Chapter 4
Interaction Models

Steffen L. Lauritzen

The articles in this bundle are all associated with the notion of *interaction* and represent the genesis of the subject of graphical models in its modern form, the origins of these being traceable back to Gibbs [11] and Wright [30] and earlier.

Around 1976, Terry was fascinated by the notion of conditional independence, along the lines later published in Dawid [6, 7]. In 1976, Terry invited me to Perth and we were running a daily research seminar with the theme of studying similarities and differences between Statistics and Statistical Mechanics. In particular, we wondered what the relations were between notions of interaction as represented in linear models, in multi-dimensional contingency tables, and in stochastic models for particle systems; in addition, the purpose was also to understand what was the relation between these concepts and conditional independence.

As we discovered that these were all essentially the same concepts, the similarity being obscured by very different traditions of notation, the term graphical model was coined. Our findings, also obtained in collaboration with John Darroch, were collected in Darroch et al. [4], and later expanded and published in Speed [24], Darroch et al. [5], and Darroch and Speed [3] as well as Lauritzen et al. [19] and to some extent Speed [25], the latter giving an overview of a number of different variants and proofs of what has become known as the Hammersley–Clifford theorem [14, 2].

Of these articles, Darroch et al. [5] rather quickly had a seminal impact and a small community of researchers in the area of graphical models gradually emerged. In a certain sense, the article does not contain much formally new material (if any at all), but for the first time a simple, visual description and interpretation of the class of log-linear models [12, 13], which otherwise could seem obscure, was available. The interpretation of a subclass of the models in terms of conditional independence had an immediate intuitive appeal. In addition, the article identified and emphasized models represented by *chordal* or *triangulated* graphs as those where estimation

S.L. Lauritzen
Department of Statistics, University of Oxford, United Kingdom
e-mail: steffen@stats.ox.ac.uk

S. Dudoit (ed.), *Selected Works of Terry Speed*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4614-1347-9_4,

and other issues had a particularly simple solution, the combinatorial theory of these graphs being further studied in Lauritzen et al. [19].

Darroch and Speed [3] studied the notion of interaction from an algebraic point of view in terms of fundamental decompositions of the linear space of functions on a product of finite sets; indeed it essentially but implicitly uses the fundamental decomposition of this space into irreducible components which are stable under a product of symmetric groups [9] and thus gives an elegant algebraic perspective on the Hammersley–Clifford theorem.

Towards the end of 1976, Terry serendipitously came across Wermuth [29], which identified that a completely analogous theory could be developed for the Gaussian case, with chordal graphs playing essentially the same role as in the case of log-linear models; indeed, Dempster [8] had developed the basic computational and statistical theory for these under the name of models for *covariance selection*. This fact and the corresponding interpretation was emphasized and discussed in Darroch et al. [4] as well as in Speed [24, 25], but received otherwise relatively little attention at the time. Gaussian graphical models have had a remarkable renaissance in connection with the modern analysis of high-dimensional data, for example concerning gene expression [10, 23]. Out of this early work with Gaussian graphical models grew also the article by Speed and Kiiveri [26], which describes and unifies a class of iterative algorithms for fitting Gaussian graphical models of which special cases previously had been considered by e.g. Dempster [8]. Essentially, there are two fundamental types, of which one initially uses the estimate under no restrictions and iteratively ensures that restrictions of the model are satisfied; the other type initially uses a trivial estimator and iteratively ensures that the likelihood equations are satisfied. The article elegantly shows that an abundance of hybrids of these algorithms can be constructed and gives a unified proof of their convergence.

The last two articles [16, 17], represent the genesis of what today is probably the most prolific and well-known type of graphical models; these are based on *directed* acyclic graphs and admitting interpretation in *causal* terms similar to that of *structural equation models* [1]. At the time when these articles appeared they were (undeservedly) largely ignored both by the statistical and structural equation communities. Graphical models based on directed acyclic graphs—now mostly known as *Bayesian networks* [21]—have an unquestionable prominence in current scientific literature, but the surge of interest in these models was in particular generated by the prolific research activities in computer science, where work such as, for example, Lauritzen and Spiegelhalter [18], Pearl [22], Spirtes et al. [27], Heckerman et al. [15], and Pearl [20] established these models as objects worthy of intense study. In retrospect, it is clear that the global Markov property defined in Kiiveri et al. [17] was not the optimal one as there are independence relations true in any Bayesian network that cannot be derived from it, but fundamentally this article establishes the correct class of directed Markov models for the first time and thus yields a conditional independence perspective on structural equation models, as later elaborated, for example by Spirtes et al. [28].

References

[1] K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley and Sons, New York, 1989.

[2] P. Clifford. Markov random fields in statistics. In G. R. Grimmett and D. J. A. Welsh, editors, *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, pages 19–32. Oxford University Press, 1990.

[3] J. N. Darroch and T. P. Speed. Additive and multiplicative models and interactions. *Ann. Stat.*, 11:724–738, 1983.

[4] J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Log-linear models for contingency tables and Markov fields over graphs. Unpublished manuscript, 1976.

[5] J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. *Ann. Stat.*, 8:522–539, 1980.

[6] A. P. Dawid. Conditional independence in statistical theory (with discussion). *J. Roy. Stat. Soc. B*, 41:1–31, 1979.

[7] A. P. Dawid. Conditional independence for statistical operations. *Ann. Stat.*, 8: 598–617, 1980.

[8] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[9] P. Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1988.

[10] A. Dobra, C. Hans, B. Jones, J. R. Nevins, and M. West. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90:196–212, 2004.

[11] W. Gibbs. *Elementary Principles of Statistical Mechanics*. Yale University Press, New Haven, Connecticut, 1902.

[12] L. A. Goodman. The multivariate analysis of qualitative data: Interaction among multiple classifications. *J. Am. Stat. Assoc.*, 65:226–256, 1970.

[13] S. J. Haberman. *The Analysis of Frequency Data*. University of Chicago Press, Chicago, 1974.

[14] J. M. Hammersley and P. E. Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, 1971.

[15] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.*, 20:197–243, 1995.

[16] H. Kiiveri and T. P. Speed. Structural analysis of multivariate data: A review. In S. Leinhardt, editor, *Sociological Methodology*. Jossey-Bass, San Francisco, 1982.

[17] H. Kiiveri, T. P. Speed, and J. B. Carlin. Recursive causal models. *J. Aust. Math. Soc. A*, 36:30–52, 1984.

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

[19] S. L. Lauritzen, T. P. Speed, and K. Vijayan. Decomposable graphs and hypergraphs. *J. Aust. Math. Soc. A*, 36:12–29, 1984.

[20] J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK, 2000.
[21] J. Pearl. Fusion, propagation and structuring in belief networks. *Artif. Intell.*, 29:241–288, 1986.
[22] J. Pearl. Probabilistic Inference in Intelligent Systems. Morgan Kaufmann Publishers, San Mateo, CA, 1988.
[23] J. Schäfer and K. Strimmer. An empirical-Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21:754–764, 2005.
[24] T. P. Speed. Relations between models for spatial data, contingency tables and Markov fields on graphs. *Adv. Appl. Prob.: Supplement*, 10:111–122, 1978.
[25] T. P. Speed. A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhyā Ser. A*, 41:184–197, 1979.
[26] T. P. Speed and H. Kiiveri. Gaussian Markov distributions over finite graphs. *Ann. Stat.*, 14:138–150, 1986.
[27] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Springer-Verlag, New York, 1993. Reprinted by MIT Press.
[28] P. Spirtes, T. S. Richardson, C. Meek, R. Scheines, and C. Glymour. Using path diagrams as a structural equation modeling tool. *Sociol. Method. Res.*, 27:182–225, 1998.
[29] N. Wermuth. Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32:95–108, 1976.
[30] S. Wright. The method of path coefficients. *Ann. Math. Statist.*, 5:161–215, 1934.

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 $I = \prod_{\gamma \in C} I_\gamma$ 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:

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

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.

MARKOV FIELDS AND LOG-LINEAR MODELS

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 $\mathcal{C}$.

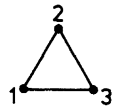
Let us examine the restrictions on our interactions given by this generating class. By the definition of a hierarchical model we have $\xi_a \equiv 0$ *unless a is contained in a maximal complete subset*, i.e., *unless a is a complete subset*. In other words, the set of probabilities P in our model is *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.

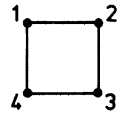
A generating class is, in LSV, called a *generating class hyper graph* (g.c. hypergraph). The first-order interaction graph of a generating class is called the *2-section of the* g.c. *hypergraph*.

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

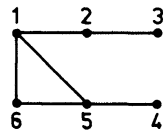
$$P(\mathbf{i}) = \frac{\Pi_{t=1}^{k} P(\mathbf{i}_{a_t})}{\Pi_{t=2}^{k} P(\mathbf{i}_{b_t})}$$

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 $\mathcal{C}$.

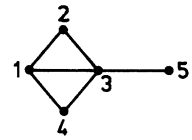
Theorem 2 in LSV also characterises decomposable graphs by a combinatorial property involving a certain counting index. Since this index is involved fundamentally in the estimation formula, we shall discuss this in the coming section.

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

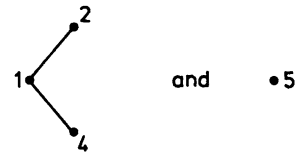
maximum likelihood estimate $\hat{P}(\mathbf{i})$ of $P(\mathbf{i})$. In LSV a related quantity is defined. Whereas the adjusted replication number is defined recursively, this index is defined directly.

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:

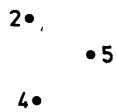


Consider this graph and let $d = \{3\}$. By removing d we get the following connected components:

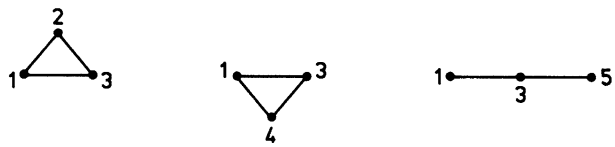


Readjoining d to these components we get the pieces:

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



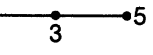
and thus pieces



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 $\leqslant 4$. 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.

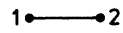
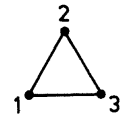
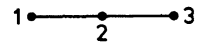
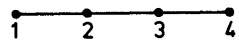
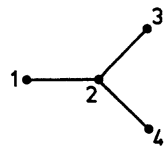
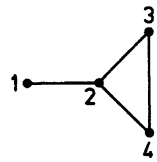
Giving the various interpretations in terms of conditional independence we shall use the notation of Goodman (1970), e.g.,

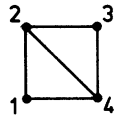
$$[1 \otimes 2|3]$$

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 $\geqslant 4$-cycle.

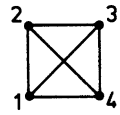
TABLE 1

Decomposable models of dimension less than or equal to five.

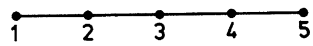
graph	interpretation
	unrestricted
	unrestricted
	unrestricted
	$[1 \otimes 3\|2]$
	$[1 \otimes 3, 4\|2] \cap [1, 2 \otimes 4\|3]$
	$[1 \otimes 3 \otimes 4\|2]$
	$[1 \otimes 3, 4\|2]$



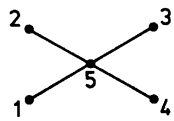
[1 ⊗ 3|2, 4]



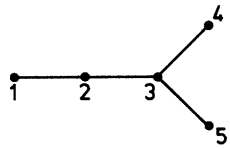
unrestricted



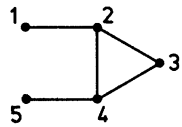
[1 ⊗ 3, 4, 5|2], etc.



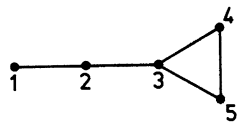
[1 ⊗ 2 ⊗ 3 ⊗ 4|5]



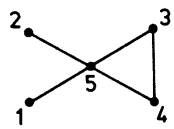
[1 ⊗ 3, 4, 5|2] ∩ [1, 2 ⊗ 4 ⊗ 5|3]



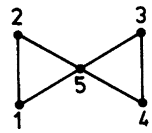
[1 ⊗ 5 ⊗ 3|2, 4] ∩ [1 ⊗ 3, 4, 5|2]
∩[5 ⊗ 1, 2, 3|4]



[1, 2 ⊗ 4, 5|3] ∩ [1 ⊗ 3, 4, 5|2]

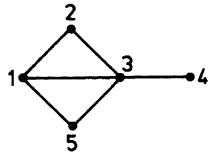


[1 ⊗ 2 ⊗ 3, 4|5]

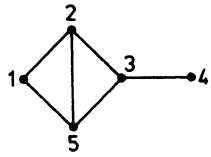


[1, 2 ⊗ 3, 4|5]

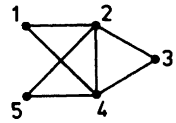
MARKOV FIELDS AND LOG-LINEAR MODELS



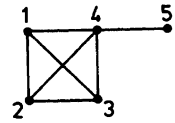
$[2 \otimes 5 \otimes 4|1, 3] \cap [1, 2, 5 \otimes 4|3]$



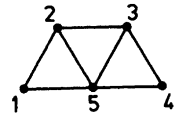
$[1, 2, 5 \otimes 4|3] \cap [1 \otimes 3, 4|2, 5]$



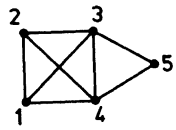
$[1 \otimes 3 \otimes 5|2, 4]$



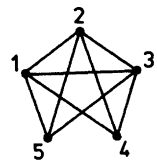
$[1, 2, 3 \otimes 5|4]$



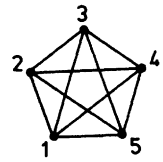
$[1 \otimes 3, 4|2, 5] \cap [1, 2 \otimes 4|3, 5]$



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



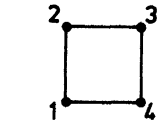
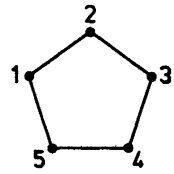
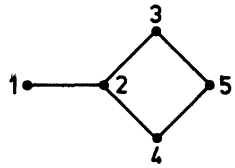
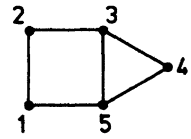
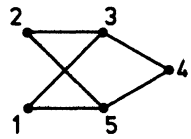
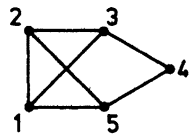
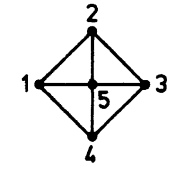
$[4 \otimes 5|1, 2, 3]$



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 ⊗ 3|2, 4] ∩ [2 ⊗ 4|1, 3]
	{1, 2, 3, 4, 5}	[1, 2 ⊗ 4|3, 5], etc.
	{2, 3, 4, 5}	[1, 2 ⊗ 5|3, 4] ∩ [1 ⊗ 3, 4, 5|2] ∩[3 ⊗ 1 ⊗ 4|2, 5]
	{1, 2, 3, 5}	[1, 2 ⊗ 4|3, 5] ∩ [1 ⊗ 3, 4|2, 5] ∩[2 ⊗ 4, 5|1, 3]
	{1, 3, 4, 5} {2, 3, 4, 5} and {1, 2, 3, 5}	[1 ⊗ 2 ⊗ 4|3, 5] ∩[3 ⊗ 5|1, 2, 4]
	{1, 3, 4, 5} and {2, 3, 4, 5}	[1, 2 ⊗ 4|3, 5] ∩[3 ⊗ 5|1, 2, 4]
	{1, 2, 3, 4}	[1 ⊗ 3|2, 4, 5] ∩[2 ⊗ 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 $\sum_{i=0}^{n}\binom{n}{i}2^{\binom{i}{2}}$. 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 $n = 5$.

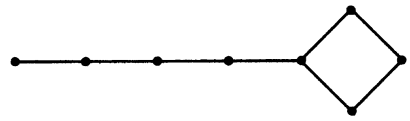
TABLE 3
Number of models of given type.

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.

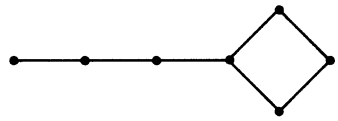
Searching for models. The graphical models are primarily relevant for the analysis of contingency tables of rather high dimension where it is difficult a priori to have very precise ideas about the relevant models and where one initially is looking for possible conditional independence among factors. 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:

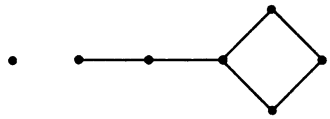


The model is not decomposable because of the 4-cycle to the right. On the other hand, the nondecomposability is isolated to that region. So, in fact, numerical iteration is only needed to find the marginal estimates in the table corresponding to these four factors. The estimate for the entire table can 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

The Annals of Statistics
1983, Vol. 11, No. 3, 724–738

ADDITIVE AND MULTIPLICATIVE MODELS AND INTERACTIONS

BY J. N. DARROCH AND T. P. SPEED

The Flinders University of South Australia and C.S.I.R.O.

A unified treatment is given of the classical additive models for complete factorial experiments and of multiplicative models and Lancaster-additive models for multi-dimensional contingency tables. The models are characterised by properties of being simplest subject to having a prescribed set of marginals. It is shown that, by using averaging operators and the notion of a generalised interaction, the interaction properties of these models can be derived very simply.

1. Introduction. Interaction models provide simplified structures for the arrays of unknown parameters which arise in *factorial experiments* and in multidimensional *contingency tables*. These two fields of application will be considered side by side, rather more attention being given to contingency tables.

In a factorial experiment there are, say, s factors $A_1, \cdots, A_s$ and a single response y. If the factors have $r_1, \cdots, r_s$ levels there are $r_1 \times \cdots \times r_s$ different combinations of levels called *cells*. The expected value $Ey = \eta$ of the response varies from cell to cell and inferential attention is focused on the array η of the $r_1 \times \cdots \times r_s$ values of η.

In a *pure response* s-dimensional contingency table there are s categorical variables $X_1, \cdots, X_s$ taking $r_1, \cdots, r_s$ values. This time the unknown parameter at each cell is the probability p of that particular combination of response values. The following discussion also applies to s-dimensional contingency tables in which some of the dimensions correspond to factors and the remainder to responses. The probability p is then the probability of the response values given the factor levels. There remains one further model for contingency tables. In it the $r_1 \times \cdots \times r_s$ frequencies are independent Poisson variables and the theory of this paper is applied to the array of their mean values μ.

The standard models for η, p, μ or some function of them are defined by *linear subspaces* of $\mathbb{R}^{\mathscr{I}}$, where

$$\mathscr{I} = \{i = (i_1, \cdots, i_s) : 1 \leq i_\sigma \leq r_\sigma, \sigma = 1, \cdots, s\}.$$

They are usually obtained by introducing a system of *interactions* and then requiring that a subset of these interactions vanish. This may be quite appropriate with additive models for factorial experiments, where the individual interactions can have a practical interpretation, but it is not necessarily so with multiplicative models for contingency tables. One of the aims of this paper is to give a simple account of an alternative approach in which we define models first (Section 2) and interactions later (Section 4). In doing so we take the opportunity to compare and contrast additive and multiplicative models, and to note the similarities and differences between two widely used parametrizations.

There is a certain amount of overlap in subject matter between this paper and the work of Haberman (1974, 1975) but the mathematical treatments of the common material are substantially different. Andersen (1975) gives a very clear summary of the general properties of interaction subspaces, applicable either to additive or multiplicative models, whilst other general treatments are by Mann (1949), Good (1958, 1963), Kurkjian and Zelen (1962), Grizzle, Starmer and Koch (1969), Goodman (1970) and Davidson (1973). Writings which concentrate upon multiplicative models for probabilities include several books: Haberman (1974, 1978, 1979), Bishop, Fienberg and Holland (1975), Fienberg (1977),

Received February 1980; revised December 1982.
AMS 1980 *subject classifications*. 62E10, 62E30.
Key words and phrases. Generalised interaction, Lancaster model, linear model log-linear model.

Gokhale and Kullback (1978), and Plackett (1981). Lancaster's theory of interaction and generalised correlation can be found in his book (1969), although the formulation given here (for finitely-valued random variables) is slightly different from his, being chosen to facilitate comparisons with other models. Further literature references are given in the body of the paper.

Inference matters are not discussed apart from a few comments on least-squares, sufficient reductions and maximum likelihood estimation. There is also no discussion of experimental design questions. A number of the results in this paper are new but in general the emphasis is on unifying existing results and on proving them by elementary methods.

2. Models and marginals. In this section we introduce the models which will be the main topic of the paper. The s factors or responses will be labeled by elements of $S = \{1, 2, \cdots, s\}$, subsets of which will be denoted by a, b, c, d. As in the introduction $\sigma \in S$ is supposed to have r_σ values (levels or response categories), and we write $\mathscr{I}$ for the set of cells i; precisely $\mathscr{I} = \{i = (i_\sigma) : i \le i_\sigma \le r_\sigma, \sigma \in S\}$. More generally we write i_a for the sub-tuple $i_a = (i_\sigma : \sigma \in a)$, $a \subseteq S$.

2.1 *The models.* Let $\mathscr{A}$ be a collection of subsets of S. The linear subspace $\Omega_\mathscr{A}$ of $\Omega = \mathbb{R}^\mathscr{I}$ is defined by the property that the function $f = (f(i) : i \in \mathscr{I})$ belongs to $\Omega_\mathscr{A}$ if and only if

$$f(i) = \sum_{a \in \mathscr{A}} \lambda_a(i_a) \tag{2.1}$$

for some functions $\{\lambda_a : a \in \mathscr{A}\}$. Having defined $\Omega_\mathscr{A}$, the *model* $M_\mathscr{A}$ for f is simply the property that f belongs to $\Omega_\mathscr{A}$. The collection $\mathscr{A}$ is called the *generating class* of the model. Given $\mathscr{A}$ let $\mathscr{A}^*$ denote the sub-collection of elements of $\mathscr{A}$ which are maximal with respect to inclusion. It is clear that $M_{\mathscr{A}^*}$ is the same model as $M_\mathscr{A}$ because if $b \subseteq a$, then $\lambda_a(i_a) + \lambda_b(i_b) = \mu_a(i_a)$. Whilst it is economical in practice to work with $\mathscr{A}^*$, the theory does not require us to do so.

EXAMPLE 2.1. All our examples will have $s \le 4$ and for convenience we will write i, j, k and l instead of i_1, i_2, i_3 and i_4. Whenever no confusion is possible, we will use subscripts and omit the set describing the relevant indices. Thus we will write λ_{ijk} instead of $\lambda_{(1,2,3)}(i_1, i_2, i_3)$.

Suppose that $s = 3$ and $\mathscr{A} = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$. Then $\Omega_\mathscr{A}$ consists of all arrays $f = (f_{ijk})$ representable in the form

$$f_{ijk} = \alpha_{ij} + \beta_{jk} + \gamma_{ki}$$

for some arrays (α_{ij}), (β_{jk}) and (λ_{ki}). □

Of the following interpretations of $M_\mathscr{A}$, the first is applicable mainly to a factorial experiment with observations $y = (y(i) : i \in \mathscr{I})$ and expected values $\eta = (\eta(i) : i \in \mathscr{I})$. The others are applicable to a contingency table with cell frequencies $n = (n(i) : i \in \mathscr{I})$ and probabilities $p = (p(i) : i \in \mathscr{I})$ or expected frequencies $\mu = (\mu(i) : i \in \mathscr{I})$.

Additive model: $\eta \in \Omega_\mathscr{A}$.
Multiplicative model: $\log p \in \Omega_\mathscr{A}$, p positive.
or $\log \mu \in \Omega_\mathscr{A}$, μ positive.
Lancaster-additive model I: $p/q \in \Omega_\mathscr{A}$.
Lancaster-additive model II: $P/Q \in \Omega_\mathscr{A}$.

Here the function $f = \log p$ is defined by $f(i) = \log p(i)$ whilst $f = p/q$ means $f(i) = p(i)/q(i)$ where $q(i) = p_1(i_1) \cdots p_s(i_s)$ is the product of the one-dimensional marginal probabilities from p. Finally $f = P/Q$ means $f(i) = P(i)/Q(i)$ where $P(i) = \sum_{j \le i} p(i)$ and similarly for Q, where $j \le i$ means $j_\sigma \le i_\sigma$, $\sigma = 1, \cdots, s$. Additive and multiplicative models

are commonly called *linear* and *log-linear* models, respectively. The general results below apply also to any generalised linear model; see Nelder and Wedderburn (1972), Baker and Nelder (1978).

Why should we study additive, multiplicative and Lancaster-additive models? In the first place, the way in which they combine linearity and economy has an obvious appeal. Less obvious is that they can be characterised by attractive properties relating them to their $\mathscr{A}$-marginal functions; these are given in the following section. Their best-known properties are the no-interaction ones by which they are usually characterised, and these are given in Section 4.

Suppose that f is known or assumed to satisfy $M_{\mathscr{A}}$ so that $f(i)$ is representable as the sum of parameters $\lambda_a(i_a)$. Leaving aside the trivial case when the generating class $\mathscr{A}$ contains only one element, it is always possible to choose more than one parametric representation of f. That is the parameters $\lambda_a(i_a)$ are not uniquely determined by f. The extent to which they are unique is discussed in Section 4.

Generally speaking, the parameters $\lambda_a(i_a)$ have little more than a mathematical existence but, on rare occasions, they also have a physical meaning.

EXAMPLE 2.2. Let i index the cities of a country, let j index age-categories of brides and let k index age-categories of bridegrooms. Let μ_{ijk} be the expected number of marriages, in a given year, in city i between brides of age j and bridegrooms of age k. Then

$$\mu_{ijk} = M_{ij}N_{ik}\rho_{ijk},$$

M_{ij}, N_{ik} being the numbers of eligible women of age j, men of age k in city i at the beginning of the year, and where ρ_{ijk} is the rate of marriages in city i between women of age j and men of age k. It may be very reasonable to assume that $\rho_{ijk} = \rho_{jk}$ so that

$$\log \mu_{ijk} = \log M_{ij} + \log \rho_{jk} + \log N_{ik}.$$

Thus we have an instance of the model of Example 2.1 in which the parameters (α_{ij}), (β_{jk}), (γ_{ik}) can be given a physical interpretation. □

2.2 *Marginals.* For an arbitrary element $\xi = (\xi(i) : i \in \mathscr{I}) \in \Omega$ and a subset $a \subseteq S$ we write $\xi_a(i_a) = \sum_{i_{a'}} \xi(i)$, the sum being over all $i_\sigma = 1, \cdots, r_\sigma$, $\sigma \in a' = S - a$, and call ξ_a the (unweighted) *a-marginal* of ξ. The $\mathscr{A}$-marginals of ξ are $\{\xi_a : a \in \mathscr{A}\}$. Now let $m(i)$ be a positive weight attached to cell i, where $\sum_i m(i) = 1$. It is necessary in much of what follows to work with weight functions different from the uniform weight function $m(i) = (\Pi_\sigma r_\sigma)^{-1}$. We define the ($m$-weighted) *a-marginal mean* $\bar{\eta}_a$ of $\eta \in \Omega$ by

$$\bar{\eta}_a(i_a) = \frac{1}{m_a(i_a)} \sum_{i_{a'}} m(i)\eta(i), \tag{2.2}$$

and the *$\mathscr{A}$-marginal* means of η are $\{\bar{\eta}_a : a \in \mathscr{A}\}$. The ($m$-weighted) *inner product* $\langle \xi, \eta \rangle_m$ of $\xi, \eta \in \Omega$ is defined by

$$\langle \xi, \eta \rangle_m = \sum_i m(i)\xi(i)\eta(i), \tag{2.3}$$

and its associated *norm* (length) is $\| \xi \|_m = \{\langle \xi, \xi \rangle_m\}^{1/2}$.

Additive models. In terms of these notions we can now characterise the additive model $M_{\mathscr{A}}$. We begin with a lemma.

LEMMA 2.1. *Fix ξ, $\eta_0 \in \Omega$ and consider the set of all η with the same $\mathscr{A}$-marginal means as η_0 and the squared distance $\| \eta - \xi \|_m^2$ of each such η from ξ. Suppose that, in this set, there exists η_1 satisfying $\eta_1 - \xi \in \Omega_{\mathscr{A}}$. Then η_1 uniquely minimizes $\| \eta - \xi \|_m^2$.*

PROOF. The condition that η_0 have the same $\mathscr{A}$-marginal means is $\eta - \eta_0 \perp_m \Omega_{\mathscr{A}}$, where orthogonality $\perp_m$ is with respect to the inner product (2.3). Therefore $\eta - \eta_1 \perp_m \Omega_{\mathscr{A}}$ since

$\eta - \eta_0 \perp_m \Omega_{\mathscr{A}}$ and $\eta_1 - \eta_0 \perp_m \Omega_{\mathscr{A}}$. But $\eta_1 - \xi \in \Omega_{\mathscr{A}}$ and so $\langle \eta - \eta_1, \eta_1 - \xi \rangle_m = 0$. Rearrangement gives

$$\|\eta - \xi\|_m^2 - \|\eta_1 - \xi\|_m^2 = \|\eta - \eta_1\|_m^2$$

which establishes the truth of the Lemma. □

Note that if e is the unit function $e(i) = 1$ and $\xi = ke$, k constant, then $\xi \in \Omega_{\mathscr{A}}$ and it seems appropriate to describe ξ as *uniform*. The characterisation of the additive model can now be stated: any $\eta \in \Omega_{\mathscr{A}}$ is *simplest* in the sense that it is *closest* to being uniform amongst all arrays with the same $\mathscr{A}$-marginal means. Closeness is measured by $\|\cdot\|_m^2$ and simplest means that ξ is uniform. There is thus a separate characterisation for each positive weight function m.

The above discussion has not involved the question of existence, given η_0, ξ, of η_1 satisfying

$$\eta_1 - \eta_0 \perp_m \Omega, \quad \eta_1 - \xi \in \Omega_{\mathscr{A}} \tag{2.4}$$

but this question is well-known to have an affirmative answer. For $\eta_1 - \xi$ is the projection of $\eta_0 - \xi$ onto $\Omega_{\mathscr{A}}$ orthogonal with respect to $\langle \cdot, \cdot \rangle_m$; equivalently, $\eta_0 - \eta_1$ is the orthogonal projection of $\eta_0 - \xi$ onto $\Omega_{\mathscr{A}}^{\perp}$, the orthogonal complement of $\Omega_{\mathscr{A}}$.

Multiplicative models. The analogous characterisation of the multiplicative model $M_{\mathscr{A}}$, which is due to Good (1963) and Ku and Kullback (1968), closely resembles the previous one. Let the (unweighted) $\mathscr{A}$-marginals of the probability p be fixed at those of p_0 and measure the difference between p and a positive probability π by the Kullback discriminatory information

$$K(p, \pi) = \sum_i p(i) \log p(i)/\pi(i) = \langle p, \log p/\pi \rangle \tag{2.5}$$

where $\langle \xi, \eta \rangle = \sum_i \xi(i)\eta(i)$ is the unweighted inner product.

LEMMA 2.2. *Suppose that, among all p with the same $\mathscr{A}$-marginals as p_0, there exists p_1 satisfying $\log p_1/\pi \in \Omega_{\mathscr{A}}$. Then p_1 uniquely minimises $K(p, \pi)$.*

PROOF. Since $p - p_0 \perp \Omega_{\mathscr{A}}$, $p_1 - p_0 \perp \Omega_{\mathscr{A}}$ and $\log p_1/\pi \in \Omega_{\mathscr{A}}$, we deduce that $\langle p - p_1, \log p_1/\pi \rangle = 0$. Rearranging this gives $\langle p, \log p/\pi \rangle - \langle p_1, \log p_1/\pi \rangle = \langle p, \log p/p_1 \rangle$, i.e. $K(p, \pi) - K(p_1, \pi) = K(p, p_1)$, from which the lemma follows. □

Taking π to be the uniform probability function gives the following characterisation: any p satisfying the multiplicative model $\log p \in \Omega_{\mathscr{A}}$ is simplest in the sense that it maximises $-\sum_i p(i) \log p(i)$ among all probabilities having the same $\mathscr{A}$-marginals. Assuming that $\cup \{a : a \in \mathscr{A}\} = S$, we may take $\pi = q_0$, the product of the one-dimensional marginals of p_0 and obtain the conclusion that any p satisfying the multiplicative model $M_{\mathscr{A}}$ is closest to being independent amongst all probabilities with the same $\mathscr{A}$-marginals, closeness being measured by K.

The existence of p_1 satisfying

$$p_1 - p_0 \perp \Omega_{\mathscr{A}}, \quad \log p_1/\pi \in \Omega_{\mathscr{A}} \tag{2.6}$$

is assured provided that the $\mathscr{A}$-marginals of p_0 admit a positive probability, see Haberman (1974), Barndorff-Nielsen (1978). Darroch and Ratcliff (1972) proved that, with this proviso and for any subspace ω of Ω, it is possible to construct p_1 given p_0, π and ω by generalised iterative scaling. When $\omega = \Omega_{\mathscr{A}}$ then iterative proportional scaling can be used.

Lancaster-additive Model I. The results concerning additive models can be adapted to provide a characterisation of the Lancaster-additive model I and because it is very similar to the two preceding ones, we only give a brief outline.

Suppose that the unweighted $\mathscr{A}$-marginals of p are held fixed at those of p_0, and that

$\cup \{a: a \in \mathscr{A}\} = S$. Then all of the univariate marginals p_σ are also held fixed and so too is $q = \Pi p_\sigma$, equal to q_0 say. If we put $m = q_0$ and $\eta = p/q = p/q_0$ in (2.2) we find that holding p_a fixed is equivalent to holding $\bar{\eta}_a$ fixed. With $\xi = e$ the difference $\| \eta - \xi \|_m^2$ simplifies to

$$\phi^2(p, q_0) = \sum_i [p(i) - q_0(i)]^2 / q_0(i), \tag{2.7}$$

the Pearson Chi squared measure of difference between p and q_0. Lemma 2.1 may be translated to apply in this context and using that we obtain the following characterisation: any p satisfying the Lancaster-additive model I with $\Omega_{\mathscr{A}}$ is simplest in the sense of being closest to independent among all probabilities having the same $\mathscr{A}$-marginals, closeness being measured by ϕ^2.

The equations that p_1 satisfies are

$$p_1 - p_0 \perp \Omega_{\mathscr{A}}, \quad p_1/q_0 \in \Omega_{\mathscr{A}} \tag{2.8}$$

where $\perp$ here denotes orthogonality with respect to the unweighted inner product. The existence of p_1 given p_0, that is, of a probability function having prescribed $\mathscr{A}$-marginals and satisfying the Lancaster additive model I is not now guaranteed; see Darroch (1974) for a counter-example when $\mathscr{A} = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$, and for further comparisons between these models and the analogous multiplicative models.

2.3 *Fitting the models.* Let us suppose that data $y = (y(i): i \in \mathscr{I})$ from a factorial experiment has a normal distribution with mean $\eta \in \Omega_{\mathscr{A}}$ and covariance matrix $\sigma^2 \operatorname{diag}(m)^{-1}$, the diagonal matrix with value $m(i)^{-1}$ in the ith position. Then a *sufficient* reduction of y is to the pair $(\eta_1, \| \eta_1 - \eta_0 \|_m^2)$ where η_1, the projection of $y = \eta_0$ onto $\Omega_{\mathscr{A}}$ orthogonal with respect to $\langle \cdot, \cdot \rangle_m$, satisfies (2.4) with $\xi = 0$. We have already seen that η_1 is completely determined by its $\mathscr{A}$-marginal means, and these coincide with those of y. If we further suppose that m is *completely multiplicative* in that it can be written

$$m(i) = k \prod_\sigma m_\sigma(i_\sigma),$$

where for each $\sigma \in S$, $m_\sigma(i_\sigma) \geq 0$, $\sum_{i_\sigma} m_\sigma(i_\sigma) = 1$ and k is a constant, then we can express η_1 in terms of the $\mathscr{A}$-marginal means of y via formula (3.6) below. Thus (when m is completely multiplicative) the set of $\mathscr{A}$-marginal means is not only a sufficient reduction of y under the *additive model* $M_{\mathscr{A}}$, but also there is a closed-form solution of the least-squares (= maximum likelihood) estimation problem.

We turn now to the contingency table $n = (n(i): i \in \mathscr{I})$, supposing that n has a multinomial distribution with probability parameter p satisfying the *multiplicative model* $M_{\mathscr{A}}$ and total sample size $N = \sum_i n(i)$. The (unweighted) $\mathscr{A}$-marginal totals $\{n_a: a \in \mathscr{A}\}$ constitute a sufficient reduction of n and, provided these marginals admit a positive table, the log-likelihood $\langle n, \log p \rangle$ is maximised, or $K((1/N)n, p)$ is minimised, subject to $\log p/\pi \in \Omega_{\mathscr{A}}$ (normally π is uniform) when $p = p_1$ satisfies (2.6) with $p_0 = (1/N)n$. That these equations give the unique maximum likelihood solution is immediately verified on noting that $\langle \log p_1 - \log p, (1/N)n - p_1 \rangle = 0$ and on rearranging the term on the left-hand side of this equation to give $K((1/N)n, p) - K((1/N)n, p_1) = K(p_1, p)$. As was noted in 2.2 above, the equations (2.6) can be solved by the well-known iterative proportional scaling procedure.

To our knowledge there is no exact maximum-likelihood theory for the fitting of Lancaster-additive multinomial models to contingency tables, although a number of authors have discussed asymptotic theory for likelihood-ratio tests under the independence alternative, see Lancaster (1969) for details.

3. Generalised interactions. Denote by $M_{\mathscr{A}}$ the model for $f = (f(i): i \in \mathscr{I})$ defined by

$$M_{\mathscr{A}}: f \in \Omega_{\mathscr{A}}.$$

MODELS AND INTERACTIONS

The function f will be variously interpreted as η, $\log p$, p/q or P/Q. In 3.2 below $M_{\mathscr{A}}$ will be formulated as imposing *zero generalised $\mathscr{A}$-interaction*, where generalised interactions are defined very simply by repeatedly averaging over the values $f(i)$ of f.

3.1 *Averaging operators.* Let w_σ be a weight function defined on $\{1, 2, \cdots, r_\sigma\}$, i.e. $\sum_{i_\sigma} w_\sigma(i_\sigma) = 1$. The numbers $w_\sigma(i_\sigma)$ will be thought of as non-negative although there is no strict need for them to be so. Write $S - \{\sigma\} = S - \sigma$. Then the *averaging operator* $T_{S-\sigma}$ operating on f is defined by

$$(T_{S-\sigma}f)(i) = \sum_{i_\sigma} w_\sigma(i_\sigma) f(i).$$

Thus $T_{S-\sigma}$ takes weighted averages over the σth coordinate and leaves a function which depends on i through $i_{S-\sigma}$ only. For $a \subseteq S$ let T_a be the operator which takes averages over all coordinates with indices in $a' = S - a$. In other words,

$$T_a = \Pi_{\sigma \in a'} T_{S-\sigma}.$$

For example, if $S = \{1, 2, 3\}$, then $T_{\{1\}} = T_{S-2} T_{S-3} = T_{\{1,3\}} T_{\{1,2\}}$. When $a = S$ we define $T_S = I$, the *identity* operator. An alternative definition of T_a is possible via (2.2): $T_a \eta = \bar{\eta}_a$ where this average is weighted with respect to the completely multiplicative weight function $w(i) = \Pi_{\sigma \in S} w_\sigma(i_\sigma)$. It is immediate that T_a is a linear operator on Ω, that $T_a^2 = T_a$ and, more generally that

$$T_a T_b = T_b T_a = T_{ab} \tag{3.1}$$

where for $a, b \subseteq S$ we write $a \cap b = ab$.

Two particular weight functions w are of special interest. One is the *uniform weight function* defined by

$$w_\sigma(i_\sigma) = 1/r_\sigma.$$

The other is the *substitution weight function* defined by

$$w_\sigma(i_\sigma) = \begin{cases} 0 & \text{if} \quad i_\sigma \neq r_\sigma, \\ 1 & \text{if} \quad i_\sigma = r_\sigma. \end{cases}$$

The resulting substitution operator T_a has the defining property $(T_a f)(i) = f(i_a r_{a'})$ where $j = i_a r_{a'}$ denotes the cell with $j_\sigma = i_\sigma$ if $\sigma \in a$ and $j_\sigma = r_\sigma$, if $\sigma \in a'$. Thus T_a substitutes r_σ for i_σ, $\sigma \in a'$. Of course any other fixed reference cell could be used instead of r. It will be convenient to denote $f(i_a r_{a'})$ by $f_a^r(i_a)$.

EXAMPLE 3.1. Let $s = 4$ and $a = \{1, 2\}$. When w is the uniform weight function the transformation $f \to T_a f$ replaces f_{ijkl} by $f_{ij\cdot\cdot}$ where, as usual, $\cdot$ denotes uniform average. When w is the substitution weight function f_{ijkl} is replaced under T_a by $f_{ijr_3r_4}$. □

Much of the theory in this paper is obtained using only the simple algebraic equipment of averaging operators. The same ground may be covered using sums and products of linear subspaces and their orthogonal projections. Little will be said about this approach here because it is part of this paper's aim to demonstrate the feasibility of the more elementary approach. It will suffice to show that T_a is an orthogonal projection operator.

We have already noted that $T_a^2 = \mathrm{T}_a$ and so T_a is a projection operator. Since $T_a f = f$ iff $f(i) = \lambda(i_a)$ it follows that T_a projects onto the subspace Ω_a of Ω defined by this property. Further, T_a is self-adjoint with respect to $\langle \cdot, \cdot \rangle_w$ since

$$\begin{aligned} \langle f, T_a g \rangle_w &= \sum_i w(i) f(i) [\sum_{i_{a'}} w_{a'}(i_{a'}) g(i)] \\ &= \sum_{i_a} w_a(i_a) [\sum_{i_{a'}} (i_{a'}) f(i)] [\sum_{i_{a'}} w_{a'}(i_{a'}) g(i)] = \langle T_a f, g \rangle_w. \end{aligned}$$

Finally, T_a is orthogonal with respect to $\langle \cdot , \cdot \rangle_w$ because $\langle (I - T_a)f, T_a f\rangle_w = \langle T_a(I - T_a)f, f\rangle_w = \langle 0, f\rangle_w = 0$.

3.2 *Zero generalised interaction.* Given a generating class $\mathscr{A}$ of subsets of S, define the *generalised $\mathscr{A}$-interaction operator* $I - T_{\mathscr{A}}$ by

$$I - T_{\mathscr{A}} = \Pi_{a\in\mathscr{A}}(I - T_a). \tag{3.2}$$

By (3.1) the terms on the right-hand side of (3.2) can be multiplied together in any order and so, on expanding it, we find

$$T_{\mathscr{A}} = \sum_a T_a - \sum_{a\neq b}\sum T_{ab} + \cdots \mp T_{\cap\mathscr{A}} \tag{3.3}$$

where the sums are over all $a \in \mathscr{A}$, distinct pairs $a, b \in \mathscr{A}$, etc. Another useful expression for $T_{\mathscr{A}}$ results from ordering the elements of $\mathscr{A}$ as $a_1, a_2, \cdots, a_m$, namely

$$T_{\mathscr{A}} = T_{a_1} + (I - T_{a_1})T_{a_2} + \cdots + \Pi_{l<m}(I - T_{a_l})T_{a_m}. \tag{3.4}$$

PROPOSITION 3.1. *The function f satisfies $M_{\mathscr{A}}$ if and only if*

$$T_{\mathscr{A}} f = f. \tag{3.5}$$

PROOF. If f satisfies $M_{\mathscr{A}}$ then for some functions $\{\lambda_a : a \in \mathscr{A}\}$ we can write $f = \sum_{a\in\mathscr{A}}\lambda_a$. Now $(I - T_a)\lambda_a = 0$ for each $a \in \mathscr{A}$, and so it follows that $\Pi_{a\in\mathscr{A}}(I - T_a)\sum_{a\in\mathscr{A}}\lambda_a = 0$; that is, $T_{\mathscr{A}} f = f$.

Conversely, if $T_{\mathscr{A}} f = f$ then, by (3.4),

$$f = T_{a_1} f + (I - T_{a_1})T_{a_2} f + \cdots + \Pi_{l<m}(I - T_{a_l})\cdot T_{a_m} f$$

which is of the form $\sum_{a\in\mathscr{A}}\lambda_a$. □

Since the $\{T_a\}$ are orthogonal projections onto the subspaces $\{\Omega_a\}$, it follows that $T_{\mathscr{A}}$ is the orthogonal projection onto $\Omega_{\mathscr{A}} = \sum_{a\in\mathscr{A}}\Omega_a$, although we do not use this fact in what follows.

The proposition formulates $M_{\mathscr{A}}$ as imposing zero generalised $\mathscr{A}$-interaction, in that $(I - T_{\mathscr{A}})f = 0$.

As foreshadowed in Section 2.3 above, when the weight function is completely multiplicative we have an explicit formula for an element satisfying the *additive model* $M_{\mathscr{A}}$ in terms of its $\mathscr{A}$-marginal means, namely

$$\eta = \sum_a \bar{\eta}_a - \sum_{a\neq b}\sum \bar{\eta}_{ab} + \cdots \mp \bar{\eta}_{\cap\mathscr{A}}. \tag{3.6}$$

This result is an immediate consequence of (3.3) as soon as we recall that $T_a\eta = \bar{\eta}_a$. Using the substitution weight function we obtain the following special case of (3.6).

$$\eta(i) = \sum_a \eta^r_a(i_a) - \sum_{a\neq b}\sum \eta^r_{ab}(i_{ab}) + \cdots \mp \eta^r_{\cap\mathscr{A}}(i_{\cap\mathscr{A}}).$$

From $(I - T_{\mathscr{A}})\log p = 0$ when $T_{\mathscr{A}}$ is based upon the substitution weight function, the multiplicative model is seen to be expressible as

$$\frac{p(i)\cdot \prod_{a\neq b}\prod p^r_{ab}(i_{ab}) \quad \cdots}{\Pi_a p^r_a(i_a)\cdot \prod_{a\neq b\neq c}\prod\prod p^r_{abc}(i_{abc}) \quad \cdots}\ [p^r_{\cap\mathscr{A}}(i_{\cap\mathscr{A}})]^{\pm 1} = 1. \tag{3.7}$$

The left-hand side of (3.7) is a *generalised cross product ratio.*

EXAMPLE 3.2. As in Example 2.1 let $\mathscr{A} = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$. Then (3.3) becomes

$$T_{\mathscr{A}} = T_{\{1,2\}} + T_{\{2,3\}} + T_{\{3,1\}} - T_{\{1\}} - T_{\{2\}} - T_{\{3\}} + T_{\phi}.$$

MODELS AND INTERACTIONS

Using the uniform weight function, (3.6) expresses $M_{\mathscr{A}}$ as the familiar

$$\eta_{ijk} = \eta_{ij\cdot} + \eta_{\cdot jk} + \eta_{i\cdot k} - \eta_{i\cdot\cdot} - \eta_{\cdot j\cdot} - \eta_{\cdot\cdot k} + \eta_{\cdot\cdot\cdot}$$

while (3.7) becomes the equally familiar cross-product ratio formulation of no three-dimensional interaction, namely

$$\frac{p_{ijk}p_{ir_2r_3}p_{r_1jr_3}p_{r_1r_2k}}{p_{ijr_3}p_{r_1jk}p_{ir_2k}p_{r_1r_2r_3}} = 1. \qquad \square$$

Alternative formulations of the *Lancaster-linear models* $M_{\mathscr{A}}$: $p/q \in \Omega_{\mathscr{A}}$ and $P/Q \in \Omega_{\mathscr{A}}$, will now be given. First choose $w_\sigma = p_\sigma$. Then

$$T_a \frac{p(i)}{q(i)} = \sum_{i_{a'}} q_{a'}(i_{a'}) \frac{p(i)}{q(i)} = \frac{1}{q_a(i_a)} \sum_{i_{a'}} p(i) = \frac{p_a(i_a)}{q_a(i_a)}.$$

Applying (3.3) the Lancaster-additive Model I is seen to be expressible was

$$\frac{p}{q} = \sum_a \frac{p_a}{q_a} - \sum_{a\neq b}\sum \frac{p_{ab}}{q_{ab}} + \cdots \mp \frac{p_{\cap\mathscr{A}}}{q_{\cap\mathscr{A}}}. \tag{3.8}$$

Turning now to the Lancaster-additive model II, let T_a be based on the substitution weight function. Then

$$T_a \frac{P(i)}{Q(i)} = \frac{P(i_a r_{a'})}{Q(i_a r_{a'})} = \frac{P_a(i_a)}{Q_a(i_a)}.$$

Consequently the model here is

$$\frac{P}{Q} = \sum_a \frac{P_a}{Q_a} - \sum_{a\neq b}\sum \frac{P_{ab}}{Q_{ab}} + \cdots \mp \frac{P_{\cap\mathscr{A}}}{Q_{\cap\mathscr{A}}}. \tag{3.9}$$

It is now easy to see that the two Lancaster-additive models are equivalent. After multiplication of (3.8) by $q(i)$ and (3.9) by $Q(i)$, each term in (3.9) is seen to be the distribution function of the corresponding term in (3.8).

3.3 *Marginals and generalised interactions.* A by-product of the model characterisations of 2.2 above is that, given $f \in \Omega_{\mathscr{A}}$, where f is η, $\log p$, p/q or P/Q, f is uniquely determined by its $\mathscr{A}$-marginals, suitably interpreted as weighted means or unweighted sums. This is a special case of the result which we now prove that given its $\mathscr{A}$-marginals and its generalised $\mathscr{A}$-interaction, f is uniquely determined.

There is almost nothing in the proof for η, p/q, P/Q. Thus, defining $T_{\mathscr{A}}$ with respect to any completely multiplicative weight function w, we can write $\eta = T_{\mathscr{A}}\eta + (I - T_{\mathscr{A}})\eta$ as the sum of the expansion (3.6), involving its $\mathscr{A}$-marginals, and its generalised $\mathscr{A}$-interaction. Similarly for p/q, except that we now define $T_{\mathscr{A}}$ with respect to $w = q$ and use (3.8), and for P/Q where the substitution operators are used.

There is no explicit demonstration of this uniqueness result for $\log p$ and it has to be proved using Lemma 2.2. Let us suppose that p is a positive probability and that $(I - T_{\mathscr{A}})\log p = u$. Define $\pi = k \exp u$ where k is the normalising constant making $\sum_i \pi(i) = 1$. Then $T_{\mathscr{A}}\log p/\pi = T_{\mathscr{A}}(\log p - \log k - u) = T_{\mathscr{A}}\log p - \log k = \log p - \log k - u$ by the definition of u and the fact that $T_{\mathscr{A}}u = 0$. But this means that $\log p/\pi \in \Omega_{\mathscr{A}}$ and by Lemma 2.2 there is only one p with this property having given $\mathscr{A}$-marginal sums, provided only that these marginals admit a positive probability.

A postscript on this result is the following: it does not matter which (completely multiplicative) weight function w is used to define the generalised $\mathscr{A}$-interaction function $(I - T_{\mathscr{A}})f$ because $(I - T_{\mathscr{A}})f$ defined with respect to one weight function is recoverable from $(I - T_{\mathscr{A}})f$ defined with respect to another. For, if $\{T_a\}$ and $\{\tilde{T}_a\}$ are defined with respect to w and $\tilde{w}$, we see from $\tilde{T}_{\mathscr{A}}T_a = T_a$, $a \in \mathscr{A}$, and (3.3) that $\tilde{T}_{\mathscr{A}}T_{\mathscr{A}} = T_{\mathscr{A}}$, i.e. that

$$(I - \tilde{T}_{\mathscr{A}})(I - T_{\mathscr{A}})f = (I - \tilde{T}_{\mathscr{A}})f.$$

Incidentally, this identity shows directly why $T_{\mathscr{A}}f = f$ iff $\tilde{T}_{\mathscr{A}}f = f$, a fact implicit in Proposition 3.1.

4. Interactions.

4.1 *Interaction operators.* In the previous section we saw that, given a weight function w and averaging operators T_a, the operators $T_{\mathscr{A}}$ and $I - T_{\mathscr{A}}$ arise naturally from consideration of the model $M_{\mathscr{A}}$. In the particular case $\mathscr{A} = \{S - \sigma: \sigma \in S\}$ the operator $I - T_{\mathscr{A}} = \Pi_{\sigma\in S}(I - T_{S-\sigma})$ will be denoted by U_S and called the S-interaction operator. Thus

$$U_S = \Pi_{\sigma\in S}(T_S - T_{S-\sigma}).$$

The definition is now extended to cover any subset b of S. Define $U_\phi = T_\phi$ and, otherwise

$$U_b = \Pi_{\sigma\in b}(T_b - T_{b-\sigma}).$$

The operator U_b will be called the b-interaction operator. Alternative ways of writing it are easily seen to be

$$U_b = \Pi_{\sigma\in b}(I - T_{b-\sigma}).\ T_b, \tag{4.1}$$

$$U_b = \Pi_{\sigma\in b}(I - T_{S-\sigma}).\ \Pi_{\sigma\in b'}T_{S-\sigma}, \tag{4.2}$$

$$U_b = \textstyle\sum_{c\subseteq b}(-1)^{|b-c|}T_c. \tag{4.3}$$

EXAMPLE 4.1. Again let $s = 3$. The interaction operator $U_{\{1,2,3\}}$ is identical to the operator $I - T_{\mathscr{A}}$, with $\mathscr{A} = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$, discussed in Example 3.2. The interaction operator $U_{\{1,2\}}$ is expressible in various ways as

$$\begin{aligned} U_{\{1,2\}} &= (I - T_{\{2\}})(I - T_{\{1\}})T_{\{1,2\}} = (I - T_{\{2,3\}})(I - T_{\{1,3\}})T_{\{1,2\}} \\ &= T_{\{1,2\}} - T_{\{1\}} - T_{\{2\}} + T_\phi. \end{aligned}$$

Thus, for the uniform weight function,

$$(U_{\{1,2\}}\eta)_{ijk} = \eta_{ij\cdot} - \eta_{i\cdot\cdot} - \eta_{\cdot j\cdot} + \eta_{\cdots}. \quad \square$$

Interactions are usually introduced recursively and their recursive structure is clearly seen in the interaction operators. For example, when $s = 3$,

$$U_{\{1,2,3\}} = (I - T_{\{2,3\}})(I - T_{\{1,3\}}) - (I - T_{\{2,3\}})(I - T_{\{1,3\}})T_{\{1,2\}}.$$

The second term on the right side is $U_{\{1,2\}}$ and gives $\{1, 2\}$ interactions averaged over k. The first term gives $\{1, 2\}$ interactions within each level k. Thus $\{1, 2, 3\}$ interactions are clearly seen to be differences of $\{1, 2\}$ interactions.

Some basic results about interaction operators are collected together in the following lemma.

LEMMA 4.1 (i) $T_aU_b = 0$ *if* $b \not\subset a$.

(ii) $\sum_{i_\sigma} w_\sigma(i_\sigma)U_bf(i) = 0$ *if* $\sigma \in b$.

(iii) $T_aU_b = U_b$ *if* $b \subseteq a$.

(iv) $\sum_{b\subseteq a} U_b = T_a$.

(v) $U_b^2 = U_b$.

(vi) $U_aU_b = 0$ *if* $a \neq b$.

(vii) *Let* $b_1, \cdots, b_m$ *be distinct sets. Then*

$\sum_j k_jU_{b_j}f = 0$ *implies that* $k_jU_{b_j}f = 0$ *for all* j.

(viii) U_b *is self-adjoint with respect to the inner product* $\langle \cdot, \cdot\rangle_w$.

PROOF (i) Choose $\tau \in b - a$. Since $T_a = \Pi_{\sigma \in a'} T_{S-\sigma}$, it follows that $T_{S-\tau}(I - T_{S-\tau})$ is a factor of $T_a U_b$.

(ii) By (i) $T_{S-\sigma} U_b = 0$, $\sigma \in b$.

(iii) Apply (4.1) and (3.1).

(iv) First consider $a = S$. By (4.2) $\sum_{b \subseteq S} U_b = \sum_{b \subseteq S} [\Pi_{\sigma \in b}(I - T_{S-\sigma}) \Pi_{\sigma \in b'} T_{S-\sigma}]$ $= \Pi_{\sigma \in S}[(I - T_{S-\sigma}) + T_{S-\sigma}] = \Pi_{\sigma \in S} I = I$. Having established that $\sum_{b \subseteq S} U_b = I$, we now multiply by T_a to get $\sum_{b \subseteq S} T_a U_b = T_a$. Application of (i) and (iii) now gives (iv).

(v) U_b is a product of idempotent operators which commute and hence is itself idempotent.

(vi) Choose $\tau \in (b - a) \cup (a - b)$ and reason as in the proof of (i).

(vii) Multiply $\sum_j k_j U_b f = 0$ by U_b and apply (v) and (vi).

(viii) By (4.3) U_b is a linear combination of operators which are self-adjoint. □

We note that U_b is an orthogonal projection operator because it is idempotent and self-adjoint. Further $U_b f(i) = g(i_b)$ say and for each $\sigma \in b$, $\sum_{i_\sigma} w_\sigma^*(i_\sigma) g(i_b) = 0$. Moreover, if f is a function satisfying (a) $f(i) = h(i_b)$ and (b) $\sum_{i_\sigma} w_\sigma(i_\sigma) f(i) = 0$ for all $\sigma \in b$, then, by (4.1), $U_b f = f$. Thus U_b is the orthogonal projection operator onto the subspace Θ_b of all functions satisfying (a) and (b), although we will not use this interpretation in the sequel.

4.2 *Hierarchical no-interaction models.* Let the *closure* $\bar{\mathscr{A}}$ of a generating class $\mathscr{A}$ be defined by

$$\bar{\mathscr{A}} = \{b: b \subseteq a \quad \text{for some} \quad a \in \mathscr{A}\}.$$

The complement of $\bar{\mathscr{A}}$ is

$$\bar{\mathscr{A}}' = \{b: b \not\subseteq a \quad \text{for all} \quad a \in \mathscr{A}\}.$$

Note that the class $\bar{\mathscr{A}}'$ is hierarchical. That is, if $b_1 \in \bar{\mathscr{A}}'$ and $b_2 \supseteq b_1$, then $b_2 \in \bar{\mathscr{A}}'$.

PROPOSITION 4.1. $T_{\mathscr{A}} = \sum_{b \in \bar{\mathscr{A}}} U_b$.

PROOF. It is easier to prove that

$$I - T_{\mathscr{A}} = \sum_{b \in \bar{\mathscr{A}}'} U_b \tag{4.4}$$

from which the proposition follows. But this is a direct consequence of our definitions and Lemma 4.1. For

$$\begin{aligned} I - T_{\mathscr{A}} &= \Pi_{a \in \mathscr{A}}(I - T_a) \text{ by the definition (3.2)} \\ &= \Pi_{a \in \mathscr{A}}(\textstyle\sum_{b \not\subseteq a} U_b) \text{ by (iv) of Lemma 4.1} \\ &= \textstyle\sum_{b \in \bar{\mathscr{A}}'} U_b \text{ by (v) and (vi) of Lemma 4.1,} \end{aligned}$$

and the definition of $\bar{\mathscr{A}}'$. □

Thus the model $M_{\mathscr{A}}$ for f may now be expressed as

$$f(i) = \sum_{b \in \bar{\mathscr{A}}} U_b f(i) \tag{4.5}$$

or as

$$U_b f(i) = 0 \quad \text{for all} \quad b \in \bar{\mathscr{A}}'. \tag{4.6}$$

Formula (4.5) follows immediately from Proposition 4.1 and formula (4.6) by application of (iv) with $a = S$ and (vii) of Lemma 4.1. By virtue of (4.6), $M_{\mathscr{A}}$ may be called a *hierarchical no-interaction model.* Proposition 4.1 thus provides the link with the more common approach to models and interactions which starts with interactions and then defines models by requiring that a hierarchical set of interactions are zero.

Models with equal sized generating sets are frequently used in searches for parsimonious fits to data and, for such models, there is a simple formula relating $T_{\mathscr{A}}$ to $\{T_b: b \in \bar{\mathscr{A}}\}$.

EXAMPLE 4.2. Let $s = 5$ and consider $\mathscr{A} = \{12, 13, 14, 15, 23, 24, 25, 34, 35, 45\}$ where 12 denotes $\{1, 2\}$ etc. We shall prove that

$$T_{\mathscr{A}} = [T_{12} + \cdots + T_{45}] - 3[T_1 + \cdots + T_5] + 6T_\phi. \quad \square$$

The general result is given in Proposition 4.2 below. It really belongs in Section 3 but its proof uses results of this section.

PROPOSITION 4.2. *For $0 \le t < s$ let*

$$\mathscr{A} = \mathscr{A}_t^s = \{a \subseteq S: |a| = t\}.$$

Then

$$T_{\mathscr{A}} = \sum_{u=0}^{t} (-1)^{t-u} \binom{s-u-1}{t-u} \sum_{b:|b|=u} T_b.$$

PROOF.

$$\begin{aligned} T_{\mathscr{A}} &= \sum_{b:|b|\le t} U_b = \sum_{b:|b|\le t} \prod_{\sigma\in b}(I - T_{S-\sigma}) \prod_{\sigma\in b'} T_{S-\sigma} \\ &= \sum_{u=0}^{t} \text{coefficient of } z^u \text{ in } \prod_{\sigma\in S}[z(I - T_{S-\sigma}) + T_{S-\sigma}] \\ &= \text{coefficient of } z^t \text{ in } (1-z)^{-1} \prod_{\sigma\in S}[zI + (1-z)T_{S-\sigma}] \\ &= \text{coefficient of } z^t \text{ in } (1-z)^{-1} \sum_{u=0}^{s} z^u (1-z)^{s-u} \sum_{b:|b|=u} T_b \\ &= \text{coefficient of } z^t \text{ in } \sum_{u=0}^{t} z^u (1-z)^{s-u-1} \sum_{b:|b|=u} T_b \\ &= \sum_{u=0}^{t} (-1)^{t-u} \binom{s-u-1}{t-u} \sum_{b:|b|=u} T_b. \quad \square \end{aligned}$$

4.3 *Dimensions of models.* Let us denote the rank of a linear operator T by $r(T)$, and the dimension of a subspace ω of Ω by dim ω. The following are immediate consequences of the relevant definitions.

$$\dim \Omega_a = r(T_a) = \prod_{\sigma\in a} r_\sigma.$$

$$\dim \Theta_b = r(U_b) = \prod_{\sigma\in b}(r_\sigma - 1).$$

Our next result is an immediate consequence of Propositions 3.1, 4.1, and the linearity of trace, as soon as we recall that $r(P) = \text{trace } (P)$ for a projection operator P.

PROPOSITION 4.3. (i) *For a generating class $\mathscr{A}$*

$$\dim \Omega_{\mathscr{A}} = \sum_a \prod_{\sigma\in a} r_\sigma - \sum\sum_{a\ne b} \prod_{\sigma\in ab} r_\sigma + \cdots \mp \prod_{\sigma\in\cap\mathscr{A}} r_\sigma = \sum_{b\in\bar{\mathscr{A}}} \prod_{\sigma\in b}(r_\sigma - 1).$$

(ii) *For any t satisfying $0 \le t < |S| = s$*

$$\dim \Omega_{\mathscr{A}_t^s} = \sum_{u=0}^{t} (-1)^{t-u} \binom{s-u-1}{t-u} \sum_{b:|b|=u} \prod_{\sigma\in b} r_\sigma.$$

4.4 *Discussion.* As an illustration of the use of the interaction operators with the *additive model* $M_{\mathscr{A}}$, consider the following simple method of deriving the least-squares estimates of the interactions $U_b\eta(i)$ of η when we have data $y = (y(i,j): j = 1, \cdots, n(i), i \in \mathscr{I})$ with $n(i)$ observations made on cell i, and the cell frequencies $n(i)$ are proportional, that is, completely multiplicative

$$n(i) = \prod_\sigma n_\sigma(i_\sigma)/N^{s-1} \tag{4.7}$$

where $N = \sum_i n(i)$. Condition (4.7) is of course most likely to be realised in practice when $n(i)$ is constant. If we denote the mean of $y(i, j)$ over j by $y(i)$ then the sum of squared deviations of the observations from their expectations is $\sum_{i,j} (y(i, j) - \eta(i))^2 = \sum_{i,j} (y(i, j) - y(i))^2 + \sum_i n(i)(y(i) - \eta(i))^2$, so that the least value has to be found of

$$\sum_i n(i)(y(i) - \eta(i))^2 = \sum_{b \subseteq S} \sum_{i_b} n_b(i_b)(U_b y(i) - U_b \eta(i))^2. \tag{4.8}$$

Identity (4.8) follows from the calculation

$$\langle z, z \rangle_w = \langle z, \sum_b U_b z \rangle_w = \sum_b \langle z, U_b^2 z \rangle_w = \sum_b \langle U_b z, U_b z \rangle_w,$$

using Lemma 4.1 (iv), (v) and (viii), with $z = y - \eta$ and $w(i) = n(i)/N$. Identity (4.8) shows that, for any no-interaction model (hierarchical or not), the least squares estimate of $U_b \eta$ is $U_b y$ (when the cell frequencies are proportional) for every model in which this interaction is not assumed zero.

Consider now the *multiplicative model* $M_{\mathscr{A}}$, i.e. $\log p \in \Omega_{\mathscr{A}}$. Two particular weight functions have been widely used in the literature. Since Birch's (1963) paper, most authors have used the uniform weight function. In this case

$$U_b \log p(i) = \sum_{c \subseteq b} (-1)^{|b-c|} \log p_c^*(i_c)$$

where $p_c^*(i_c)$ is the *geometric mean* of all $p(j)$ for which $j_c = i_c$, and we do not find these interactions easy to interpret. The system of interactions based upon the substitution weight function does seem easier to interpret with multiplicative models and has been used to effect by Plackett (1974). It was introduced by Mantel (1966), and is used more generally in GLIM, see Baker and Nelder (1978). Here

$$U_b \log p(i) = \sum_{c \subseteq b} (-1)^{|b-c|} \log p_c^r(i_c),$$

which is the logarithm of a cross product ratio. Thus if $d = 3$ and $b = \{1, 2\}$, the cross product ratio is

$$\frac{p(i, j, r_3) p(r_1, r_2, r_3)}{p(i, r_2, r_3) p(r_1, j, r_3)}.$$

Referring back to Section 2.3 above, we now turn to what may be called the estimated model interactions $U_b \log \hat{p}$, $b \in \bar{\mathscr{A}}$, where $\hat{p}$ is the maximum likelihood estimate of p under $M_{\mathscr{A}}$. No matter which w is chosen, $U_b \log \hat{p}$ does not share the attractive properties of $U_b \hat{\eta}$ when the cell frequencies are proportional, properties which stem from the equation $U_b \hat{\eta} = U_b y$. Thus $U_b \log \hat{p}$ does not depend only on the b-marginal table n_b of $n = (n(i) : i \in \mathscr{I})$ but, in general, on all $\mathscr{A}$-marginals. (An important exception occurs when the generating class is decomposable; see Haberman (1974), Darroch, Lauritzen and Speed (1980) and Lauritzen, Speed and Vijayan (1978).) Also it changes each time a different model (that is, a different $\mathscr{A}$) is fitted. This is one of the most important differences between the additive and multiplicative models. Of course when b is one of the maximal elements of $\mathscr{A}$, that is $b \in \mathscr{A}^*$, then $U_b \log \hat{p}$ can be put to use since its magnitude, relative to its standard deviation, indicates whether or not the model obtained from $M_{\mathscr{A}}$ by putting $U_b \log p = 0$ is likely to be acceptable; see Baker and Nelder (1978).

Finally we consider the implications of Proposition 4.1 for *Lancaster-additive models.* Using the weight function q (see Section 3.2) the b-interaction for model I is

$$U_b \frac{p}{q} = \sum_{c \subseteq b} (-1)^{|b-c|} \frac{p_c}{q_c}, \tag{4.9}$$

and using the substitution weight function the b-interaction for model II is

$$U_b \frac{P}{Q} = \sum_{c \subseteq b} (-1)^{|b-c|} \frac{P_c}{Q_c}. \tag{4.10}$$

It is easy to see (cf. Section 3.2) that the two definitions of no b-interaction obtained from

(4.9) and (4.10) are equivalent to each other and to Lancaster's (1969, page 256) definition, namely

$$\Pi_{a\in b}(P_a^*(i_a) - P_a(i_a)) = 0, \tag{4.11}$$

where the P_a^* are artificial functions multiplied according to the rule

$$\Pi_{a\in c}P_a^*(i_a) = P_c(i_c).$$

Zentgraf (1975) proved that, if (4.11) holds for all b with $|b| > t$, then

$$P(i) = \sum_{u=0}^{t} (-1)^{t-u}\binom{s-u-1}{t-u} \sum_{b:|b|=u} P_b(i_b)Q_{b'}(i_{b'}). \tag{4.12}$$

This result, when combined with its converse, amounts to a special case of Proposition 4.2 above.

4.5 *A uniqueness property of interactions.* The main purpose of this paper has been to show that many general properties linking models and interactions can be easily stated and proved using interaction operators. We have seen that given any model $M_{\mathscr{A}}$ and any multiplicative weight function w there corresponds a generalized interaction operator $T_{\mathscr{A}}$, that the interaction operators U_b provide a useful way of partitioning $T_{\mathscr{A}}$ and, finally, that $M_{\mathscr{A}}$ has the "hierarchical no-interaction" property by which it is usually characterised.

We conclude by returning to a question raised in Section 2.1, namely: given that f satisfies $M_{\mathscr{A}}$, to what extent are the parameters $\lambda_a(i_a)$ uniquely determined by f? The answer, as shown in the following proposition, is that interactions and only interactions of λ_a are uniquely determined.

PROPOSITION 4.4. *Assume*

$$f(i) = \sum_{a\in\mathscr{A}} \lambda_a(i_a) \tag{4.13}$$

and let $c \in \mathscr{A}$. The extent to which λ_c is determined by f is defined by the equations

$$U_b\lambda_c(i_c) = U_bf(i) \quad \text{for all} \quad b \in \bar{\mathscr{A}} - \bar{\mathscr{C}} \tag{4.14}$$

where $\mathscr{C} = \mathscr{A} - \{c\}$ and where the U_b are defined with respect to any multiplicative weight function.

PROOF. Since

$$f(i) - \lambda_c(i_c) = \sum_{a\in\mathscr{C}} \lambda_a(i_a)$$

therefore

$$U_b(f(i) - \lambda_c(i_c)) = 0 \quad \text{for all} \quad b \in \bar{\mathscr{C}}'.$$

However $U_bf(i) = U_b\lambda_c(i_c) = 0$ for all $b \in \bar{\mathscr{A}}'$. Thus, given (4.13), λ_c certainly satisfies (4.14).

We now prove that equations (4.14) define *all* that is uniquely determined about λ_c from a knowledge of f. This is done by showing that the information about λ_c contained in (4.14) is sufficient for us to construct a λ_c, λ_c^* say, such that

$$f(i) = \lambda_c^*(i_c) + \sum_{a\in\mathscr{C}} \lambda_a(i_a).$$

Simply define

$$\lambda_c^* = \sum_{b\in\bar{\mathscr{A}}-\bar{\mathscr{C}}} U_bf.$$

Then

$$f(i) - \lambda_c^*(i_c) = \sum_{b\in\bar{\mathscr{C}}} U_bf(i)$$

and, by Proposition 4.1, the right side can be written in the form

$$\sum_{a \in \mathscr{C}} \lambda_a(i_a). \quad \square$$

EXAMPLE 2.2 (continued). We have $s = 3$ and $\mathscr{A} = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}$. Let $c = \{2, 3\}$ so that $\bar{\mathscr{A}} - \bar{\mathscr{C}} = \{\{2, 3\}\}$. Using the substitution weight function for convenience, we find that the total information about the marriage rates ρ_{jk} that can be determined from a knowledge of the expected numbers of marriages μ_{ijk} is contained in the equations

$$\frac{\rho_{jk}\rho_{r_2r_3}}{\rho_{jr_3}\rho_{r_2k}} = \frac{\mu_{r_1jk}\mu_{r_1r_2r_3}}{\mu_{r_1jr_3}\mu_{r_1r_2k}}.$$

Likewise, all that can be determined about the numbers M_{ij} of eligible women is contained in the equations

$$\frac{M_{ij}M_{r_1r_2}}{M_{ir_2}M_{r_1j}} = \frac{\mu_{ijr_3}\mu_{r_1r_2r_3}}{\mu_{ir_2r_3}\mu_{r_1jr_3}}. \quad \square$$

Acknowledgement. We are grateful to the referees and an Associate Editor for their helpful comments.

REFERENCES

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

BAKER, R. J. and NELDER, J. A. (1978). *The GLIM System, Release* 3. *Generalised Linear Interactive Modelling.* The Numerical Algorithms Group, Oxford.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families.* Wiley, Chichester.

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. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* M.I.T. Press, Cambridge, Mass.

DARROCH, J. N. (1974). Multiplicative and additive interactions in contingency tables. *Biometrika* **61** 207–214.

DARROCH, J. N., LAURITZEN, S. L. and SPEED, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8** 522–539.

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

DAVIDSON, R. (1973). Determination of confounding. *Stochastic Analysis.* Edited by D. G. Kendall and E. P. Harding. Wiley, New York.

FIENBERG, S. E. (1977). *The Analysis of Cross-Classified Data.* M.I.T. Press, Cambridge, Mass.

GOKHALE, D. V. and KULLBACK, S. (1978). *The Information in Contingency Tables.* Dekker, New York.

GOOD, I. J. (1958). The interaction algorithm and practical Fourier analysis. *J. Roy. Statist. Soc. Ser. B* **20** 361–372.

GOOD, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contengency tables. *Ann. Math. Statist.* **34** 911–934.

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

GRIZZLE, J. E., STARMER, C. F. and KOCH, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25** 489–504.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data.* Univ. of Chicago Press, Chicago, Il.

HABERMAN, S. J. (1975). Direct products and linear models for complete factorial tables. *Ann. Statist.* **3** 314–333.

HABERMAN, S. J. (1978). *Analysis of Qualitative data,* Vol. 1. Academic, New York.

HABERMAN, S. J. (1979). *Analysis of Qualitative data,* Vol. 2. Academic, New York.

KU, H. H. and KULLBACK, S. (1968). Interaction in multidimensional contingency tables: an information theoretic approach. *J. Res. Nat. Bur. of Stand.* **72B** 159–199.

KURKJIAN, B., and ZELEN, M. (1962). A calculus for factorial arrangements. *Ann. Math. Statist.* **33** 600–619.

LANCASTER, H. O. (1969). *The Chi-Squared Distribution.* Wiley, New York.

LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1978). *Decomposable Graphs and Hypergraphs.* Preprint No. 9. Institute of Mathematical Statistics, University of Copenhagen.

MANN, H. B. (1949). *Analysis and Design of Experiments.* Dover, New York.
MANTEL, N. (1966). Models for complex contingency tables and polychotomous dosage response curves. *Biometrics* **22** 83–95.
NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalised linear models. *J. Roy. Statist. Soc. Ser.* A **135** 370–384.
PLACKETT, R. L. (1981). *Analysis of Categorical Data*, 2nd Ed. Griffin, London.
ZENTGRAF, R. (1975). A note on Lancaster's definition of higher-order interactions. *Biometrika* **62** 375–378.

SCHOOL OF MATHEMATICAL SCIENCES,
THE FLINDERS UNIVERSITY OF SOUTH AUSTRALIA,
ADELAIDE. S.A. 5042,
AUSTRALIA.

D.M.S., C.S.I.R.O.,
P.O. BOX 1965,
CANBERRA. A.C.T. 2601,
AUSTRALIA.

The Annals of Statistics
1986, Vol. 14, No. 1, 138–150

GAUSSIAN MARKOV DISTRIBUTIONS OVER FINITE GRAPHS

BY T. P. SPEED AND H. T. KIIVERI

CSIRO Division of Mathematics and Statistics, Canberra and Perth, Australia

Gaussian Markov distributions are characterised by zeros in the inverse of their covariance matrix and we describe the conditional independencies which follow from a given pattern of zeros. Describing Gaussian distributions with given marginals and solving the likelihood equations with covariance selection models both lead to a problem for which we present two cyclic algorithms. The first generalises a published algorithm for covariance selection whilst the second is analogous to the iterative proportional scaling of contingency tables. A convergence proof is given for these algorithms and this uses the notion of I-divergence.

1. Introduction. Most modelling of jointly Gaussian (normal) random variables involves the specification of a structure on the mean and the *covariance* matrix K. However, models which specify structure on K^{-1} have also been developed, although they are seemingly less popular. Our interest in this paper focuses on the covariance selection models, introduced by Dempster (1972) and studied by Wermuth (1976a, b), in which certain elements of K^{-1} are assumed to be zero.

In Section 2 we show how zeros in K^{-1} correspond to conditional independence statements and characterise all such statements consequent upon a given pattern of zeros. The characterisation is achieved by associating a simple graph [Behdzad et al. (1979)] with the elements of K^{-1} and providing rules for reading the graph. The results are a direct analogue of those given in Darroch et al. (1980) for contingency table models; see also Speed (1979).

The likelihood equations for covariance selection models lead naturally to a consideration of the problem of finding Gaussian distributions with prescribed margins. The results in Sections 3 and 4 provide a solution to this problem and a general algorithm for constructing the required distributions is given. Two special cases of this algorithm are considered. The first one is a generalisation of an algorithm in Wermuth and Scheidt (1977) whilst the second one has properties analogous to iterative proportional scaling for contingency tables [Haberman (1974)]. The notion of I-divergence [Csiszár (1975)] or discrimination information in the terminology of Kullback (1959), plays an important role in the convergence proof of this algorithm.

Finally, in Section 5 we show how the I-divergence geometry of Csiszár (1975) provides a framework in which both algorithms can be seen to be an iterated sequence of I-projections.

Received November 1983; revised September 1985.

AMS 1980 *subject classifications.* Primary 62F99; secondary 60K35.

Key words and phrases. Conditional independence, Markov property, simple graph, covariance selection, I-divergence geometry.

GAUSSIAN MARKOV DISTRIBUTIONS OVER FINITE GRAPHS

2. Conditional independence for Gaussian random variables. In the following we consider a random vector $\mathbf{X}$ having a Gaussian distribution with mean $\mathbf{0}$ and positive definite covariance matrix K. The components of $\mathbf{X}$ will be indexed by a finite set C and for $a \subset C$ we write $\mathbf{X}_a$ for the subset of the components of $\mathbf{X}$ indexed by a, namely $(X_\gamma\colon\ \gamma \in a)$. The covariance matrix $K = (K(\alpha, \beta)\colon\ \alpha, \beta \in C)$ on C is defined by $K(\alpha, \beta) = \mathbb{E}\{X_\alpha X_\beta\}$, $\alpha, \beta \in C$, where $\mathbb{E}$ denotes expected value. For subsets $a, b \subseteq C$, $K_{a,b} = \{K(\alpha, \beta)\colon\ \alpha \in a, \beta \in b\}$ denotes the cross covariance matrix of $\mathbf{X}_a$ and $\mathbf{X}_b$. When $a = b$ we write K_a instead of $K_{a,a}$. Note that care must be taken to distinguish between K_a^{-1} and $(K^{-1})_a$. The density $p(\mathbf{x})$ of $\mathbf{X}$ is, of course,

$$(1) \qquad p(\mathbf{x}) = (2\pi)^{-|C|/2}(\det K)^{-1/2}\exp\{-\tfrac{1}{2}\mathbf{x}^T K^{-1}\mathbf{x}\}, \qquad x \in \mathbb{R}^{|C|},$$

where $|\cdot|$ denotes the cardinality of the argument. Marginal densities are subscripted by their defining sets, e.g., $p_a(\mathbf{x}_a)$ or simply p_a, refers to the marginal density of $\mathbf{X}_a$, where a is an arbitrary subset of C.

Proposition 1 relates the conditional independence of two components of $\mathbf{X}$ to the structure of K. In the proposition and following we abbreviate the set intersection $a \cap b$ to ab and write $a \setminus b$ for the complement of b in a. The set $C \setminus b$ will be denoted b'.

PROPOSITION 1. *For subsets a, b of C with $a \cup b = C$ the following statements are equivalent.*

(i) $K_{a,b} = K_{a,ab}K_{ab}^{-1}K_{ab,b}$.
(i′) $K_{a\setminus b, b\setminus a} = K_{a\setminus b, ab}K_{ab}^{-1}K_{ab, b\setminus a}$.
(ii) $(K^{-1})_{a\setminus b, b\setminus a} = 0$.
(iii) $\mathbf{X}_a$ *and* $\mathbf{X}_b$ *are conditionally independent given* $\mathbf{X}_{ab}$.

PROOF. (i) and (i′) are easily seen to be equivalent by partitioning the rows of K over $a \setminus b$ and ab and the columns over $b \setminus a$ and ab. By partitioning over $a \setminus b$, $b \setminus a$, and ab, a straightforward use of the expression for the inverse of a partitioned matrix [Rao (1973, page 33)] proves that (i′) is equivalent to (ii). The standard formula (2) for the conditional covariance matrix gives the connection between (iii) and (i′),

$$(2) \qquad \operatorname{cov}(\mathbf{X}_{a\setminus b}, \mathbf{X}_{b\setminus a}|\mathbf{X}_{ab}) = K_{a\setminus b, b\setminus a} - K_{a\setminus b, ab}K_{ab}^{-1}K_{ab, b\setminus a}. \quad \square$$

A useful special case of the above proposition is the following corollary, given by Wermuth (1976a).

COROLLARY 1. *For distinct elements α, β of C, X_α and X_β are conditionally independent given $X_{\{\alpha,\beta\}'}$ iff $K^{-1}(\alpha, \beta) = 0$.*

PROOF. Put $a = C \setminus \{\alpha\} = \{\alpha\}'$ and $b = \{\beta\}'$ in Proposition 1. $\square$

Having shown that zeros in K^{-1} correspond to conditional independence statements we now describe all such statements which follow from a given

pattern of zeros in K^{-1}. To do this we associate a simple undirected graph with the pattern of zeros and then give rules for reading the graph to obtain the independence relations.

To begin, some graph-theoretic notation and definitions are needed; for a general reference see Behdzad et al. (1979). Our simple undirected graph will be denoted by $\mathbf{C} = (C, E(C))$ where C is the vertex set, and $E(\mathbf{C})$ the edge set which consists of unordered pairs of distinct vertices. Pairs of vertices $\{\alpha, \beta\} \in E(\mathbf{C})$ are said to be *adjacent*. A maximal set of (≥ 2) vertices for which every pair is adjacent is called a *clique*. For any vertex γ we write $\partial\gamma = \{\alpha: \{\alpha, \gamma\} \in E(\mathbf{C})\}$ for the set of neighbours of γ. We also write $\bar{\gamma} = \gamma \cup \partial\gamma$.

An important notion is the separation of sets of vertices in $\mathbf{C}$. To define this we first need to define a *chain* which is a sequence $\gamma = \gamma_0, \gamma_1, \ldots, \gamma_m = \beta$ of vertices such that $\{\gamma_l, \gamma_{l+1}\} \in E(\mathbf{C})$ for $l = 0, 1, \ldots, m-1$. If $\gamma_0 = \gamma_m$ the chain is called a *cycle*. Two sets of vertices a, b are said to be separated by a third set d if every chain connecting an $\alpha \in a$ to a $\beta \in b$ intersects d.

The graph $\mathbf{C}$ is said to be *triangulated* [see Lauritzen et al. (1984)] iff all cycles $\gamma_0, \gamma_1, \ldots, \gamma_p = \gamma_0$ of length $p \geq 4$ possess a chord, where a *chord* is an edge connecting two nonconsecutive vertices of the cycle.

Finally, the graph $\tilde{\mathbf{C}}$ complementary to $\mathbf{C}$ has vertex set C and edge set $E(\tilde{\mathbf{C}})$ with the property that $\{\alpha, \beta\} \in E(\tilde{\mathbf{C}})$ iff $\alpha \neq \beta$ and $\{\alpha, \beta\} \notin E(\mathbf{C})$. Example 1 illustrates these ideas.

EXAMPLE 1. The graph $\mathbf{C}$ with vertex set $\{1,2,3,4\}$ and edge set $\{\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{3,4\}\}$ could be depicted as in Figure 1. For this graph the set of neighbours of 1 is $\{2,3,4\}$; the cliques are $\{1,2,3\}$, $\{1,3,4\}$; a chain from $\{2\}$ to $\{4\}$ is 2, 3, 1, 4 and $\{2\}$ is separated from $\{4\}$ by $\{1,3\}$. Figure 2 shows the complementary graph.

As it stands the graph in Figure 1 is triangulated. However, if the edge $\{1,3\}$ were removed we would have the simplest example of a nontriangulated graph.

The characterisation of all conditional independence relations consequent upon a given pattern of zeros in K^{-1} is presented in Proposition 2.

PROPOSITION 2. *Let* $\mathbf{C}$ *be a simple graph with vertex set* C *indexing the Gaussian random variables* $\mathbf{X}$. *Then the following are equivalent.*

(i) $K^{-1}(\alpha, \beta) = 0$ *if* $\{\alpha, \beta\} \notin E(\mathbf{C})$ *and* $\alpha \neq \beta$;

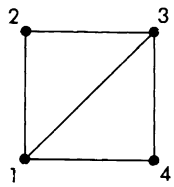


FIG. 1

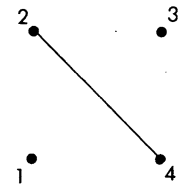


FIG. 2

GAUSSIAN MARKOV DISTRIBUTIONS OVER FINITE GRAPHS

The local Markov property:

(ii) *For every* $\gamma \in C$, X_γ *and* $\mathbf{X}_{\{\gamma\}'}$ *are conditionally independent given* $\mathbf{X}_{\partial\gamma}$;

The global Markov property:

(iii) *For every* a, b *and* d *with* d *separating* a *from* b *in* $\mathbf{C}$, $\mathbf{X}_a$ *and* $\mathbf{X}_b$ *are conditionally independent given* $\mathbf{X}_d$.

PROOF. To show the equivalence of (i) and (ii) we note that (i) is equivalent to $K^{-1}(\gamma, \overline{\{\gamma\}}') = 0$. Putting $a = \overline{\{\gamma\}}$ and $b = \{\gamma\}'$ in Proposition 1 then proves the result.

The equivalence of (i) and (iii) for the case $a \cup b \cup d = C$ follows in a similar way if we put "a" $= a \cup d$ and "b" $= b \cup d$ in Lemma 1. When $a \cup b \cup d \neq C$ a simple maximality argument as in Vorobev (1963) shows that maximal sets a^*, b^* exist such that $a \subseteq a^*$, $b \subseteq b^*$, $a^* \cup b^* \cup d = C$, and a^* is separated from b^* by d. Proposition 1 then gives us $p = p_{a^*} p_{b^*}/p_d$ and integration to obtain the marginal density of $\mathbf{X}_{a \cup b \cup d}$ shows that (i) implies (iii).

The implication in the reverse direction follows on noting that if $(\alpha, \beta) \notin E(\mathbf{C})$ then α, β are separated by $\{\alpha, \beta\}'$. Hence by (iii) X_α and X_β are conditionally independent given $X_{\{\alpha,\beta\}'}$ and Corollary 1 shows that $K^{-1}(\alpha, \beta) = 0$. □

The results of Proposition 2 are illustrated in Example 2.

EXAMPLE 2. Suppose K^{-1} has the following pattern with $*$ denoting a nonzero element:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} * & * & 0 & 0 & 0 \\ * & * & * & 0 & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & * & * & * & * \end{bmatrix} \end{array}.$$

Then the corresponding graph $\mathbf{C}$ would be as shown in Figure 3. If we put $\gamma = \{2\}$, $\partial\gamma = \{1,3,5\}$, and use the local Markov property we deduce that X_2 and X_4 are conditionally independent given $\mathbf{X}_{\{1,3,5\}}$. Similarly with $a = \{1\}$, $b = \{4\}$, and $d = \{2\}$, the global Markov property can be used to assert that X_1 and X_4 are conditionally independent given X_2.

3. Gaussian Markov distributions with prescribed marginals. In this section we consider the problem of finding a Gaussian probability measure with prescribed marginals, i.e., we seek a joint probability density p whose marginals

$$p_{c_1}, \ldots, p_{c_n} \tag{3}$$

are known beforehand, $c_1, \ldots, c_n$ being proper subsets of C. (The notation is explained after (1) above.) Clearly if our marginal specifications are consistent it is necessary to give only the maximal c_i in (3).

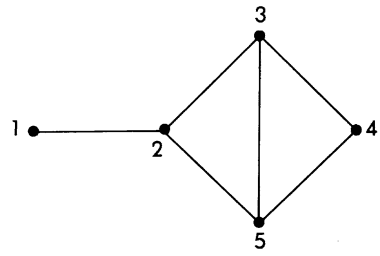


FIG. 3

As motivation for this problem consider the following. Suppose we have n independent and identically distributed observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from (1) and we wish to find a maximum likelihood estimate of K subject to certain elements of K^{-1} being zero. When written in our notation, the likelihood equations for such a model (Dempster, 1972) are:

$$\begin{aligned} K(\alpha, \beta) &= S(\alpha, \beta) \quad \text{if } \{\alpha, \beta\} \in E(\mathbf{C}) \text{ or } \alpha = \beta, \\ K^{-1}(\alpha, \beta) &= 0 \qquad\quad \text{if } \{\alpha, \beta\} \notin E(\mathbf{C}) \text{ and } \alpha \neq \beta, \end{aligned} \tag{4}$$

where $nS = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. The first equation in (4) is easily shown to be equivalent to

$$K_c = S_c \quad \text{if } c \in \mathscr{C}(\mathbf{C}), \tag{4'}$$

where $\mathscr{C}(\mathbf{C})$ is the class of cliques of $\mathbf{C}$. Since a Gaussian distribution with mean zero is completely specified by its covariance matrix, (4') amounts to specifying the marginal distributions p_c for $c \in \mathscr{C}(\mathbf{C})$.

Theorem 1 can be used to describe the class of Gaussian measures with prescribed margins.

THEOREM 1. *Given positive definite matrices L and M defined on the vertices C of a graph $\mathbf{C} = (C, E(\mathbf{C}))$ there exists a unique positive definite matrix K such that*

(i) $K(\alpha, \beta) = L(\alpha, \beta)$ *if* $\{\alpha, \beta\} \in E(\mathbf{C})$ *or* $\alpha = \beta$,
(ii) $K^{-1}(\alpha, \beta) = M(\alpha, \beta)$ *if* $\{\alpha, \beta\} \notin E(\mathbf{C})$ *and* $\alpha \neq \beta$.

Equivalently

(i') $K_c = L_c$ *if* $c \in \mathscr{C}(\mathbf{C})$;
(ii') $K^{-1}(\tilde{c}, \tilde{c})$ *and* $M(\tilde{c}, \tilde{c})$ *agree except on the diagonals,* $\tilde{c} \in \mathscr{C}(\tilde{\mathbf{C}})$.

PROOF. The equivalence of (i) and (i') follows from the relation

$$E(\mathbf{C}) = \bigcup_{c \in \mathscr{C}(\mathbf{C})} \bigcup_{\{\alpha, \beta\} \subseteq c} \{\alpha, \beta\}. \tag{5}$$

Replacing $\mathbf{C}$ by $\tilde{\mathbf{C}}$ in (5) enables the equivalence of (ii) and (ii′) to be demonstrated.

The main result of Theorem 1 can be established using the theory of exponential families [Barndorff-Nielsen (1978), Johansen (1979)] and such a proof is sketched by Dempster (1972, Appendixes A and B).

The results in Section 4 will show how to generate a sequence of matrices converging to the K of Theorem 1 and thus provide an alternative proof. We prefer this proof as it provides a basis for simple numerical algorithms which do not require Newton–Raphson type iterations or storage of large matrices to compute K. □

Replacing the L in Theorem 1 by the sample covariance matrix and setting $M = I$ shows that the estimation problem for covariance selection models has a well defined solution. When $M = I$, the K in Theorem 1 gives the Gaussian distribution with maximum entropy satisfying (i) or (i′) [see Dempster (1972)].

Note that varying the M in Theorem 1 gives the family of distributions with margins prescribed by L_c, $c \in \mathscr{C}(\mathbf{C})$.

In the next section we will make use of the notion of the *I-divergence* of two positive definite matrices. This is defined by

$$\mathscr{I}(P|R) = -\tfrac{1}{2}\{\log\det(PR^{-1}) + \operatorname{tr}(I - PR^{-1})\}. \tag{6}$$

The definition (6) results from evaluating the discrimination information measure of Kullback (1959), namely $\int p(\mathbf{x})\log\{p(\mathbf{x})/r(\mathbf{x})\}\,d\mathbf{x}$ for the two Gaussian distributions with densities $p(\mathbf{x})$, $r(\mathbf{x})$ defined by covariance matrices P, R. When it exists, the I-divergence behaves somewhat like a norm on a space of probability measures (Csiszár, 1975), although it is not.

Some properties of (6) which we will use later are given in Lemma 1. We write $\mathscr{P}$ for the set of $|C| \times |C|$ positive definite matrices and regard this as a (convex) subset of $\mathbb{R}^q$ where $q = |C|^2$. In the following a set of unordered pairs of (not necessarily distinct) elements of C will be denoted by E.

LEMMA 1. *The I-divergence $\mathscr{I}(\cdot|\cdot)$ has the following properties.*

(i) *If $P, R \in \mathscr{P}$, $\mathscr{I}(P|R) \geqq 0$ with equality iff $P = R$.*

(ii) *Given $P, R \in \mathscr{P}$, if there exists a $Q \in \mathscr{P}$ such that*

(a) *$Q(\alpha, \beta) = P(\alpha, \beta)$ if $(\alpha, \beta) \in E$, and*

(b) *$Q^{-1}(\alpha, \beta) = R^{-1}(\alpha, \beta)$ if $(\alpha, \beta) \notin E$, then*

$$\mathscr{I}(P|R) = \mathscr{I}(P|Q) + \mathscr{I}(Q|R). \tag{7}$$

If such a Q exists it is unique.

(iii) *If $\{K_n\}$ and $\{L_n\}$ are sequences contained in compact subsets of $\mathscr{P}$ then $\mathscr{I}(K_n|L_n) \to 0$ implies $K_n - L_n \to 0$.*

PROOF. The first assertion is a well known property of the Kullback information measure so we focus on (ii) and (iii).

(ii) A simple calculation shows that for $Q \in \mathscr{P}$

$$\mathscr{I}(P|Q) + \mathscr{I}(Q|R) = \mathscr{I}(P|R) - \tfrac{1}{2}\operatorname{tr}\{(Q - P)\Delta\}, \tag{8}$$

where $\Delta = Q^{-1} - R^{-1}$. Conditions (a) and (b) then ensure that the trace term in (8) is zero.

To prove uniqueness suppose Q_1 and Q_2 satisfy (a) and (b) of (ii). Then setting $P = R = Q_1$ shows that

$$\mathscr{I}(Q_1|Q_1) = \mathscr{I}(Q_1|Q_2) + \mathscr{I}(Q_2|Q_1)$$

and since I-divergences are positive unless both arguments are equal we must have $Q_1 = Q_2$.

(iii) Suppose $\mathscr{I}(K_n|L_n) \to 0$ but $K_n - L_n \not\to 0$. Then there exist convergent subsequences $K_{n'} \to K$ and $L_{n'} \to L$ with $K \neq L$. By continuity $\mathscr{I}(K_{n'}|L_{n'}) \to \mathscr{I}(K|L) \neq 0$, which is a contradiction. □

4. Algorithms. This section develops two algorithms for constructing the K of Theorem 1. The first algorithm preserves (i′) of Theorem 1 throughout the iterations and cycles through $\tilde{c} \in \mathscr{C}(\tilde{\mathbf{C}})$ forcing the off-diagonal elements of $K^{-1}(\tilde{c}, \tilde{c})$ to zero. The second algorithm preserves (ii′) whilst forcing $K_c = L_c$ as it cycles through $c \in \mathscr{C}(\mathbf{C})$. Both of these algorithms are special cases of a more general cyclic algorithm and we begin by presenting this algorithm. Throughout the discussion $E_1, E_2, \ldots, E_m$ denote sets of unordered pairs of (not necessarily distinct) elements of C whose union is denoted by E.

4.1. A general cyclic algorithm. The general cyclic algorithm is designed to solve the following problem. Given $G, H \in \mathscr{P}$ find an $F \in \mathscr{P}$ with the property that

$$F(\alpha, \beta) = G(\alpha, \beta) \quad \text{if } (\alpha, \beta) \in E, \tag{9}$$

$$F^{-1}(\alpha, \beta) = H(\alpha, \beta) \quad \text{if } (\alpha, \beta) \notin E. \tag{10}$$

The algorithm is defined as follows. Generate a sequence $\{F_n\}$ of positive definite matrices satisfying $F_0 = H^{-1}$ and, for $n \geqq 1$,

$$F_n(\alpha, \beta) = G(\alpha, \beta) \qquad \text{if } (\alpha, \beta) \in E_{n'}, \tag{9′}$$

$$F_n^{-1}(\alpha, \beta) = F_{n-1}^{-1}(\alpha, \beta) \quad \text{if } (\alpha, \beta) \notin E_{n'}, \tag{10′}$$

where $n' = n(\text{mod } m)$. Basically the idea is to maintain (10) throughout the sequence whilst cycling through the E_m and forcing (9). The crucial step in the algorithm involves going from F_{n-1} to F_n. Assuming for the moment that this step can be performed, a convergence proof for this algorithm, modelled upon that found in Csiszár (1975, Theorem 3.2), is given in Proposition 3. The two algorithms to be discussed are examples for which the sequence $\{F_n\}$ can be easily constructed. We write $\mathbb{N}$ for the set of nonnegative integers.

PROPOSITION 3. *The sequence $\{F_n\}$ generated by the general cyclic algorithm converges to the unique $F \in \mathscr{P}$ with the properties* (9) *and* (10).

PROOF. By (ii) of Lemma 1 we can write for $r \geqq 1$

$$\mathscr{I}(G|F_{r-1}) = \mathscr{I}(G|F_r) + \mathscr{I}(F_r|F_{r-1}). \tag{11}$$

Summing relations of the form (11) over r gives for $u \geqq 1$

$$\mathscr{I}(G|F_0) = \mathscr{I}(G|F_u) + \sum_{r=1}^{u} \mathscr{I}(F_r|F_{r-1}) \tag{12}$$

and from (12) we deduce that

$$\{F_n\} \in \{F\colon \mathscr{I}(G|F) \leqq \mathscr{I}(G|F_0)\} = A \quad \text{(say)}. \tag{13}$$

The set A is compact since $\mathscr{I}(G|F)$ is strictly convex (as a function of F^{-1}) with a unique minimum. From (12) it also follows that

$$\sum_{r=1}^{u} \mathscr{I}(F_r|F_{r-1}) \leqq \mathscr{I}(G|F_0). \tag{14}$$

Hence $\sum_{r=1}^{\infty} \mathscr{I}(F_r|F_{r-1})$ is convergent and $\mathscr{I}(F_r|F_{r-1}) \to 0$ as $r \to \infty$.

Now by (13) the vector sequence $\{F_{sm+1}, F_{sm+2}, \ldots, F_{sm+m})\colon s \geqq 0\}$ has a convergent subsequence, defined by $s \in \mathbb{N}_1 \subseteq \mathbb{N}$, with limit $(F_1^*, F_2^*, \ldots, F_m^*)$ say. For any $2 \leqq t \leqq m$ we can write

$$(F_t - F_{t-1}) = (F_t - F_{sm+t}) + (F_{sm+t} - F_{sm+t-1}) + (F_{sm+t-1} - F_{t-1}). \tag{15}$$

Letting $s \in \mathbb{N}_1 \to \infty$ and using (iii) of Lemma 1 with $L_n = K_{n-1}$ shows that $F_1^* = F_2^* = \cdots F_m^* = F$ (say). Note that (10) holds for each F_r and hence for the limit F. Similarly for each $s \in \mathbb{N}_1$ and t, $F_{sm+t}(\alpha, \beta) = G(\alpha, \beta)$ if $(\alpha, \beta) \in E_t$, so the same property holds for the limit F, i.e., (9) holds.

A similar argument for any other convergent subsequence shows that the limit point satisfies (9) and (10) of our proposition. Lemma 1, part (ii) then establishes that all convergent subsequences have the same limit and hence $\{F_n\}$ converges. □

The next lemma enables sequences $\{F_n\}$ satisfying (9′) or (10′) to be constructed when either

$$E_i = \{(\alpha, \beta)\colon \alpha, \beta \in a_i \subseteq C\} \tag{16}$$

or

$$E_i = \{(\alpha, \beta)\colon \alpha, \beta \in a_i \subseteq C, \alpha \neq \beta\}. \tag{17}$$

LEMMA 2. *Suppose Q, R, and $B \in \mathscr{P}$. Then*

(i) *for $a \subseteq C$ the matrix*

$$Q^{-1} = R^{-1} + \begin{bmatrix} B_a^{-1} - R_a^{-1} & 0 \\ 0 & 0 \end{bmatrix} \tag{18}$$

is positive definite and satisfies

(a) $Q(\alpha, \beta) = B(\alpha, \beta)$ *if* $\alpha \in a$ *and* $\beta \in a$; *and*

(b) $Q^{-1}(\alpha, \beta) = R^{-1}(\alpha, \beta)$ *if* $\alpha \notin a$ *or* $\beta \notin a$.

(ii) *The matrix Q is given by*

$$Q = \begin{bmatrix} B_a & B_a R_a^{-1} R_{a,a'} \\ R_{a',a} R_a^{-1} B_a & R_{a'} - R_{a',a} R_a^{-1}(I - B_a R_a^{-1}) R_{a,a'} \end{bmatrix} \tag{19}$$

(iii) *We have the expression*:

$$\mathcal{I}(Q|R) = -\tfrac{1}{2}\{\log\det B_a R_a^{-1} + \operatorname{tr}(I_a - B_a R_a^{-1})\}. \tag{20}$$

PROOF. (i) We use the density scaling of Kullback (1968). In the Gaussian case, given densities $b(\mathbf{x})$ and $r(\mathbf{x})$ corresponding to positive definite matrices B and R, scaling so that $r_a(\mathbf{x}_a)$ agrees with $b_a(\mathbf{x}_a)$ corresponds to computing

$$q(\mathbf{x}) = \frac{r(\mathbf{x}) b_a(\mathbf{x}_a)}{r_a(\mathbf{x}_a)}. \tag{21}$$

Expanding the right-hand side of (21) gives

$$\begin{aligned} q(\mathbf{x}) = (2\pi)^{-|C|/2} &\left(\frac{\det R \det B_a}{\det R_a}\right)^{-1/2} \\ &\times \exp\left\{-\frac{1}{2}\mathbf{x}^T\left[R^{-1} + \begin{pmatrix} B_a^{-1} - R_a^{-1} & 0 \\ 0 & 0 \end{pmatrix}\right]\mathbf{x}\right\}, \end{aligned} \tag{22}$$

which by (18) is just

$$(2\pi)^{-|C|/2}(\det Q)^{-1/2}\exp\{-\tfrac{1}{2}\mathbf{x}^T Q^{-1}\mathbf{x}\}. \tag{23}$$

The properties (a) and (b) are now immediate. A direct proof using matrix algebra can also be given.

The proofs of (ii) and (iii) are straightforward so we omit them. □

The two algorithms discussed below correspond to choosing the a_i in (16) and (17) to be the cliques of $\mathbf{C}$ or $\tilde{\mathbf{C}}$, respectively. In the following we will abbreviate the class of cliques of $\mathbf{C}$ by $\mathscr{C}$ and the class of cliques of $\tilde{\mathbf{C}}$ by $\tilde{\mathscr{C}}$. The notation $\operatorname{diag}(A)$ refers to a diagonal matrix whose diagonals are the same as those of A.

4.2. The first cyclic algorithm. List the cliques of the complementary graph $\tilde{\mathbf{C}}$ as $\tilde{c}_1, \ldots, \tilde{c}_m$ and generate a sequence $\{K_u\}$ as follows: $K_0 = L$; for $s \in \mathbb{N}$, $1 \leqq t \leqq m$, $K_{sm+t} = Z_t(K_{sm+t-1})$, where $Z_t(K) = Q^{-1}$, Q being the matrix (18) of Lemma 2 with $R = K^{-1}$, $a = \tilde{c}_t$, and $B_a = \operatorname{diag}((K^{-1})_a^{-1})^{-1}$. The fact that this sequence converges to the required matrix K when $M = I$ follows from Proposition 3 on replacing a_i in (17) by $\tilde{c}_i$ and making the identifications $F_n = K_n^{-1}$, $G = M$, and $H = L$. It does not seem possible to give an explicit expression for B_a in the case when $M \neq I$.

For this algorithm the elements of the sequence $\{K_n\}$ are fixed over $\mathscr{C}$ whilst the elements of $\{K_n^{-1}\}$ vary over $\tilde{\mathscr{C}}$. From a computational point of view it is not necessary to compute the sequence $\{K_n\}$ by inverting K_n^{-1} at each step. The expression (18) provides a simple updating formula for K_n given K_{n-1}. Hence it

is only necessary to invert $|\tilde{c}| \times |\tilde{c}|$ positive definite matrices when cycling through $\tilde{c} \in \tilde{\mathscr{C}}$.

The cyclic algorithm of Wermuth and Scheidt (1977) is also a special case of the general algorithm. Instead of using the cliques of $\tilde{\mathbf{C}}$ these authors cycle through the edges $\{\alpha, \beta\} \in E(\tilde{\mathbf{C}})$. The 2×2 matrix inversions required are explicitly performed and used to give a simple updating formula. Their algorithm is defined in the same way as above but they have $a \in E(\tilde{\mathbf{C}})$ and

$$B_a = \delta \begin{bmatrix} w^{-1} & 0 \\ 0 & u^{-1} \end{bmatrix},$$

where

$$(K^{-1})_a = \begin{bmatrix} u & v \\ v & w \end{bmatrix}$$

and $\delta = uw - v^2$. It is easily seen that at each step the current value of $K(\alpha, \beta)$ is changed by $-v/\delta$ so that $K^{-1}(\alpha, \beta) = 0$. A computer program for performing the adjustments is given in Wermuth and Scheidt's paper.

4.3. The second cyclic algorithm. Enumerate the cliques of $\mathbf{C}$ as $c_1, c_2, \ldots, c_m$ and define a sequence $\{K_r\}$ as follows: $K_0 = M^{-1}$; for $s \geqq 0$, $1 \leqq t \leqq m$, $K_{sm+t} = Y_t(K_{sm+t-1})$, where $Y_t(K) = Q$, Q being the matrix (6) of Lemma 1 with $R = K$, $a = c_t$, and $B = L$. Making the identifications $a_i = c_i$ in (16) and $F_n = K_n$, $G = L$, and $H = M$ in Proposition 3 shows that the second algorithm converges to the K of Theorem 1. This result also gives an alternative proof of Theorem 1. Note that $\{K_n^{-1}\}$ is held fixed over $\tilde{\mathscr{C}}$ whilst $\{K_n\}$ varies over $\mathscr{C}$.

That this second algorithm is analogous to iterative proportional scaling for contingency tables should be clear. At each step we "scale" the current covariance matrix to match the relevant "margin" L_c. We can also connect this algorithm with a general procedure in Kullback (1968) where, however, the proofs are incomplete. Using our notation, Kullback's procedure can be described as follows. Given the required marginal densities $g_{c_1}, \ldots, g_{c_m}$ and an initial density $\pi(\mathbf{x})$ construct the sequence $\{f_n\}$ (assumed to exist) defined by

$$f_0(\mathbf{x}) = \pi(\mathbf{x}),$$

and for $s \geqq 0$, $1 \leqq t \leqq m$

$$f_{sm+t}(\mathbf{x}) = \frac{f_{sm+t-1}(\mathbf{x}) g_{c_t}(\mathbf{x}_{c_t})}{(f_{sm+t-1})_{c_t}(\mathbf{x}_{c_t})}.$$

Note that this simply amounts to scaling the previous density to ensure the desired marginals and this is how we obtain the matrix Q of Lemma 2. Hence the second cyclic algorithm is a Gaussian version of Kullback's general procedure. It can also be shown to be a cyclic ascent algorithm.

4.4. Finite termination. When the graph $\mathbf{C}$ is triangulated and $M = I$ the second cyclic algorithm converges after one cycle if the cliques are suitably ordered. This result is completely analogous to the one cycle convergence of

iterative proportional scaling for contingency tables when the generating class is decomposable [see Haberman (1974, Chapter 5)].

To demonstrate the result we need the following two lemmas. Without loss of generality we assume that the graph $\mathbf{C}$ is connected.

LEMMA 3. *If* $\mathbf{C}$ *is triangulated then there exists an enumeration* $c_1, \ldots, c_m$ *of the cliques such that for* $i = 2, \ldots, m$

$$c_i \setminus \bigcup_{l=1}^{i-1} c_l \neq \varnothing. \tag{24}$$

PROOF. The result is obtained by successively removing detachable cliques from $\mathbf{C}$ [see Lauritzen et al. (1984)]. □

Note that (24) states that for each i the clique c_i contains a vertex not in c_l for $l = 1, \ldots, i-1$.

The second lemma gives an expression for the determinant of the matrix K in Proposition 1 which is useful in proving the finite termination of the second algorithm.

LEMMA 4. *Suppose* $K \in \mathscr{P}$ *and* $K^{-1}_{a\setminus b,\, b\setminus a} = 0$ *for* a, b *with* $a \cup b = C$. *Then*

$$\det K = (\det K_a)(\det K_b)/\det K_{ab}. \tag{25}$$

PROOF. Note that (iii) of Proposition 1 implies $p = p_a p_b / p_{ab}$. Evaluation at $x = 0$ then gives the result. □

PROPOSITION 4. *If the cliques of* $\mathbf{C}$ *are ordered as in Lemma* 3 *and we start the second cyclic algorithm with* $K_0 = I$, *then*

(i) $(K_m)_c = L_c$ *for* $c \in \mathscr{C}$;
(ii) $(K_m^{-1})_{\tilde{c}}$ *is diagonal for* $\tilde{c} \in \tilde{\mathscr{C}}$.

PROOF. We will prove that $\mathscr{I}(K|K_m) = 0$ where K is the unique matrix of Theorem 1 with $M = I$. This will follow directly from (12) provided we can show that

$$\mathscr{I}(K|I) = \sum_{i=1}^{m} \mathscr{I}(K_i|K_{i-1}) \tag{26}$$

and we prove this by induction on m, the number of cliques. It is clearly true for $m = 1$ and so we assume that it is true for all $m \leqq q$ where $q \geqq 1$. If we can prove

$$\mathscr{I}(K|I) = \mathscr{I}(K_{q+1}|K_q) + \mathscr{I}(K_{\bar{c}}|I_{\bar{c}}), \tag{27}$$

where $\bar{c} = \cup_{i=1}^{q} c_i$, then (26) will follow for $m = q + 1$; q steps of the second algorithm starting from $K_0 = I$ generate matrices having the form

$$K_i = \begin{bmatrix} I & 0 \\ 0 & \tilde{K}_i \end{bmatrix}, \qquad i = 1, \ldots, q,$$

where $\tilde{K}_i$ is $|\bar{c}| \times |\bar{c}|$ and from the inductive hypothesis

$$\mathcal{I}(K_{\bar{c}}|I_{\bar{c}}) = \sum_1^q \mathcal{I}(\tilde{K}_i|\tilde{K}_{i-1}) = \sum_1^q \mathcal{I}(K_i|K_{i-1}).$$

Turning now to the proof of (27) we remark that it follows from Lemma 4 with $a = c_{q+1}$ and $b = \bar{c}$, the relationship (20) with $Q = K_{q+1}$, $R = K_q$, and $a = c_{q+1}$ as before, and the fact that

$$(K_q)_a = \begin{bmatrix} I & 0 \\ 0 & L_{ab} \end{bmatrix}.$$

The log det terms in the definition of $\mathcal{I}$ match up by Lemma 4 and the trace terms correspond by (20) and the fact just noted. □

We conclude this section with a few remarks comparing the two algorithms. When $M = I$, the main drawback of the first algorithm is the need to invert L at the beginning. It is possible that a numerical inversion of L could be difficult or impossible yet the second algorithm would work. This problem aside, it should be clear that the choice of which algorithm is to be favoured in any given situation is very much dependent on the number and sizes of the cliques in $\mathcal{C}$ and $\tilde{\mathcal{C}}$. However, if $\mathbf{C}$ is triangulated and $M = I$, the finite termination property of the second algorithm makes it attractive.

5. Some comments about the geometry. To give a geometric interpretation of the two algorithms it is convenient to define the "subspaces" $\mathcal{P}_{L,c} = \{P \in \mathcal{P}\colon P_c = L_c\}$, $\mathcal{Q}_{M,\tilde{c}} = \{Q \in \mathcal{P}\colon (Q^{-1})_{\tilde{c}}$ agrees with $M_{\tilde{c}}$ except on the diagonal$\}$, and $\mathcal{P}_{L,\mathcal{C}} = \cap\{\mathcal{P}_{L,c}\colon c \in \mathcal{C}\}$, $\mathcal{Q}_{M,\tilde{\mathcal{C}}} = \cap\{\mathcal{Q}_{M,\tilde{c}}\colon \tilde{c} \in \tilde{\mathcal{C}}\}$.

Equation (7) bears a resemblance to Pythagoras' theorem and clearly for all $P \in \mathcal{P}_{L,c}$ we have $\mathcal{I}(P|R) \geq \mathcal{I}(Q|R)$ with equality iff $Q = P$. Hence one can call the matrix Q the I-projection of R on to $\mathcal{P}_{L,c}$ [see Csiszár (1975)].

Viewing the adjustment defined by Q in Lemma 2 as an I-projection we can give an interpretation of the two cyclic algorithms as follows.

The first algorithm begins with a $K_0 \in \mathcal{P}_{L,\mathcal{C}}$ and cycles through $\tilde{c} \in \tilde{\mathcal{C}}$, I-projecting the current estimate of K onto $\mathcal{P}_{L,\mathcal{C}} \cap \mathcal{Q}_{I,\tilde{c}}$ in order to obtain the required element in $\mathcal{P}_{L,\mathcal{C}} \cap \mathcal{Q}_{I,\tilde{\mathcal{C}}}$. The fact that we are I-projecting follows from (ii) of Lemma 1. Using this, for all $K \in \mathcal{Q}_{I,c}$ we have

$$\mathcal{I}(K^{-1}|R^{-1}) = \mathcal{I}(K^{-1}|Q) + \mathcal{I}(Q|R^{-1})$$

or equivalently

$$\mathcal{I}(R|K) = \mathcal{I}(Q^{-1}|K) + \mathcal{I}(R|Q^{-1}),$$

and so $\mathcal{I}(R|K) \geq \mathcal{I}(R|Q^{-1})$ for all $K \in \mathcal{Q}_{I,c}$ with equality iff $K = Q^{-1}$.

For the second algorithm we begin with $K_0 \in \mathcal{Q}_{M,\tilde{\mathcal{C}}}$ and cycle through $c \in \mathcal{C}$, I-projecting the current estimate K onto $\mathcal{Q}_{M,\tilde{\mathcal{C}}} \cap \mathcal{P}_{L,c}$.

Both of the above algorithms are analogous to computing the projection onto the intersection of nonorthogonal (linear) subspaces by successively projecting onto each subspace [see for example von Neumann (1950, Chapter 13)].

Acknowledgment. The referees made many valuable suggestions and are warmly thanked for their contribution.

REFERENCES

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester.

BEHDZAD, M., CHARTRAND, B. and LESNIAK-FOSTER, L. (1979). *Graphs and Digraphs*. Prindle, Weber and Schmidt, Boston.

CSISZÁR, I. (1975). *I*-divergence geometry of probability distributions and minimisation problems. *Ann Probab.* **3** 146–158.

DARROCH, J. N., LAURITZEN, S. L. and SPEED, T. P. (1980). Log-linear models for contingency tables and Markov fields over graphs. *Ann. Statist.* **8** 522–539.

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

HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. Univ. Chicago Press.

JOHANSEN, S. (1979). *Introduction to the Theory of Regular Exponential Families*. Lecture Notes 3. Institute of Mathematical Statistics, Univ. Copenhagen.

KIIVERI, H. T. and SPEED, T. P. (1982). Structural analysis of multivariate data: a review. *In Sociological Methodology 1982* (S. Leinhardt, ed.). Jossey-Bass, San Francisco.

KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

KULLBACK, S. (1968). Probability densities with given marginals. *Ann. Math. Statist.* **39**: 1236–1243.

LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1984). Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. Ser. A* **36** 12–29.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*. 2nd ed. Wiley, New York.

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

VON NEUMANN, J. (1950). *Functional Operators: The Geometry of Orthogonal Spaces* **2**. Princeton Univ. Press.

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

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. AND SCHEIDT, E. (1977). Fitting a covariance selection model to a matrix. Algorithm AS105. *Appl. Statist.* **26** 88–92.

CSIRO
DIVISION OF MATHEMATICS
AND STATISTICS
GPO BOX 1965
CANBERRA, ACT 2601
AUSTRALIA

CSIRO
DIVISION OF MATHEMATICS
AND STATISTICS
PRIVATE BAG, P.O.
WEMBLEY, W.A. 6014
AUSTRALIA