The Annals of Statistics
1980, Vol. 8, No. 3, 522–539

MARKOV FIELDS AND LOG-LINEAR INTERACTION MODELS FOR CONTINGENCY TABLES[1]

BY J. N. DARROCH, S. L. LAURITZEN AND T. P. SPEED

The Flinders University of South Australia, University of Copenhagen and University of Western Australia.

We use a close connection between the theory of Markov fields and that of log-linear interaction models for contingency tables to define and investigate a new class of models for such tables, graphical models. These models are hierarchical models that can be represented by a simple, undirected graph on as many vertices as the dimension of the corresponding table. Further all these models can be given an interpretation in terms of conditional independence and the interpretation can be read directly off the graph in the form of a Markov property. The class of graphical models contains that of decomposable models and we give a simple criterion for decomposability of a given graphical model. To some extent we discuss estimation problems and give suggestions for further work.

0. Introduction and summary. In the present paper we shall utilize some close connections between the theory of Markov fields and that of log-linear interaction models to define a new class of models for multidimensional contingency tables: *graphical models*. The graphical models have two important properties:

(i) they can be represented by an undirected, finite graph with as many vertices as the table has dimensions;

(ii) they can be interpreted in terms of conditional independence (in fact, a Markov property) and the interpretation can be read directly off the graph.

This class of models is a proper subclass of the so-called *hierarchical models*, but it strictly contains the *decomposable models* (Goodman (1970, 1971), Haberman (1970, 1974), Andersen (1974)). This implies that we can give a simple, visual representation of any decomposable model, thus making the interpretation easy.

We also characterise those graphs that correspond to decomposable models, thus giving an alternative to Goodman's algorithm for checking decomposability of a given hierarchical model: first, check whether it is graphical and then, if it is, check whether the graph is decomposable, i.e., whether there are any cyclic subgraphs of length $\geqslant 4$.

In Section 1 we introduce some notation and define the various classes of models for contingency tables. In Section 2 we review some basic elements of the theory of Markov fields and Gibbs states. In Section 3 we draw together the results in these

Received November 1978; revised March 1979.

[1]This research was supported in part by the Danish Natural Science Research Council.

AMS 1970 *subject classifications.* Primary 62F99; secondary 60K35.

Key words and phrases. Contingency tables, decomposability, Gibbs states, graphical models, triangulated graphs.

two sections, define the graphical models and discuss their interpretation. Section 4 contains the arguments needed to realise that all decomposable models are graphical and we also give the characterisation of decomposable graphs. Section 5 is devoted to maximum likelihood estimation in decomposable models. Although this is completely solved by Haberman (1974) we define an index directly interpretable from the graph and show how these indices are the powers of the marginal counts in the estimation formula. A combinatorial property of this index can also be used as a characterisation of decomposable graphs. Section 6 contains a list of all graphical models of dimension less than or equal to five together with their interpretation and these are divided into decomposables and nondecomposables. This is meant to both illustrate our theory and be an analogue of the tables in Goodman (1974) with all hierarchical models of dimension less than or equal to four together with an interpretation of the decomposables among them. Finally we give some suggestions regarding the use of the models and some directions for possible further work.

The present paper is almost without proofs. Most of our results are just "translations" of results from other areas. It is somewhat technical to establish the connection between graphical models and decomposable models. In fact, in our opinion these results are of a purely graph theoretic nature and the proofs and necessary formalism to derive the results can be found in Lauritzen, Speed and Vijayan (1978).

1. Preliminaries. We shall discuss log-linear interaction models for contingency tables. Since we want to use the analogies between the theory of Markov fields and that of such models, it will be convenient to introduce a notation that makes such analogies more apparent.

We shall consider a finite set C of *classification criteria* or *factors*. For each $\gamma \in C$ we let I_γ be the set of *levels* of the criterion or factor γ. The set of *cells* in our table is the set $I = \Pi_{\gamma \in C} I_\gamma$ and a particular cell will be denoted $\mathbf{i} = (i_\gamma, \gamma \in C)$. A set of n objects is classified according to the criteria and we let the *counts* $n(\mathbf{i})$ be the number of objects in cell $\mathbf{i}$.

For $a \subseteq C$, we consider the *marginal counts* $n(\mathbf{i}_a)$. $n(\mathbf{i}_a)$ is the number of objects in the marginal cell $\mathbf{i}_a = (i_\gamma, \gamma \in a)$ and is obtained as the sum of the $n(\mathbf{i})$ for all such $\mathbf{i}$ that agree with $\mathbf{i}_a$ on the coordinates corresponding to a. In other words, $n(\mathbf{i}_a)$ are the counts in the *marginal table*, where objects only are classified according to the criteria in a. Similarly we let $P(\mathbf{i})[P(\mathbf{i}_a)]$ denote the probability that any given object belongs to the [marginal] cell $\mathbf{i}[\mathbf{i}_a]$.

We consider the classifications of the n objects as n independent observations of the distribution P such that the distribution of the counts becomes a multinomial distribution:

$$P\{N(\mathbf{i}) = n(\mathbf{i}), \mathbf{i} \in I\} = \binom{n}{n(\mathbf{i}), \mathbf{i} \in I} \Pi_{\mathbf{i} \in I} P(\mathbf{i})^{n(\mathbf{i})}.$$

The general log-linear interaction model involves specification of the above unknown distribution P as follows: firstly we expand the logarithm of P as

$$\log P(\mathbf{i}) = \Sigma_{a \subseteq C} \xi_a(\mathbf{i}_a),$$

where ξ_a are functions of $\mathbf{i}$ that only depend on $\mathbf{i}$ via the coordinates in a, i.e., through $\mathbf{i}_a$. If $a = \varnothing$, $\xi_\varnothing$ is the constant vector.

Such an expansion can be made for any P with $P(\mathbf{i}) > 0$ for all $\mathbf{i} \in I$. If we are interested in having a one-to-one correspondence between the system of functions $\{\xi_a, a \subseteq C\}$ and P, we have to introduce standardising constraints as, e.g.,

$$\forall b \subset a : \Sigma_{\{\mathbf{i}'_a : \mathbf{i}'_b = \mathbf{i}_b\}} \xi_a(\mathbf{i}'_a) \equiv 0 \quad \text{for all } \mathbf{i}_b,$$

i.e., that summation over any factor gives a zero. This is all well known and standard although the notation is slightly unusual.

The functions ξ_a are called the *interactions* among the factors in a. If $|a| = 1$ we call ξ_a the *main effect*, if $|a| = 2$ a *first-order interaction* and, in general, if $|a| = m$, ξ_a is an interaction of order $m - 1$. A general log-linear interaction model involves specifying certain of these interactions to vanish and letting the remaining interactions be arbitrary and unknown. It is usually convenient to work with a smaller class of models, the *hierarchical models*.

A hierarchical model is an interaction model where the specifications of vanishing interactions satisfy the following property: *if ξ_a is specified to vanish and $b \supseteq a$ then ξ_b is specified to vanish*. In other words, if there is no interaction among factors in a then there is no interaction of higher order involving all the factors in a.

As is easily seen and well known, a hierarchical model can be specified via a so-called *generating class* being a set $\mathcal{C}$ of pair-wise incomparable (w.r.t. inclusion) subsets of C to be interpreted as the maximal sets of permissible interactions, i.e.,

$$\xi_a \equiv 0 \text{ iff } \textit{there is no } c \in \mathcal{C} \text{ with } a \subseteq c.$$

A probability P belonging to a hierarchical model with generating class $\mathcal{C}$ is uniquely determined by the marginal probabilities given by the elements of $\mathcal{C}$. The maximum likelihood estimate of P is obtained by equating these marginal probabilities to the marginal sample proportions.

A certain subclass of hierarchical models is of special interest: the *decomposable models*, introduced by Goodman (1970, 1971) and later defined formally by Haberman (1970, 1974). Following Haberman, a generating class is *decomposable* if either it has only one element or if it can be partitioned into generating classes $\mathcal{A}$ and $\mathcal{B}$ with $\mathcal{A} \cap \mathcal{B} = \varnothing$, $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$ and such that

$$(\cup_{a \in \mathcal{A}} a) \cap (\cup_{b \in \mathcal{B}} b) = a^* \cap b^*$$

for some $a^* \in \mathcal{A}$, $b^* \in \mathcal{B}$. A slightly different definition was given by Lauritzen, Speed and Vijayan (1978) (henceforth referred to as LSV) but it is shown in the same paper that the definitions are equivalent.

As shown by Haberman (1970) these models have two fundamental properties

(i) the problem of maximum likelihood estimation has an explicit solution;

(ii) the models can be interpreted in terms of conditional independence, independence and equiprobability.

The basic idea in our work is that such an interpretation is most directly formulated as a Markov property. Goodman (1970), in fact, uses the terminology "models of Markov type" for decomposable models.

This leads us to consider Markov fields on finite graphs and from these considerations it turns out that it is natural to define a class of models, *graphical models* whose interpretation most elegantly is given as a Markov property of a certain random field associated with the model.

2. Markov fields and Gibbs states. In the theory of Markov fields, see, e.g., Kemeny, Snell and Knapp (1976), we operate with a set Γ of *sites* and here we assume Γ to be finite. Γ will correspond to the set of factors C. At each site $\gamma \in \Gamma$ there is a finite set I_γ of *elementary states*. The set $I = \prod_{\gamma\in\Gamma} I_\gamma$ is the set of *configurations*. A given configuration is denoted by $\mathbf{i} = (\mathbf{i}_\gamma, \gamma \in \Gamma)$. Further there is an undirected *graph* $\mathbf{\Gamma}$ on Γ, i.e., a pair $\mathbf{\Gamma} = (V(\mathbf{\Gamma}), E(\mathbf{\Gamma}))$ consisting of the *vertex set* $V(\mathbf{\Gamma}) = \Gamma$ and *edge set* $E(\mathbf{\Gamma})$, where $E(\mathbf{\Gamma})$ is a set of unordered pairs of distinct elements of Γ. We say that α and β are *adjacent* or *neighbours* and write $\alpha \sim \beta$ iff $\{\alpha, \beta\} \in E(\mathbf{\Gamma})$.

If $a \subseteq \Gamma$, the *boundary* of a, ∂a, is the set of vertices in $\Gamma \setminus a$ that are adjacent to some vertex in a. The *closure* of a is $a \cup \partial a$ and is denoted by $\bar{a}$. When no confusion is possible we write $\partial\alpha$, $\bar{\alpha}$ instead of $\partial\{\alpha\}$, $\overline{\{\alpha\}}$. A *complete subset* is a subset $a \subseteq \Gamma$ where all elements are mutual neighbours. A *clique* is a maximal (w.r.t. inclusion) complete subset.

We now consider a probability P on I with $P(\mathbf{i}) > 0$ for all $\mathbf{i} \in I$ and the random variables defined by coordinate projections:

$$X_\gamma(\mathbf{i}) = i_\gamma, \qquad \gamma \in \Gamma$$

and

$$X_a(\mathbf{i}) = \mathbf{i}_a \quad \text{for } a \subseteq \Gamma, \qquad a \neq \varnothing.$$

The random field $(X_\gamma, \gamma \in \Gamma)$ is said to be *Markov* w.r.t. *P and* $\mathbf{\Gamma}$ (or *P is Markov* w.r.t. $\mathbf{\Gamma}$) if one of the following four equivalent properties hold:

(i) for all $\gamma \in \Gamma$, X_γ and $X_{\Gamma\setminus\bar{\gamma}}$ are conditionally independent given $X_{\partial\gamma}$;

(ii) for all $\alpha, \beta \in \Gamma$ with $\alpha \not\sim \beta$, X_α and X_β are conditionally independent given $X_{\Gamma\setminus\{\alpha,\beta\}}$;

(iii) for all $a \subseteq \Gamma$, X_a and $X_{\Gamma\setminus\bar{a}}$ are conditionally independent given $X_{\partial a}$;

(iv) if two disjoint subsets $a \subseteq \Gamma$ and $b \subseteq \Gamma$ separated by a subset $d \subseteq \Gamma$ in the sense that all paths from a to b in $\mathbf{\Gamma}$ go via d, then X_a and X_b are conditionally independent given X_d.

That these four conditions in fact are equivalent for a probability with $P(\mathbf{i}) > 0$ is more or less well known, see, e.g., Pitman (1976) or Kemeny, Snell and Knapp (1976). It can be proved with quite elementary methods.

A *potential* is a real-valued function Φ on I of the form

$$\Phi(\mathbf{i}) = \Sigma_{a \subseteq \Gamma} \xi_a(\mathbf{i}_a)$$

where the functions ξ_a depend on $\mathbf{i}$ through $\mathbf{i}_a$ only and are called the *interaction potentials*. In fact, any real-valued function is a potential, see the remarks in the previous section, so this notion first gets interesting when we make restrictions on the ξ_a - functions.

A probability P on I is called a *Gibbs state with potential* Φ if

$$P(\mathbf{i}) = e^{\Phi(\mathbf{i})}.$$

Similarly, any probability on I with $P(\mathbf{i}) > 0$ for all $\mathbf{i}$ is a Gibbs state (with potential $\Phi(\mathbf{i}) = \log P(\mathbf{i})$). Φ is called a *nearest-neighbour potential* if it is built up from interactions only among mutual neighbours, i.e., if $\xi_a \equiv 0$ if not all vertices in a are mutual neighbours, i.e., if a is not a *complete subset* of Γ. P is called a *nearest-neighbour Gibbs state* iff P is a Gibbs state with potential Φ, where Φ is a nearest-neighbour potential.

One of the most basic results about Markov fields and nearest-neighbour Gibbs states asserts that, in fact, the two notions are identical: *P is a nearest-neighbour Gibbs state if and only if the corresponding random field is Markov*. A proof of this result can be found many places. In the case $I_\gamma = I_0$ there is, e.g., a proof in Kemeny, Snell and Knapp (1976), and the method of proof there easily extends to the case with I_γ depending on γ, see, e.g., Pitman (1976) or Speed (1976).

This theorem is in fact the key to our results: it establishes a connection between certain linear restrictions on the logarithm of a probability (being n.-n.-Gibbs) and a Markov property (an interpretation in terms of conditional independence). What remains to be done is to introduce the graphs in the contingency table framework.

3. Graphical models. Let us return to the contingency table set-up. Assume that we have given a graph $\mathbf{C}$ on our set of factors C, specified by the vertex set $V(\mathbf{C}) = C$ and edge set $E(\mathbf{C})$. Let $\mathcal{C}$ be the *cliques* of $\mathbf{C}$, i.e., the maximal complete subsets. The *graphical model* given by $\mathbf{C}$ is the hierarchical model with generating class $\mathcal{C}$. Note that $\mathcal{C}$ also uniquely defines the graph $\mathbf{C}$ by $\alpha \sim \beta$ iff $\exists c \in \mathcal{C}$ such that $\{\alpha, \beta\} \subseteq c$. In that sense our graph $\mathbf{C}$ is just another representation of the generating class $\mathcal{C}$.

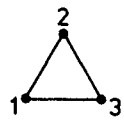
Let us examine the restrictions on our interactions given by this generating class. By the definition of a hierarchical model we have $\xi_a \equiv 0$ *unless a is contained in a maximal complete subset*, i.e., *unless a is a complete subset*. In other words, the set of probabilities P in our model is *exactly the set of nearest-neighbour Gibbs states corresponding to* $\mathbf{C}$.

Consequently, by the fundamental theorem in the previous section, we have that the probabilities P, contained in our model are exactly *those making* $(X_\gamma, \gamma \in \mathbf{C})$ *a Markov field.* It is now clear that our model is given by conditional independence constraints involved in the four equivalent formulations of the Markov property. It is thus clear that if two sets of factors are in different connected components of the graph, they are independent. If two factors are not neighbours, they are conditionally independent given the other factors. If two sets of factors a and b are separated by a set of factors d, they are conditionally independent given those in d, etc.

We should like to point out, that not all hierarchical models are of the graphical type. It is, however, still possible to associate a graph with any generating class. The graph defines the interaction structure in part.

Let $\mathcal{C}$ be a generating class and assume that $C = \cup_{c \in \mathcal{C}} c$ (this assumption is merely of technical nature). Define a graph $\mathbf{C} = (V(\mathbf{C}), E(\mathbf{C}))$ by letting $V(\mathbf{C}) = C$ and $\{\alpha, \beta\} \in E(\mathbf{C})$ if and only if $\{\alpha, \beta\} \subseteq c$ for some $c \in \mathcal{C}$. We could call this graph the *first-order interaction graph* for $\mathcal{C}$ since it has all main effects as vertices and first-order interactions as edges. It is clear, that $\mathcal{C}$ corresponds to a graphical model if and only if $\mathcal{C}$ exactly is the set of cliques of this graph. If this is the case, we shall say that $\mathcal{C}$ is a graphical generating class. If there are cliques in the graph that are not in $\mathcal{C}$, which very well can be the case, then $\mathcal{C}$ is not graphical and the interaction structure in the model is not adequately described by the graph alone. Note that these remarks imply that the interaction structure in a graphical model is *determined by the first-order interactions*, since these interactions define the graph, which, in turn, gives us its cliques and thus its interactions of higher order.

The simplest example of a hierarchical model which is not graphical is that with $C = \{1, 2, 3\}$ and $\mathcal{C} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$. Its first-order interaction graph is



i.e., the complete 3-graph. If $\mathcal{C}$ had been graphical, $\mathcal{C}$ should have been $\{\{1, 2, 3\}\}$ which is not the case. The model in question, that of vanishing second-order interaction in a three-way table, is also known as the simplest nondecomposable hierarchical model, and it is well known that it cannot be interpreted in terms of conditional independence.

In the next section we shall see that all decomposable models are graphical and characterise graphs corresponding to decomposable models.

4. Decomposable models and graphical models. Lauritzen, Speed and Vijayan (1978) (LSV) study properties of generating classes and their first-order interaction graphs, especially w.r.t. the notion of a decomposition. This is done in a purely graph-theoretic framework and they therefore use a slightly different terminology to be able to relate their results to other areas of mathematics.

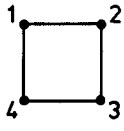
A generating class is, in LSV, called a *generating class hyper graph* (g.c. hypergraph). The first-order interaction graph of a generating class is called the *2-section of the* g.c. *hypergraph*.

Here we shall quote some of the results from LSV of importance to us. For proofs and details, the reader is referred to that paper using the "translation key" just given. Corollary 4 in LSV asserts that *any decomposable model is graphical*. This fact was noted by Andersen (1974) in a somewhat disguised form (his Theorem 5).

We are now led to the following considerations: decomposability is a property of a generating class, a property which is not too easy to get hold of and verify directly. We have just seen that any decomposable model is graphical, i.e., is very well represented by its first-order interaction graph. Then decomposability must be a property of such a graph. Theorem 2 of LSV asserts (among other things) that: *the cliques of a graph form a decomposable generating class if and only if the graph is triangulated* (i.e., contains no cycles of length $\geqslant 4$ without a chord). For the notion of a triangulated graph, see Berge (1973).

This result is definitely the main result of LSV and gives us a possibility of making an immediate visual check on the decomposability of a given graphical model, see our tables in Section 6.

Thus the smallest nondecomposable graphical generating class is given by the 4-cycle:



i.e., with $C = \{1, 2, 3, 4\}$, $\mathcal{C} = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{1, 4\}\}$. In fact, Andersen (1974) gives this example of a nondecomposable model that can be interpreted in terms of conditional independence (1 and 3 are c.i. given 2 and 4, 2 and 4 are c.i. given 1 and 3).

The Markov interpretation originally made by Goodman, Haberman etc. is along the following lines: a generating class $\mathcal{C} = \{a_1, \cdots, a_k\}$ is decomposable iff its elements can be ordered so that

$$a_t \cap (a_1 \cup \cdots \cup a_{t-1}) = a_t \cap a_{r_t},\ r_t \in \{1, \cdots, t-1\}, \qquad t = 2, \cdots, k. \tag{4.1}$$

It follows that

$$b_t = a_t \setminus (a_1 \cup \cdots \cup a_{t-1}) = a_t \setminus a_{r_t} \neq \emptyset.$$

It is easy to see that, if P is hierarchical with generating class $\mathcal{C}$, that is

$$P(\mathbf{i}) = \exp \Sigma_{t=1}^{k} \Sigma_{a \subseteq a_t} \xi_a(\mathbf{i}_a),$$

then the conditional probability

$$P(\mathbf{i}_{b_k} | \mathbf{i}_{a_1 \cup \cdots \cup a_{k-1}})$$

simplifies to $P(\mathbf{i}_{b_k}|\mathbf{i}_{c_k})$ where

$$c_t = a_t \setminus b_t = a_t \cap a_{r_t},$$

and that the marginal probability $P_{a_1 \cup \cdots \cup a_{k-1}}$ satisfies the hierarchical model with generating class $\mathcal{C} = \{a_k\}$. It follows by induction that

$$P(\mathbf{i}) = P(\mathbf{i}_{a_1}) \Pi_{t=2}^{k} P(\mathbf{i}_{b_t}|\mathbf{i}_{c_t})$$

and that the distribution of an $\mathbf{X}$ with probability P may be characterised by the sequence of Markov properties

$$\text{conditional distribution of } \mathbf{X}_{b_t} \text{ given } \mathbf{X}_{a_1 \cup \cdots \cup a_{t-1}}$$
$$= \text{conditional distribution of } \mathbf{X}_{b_t} \text{ given } \mathbf{X}_{c_t}, \qquad t = 2, \cdots, k.$$

Further, (2) may be rearranged as

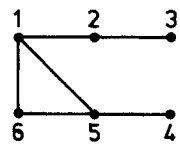
$$P(\mathbf{i}) = \frac{\Pi_{t=1}^{k} P(\mathbf{i}_{a_t})}{\Pi_{t=2}^{k} P(\mathbf{i}_{b_t})}$$

which is the explicit formula for P and includes as a special case the formula for the maximum likelihood estimate of P.

In order to arrive at this formula by the above method it is necessary to search for an ordering of the elements of $\mathcal{C}$ which satisfies (4.1). This search is helped by reference to the graph and also by the awareness that each element a_t must contain at least one element which is not in $a_1 \cup \cdots \cup a_{t-1}$. There are, generally, many orderings satisfying (4.1). Haberman proved that there are at least k by proving that any element of $\mathcal{C}$ may be chosen as initial member of some sequence. That there may be many more is illustrated by the example with $|\Gamma| = 6$ and

$$\mathcal{C} = \{\{1, 2\}, \{2, 3\}, \{4, 5\}, \{1, 5, 6\}\}$$

for which the graph is



It turns out that 14 of the $4! = 24$ possible orderings satisfy (4.1).

The description of the Markov property given by the graph seems more natural since it is immediate that the property does not involve an ordering of the elements of $\mathcal{C}$.

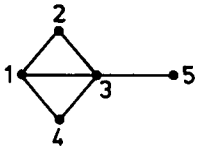
Theorem 2 in LSV also characterises decomposable graphs by a combinatorial property involving a certain counting index. Since this index is involved fundamentally in the estimation formula, we shall discuss this in the coming section.

5. The index and the estimation formula. Haberman (1974) introduces the *adjusted replication number* for subsets of sets in a generating class. In the decomposable case he shows that this number enters in the explicit formula for the

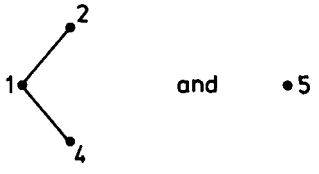
maximum likelihood estimate $\hat{P}(\mathbf{i})$ of $P(\mathbf{i})$. In LSV a related quantity is defined. Whereas the adjusted replication number is defined recursively, this index is defined directly.

Let $\mathbf{C}$ be a connected graph $(C, E(\mathbf{C}))$ and $d \subseteq C$ be a complete subset. The *pieces* of C relative to d are defined as follows: remove d from C and form the subgraph $\mathbf{C} \setminus d$ with vertices $C \setminus d$ and edges which are those in $E(\mathbf{C})$ that do not involve vertices in d. $\mathbf{C} \setminus d$ now has one or more connected components $\mathbf{A}_t$, $t \in T$, say. Let $\mathbf{C}_t$ be the subgraphs of $\mathbf{C}$ obtained by readjoining d to the subgraphs $\mathbf{A}_t$, i.e., $\mathbf{C}_t$ has vertex set $A_t \cup d$ and edges which are those in $E(\mathbf{C})$ that only involve vertices in $A_t \cup d$. $\mathbf{C}_t$, $t \in T$ are *the pieces of C relative to d.*

Probably the procedure is best illustrated by an example:

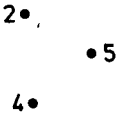


Consider this graph and let $d = \{3\}$. By removing d we get the following connected components:

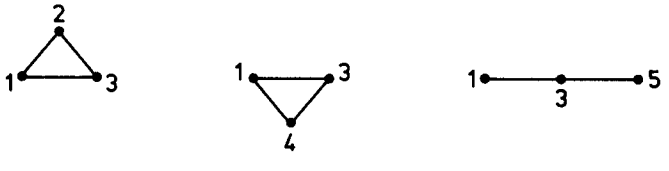


Readjoining d to these components we get the pieces:

For $d = \{1, 3\}$ we get components of $\mathbf{C} \setminus d$:



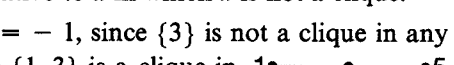
and thus pieces



Clearly, since d was complete in $\mathbf{C}$, d is complete in all the pieces $\mathbf{C}_t$, but not necessarily a clique in $\mathbf{C}_t$ (i.e., maximal).

Let $\nu(d)$ be defined as

$\nu(d) = 1 -$ the number of pieces of $\mathbf{C}$ relative to d in which d is not a clique.

In the example given above we have $\nu(\{3\}) = -1$, since $\{3\}$ is not a clique in any of the two pieces and $\nu(\{1, 3\}) = -1$ since $\{1, 3\}$ is a clique in 

but not in the two remaining pieces.

Corollary 7 of LSV asserts that *for any connected graph* $\mathbf{C}$ *we have*

$$\Sigma_{d \text{ complete}} \nu(d) \geqslant 1$$

and Theorem 2 of LSV that $\mathbf{C}$ *is decomposable if and only if equality holds*. Thus we have a combinatorial identity characterising decomposable graphs.

If $\mathbf{C}$ is not connected itself but has connected components $\mathbf{C}_t$, $t \in T$ we define an index $\nu_t(d)$ for each of the components and have that $\mathbf{C}$ *is decomposable* iff

$$\Sigma_{t \in T} \Sigma_d \nu_t(d) = |T|,$$

which is an easy consequence of the inequality.

The index is primarily a tool for revealing combinatorial properties of decomposable graphs. However, it is worth noting that this index occurs in the estimation formula.

In a decomposable, and thus graphical model the maximum likelihood estimate $\hat{P}(\mathbf{i})$ *of* $P(\mathbf{i})$ *based upon* n *independent observations, is given by*

$$\hat{P}(\mathbf{i}) = \left[\Pi_{t \in T} \Pi_d n(\mathbf{i}_d)^{\nu_t(d)}\right] / n^{|T|},$$

provided that all $n(\mathbf{i}_d)$ *are positive.* (In this formula $\nu_t(d)$ is interpreted as zero if $d \not\subseteq \mathbf{C}_t$.)

To show this result we first realise that it is enough to consider connected graphs. For the various connected components correspond to independent sets of factors and their probabilities as well as their estimates multiply. Next we see that the formula is correct for a graph with just one clique. This is clear because such a graph corresponds to an unrestricted probability and in that case we have

$$\hat{P}(\mathbf{i}) = n(\mathbf{i})/n.$$

Noting that for such a graph we have $\nu(d) = 0$ unless $d = C$ in which case $\nu(d) = 1$, we see that our formula is correct in this case.

The final step in the proof is an induction argument using two basic facts:

(i) if a generating class $\mathcal{C}$ is decomposed into $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{A} \cup \mathcal{B} = \mathcal{C}$, $\mathcal{A} \cap \mathcal{B} = \varnothing$ and $A \cap B = a^* \cap b^*$ for some $a^* \in \mathcal{A}$, $b^* \in \mathcal{B}$, where $A = \cup_{a \in \mathcal{A}} a$, $B = \cup_{b \in \mathcal{B}} b$, then

$$\hat{P}_{\mathcal{C}}(\mathbf{i}) = \frac{\hat{P}_{\mathcal{A}}(\mathbf{i}_A)\hat{P}_{\mathcal{B}}(\mathbf{i}_B)}{\hat{P}_{\{a^* \cap b^*\}}(\mathbf{i}_{a^* \cap b^*})},$$

which, e.g., follows directly from Theorem 2 of Andersen (1974);

(ii) if a generating class $\mathcal{C}$, where $\mathcal{C}$ is the maximal cliques of a connected graph **C** is decomposed as above, then both $\mathcal{A}$ and $\mathcal{B}$ are the cliques of the subgraphs **A** and **B**, these are both connected and the indices ν_A, ν_B and ν_C satisfy

$$\nu_C(d) = \nu_A(d) + \nu_B(d) \qquad \text{for } d \neq a^* \cap b^*$$
$$\nu_C(d) = \nu_A(d) + \nu_B(d) - 1 \quad \text{for } d = a^* \cap b^*.$$

This is Lemma 8 of LSV.

If we use these two facts and assume the result to be true for all graphical models with fewer than $|\mathcal{C}|$ cliques, we get

$$\hat{P}_{\mathcal{C}}(\mathbf{i}) = \frac{\hat{P}_{\mathcal{A}}(\mathbf{i}_A)\hat{P}_{\mathcal{B}}(\mathbf{i}_B)}{\hat{P}_{\{a^*\cap b^*\}}(\mathbf{i}_{a^*\cap b^*})} = \frac{\Pi_d n(\mathbf{i}_d)^{\nu_A(d)}\Pi_d n(\mathbf{i}_d)^{\nu_B(d)}}{n(\mathbf{i}_{a^*\cap b^*})}/n$$
$$= \Pi_d n(\mathbf{i}_d)^{\nu_C(d)}/n$$

where we again have let $\nu_A(d) = 0[\nu_B(d) = 0]$ if $d \not\subseteq A[d \not\subseteq B]$.

The estimation formula makes it possible for us to derive some further properties of our index. Let $n_\gamma = |I_\gamma|$ and suppose that we have $n = |I| = \Pi_{\gamma\in C} n_\gamma$ observations with exactly one observation in each cell, i.e., $n(\mathbf{i}) = 1$ for all $\mathbf{i}$. Then, clearly

$$\hat{P}(\mathbf{i}) = n^{-1}.$$

Using our formula for a connected graph C we also get

$$\hat{P}(\mathbf{i}) = n^{-1}\Pi_d n(\mathbf{i}_d)^{\nu(d)}$$
$$= n^{-1}\Pi_d(\Pi_{\gamma\notin d} n_\gamma)^{\nu(d)}$$
$$= n^{-1}\Pi_{\gamma\in C}\Pi_{d\not\ni\gamma} n_\gamma^{\nu(d)} = n^{-1}\Pi_{\gamma\in C} n_\gamma^{\Sigma_{d\subseteq C\setminus\{\gamma\}}\nu(d)}.$$

Since this expression is valid for all possible values of n_γ, we must have *for a connected, decomposable* graph **C**

$$\Sigma_{d\subseteq C\setminus\{\gamma\}}\nu(d) = 0 \quad \text{for all } \gamma \in C.$$

Since

$$\Sigma_d \nu(d) = 1 = \Sigma_{d\not\ni\gamma}\nu(d) + \Sigma_{d\ni\gamma}\nu(d),$$

we thus have, for all $\gamma \in \mathbf{C}$,

$$\Sigma_{d:\gamma\in d}\nu(d) = 1$$

for any connected, decomposable graph **C**.

A further identity is obtained by summation of the above identity for $\gamma \in C$:

$$|C| = \Sigma_{\gamma\in C}\Sigma_{d\ni\gamma}\nu(d) = \Sigma_d |d|\nu(d).$$

6. Graphical models of dimension less than or equal to five. Here, we shall give the graphical representation and the interpretation of all graphical models corresponding to an m-dimensional contingency table with $m \leqslant 5$. Apart from the

interpretation column this is just a question of listing all graphs with less than five vertices. We do this both to illustrate the material in the previous sections and as a counterpart to the tables in Goodman (1970) of all hierarchical models of dimension ≤ 4. We only list *connected* graphs since other models can be constructed by using these as connected components of other graphs. As remarked earlier, the various connected components in a graph of a graphical model correspond to independent sets of factors.

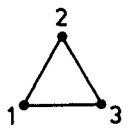
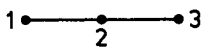
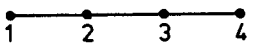
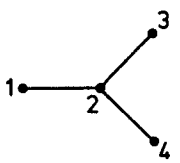
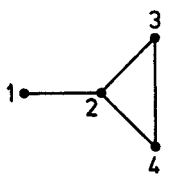
Giving the various interpretations in terms of conditional independence we shall use the notation of Goodman (1970), e.g.,

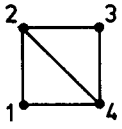
$$[1 \otimes 2|3]$$

meaning that, given 3, the factors 1 and 2 are conditionally independent. In Table 1 we list the decomposable graphical models and in Table 2 the nondecomposable models where we also indicate the critical ≥ 4-cycle.

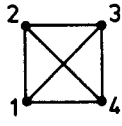
TABLE 1

Decomposable models of dimension less than or equal to five.

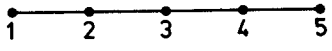
graph	interpretation
	unrestricted
	unrestricted
	unrestricted
	[1 ⊗ 3|2]
	[1 ⊗ 3, 4|2] ∩ [1, 2 ⊗ 4|3]
	[1 ⊗ 3 ⊗ 4|2]
	[1 ⊗ 3, 4|2]



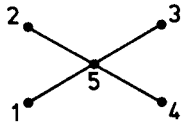
$[1 \otimes 3|2, 4]$



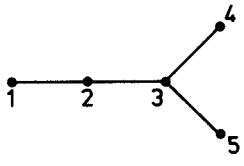
unrestricted



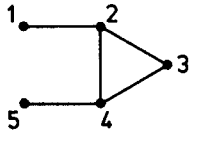
$[1 \otimes 3, 4, 5|2]$, etc.



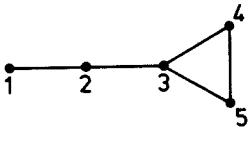
$[1 \otimes 2 \otimes 3 \otimes 4|5]$



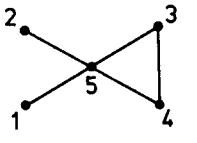
$[1 \otimes 3, 4, 5|2] \cap [1, 2 \otimes 4 \otimes 5|3]$



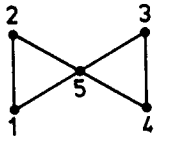
$[1 \otimes 5 \otimes 3|2, 4] \cap [1 \otimes 3, 4, 5|2]$
$\cap [5 \otimes 1, 2, 3|4]$



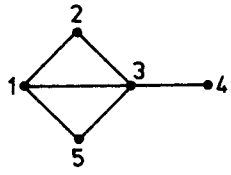
$[1, 2 \otimes 4, 5|3] \cap [1 \otimes 3, 4, 5|2]$



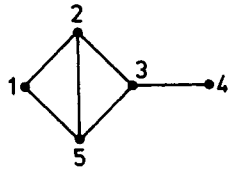
$[1 \otimes 2 \otimes 3, 4|5]$



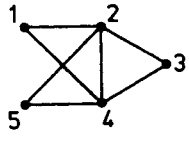
$[1, 2 \otimes 3, 4|5]$



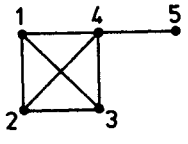
$[2 \otimes 5 \otimes 4|1, 3] \cap [1, 2, 5 \otimes 4|3]$



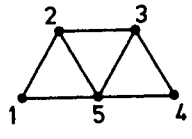
$[1, 2, 5 \otimes 4|3] \cap [1 \otimes 3, 4|2, 5]$



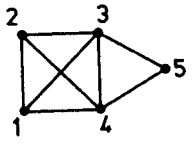
$[1 \otimes 3 \otimes 5|2, 4]$



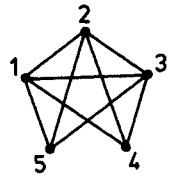
$[1, 2, 3 \otimes 5|4]$



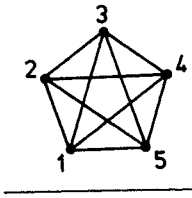
$[1 \otimes 3, 4|2, 5] \cap [1, 2 \otimes 4|3, 5]$



$[1, 2 \otimes 5|3, 4]$



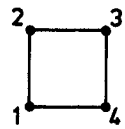
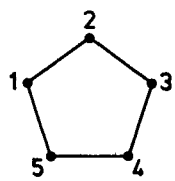
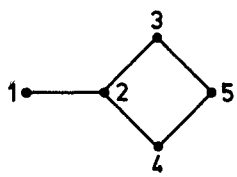
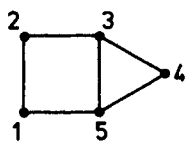
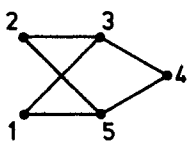
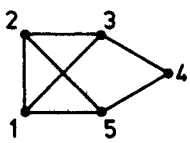
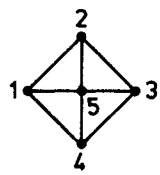
$[4 \otimes 5|1, 2, 3]$



unrestricted.

TABLE 2

Nondecomposable models that are graphical of dimension less than or equal to five.

graph	> 4-cycle	interpretation
	{1, 2, 3, 4}	$[1 \otimes 3\|2, 4] \cap [2 \otimes 4\|1, 3]$
	{1, 2, 3, 4, 5}	$[1, 2 \otimes 4\|3, 5]$, etc.
	{2, 3, 4, 5}	$[1, 2 \otimes 5\|3, 4] \cap [1 \otimes 3, 4, 5\|2] \cap [3 \otimes 1 \otimes 4\|2, 5]$
	{1, 2, 3, 5}	$[1, 2 \otimes 4\|3, 5] \cap [1 \otimes 3, 4\|2, 5] \cap [2 \otimes 4, 5\|1, 3]$
	{1, 3, 4, 5} {2, 3, 4, 5} and {1, 2, 3, 5}	$[1 \otimes 2 \otimes 4\|3, 5] \cap [3 \otimes 5\|1, 2, 4]$
	{1, 3, 4, 5} and {2, 3, 4, 5}	$[1, 2 \otimes 4\|3, 5] \cap [3 \otimes 5\|1, 2, 4]$
	{1, 2, 3, 4}	$[1 \otimes 3\|2, 4, 5] \cap [2 \otimes 4\|1, 3, 5]$.

Note that the last graph in Table 2 is *not* triangulated although it is made up by triangles. {1, 2, 3, 4} is a cyclic subgraph without a chord. Thus the term "triangulated" is a bit misleading.

The interpretation column is made to give an interpretation in usual terms. Of course other conditional independence properties can be derived from those listed using rules of conditional independence. The most accurate interpretation will always be that the model consists of all Markov fields on the given graph.

To illustrate the complexity of the various types of models we have computed the number of possible models of any given type for a given contingency table of dimension $\leqslant 5$. The number of general log-linear interaction models is equal to 2^{2^n-1}. The number of graphical models is equal to $\Sigma_{i=0}^{n}\binom{n}{i}2^{\binom{i}{2}}$. The number of decomposable models does not seem to admit an explicit formula, but can be counted using the graphs in Tables 1 and 2. To count the number of hierarchical models is tedious for $n = 5$.

TABLE 3
Number of models of given type.

type \ dimension	1	2	3	4	5
Interaction	2	8	128	32,768	2,147,483,648
Hierarchical	2	5	19	167	7,580
Graphical	2	5	18	113	1,450
Decomposable	2	5	18	110	1,233

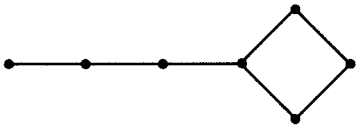
7. Some final remarks. Finally we shall give some suggestions as how to use the models and some possible directions for further work.

Searching for models. The graphical models are primarily relevant for the analysis of contingency tables of rather high dimension where it is difficult a priori to have very precise ideas about the relevant models and where one initially is looking for possible conditional independence among factors. We suggest that in such cases the graphs and their associated models be used directly in the search for possible models rather than the generating classes. It assures interpretability of any final model and it is in fact a very handy aid in visualising the features of the models. So, instead of trying gradually to remove interactions of high order, try to remove edges or throw in edges.

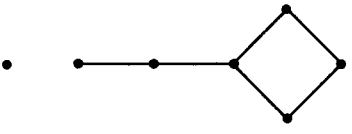
Estimation and test of hypotheses. At present, the graphs do not seem to be of great help in the numerical procedures of estimation and testing. There is something to be gained in discovering decomposability, thereby reducing the estimation problems. It might be the case that the graphs could be used in the estimation and testing problems. Consider for example the following model:

The model is not decomposable because of the 4-cycle to the right. On the other hand, the nondecomposability is isolated to that region. So, in fact, numerical iteration is only needed to find the marginal estimates in the table corresponding to these four factors. The estimate for the entire table can then be combined easily from this and an explicit formula for the marginal probability of the remaining factors using fact (i) in the proof of the basic estimation formula.

Similarly, we can get a simplification in a testing problem. Suppose that we want to find the likelihood ratio statistic for the hypothesis that the model



can be reduced to



Even though neither of the two models are decomposable, the difference between them is isolated to a decomposable region. Therefore, the likelihood ratio test statistic is nothing but that of testing independence in the two-way table involving the two factors at the left.

There is some work to be done in giving a good formulation of "local decomposability" and using such a notion in an efficient way in estimation and testing problems.

Exposition of the theory. Another possible use of the graphs is in an exposition of a theory of graphical models for contingency tables that uses the graphs *directly* instead of first relating these to generating classes and hierarchical models. This could have important pedagogical advantages.

We hope in the future to be able to give some more content to the vague remarks above.

Acknowledgments. We are grateful to M. L. Eaton, Minneapolis, for reading our manuscript and giving valuable suggestions.

REFERENCES

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

[2] BERGE, C. (1973). *Graphs and Hypergraphs*. Translated from French by E. Minieka. North-Holland, Amsterdam.

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

[4] GOODMAN, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *J. Amer. Statist. Assoc.* **66** 339–344.
[5] HABERMAN, S. J. (1970). The general log-linear model. Ph. D. thesis, Depart. Statist. Univ. Chicago.
[6] HABERMAN, S. J. (1974). *The Analysis of Frequency Data.* IMS monographs, Univ. Chicago Press.
[7] KEMENY, J. G., SNELL, J. L. and KNAPP, A. W. (1976). *Denumerable Markov Chains 2nd edition.* Springer, Heidelberg, New York, Berlin.
[8] LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1978). Decomposable graphs and hypergraphs. Preprint No. 9 Univ. Copenhagen, Inst. Math. Statist.
[9] PITMAN, J. W. (1976). Markov random fields. Lecture notes from a course given at the Univ. Copenhagen. Mimeographed.
[10] SPEED, T. P. (1976). Interaction. Unpublished manuscript.

SCHOOL OF MATH. SCIENCES
THE FLINDERS UNIVERSITY OF SOUTH AUSTRALIA
BEDFORD PARK, SOUTH AUSTRALIA 5042
AUSTRALIA 760511

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WESTERN AUSTRALIA
NEDLANDS, WESTERN AUSTRALIA 6009
AUSTRALIA

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN
5 UNIVERSITETSPARKEN, DK-2100
COPENHAGEN Ø
DENMARK