + h^c(i_a)y_c + h^d(i_{ab})y_d - \{k^{cc}(i_a)y_c^2 + 2k^{cd}(i_a)y_c y_d + k^{dd}(i_{ab})y_d^2\}/2. \tag{3}$$

Here and in the following, i_{ab} is short for $i_{\{a,b\}} = (i_a, i_b)$.

The graph is decomposable and the estimates can therefore be obtained in closed form. From Frydenberg and Lauritzen (1989) we have

$$\hat{p}(i) = \prod_{j=1}^{k} n(i_{C_j})/n(i_{S_j}) \tag{4}$$

$$\hat{h}(i) = \sum_{j=1}^{k} [\{ssd_{C_j}(i_{C_j})^{-1}s_{C_j}(i_{C_j})\}^0 - \{ssd_{S_j}(i_{S_j})^{-1}s_{S_j}(i_{S_j})\}^0] \tag{5}$$

$$\hat{K}(i) = \sum_{j=1}^{k} [n(i_{C_j})\{ssd_{C_j}(i_{C_j})^{-1}\}^0 - n(i_{S_j})\{ssd_{S_j}(i_{S_j})^{-1}\}^0], \tag{6}$$

where $\{A\}^0$ is the matrix or vector obtained by filling up coordinates with zeros until the right dimension is obtained. The sets C_j are the cliques ordered to form a perfect sequence and S_j are the separators (see subsection 2.3). In our example we have the cliques $\{a,b,d\}$, $\{a,c,d\}$ and the separator $\{a,d\}$. Recalling that in the formulae C_j is either short for $C_j \cap \Gamma$ or for $C_j \cap \Delta$, whichever appropriate, we get

$$\hat{p}(i_{ab}) = n(i_{ab})/n$$

$$\hat{h}^c(i_a) = ssd^{cc}(i_a)s_c(i_a) + ssd^{cd}(i_a)s_d(i_a)$$

$$\hat{h}^d(i_{ab}) = ssd^{dc}(i_a)s_c(i_a) + \{ssd^{dd}(i_a) - 1/ssd_d(i_a)\}s_d(i_a) + s_d(i_{ab})/ssd_d(i_{ab})$$

$$\hat{k}^{cc}(i_a) = n(i_a)ssd^{cc}(i_a)$$

$$\hat{k}^{cd}(i_a) = n(i_a)ssd^{cd}(i_a)$$

$$\hat{k}^{dd}(i_{ab}) = n(i_a)ssd^{dd}(i_a) + n(i_{ab})/ssd_d(i_{ab}) - n(i_a)/ssd_d(i_a).$$

Here ssd^{cc} is the cc-element of ssd^{-1} and so on. □

Example 4. Consider then the model given by II of Fig. 14. It is not decomposable because its graph is not triangulated. Therefore there is no explicit expression for the maximum likelihood estimates of the parameters involved. Apart from the interactions that were forbidden in model I, no interaction involving a and d is allowed whereby the log-density has the expression

$$\log f = g(i_{ab}) + h^c(i_a)y_c + h^d(i_b)y_d - \{k^{cc}(i_a)y_c^2 + 2k^{cd}y_c y_d + k^{dd}(i_b)y_d^2)/2.$$

The model has to be fitted iteratively using MIM. □

Example 5. Finally, let us consider the third model in Fig. 14. Here no interactions among a and c are allowed either, and the log-density has therefore an expression as

$$\log f = g(i_{ab}) + h^c y_c + h^d(i_b)y_d - \{k^{cc}y_c^2 + 2k^{cd}y_c y_d + k^{dd}(i_b)y_d^2\}/2.$$

This model is again decomposable. The cliques, ordered to form a perfect sequence are $\{a,b\}$, $\{b,d\}$, $\{c,d\}$ with separators $\{b\}$, $\{d\}$. We get from the formulae (4, 5, 6) that

$$\hat{p}(i_{ab}) = n(i_{ab})/n$$

$$\hat{h}^c = ssd^{cc}s_c + ssd^{cd}s_d$$

$$\hat{h}^d(i_b) = ssd^{dc}s_c + (ssd^{dd} - 1/ssd_d)s_d + s_d(i_b)/ssd_d(i_b)$$

$$\hat{k}^{cc} = n\ ssd^{cc}$$

$$\hat{k}^{cd} = n\ ssd^{cd}$$

$$\hat{k}^{dd}(i_b) = n\ ssd^{dd} + n(i_b)/ssd_d(i_b) - n/ssd_d.$$

The likelihood ratio for comparing models I and II as well as that for comparing II and III cannot be obtained in closed form, essentially because model II is not decomposable. But the likelihood ratio for comparing model III to model I cannot only be obtained in closed form but it can be partitioned into the product of those for comparing model III to II' and for comparing model II′ to model I. Here model II′ is obtained from model III by adding an edge between a and d. Each of these likelihood ratios can be calculated in appropriate marginal problems. This has been shown in Frydenberg & Lauritzen (1989) and is illustrated in Fig. 15.

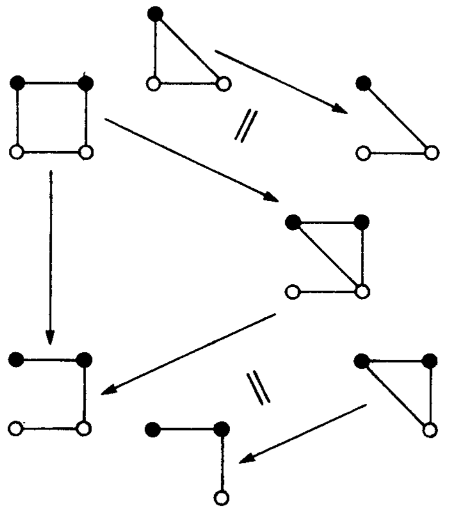

Fig. 15. Illustration of the decomposition of likelihood ratios. The likelihood ratio for the total reduction factorizes into products of likelihood ratios of the stepwise reductions indicated. Each of these are equal to the corresponding likelihood ratios for removing one edge in suitable marginal saturated models.

The result is analogous to corresponding results in the pure cases obtained by Sundberg (1975) and Wermuth (1976b), and can be used to obtain Bartlett corrections of the test statistic such as in Williams (1976) and Porteous (1985b) (see also Porteous, 1989). □

7. Homogeneous recursive models

In the present section we shall illustrate the elements involved in analysing the recursive models corresponding to directed acyclic graphs. We shall for the sake of simplicity only deal with the homogeneous models $M_H(\mathscr{G})$ although there is not such a great difference between the models in the two cases. In contrast to the previous section we do not discuss the saturated models but shall base our discussion entirely on three examples, selected to have the same corresponding undirected graphs as those in Fig. 14. The models to be discussed in the present section are shown in Fig. 16 and could for example be appropriate in a situation where two groups (I_1) of recent mothers were to be compared with respect to two measurements (Y_2, Y_3) of a pregnancy-related quantity such as, for example, the log-concentration of haemoglobin taken at two consecutive time points during pregnancy, and where one is further interested in the relation between these and the frequencies of newborn babies that are underweight (I_4).

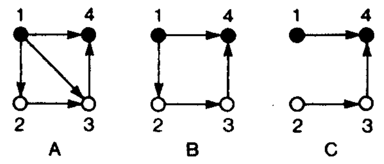

Fig. 16. Three recursive models. All models involve a logistic regression of variable 4 on variables 1 and 3 and iteration is therefore needed to maximize the joint likelihood function.

Example 6. Consider first model A of Fig. 16 which in the context given is to be interpreted by the early measurement not having any significance for predicting underweight of the baby, when the group and the later measurement is known. From Lauritzen & Wermuth (1989) we get

- I_1 has an arbitrary distribution with positive probabilities
- the conditional distribution of Y_2 given $I_1=i_1$ is $\mathscr{N}\{a_2(i_1), c_2\}$
- the conditional distribution of Y_3 given $(I_1, Y_2)=(i_1, y_2)$ is $\mathscr{N}\{a_3(i_1)+by_2, c_3\}$
- the conditional distribution of I_4, given the remaining variables has log-probabilities

$$\log p(i_4|i_1, y_{23})=\log p(i_4|i_1, y_3)=u(i_4|i_1)+v(i_4|i_1)y_3-\log \varkappa(i_1, y_3),$$

where $\varkappa$ is a normalizing constant.

This gives a joint density of the form

$$\begin{aligned}\log f&=g(i_1)+h^2(i_1)y_2+h^3(i_1)y_3-\{k^{22}(i_1)y_2^2+2k^{23}(i_1)y_2y_3+k^{33}(i_1)y_3^2\}/2\\&\quad+u(i_4|i_1)+v(i_4|i_1)y_3-\log \varkappa(i_1, y_3)\\&=g(i_{14})+h^2(i_1)y_2+h^3(i_{14})y_3-\{k^{22}(i_1)y_2^2+2k^{23}(i_1)y_2y_3+k^{33}(i_1)y_3^2\}/2-\log \varkappa(i_1, y_3)\end{aligned}$$

which should be contrasted with expression (3). The basic difference is that the normalizing constant $\varkappa$ enters into the expression. The model is fitted by fitting each of the conditional models separately. This can be done in closed form in the first three cases, where the conditional models are quite standard models and we obtain:

$$\begin{aligned}&\hat{p}(i_1)=n(i_1)/n\\&\hat{a}_2(i_1)=\bar{y}_2(i_1)\\&\quad\hat{c}_2=ssd_{22}(\{1\})/n\\&a_3(i_1)=\bar{y}_3(i_1)-ssd_{32}(\{1\})\bar{y}_2(i_1)/ssd_{22}(\{1\})\\&\quad\hat{b}=ssd_{32}(\{1\})/ssd_{22}(\{1\})\\&\quad c_3=ssd_{33}(\{1\})-ssd_{32}(\{1\})^2/ssd_{22}(\{1\}).\end{aligned}$$

The parameters of the last conditional model have to be computed iteratively and GLIM can for example be used for this purpose (as it can for the first three conditional models as well). Note that MIM cannot be used for this last part (in contrast to the first part) reflecting that the model is not equivalent to an undirected model (cf. proposition 8).

Example 7. Next consider model B that corresponds to the group being uninformative for the late quantitative measurement, when the early measurement is known. The additional restriction only involves the conditional distribution of Y_3 given (Y_2, I_1) where the intercepts (and therefore the regression lines) are now assumed identical. We get the same estimates as before, apart from

$$\hat{a}_3(i_1)=\bar{y}_3-ssd_{32}\,\bar{y}_2/ssd_{22}$$

$$\hat{b}=ssd_{32}/ssd_{22}$$

$$\hat{c}_3=ssd_{33}-ssd_{32}^2/ssd_{22},$$

where now the total sums of squares of deviations have been substituted for the "within" sums of squares of deviations. The likelihood ratio test for model B, assuming model A, is done by the usual comparison of estimates for c_3 and leads to the well-known F-test. Note that this is in spite of the fact that iteration was needed in both models A and B. This is very different from example 4. □

Example 8. Finally, consider model C of Fig. 16, having the feature that the group has no predictive value at all for the measurements. The only change in relation to the previous example concerns the conditional distribution of Y_2 given I_1 which no longer depends on the latter, i.e. we have $a_2(i_1)\equiv a_2$. Thus estimates are as before apart from

$$\hat{a}_2=s_2/n, \quad \hat{c}_2=ssd_{22}.$$

Again the likelihood ratio only involves the conditional distributions of Y_2 given I_1, and leads to the usual comparison of the "within" and "total" sums of squares of deviations $ssd_{22}(\{1\})$ and ssd_{22}, as known from one-way classification in the analysis of variance. □

Concluding this section we emphasize that although the joint likelihood could only be maximized iteratively and, in fact, no joint sufficient data reduction is available, the testing problems can be solved for some specific hypotheses. Note also that the models determined by the directed graphs in the examples above are very different from their undirected counterparts, not only in interpretation but also in the way they predict data behaviour.

8. Graphical chain models

General graphical chain models determined by $M(\mathscr{G})$ or $M_H(\mathscr{G})$ where $\mathscr{G}$ is a marked chain graph lead to likelihoods that factorize into likelihoods of the variables in each chain component given the parents determined by the conditional densities $f_{\omega|\mathrm{pa}(\omega)}\in M(\mathscr{G}_{\tilde{\omega}\cup\overline{\mathrm{pa}(\omega)}})^{\overline{\mathrm{pa}(\omega)}}$. In general each of these likelihood functions are exponential family likelihoods and standard theory (Barndorff-Nielsen, 1978) therefore applies but currently no suitable software is available for maximizing the likelihood functions in general. In special cases we can, however, exploit proposition 5 of Frydenberg & Lauritzen (1989) to deduce the following. Let $\tilde{f}_{\{\omega\cup\overline{\mathrm{pa}(\omega)}\}}$ denote the maximum likelihood estimate of the density in the model $M(\mathscr{G}_{\tilde{\omega}\cup\overline{\mathrm{pa}(\omega)}})$ for the marginal data $x_{\mathrm{cl}(\omega)}$. Then we have

Proposition 12

Let $M(\mathscr{G})$ be a graphical chain model such that all chain components ω satisfy

$$\omega\subseteq\Gamma\vee\mathrm{pa}(\omega)\subseteq\Delta.$$

Then the maximum likelihood estimate of the joint density f exists if and only if $\tilde{f}_{\{\omega\cup\overline{\mathrm{pa}(\omega)}\}}$ exist for all ω in which case we have

$$\tilde{f}=\prod_{\omega\in\Omega}\tilde{f}_{\{\omega\cup\overline{\mathrm{pa}(\omega)}\}}/\tilde{f}_{\overline{\{\mathrm{pa}(\omega)\}}}.$$

Proof. If the condition is satisfied, $\{\emptyset,\omega,\mathrm{pa}(\omega)\}$ is a decomposition of $\mathscr{G}_{\tilde{\omega}\cup\overline{\mathrm{pa}(\omega)}}$. From proposition 5 of Frydenberg and Lauritzen (1989) we then have that the maximum likelihood estimate of $f_{\omega|\mathrm{pa}(\omega)}$ in the conditional model $M(\mathscr{G}_{\tilde{\omega}\cup\overline{\mathrm{pa}(\omega)}})^{\overline{\mathrm{pa}(\omega)}}$ can be obtained by conditioning in the estimate of $f_{\omega\cup\overline{\mathrm{pa}(\omega)}}$ in the model $M(\mathscr{G}_{\tilde{\omega}\cup\overline{\mathrm{pa}(\omega)}})$ and that this is equal to the ratio between $\tilde{f}_{\{\omega\cup\overline{\mathrm{pa}(\omega)}\}}$ and $\tilde{f}_{\overline{\{\mathrm{pa}(\omega)\}}}$, which gives the result. □

In the pure case, all variables are of the same type and the condition is automatically satisfied and proposition 12 applies. The denominators correspond to saturated models and can be obtained explicitly, whereas iteration might be needed to obtain the denumerators if the corresponding part of the model does not have a decomposable graph. Again, MIM can be applied for this purpose. The general formula for the estimate of the inverse covariance matrix in the pure continuous case becomes:

$$\hat{K}=\sum_{\omega\in\Omega}(\tilde{K}_{\mathrm{cl}(\omega)})^0-n\{(ssd_{\mathrm{pa}(\omega)})^{-1}\}^0, \tag{7}$$

where $\tilde{K}_{\mathrm{cl}(\omega)}$ is the inverse covariance in $\tilde{f}_{\{\omega\cup\overline{\mathrm{pa}(\omega)}\}}$.

In the examples of this section we limit ourselves to discuss graphical chain models where all variables are continuous. This means that models are automatically homogeneous and that the distributional assumptions have a particular simple expression being multivariate Gaussian with an unknown mean. Thus these models are all models for the covariance structure of a multivariate normal distribution. In this particular case, all models can alternatively be defined through a particular type of linear structural equations, as shown by Wermuth (1988). We define a *recursive path analysis system* as a system of linear structural equations

$$\Lambda Y=U$$

where $Y=Y_V$ is the vector of variables, partitioned into groups $V(1)\cup\cdots\cup V(T)$ (corresponding to the dependence chain associated with the model), Λ is—when partitioned accordingly—an upper block-triangular matrix of coefficients with positive definite symmetric matrices in the blocks along the diagonal, U is a vector of residuals with its covariance matrix Φ being block-diagonal and with the blocks in its diagonal *equal to the corresponding blocks in* Λ. In this way, Λ and the covariance matrix of Σ of Y are in one-to-one correspondence.

Wermuth (1988) shows that the elements of Λ then can be interpreted as appropriate particular partial concentrations, whereby the conditional independence restrictions of a graphical chain model amount to specifying the elements in Λ to be zero for all corresponding missing edges in the chain graph.

We shall as examples consider the models shown in Fig. 17, but it should be noted that the interpretation of pictures similar to these given by linear structural equations such as in the LISREL models without latent variables (Jöreskog, 1973, 1977, 1981), is different from ours in general, if no special structure in the equations is assumed. This is, for example, reflected through the occurrence of identifiability problems in connection with LISREL models.

Example 9. The model in 1 of Fig. 17 is by proposition 9 equal to its undirected counterparts, i.e. covariance selection model, and has the conditional independence restriction $A\perp\!\!\!\perp b\,|\,\{a, B\}$. Thus the covariance matrix can be estimated explicitly by using equation (6).

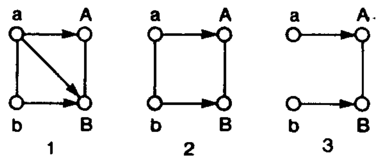

Fig. 17. Three chain models for continuous variables. The first model has an explicit solution whereas the two latter models need iteration to maximize the joint likelihood function.

The representation of the model as a recursive path analysis system is given by

$$\begin{pmatrix} \lambda_{AA} & \lambda_{AB} & \lambda_{Aa} & 0 \\ \lambda_{BA} & \lambda_{BB} & \lambda_{Ba} & \lambda_{Bb} \\ 0 & 0 & \lambda_{aa} & \lambda_{ab} \\ 0 & 0 & \lambda_{ba} & \lambda_{bb} \end{pmatrix} \begin{pmatrix} Y_A \\ Y_B \\ Y_a \\ Y_b \end{pmatrix} = \begin{pmatrix} U_A \\ U_B \\ U_a \\ U_b \end{pmatrix} \tag{8}$$

where $\lambda_{AB}=\lambda_{BA}$, $\lambda_{ab}=\lambda_{ba}$ and (U_A, U_B) and (U_a, U_b) are independent with covariance matrices given by the blocks along the diagonal of Λ.

□

Example 10. Model 2 of Fig. 17 has a representation in structural equations as

$$\begin{pmatrix} \lambda_{AA} & \lambda_{AB} & \lambda_{Aa} & 0 \\ \lambda_{BA} & \lambda_{BB} & 0 & \lambda_{Bb} \\ 0 & 0 & \lambda_{aa} & \lambda_{ab} \\ 0 & 0 & \lambda_{ba} & \lambda_{bb} \end{pmatrix} \begin{pmatrix} Y_A \\ Y_B \\ Y_a \\ Y_b \end{pmatrix} = \begin{pmatrix} U_A \\ U_B \\ U_a \\ U_b \end{pmatrix},$$

i.e. equation (8) modified by also assuming $\lambda_{Ba}=0$. The model is by proposition 9 equivalent to the corresponding (non-decomposable) model where no arrows are present. Thus iteration is needed to calculate maximum likelihood estimates, but the program MIM can be used.

This particular model has been used in an analysis by Wermuth (1989) to describe data collected by C. D. Spielberger on the personality characteristics of students. The variables A and B were state anxiety and anger, whereas a and b were trait anxiety and anger, all variables measured on a quantitative scale through psychological tests.

Example 11. Finally we consider the third model in Fig. 17. As structural equations it is given as

$$\begin{pmatrix} \lambda_{AA} & \lambda_{AB} & \lambda_{Aa} & 0 \\ \lambda_{BA} & \lambda_{BB} & 0 & \lambda_{Bb} \\ 0 & 0 & \lambda_{aa} & 0 \\ 0 & 0 & 0 & \lambda_{bb} \end{pmatrix} \begin{pmatrix} Y_A \\ Y_B \\ Y_a \\ Y_b \end{pmatrix} = \begin{pmatrix} U_A \\ U_B \\ U_a \\ U_b \end{pmatrix}.$$

Thus here the coefficient $\lambda_{ab}(=\lambda_{ba})$ has also been set to zero, reflecting the *marginal* independence of a and b. This model is *not* equivalent to its undirected counterpart. Iteration, combined with the formula (7), is needed to calculate the estimates. The joint concentration

matrix is estimated by

$$\hat{K}=\tilde{K}-n\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & ssd^{aa}_{\{ab\}} & ssd^{ab}_{\{ab\}} \\ 0 & 0 & ssd^{ba}_{\{ab\}} & ssd^{bb}_{\{ab\}} \end{pmatrix}+\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & n/ssd_{aa} & 0 \\ 0 & 0 & 0 & n/ssd_{bb} \end{pmatrix},$$

where $\tilde{K}$ is the estimate of K in example 10. But note that the likelihood ratio for comparing model 3 to model 2 reduces to that of marginal independence between a and b and leads to a standard test.

□

9. Problems for future research

This paper will be concluded by pointing out some points that, in my opinion, are particularly needing further development.

- Reliable and flexible software needs to be developed that handles models where response structures can be specified in a flexible manner and such that the computer can assist the user in interpretating the models.
- Algorithms for estimating in general chain models need to be developed to be used as elements in such software. Here theoretical research is needed.
- Modern methods and techniques for model diagnostics in the spirit of Atkinson (1985), Pregibon (1981), etc., shall be incorporated in such software with user-friendly interactive graphical interface.
- Practical experience with the application of models needs to be achieved. This has so far only been possible to a limited extent because of lack of suitable software.
- Theoretical work concerning the approximate distribution of estimates and test statistics have to be developed, based partially on asymptotic theory for maximum likelihood in exponential families. Structures of decomposability and partitioning of tests have to be maximally exploited to obtain as explicit results as possible.

Acknowledgements

The present paper was essentially completed while the author was visiting the Statistical Laboratory, University of Cambridge. The author is indebted to the Laboratory for hospitality and to the Danish Natural Science Research Council for financial support.

References

Andersen, A. H. (1974). Multidimensional contingency tables. *Scand. J. Statist.* **1**, 115–127.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*, 2nd edn. Wiley, New York.

Asmussen, S. & Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70**, 567–578.

Atkinson, A. C. (1985). *Plots, transformations and regression*. Clarendon Press, Oxford.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. Wiley, New York.

Beeri, C., Fagin, R., Maier, D., Mendelzon, A., Ullman, J. & Yannakakis, M. (1981). Properties of acyclic database schemes. In *Proc. 13th Annual ACM Symp. on the Theory of Computing*, Milwaukee. Assoc. Comput. Mach., New York.

Beeri, C., Fagin, R., Maier, D. & Yannakakis, M. (1983). On the desirability of acyclic database schemes. *J. Assoc. Comput. Mach.* **30**, 479–513.

Berge, C. (1973). *Graphs and hypergraphs* (transl. from French by E. Minieka). North-Holland, Amsterdam.

Birch, M. W. (1963). Maximum-likelihood in three way contingency tables. *J. Roy. Statist. Soc. Ser. B* **25**, 220–233.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice.* MIT Press, Cambridge, Mass.

Blalock, H. M., Jr (ed.) (1971). *Causal models in the social sciences.* Aldine-Atheston, Chicago.

Darroch, J. N., Lauritzen, S. L. & Speed, T. P. (1980). Markov-fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–539.

Darroch, J. N. & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43**, 1470–1480.

Darroch, J. N. & Speed, T. P. (1983). Additive and multiplicative models and interactions. *Ann. Statist.* **11**, 724–738.

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

Dawid, A. P. (1980). Conditional independence for statistical operations. *Ann. Statist.* **8**, 598–617.

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

Diestel, R. (1987). Simplicial decompositions of graphs—some uniqueness results. *J. Combin. Theory B* **42**, 133–145.

Dirac, G. A. (1961). On rigid circuit graphs. *Abh. Math. Sem. Univ. Hamburg* **25**, 71–76.

Edwards, D. (1987). A guide to MIM. Res. Rep. 1. Statistical Research Unit, Copenhagen.

Edwards, D. (1989a). Graphical modelling in multivariate analysis. *Proc. 1st Int. Conf. on Statistical Computing,* Izmir, Turkey (to appear).

Edwards, D. (1989b). Hierarchical interaction models (with discussion). *J. Roy. Statist. Soc. Ser. B* **51** (to appear).

Edwards, D. & Havránek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72**, 339–351.

Edwards, D. & Havránek, T. (1987). A fast model selection procedure for large families of models. *J. Amer. Statist. Assoc.* **82**, 205–211.

Edwards, D. & Kreiner, S. (1983). The analysis of contingency tables by graphical models. *Biometrika* **70**, 553–562.

Frydenberg, M. (1986). Blandede interaktionsmodeller, kausale modeller, kollapsibilitet og estimation [in Danish]. Thesis, University of Aarhus, Statistiske Interna No. 42.

Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *Am. Statist.* **18**, (to appear).

Frydenberg, M. (1989). The chain graph Markov property. Res. Rep. No. 186, Dept. Theory Stat., University of Aarhus.

Frydenberg, M. & Edwards, D. (1989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. *Comput. Statist. Data Anal.* (to appear).

Frydenberg, M. & Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed interaction models. *Biometrika* **76** (to appear).

Gavril, T. (1972). Algorithms for minimum coloring, maximum clique, minimum coloring by cliques and maximum independent set of a graph. *SIAM J. Comput.* **1**, 180–187.

Geiger, D. & Pearl, J. (1988). On the logic of influence diagrams. *Proc. 4th Workshop on Uncertainty in Artificial Intelligence,* Minneapolis, Minn., 136–147

Gibbs, W. (1902). *Elementary principles of statistical mechanics.* Yale University Press.

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

Golumbic, M. C. (1980). *Algorithmic graph theory and perfect graphs.* Academic Press, London.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interaction among multiple classifications. *J. Amer. Statist. Assoc.* **65**, 226–256.

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

Haberman, S. J. (1974). *The analysis of frequency data.* University of Chicago Press.

Holland, P. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81**, 945–960.

Jöreskog, K. G. (1973). Analysis of covariance structures. *Proc. 3rd. Symp. Mult. Anal.,* Dayton, Ohio, 1972 (ed. P. R. Krishnaiah), pp. 263–285. Academic Press, New York.

Jöreskog, K. G. (1977). Structural equation models in the social sciences: specification, estimation and testing. In P. R. Krishnaiah (ed.), *Applications of statistics,* 267–287. North-Holland, Amsterdam.

Jöreskog, K. G. (1981). Analysis of covariance structures. *Scand. J. Statist.* **8**, 65–92.

Kellerer, H. G. (1964a). Maßtheoretische Marginalprobleme. *Math. Ann.* **153**, 168–198.

Kellerer, H. G. (1964b). Verteilungsfunktionen mit gegebenen Marginalverteilungen. *Z. Wahrsch. Verw. Gebiete* **3**, 247–270.

Kiiveri, H. & Speed, T. P. (1982). Structural analysis of multivariate data: a review. In S. Leinhardt (ed.), *Sociological methodology.* Jossey-Bass, San Francisco.

Kiiveri, H., Speed, T. P. & Carlin, J. B. (1984). Recursive causal models. *J. Austral. Math. Soc. A* **36**, 30–52.

Lauritzen, S. L. (1982). *Lectures on contingency tables,* 2nd edn. Aalborg University Press.

Lauritzen, S. L. (1985). Test of hypotheses in decomposable mixed interaction models. Res. Rep. R-85-11, Inst. Elec. Sys., University of Aalborg.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H.-G. (1989). Independence properties of directed Markov fields. *Networks* (to appear).

Lauritzen, S. L., Speed, T. P. & Vijayan, K. (1984). Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. A* **36**, 12–29.

Lauritzen, S. L. & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50**, 157–224.

Lauritzen, S. L. & Wermuth, N. (1984). Mixed interaction models. Res. Rep. R-84-8. Inst. Elec. Sys., University of Aalborg.

Lauritzen, S. L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31–57.

Leimer, H.-G. (1985). Strongly decomposable graphs and hypergraphs. Thesis. Ber. z. Stochastik u. verw. Geb. 85-1, University of Mainz.

Leimer, H.-G. (1989a). Triangulated graphs with marked vertices. In *Graph theory in memory of G. A. Dirac* (eds L. D. Andersen *et al.*). *Ann. Discrete Math.* **41**, 311–324.

Leimer, H.-G. (1989b). Optimal decomposition by complete separators. *Discrete Math.* (to appear).

Olkin, I. & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat.* **32**, 448–465.

Parter, S. (1961). The user of linear graphs in Gauss elimination. *SIAM Rev.* **3**, 119–130.

Pearl, J. (1986). A constraint propagation approach to probabilistic reasoning. In L. M. Kanal & J. Lemmer (eds), *Uncertainty in artificial intelligence,* 357–370. North-Holland, Amsterdam.

Pearl, J. (1988). *Probabilistic inference in intelligent systems.* Morgan Kaufmann, San Mateo.

Pearl, J. & Paz, A. (1986). Graphoids: A graph-based logic for reasoning about relevancy relations. *Proc. European Conf. on Artificial Intelligence,* Brighton, United Kingdom.

Pearl, J. & Verma, T. (1987). The logic of representing dependencies by directed graphs. *Proc. Amer. Assoc. Art. Intell. Conf.,* Seattle, Washington, 1987.

Porteous, B. T. (1985a). Properties of log-linear and covariance selection models. Ph.D. thesis, University of Cambridge.

Porteous, B. T. (1985b). Improved likelihood ratio statistics for covariance selection models. *Biometrika* **72**, 473–475.

Porteous, B. T. (1989). Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic χ^2-distribution. *Ann. Statist.* **17** (to appear).

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705–724.

Rose, D. J. (1970). Triangulated graphs and the elimination process. *J. Math. Anal. Appl.* **32**, 597–609.

Smith, J. Q. (1989). Influence diagrams for statistical modelling. *Ann. Statist.* **17**, 654–672.

Speed, T. P. (1979). A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhya A* **41**, 184–197.

Speed, T. P. & Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138–150.

Sundberg, R. (1975). Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scand. J. Statist.* **2**, 71–79.

Verma, T. (1988). Causal networks: semantics and expressiveness. *Proc. 4th Workshop on Uncertainty in Artificial Intelligence,* Minneapolis, Minn., 352–359.

Vorob'ev, N. N. (1962). Consistent families of measures and their extensions. *Theory Probab. Appl.* **7**, 147–163.

Vorob'ev, N. N. (1963). Markov measures and Markov extensions. *Theory Probab. Appl.* **8**, 420–429.

Vorob'ev, N. N. (1967). Coalition games. *Theory Probab. Appl.* **12**, 250–266.

Wagner, K. (1937). Über eine Eigenschaft der ebenen Komplexe. *Math. Ann.* **114**, 570–590.

Wermuth, N. (1976a). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108.

Wermuth, N. (1976b). Model search among multiplicative models. *Biometrics* **32**, 253–263.

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

Wermuth, N. (1988). Block-recursive linear regression equations. Ber. z. Stoch. u. verw. Geb., University of Mainz.

Wermuth, N. (1989). Introduction to the use of graphical chain models. Ber. z. Stoch. u. verw. Geb., University of Mainz.

Wermuth, N. & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–552.

Wermuth, N. & Lauritzen, S. L. (1989). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. Ser. B* **51** (to appear).

Whittaker, J. (1982). GLIM syntax and simultaneous tests for graphical log-linear models. In R. Gilchrist (ed.), GLIM 82. *Lecture Notes in Statistics* **14**, 98–108. Springer Verlag, Berlin.

Whittaker, J. (1984). Fitting all possible decomposable and graphical models to multiway contingency tables. In T. Havránek (ed.), COMPSTAT 84, 98–108. Physica Verlag, Vienna.

Williams, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika* **63**, 33–37.

Wold, H. D. A. (1954). Causality and econometrics. *Econometrica* **22**, 162–177.

Wold, H. D. A. (1960). A generalization of causal chain models. *Econometrica* **28**, 443–463.

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

Wright, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics* **8**, 239–255.

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

Received June 1988, in final form May 1989

Steffen L. Lauritzen, Institute for Electronic Systems, University of Aalborg, Strandvejen 19, DK-9000 Aalborg, Denmark

DISCUSSION OF S. L. LAURITZEN'S PAPER

ANDERS HOLST ANDERSEN
University of Aarhus

I will start these comments by expressing my thanks to Steffen Lauritzen for giving this review of the theory of mixed graphical association models. These are important models for describing the association between a set of discrete and continuous variables because of their nice mathematical and statistical properties, but I think other models should be included in the analysis of association between variables.

In the analysis of a contingency table, not only graphical models but also hierarchical models should be considered. First, among the hierarchical models fitting a set of data there is in general a parsimonious one compared with the graphical models fitting the data. Secondly, the hierarchical models are used to increase the power of the test of no association, especially in case of only low order interactions present. As an example, consider a three-dimensional contingency table with variables A, B and C. The test of no association between A and B given C is performed in two steps. First, one tests the hypothesis of no ABC interaction and secondly, under the model of no ABC interaction, one tests that no AB interaction is present. The model with no ABC interaction is not graphical but hierarchical.

For the mixed case Edwards (1989b) defined the hierarchical mixed interaction models, which are natural extensions of the hierarchical models in the pure discrete case. In comparison with the mixed association models the class of possible restrictions on the discrete, the mixed linear and the mixed quadratic interactions is enlarged. For instance, in these models it is possible to specify how the covariances of the continuous variables depend on the discrete variables. An example of a mixed hierarchical interaction model, which is not graphical, is a two-way analysis of variance model with a random number of observations in each cell, specifying the full model for the two discrete variables and an additive model for the mean value of the continuous variable given the discrete variables. The advantages of having the hierarchical models in the discrete case are also present in the mixed case. The program MIM (Edwards, 1987) can be used to estimate the parameters in the mixed hierarchical interaction models.

DAVID EDWARDS
Novo Industry A/S, Copenhagen

I would like to compliment Professor Lauritzen on a concise and thorough survey of most of what is presently known about graphical models for mixed discrete and continuous variables. It provides a very valuable summary of the current state of the subject.

The graphical interaction models described in section 4.2 can (Edwards, 1989) easily be imbedded in a more general class, analogous to and including hierarchical log-linear models for discrete variables. A simple example of a hierarchical, non-graphical model is that of "no interaction" in a two-way layout. I would suggest that if emphasis were given to usefulness, rather than mathematical nicety, then the hierarchical models would receive more attention than they are given in the present account.

In a way the term *graphical model* is misleading: the conditional independence structure that holds under a hierarchical model can be represented by a conditional independence graph, just as with graphical models: the hierarchical models are in this sense just as graphical as the graphical ones. One loses the unique correspondence between the model and the graph, but not the ability of the graph to summarize fundamental properties of the model. It would be unfortunate if the expression *graphical modelling* comes to be understood as *applying graphical models* rather than *modelling using conditional independence graphs as a central tool.*

One loose end as I see it concerns incomplete tables (incomplete factorial designs). The theoretical development seems to require complete tables with positive cell probabilities but surely some things carry through when this condition is relaxed. Perhaps Professor Lauritzen could comment on this.

KARL G. JÖRESKOG
University of Uppsala

I congratulate Dr Lauritzen on his very impressive, lucid, and comprehensive paper. As a theory for analysis of multivariate data, his approach in several ways resembles the LISREL methodology which is widely used in the social and behavioural sciences (see Jöreskog &

Sörbom, 1988, and bibliography therein). It would therefore seem appropriate to focus my discussion on the similarities and differences between these two approaches. There are differences in scope and limitations, in the way graphs are used and interpreted, and in distributional assumptions for estimation and testing.

Scope and limitations

LISREL is primarily designed for models with latent (unobserved or unobservable) variables but can also handle the case when all variables are directly observed. Lauritzen does not deal with latent variables.

LISREL makes a distinction between jointly independent (exogenous) and jointly dependent (endogenous) variables. Lauritzen appears to treat all variables symmetrically.

In LISREL one can have models with feedback loops or *independent systems*, often used in econometrics. Such models are impossible in the Lauritzen approach because they cannot be specified in terms of conditional independence statements.

LISREL was originally designed for continuous variables having approximately normal distributions but has recently been extended so that also ordinal variables and mixtures of ordinal and continuous variables can be used (see Jöreskog & Sörbom, 1986). LISREL deals with ordinal variables by quantifying them in some way. For example, it may be assumed that there is a standard normal variable underlying each ordinal variable. One can then estimate a *polychoric* correlation for two ordinal variables and a *polyserial* correlation for an ordinal and a continuous variable. The main focus of Lauritzen's paper is on the analysis of models involving both discrete (nominal) and continuous variables. For such models, his approach is very general in that higher-order interactions may be tested. LISREL can only handle first- and second-order interactions. LISREL handles nominal variables by assuming independent groups of observations for each category combination which is not feasible if there are many categories. On the other hand, Lauritzen does not mention ordinal variables. Can his approach be extended to handle ordinal variables?

Graphs

In Lauritzen's paper, graphs and graph theory play a central role in presenting models. These graphs are supposed to represent statements about conditional independence. At least in the case when all variables are continuous, I fail to see why this is a useful approach in practice. By contrast, in the LISREL methodology, graphs—so-called *path diagrams*—are used to represent actual postulated relationships. It seems to me that most researchers are interested in the estimated relationship itself and the strength of the relationship, rather than whether or not certain variables are conditionally independent given certain other variables.

Rules for path diagrams have been given by Jöreskog & Sörbom (1988). A typical LISREL path diagram is shown in Fig. 1. This represents three linear relationships, one for each y-variable:

$$y_1=\gamma_{11}x_1+\gamma_{12}x_2+\gamma_{13}x_3+\zeta_1$$

$$y_2=\gamma_{23}x_3+\beta_{21}y_1+\zeta_2$$

$$y_3=\gamma_{32}x_2+\beta_{31}y_1+\beta_{32}y_2+\zeta_3.$$

Here the lack of arrows from x_1 and x_2 to y_2 indicates that these variables are not included in the relationship for y_2. Note also that the random disturbance terms ζ_1, ζ_2, and ζ_3 are included in

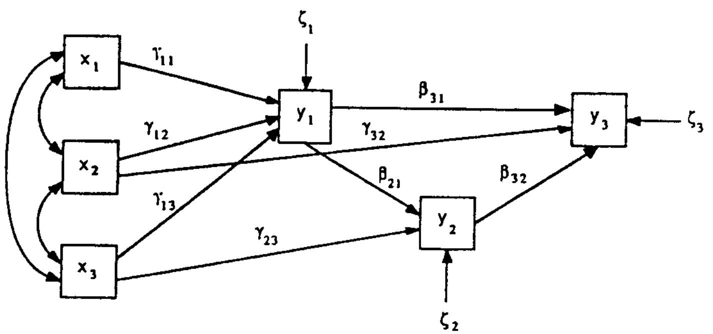

Fig. 1

the path diagram and the lack of connections between them indicates that they are mutually uncorrelated (independent if normality is assumed).

Using Lauritzen's terminology, the model expresses the two conditional independence statements

(1) y_2 is conditionally independent of x_1 and x_2 for given x_3 and y_1.
(2) y_3 is conditionally independent of x_1 and x_3 for given x_2, y_1 and y_2.

A Lauritzen graph for this model will look quite similar to Fig. 1 but the disturbance terms will not be included and the two-way arrows will be just lines.

There are models in which the two kinds of graphs are quite different. For example, Fig. 2, which expresses that 2 and 3 are conditionally independent for given 1 and 4, would probably be drawn as a path diagram as in Fig. 3. It is seen that the two graphs in Figs 2 and 3 are quite different although they represent the same conditional independence statement.

There are also graphs which do not correspond to any LISREL model. Graph 2 in Fig. 17 is such a graph. The two conditional independence statements implied by this graph correspond to two conflicting LISREL models.

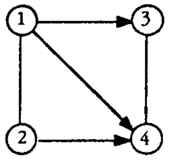

Fig. 2

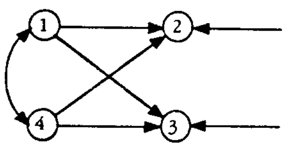

Fig. 3

Distributional assumptions

Lauritzen makes the assumption that the distribution of all continuous variables is multivariate normal for each category combination of the nominal variables. However, for the model of Fig. 1, LISREL does not make any distributional assumptions about the x-variables. It is sufficient to assume normality of y for given x. In LISREL 7, it is also possible to deal with various non-normal y-variables using asymptotically distribution-free methods developed by Browne (1984). How does Lauritzen propose to deal with variables such as age, income, and education, commonly used in the social sciences? These variables are neither nominal nor normally distributed.

References

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *Br. J. Math. Statist. Psychol.* **37**, 62–83.

Jöreskog, K. G. & Sörbom, D. (1986). *PRELIS: A program for multivariate data screening and data summarization. A preprocessor for LISREL.* Scientific Software, Inc., Mooresville, Indiana.

Jöreskog, K. G. & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications.* SPSS Publications, Chicago.

SØREN JOHANSEN
University of Copenhagen

In econometrics one considers dynamic models, where arrows could be used to suggest a dependence stretching one or more time periods ahead. Can the formulation presented in your paper be extended to dynamic modelling?

It would be important to formulate concepts like weak and strong exogeneity, endogenous and exogenous variables, aggregation, recursive systems, identification of structural equations, etc., in terms of graphs and arrows (see for instance Hendry & Richard, 1983).

This would be interesting not only because it would extend the usefulness of the graph formulation but also point at some new models that could be of interest to econometricians, in particular the models that allow interaction between discrete as well as Gaussian variables.

References

Hendry, D. F. & Richard, J.-F. (1983). The econometric analysis of economic time series. *Internat. Statist. Rev.* **51**, 111–163.

REPLY TO THE DISCUSSION

Let me first thank the discussants for their comments which both gives the reader a chance to see some parts of the paper in a different light and also gives me the opportunity to expound a few important points.

All discussants point directly or indirectly to the fact that the model class, such as I have here described it, is limited and not flexible enough. I guess I have to agree on that although it certainly would have made exposition more difficult to cover more ground than what I already did.

As emphasized by A. H. Andersen and D. Edwards, it is desirable to include in the model class the *hierarchical* interaction models such as those introduced by Edwards (1989b). This gives both greater flexibility and parsimony and higher power for testing model fit. Further, it includes into the class of models some classical models that are well understood and that would otherwise be excluded. It should, however, not be forgotten that hierarchical models in general are more sophisticated and more difficult to interpret. In particular, the independence graph (representing the smallest graphical model containing the given hierarchical model) is a powerful tool in revealing part of that interpretation. I still think that in a preliminary, exploratory phase of data analysis it would be appropriate to concentrate on graphical models and then when analysing more closely, hierarchical models will play their natural part. In particular the class of *hierarchical chain models* needs to be properly described and its statistical theory investigated.

Continuing the theme above it is clear that, as K. Jöreskog points out, it would be natural to extend the models to deal also with *ordinal* variables which are so typical in many statistical problems in the social sciences. This extension has recently been made for the LISREL models. Here I can only say that the graphical models described in the present paper are in their infancy and far from being fully developed. A similar comment applies to wishes to deal with other continuous distributions than the Gaussian and latent variables. However, the latent variable problem is different since this is not a question of "missing models" but rather a question of treating the statistical problems with missing data. I am sure that this will be treated in the near future. I must say that I believe the graphical models to have a strong potential, partly because of the clarity in their interpretation and partly because they rest on a mathematical firm foundation.

It is an interesting challenge to apply the ideas of graphical modelling to the theory and analysis of dynamic economic time series such as suggested by S. Johansen. Such a development would certainly be worthy of its own full research paper, but let me try to indicate some answers through a simple example.

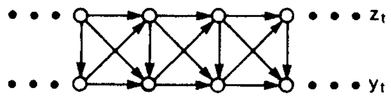

Fig. D1. Graph of a dynamic model of order one with one series of exogenous variables z_t and one endogenous y_t.

Fig. D1 is supposed to represent a simple dynamic model of order one, where one would normally think of the upper z-variables as exogenous and the y-variables as endogenous and time moves from left to right. Typically a system like that would (omitting constant effects) be represented by a system of structural equations as

$$y_t=\alpha y_{t-1}+\beta z_t+\gamma z_{t-1}+\varepsilon_t^y$$

$$z_t=\phi y_{t-1}+\psi z_{t-1}+\varepsilon_t^z.$$

Here it is important to emphasize that the equations are not uniquely determined from the distributional properties of the time series, and therefore not from the graph either. The property of weak exogeneity is a property of the relation between the model and the equation and not of the model itself. It is, in this particular instance, equivalent to the independence of the errors and can be formulated as something like "the parameters can meaningfully be attached to the arrows in the picture"; see Wermuth (1988) for a discussion of the relation between recursive systems, graphical models and structural equations.

Strong exogeneity is a combination of weak exogeneity and Granger non-causality. The latter has a simple interpretation as all edges from bottom to top missing, such as in Fig. D2.

When this applies (and only then) one can substitute the arrows within z-variables and within y-variables with lines without changing the conditional independence restrictions such as to obtain Fig. D3. This means that the distinction between exogenous and endogenous variables becomes the same as the distinction between chain components in the chain models.

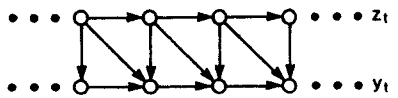

Fig. D2. Graph of a dynamic model of order one showing Granger non-causality.

The interpretation of the pictures have to be supplemented with the assumption of stationary dynamics, i.e. that the parameters in the conditional distribution of any variable given its "past" are unchanged in time. It is no problem to make more complex dependence structures and to include the possibility of both exogenous and endogenous discrete variables.

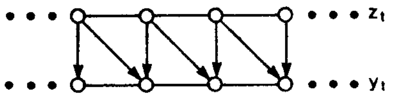

Fig. D3. Graph of a model equivalent to a dynamic model with Granger non-causality. Strong exogeneity shows up as the entire series z_t can be considered exogenous to y_t.

Finally, let me turn to a few specific points raised by K. Jöreskog. It is not true that I do not distinguish between exogenous and endogenous variables. The chain components correspond to several different levels of exogeneity and endogeneity rather than just two. Also I do not necessarily have to assume a particular distribution for the exogenous variables. One can analyse data conditionally on the exogenous variables, and this will indeed often be done. But the chain models permit the exogenous variables to be analysed separately, and this is sometimes of interest. On the relationship between the LISREL graphs and the graphs in the present paper: yes, they can be different but also quite different graphs can describe the same kind of data behaviour. Here it is important to distinguish between the graph describing something in the subject-matter context and describing a particular statistical model (see Wermuth & Lauritzen, 1989).

It is true that researchers are interested in relationships between variables, but whether or not these relationships are well represented by structural equations is another matter. Tradition and training certainly plays a strong role here and graphical modelling could provide an alternative.

Permit me in conclusion to express the hope that future research and experience with the theory and application of the models described in this and related papers will enhance a better understanding of the problems associated with structural multivariate analysis.