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.

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,

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

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* $\alpha \neq \beta$;

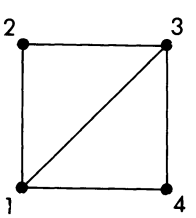

FIG. 1

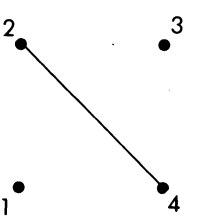

FIG. 2

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).

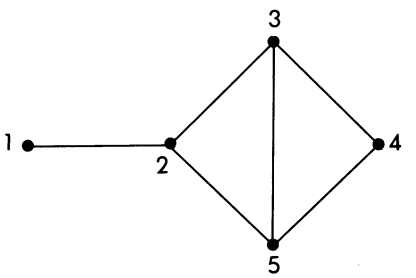

FIG. 3

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 $\bar{c} = \bigcup_{i=1}^{q} c_i$, 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 $\mathscr{C}$ and $\tilde{\mathscr{C}}$. 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" $\mathscr{P}_{L,c} = \{P \in \mathscr{P}: P_c = L_c\}$, $\mathscr{Q}_{M,\tilde{c}} = \{Q \in \mathscr{P}: (Q^{-1})_{\tilde{c}}$ agrees with $M_{\tilde{c}}$ except on the diagonal$\}$, and $\mathscr{P}_{L,\mathscr{C}} = \cap\{\mathscr{P}_{L,c}: c \in \mathscr{C}\}$, $\mathscr{Q}_{M,\tilde{\mathscr{C}}} = \cap\{\mathscr{Q}_{M,\tilde{c}}: \tilde{c} \in \tilde{\mathscr{C}}\}$.

Equation (7) bears a resemblance to Pythagoras' theorem and clearly for all $P \in \mathscr{P}_{L,c}$ we have $\mathcal{I}(P|R) \geqq \mathcal{I}(Q|R)$ with equality iff $Q = P$. Hence one can call the matrix Q the I-projection of R on to $\mathscr{P}_{L,c}$ [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 $K_0 \in \mathscr{P}_{L,\mathscr{C}}$ and cycles through $\tilde{c} \in \tilde{\mathscr{C}}$, I-projecting the current estimate of K onto $\mathscr{P}_{L,\mathscr{C}} \cap \mathscr{Q}_{I,\tilde{c}}$ in order to obtain the required element in $\mathscr{P}_{L,\mathscr{C}} \cap \mathscr{Q}_{I,\tilde{\mathscr{C}}}$. The fact that we are I-projecting follows from (ii) of Lemma 1. Using this, for all $K \in \mathscr{Q}_{I,c}$ 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 $\mathcal{I}(R|K) \geqq \mathcal{I}(R|Q^{-1})$ for all $K \in \mathscr{Q}_{I,c}$ with equality iff $K = Q^{-1}$.

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

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