 Published by Blackwell Publishers Ltd., 108 Cowley Road, Oxford OX4 1JF and 238 Main Street, Cambridge, MA 02142, USA. Vol. 23: 105–116, 1996

On a Dualization of Graphical Gaussian Models

GÖRAN KAUERMANN

Technical University of Berlin

ABSTRACT. Graphical Gaussian models as defined by Speed & Kiiveri (1986) present the conditional independence structure of normally distributed variables by a graph. A similar approach was recently motivated by Cox & Wermuth (1993) who introduced graphs showing the marginal independence structure. The interpretation of a graph in terms of conditional independence relations is based on the definition of a pairwise, local and global Markov property respectively, which are equivalent in the normal distribution. Similar definitions can be formulated for the interpretation of graphs in terms of marginal independencies. Their equivalence is proven in the normal distribution. Frydenberg (1990a) discusses equivalence statements between the graphical approach and the concept of a cut in exponential families (Barndorff-Nielsen, 1978). In this paper, similar relations are shown for the normal distribution and graphical models for marginal independencies. Parameter estimation in graphical models with marginal independence interpretation is achieved by the dual likelihood concept, which shows interesting relations to results available for maximum likelihood estimation in graphical Gaussian models for conditional independence.

Key words: conditional independence, dual likelihood theory, exponential family, graphical model, marginal independence, normal distribution, parallel foliations in exponential families

1. Introduction

Graphical Gaussian models as introduced by Speed & Kiiveri (1986) result from covariance selection models defined by Dempster (1972) (see also Lauritzen & Wermuth, 1989; Wermuth & Lauritzen, 1990). They deal with the presentation of conditional independencies in multivariate normally distributed variables by a graph. The vertices in the corresponding graph relate to the variables, missing edges represent conditional independencies. A recent approach by Cox & Wermuth (1993) considers independence structures in margins of a multivariate distribution. They introduce undirected graphs representing marginal independencies in a set of variables. To distinguish between the two graphical representations, Cox & Wermuth (1993) used dashed edges to describe the marginal independence structure, while graphs with solid edges represent the conditional independencies. The first graphs are named covariance graphs whereas the second are called concentration graphs. Take Fig. 1 as an example.

The exact interpretation of a graph in terms of conditional independence relations results from the definitions of pairwise, local and global Markov properties as given by Frydenberg (1990a, b). These properties are equivalent for distributions that fulfill $A \perp B \mid C \wedge A \perp C \mid B \Rightarrow A \perp (B \cup C)$ for disjoint sets A, B and C of variables, with symbol $\perp$ denoting independence. Here, we present dual versions of these Markov properties which allow exact interpretations of undirected graphs showing the marginal independence structure, i.e. covariance graphs as introduced by Cox & Wermuth (1993). We derive conditions, fulfilled by the normal distribution, which yield the equivalence of the presented Markov properties.

Frydenberg (1990a) proves equivalence statements between the graphical approach and the concept of a cut in exponential families (Barndorff-Nielsen, 1978), equivalently defined as θ-parallel foliation in Barndorff-Nielsen & Blæsild (1983). In the normal distribution, we find

Fig. 1. (a) Concentration graph representing the conditional independence statement $X \perp V \mid WY$, $W \perp Y \mid XV$ and $V \perp W \mid XY$. (b) Covariance graph representing the marginal independencies $X \perp V$, $W \perp Y$ and $V \perp W$.

similar results coherent with a τ-parallel foliation (Barndorff-Nielsen & Blæsild, 1983) and graphical models for marginal independencies.

The focus of the last section of the paper is on parameter estimation in graphical Gaussian models. This topic is largely developed for models with conditional independence interpretation (Speed & Kiiveri, 1986), where interesting mathematical relationships appear between the graphical concept and maximum likelihood estimation (Frydenberg & Lauritzen, 1989). In graphical models for marginal independencies we suggest a dual maximum likelihood estimation as treated by Efron (1978), Brown (1986) or Christensen (1989). This yields remarkable similarities to the classical graphical models. Consequently, results available for maximum likelihood estimation in graphical models for conditional independencies can be transferred to dual maximum likelihood estimation in graphical models for marginal independencies. This property is illustrated with an example.

Notations from graph theory

We briefly introduce some necessary notation from graph theory.

An *undirected graph* $\mathcal{G} = (V, E)$ is defined by the finite set V of vertices and the set $E \subset V \times V$ of edges. For every $(\alpha, \beta) \in E$ we have $(\beta, \alpha) \in E$. A *sub-graph* $\mathcal{G}_A$ is defined by $\mathcal{G}_A = (V \cap A, E \cap \{(V \cap A) \times (V \cap A)\})$. A graph is called *complete*, if $E = V \times V$. Two vertices α and β are called *adjacent*, if $(\alpha, \beta) \in E$ (and $(\beta, \alpha) \in E$). Two disjoint subsets A and B of vertices are adjacent, if at least two adjacent vertices $\alpha \in A$ and $\beta \in B$ exist. The *boundary* $bd(\alpha)$ of a vertex α is given by the set of adjacent vertices of α, the boundary of a set A of vertices is defined by $bd(A) = \{\beta: \beta \in bd(\alpha) \text{ for } \alpha \in A\} \backslash A$. A *path* from α to β is defined by a sequence of pairwise different vertices $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_n = \beta$, such that $(\alpha_i, \alpha_{i+1}) \in E$ for all $0 \leqslant i \leqslant n-1$. If $\alpha = \beta$ and $n \geqslant 3$, such a path is called a *cycle*. A cycle is *chordless* if only consecutive elements are joined with edges. A graph $\mathcal{G}$ is called *decomposable* if it has no chordless cycles of length $n \geqslant 4$. A decomposable graph is either complete or there exists a decomposition (A, B, C) of $\mathcal{G}$, that is we find three disjoint sets A, B, C with $B \neq \emptyset$, $C \neq \emptyset$ and $V = A \cup B \cup C$, such that $\mathcal{G}_A$ is complete and B is not adjacent to C. We call B and C *separated* by A in $\mathcal{G}$ if every path from B to C in $\mathcal{G}$ has at least one element in A. A graph $\mathcal{G}$ is called *connected* if there is a path between every two disjoint vertices in the graph, otherwise the graph is called unconnected.

Finally, the following definition is essential for the content of the paper. We call $V \backslash A$ a *simplicial collection* in a graph $\mathcal{G}$ if the boundary of every connected component in $\mathcal{G}_{V \backslash A}$ is complete in $\mathcal{G}$. The content of this definition is helpfully clarified with coroll. 2.5 of Asmussen & Edwards (1983):

$V\backslash A$ is a simplicial collection in $\mathcal{G}$ if and only if the following implication is true: If C separates A_1 and A_2 in $\mathcal{G}_A$ then C separates A_1 and A_2 in $\mathcal{G}$.

2. Markov properties

Markov properties for graphical models with conditional independence interpretation are studied by Frydenberg (1990b). He gives the following three ways of reading conditional independence restrictions from a graph $\mathcal{G}$.

Definition 2.1
A family $\mathcal{P}$ of distributions over the set V of variables is called
(i) *pairwise θ–$\mathcal{G}$ Markov*, if $\alpha \perp \beta \mid V\backslash\{\alpha, \beta\}$ for all α and β not adjacent in $\mathcal{G}$,
(ii) *local θ–$\mathcal{G}$ Markov*, if $\alpha \perp V\backslash\{bd(\alpha), \alpha\} \mid bd(\alpha)$ for all vertices α of $\mathcal{G}$,
(iii) *global θ–$\mathcal{G}$ Markov*, if $B \perp C \mid A$ whenever B and C are separated by A in $\mathcal{G}$.

The pairwise property above underlies the interpretation of concentration graphs in Cox & Wermuth (1993) (see Fig. 1a). For distributions which fulfill the implication

$$A \perp B \mid C \quad \text{and} \quad A \perp C \mid B \Rightarrow A \perp (B \cup C) \tag{1}$$

for disjoint subsets A, B and C of V, we find the equivalence of the three concepts, as given in the following proposition (Pearl & Paz, 1986; see also Frydenberg, 1990b). Note that property (1) holds particularly for distributions with strictly positive joint density for all variables (Dawid, 1979).

Proposition 2.1
Let $\mathcal{P}$ be a family of distributions that fulfills (1). *The following statements are then equivalent: (i) $\mathcal{P}$ fulfills the pairwise θ–$\mathcal{G}$ Markov property; (ii) $\mathcal{P}$ fulfills the local θ–$\mathcal{G}$ Markov property; (iii) $\mathcal{P}$ fulfills the global θ–$\mathcal{G}$ Markov property.*

In order to interpret a graph $\mathcal{G}$ in terms of marginal independence relations, as suggested by Cox & Wermuth (1993), we introduce the following Markov properties.

Definition 2.1
A family $\mathcal{P}$ of distributions over the set V of variables is called
(i) *pairwise τ–$\mathcal{G}$ Markov*, if $\alpha \perp \beta$ for all α and β not adjacent in $\mathcal{G}$,
(ii) *local τ–$\mathcal{G}$ Markov*, if $\alpha \perp V\backslash\{\alpha, bd(\alpha)\}$ for all vertices α of $\mathcal{G}$,
(iii) *global τ–$\mathcal{G}$ Markov*, if $B \perp C \mid A$ whenever B and C are separated by $D = V\backslash\{A, B, C\}$ in $\mathcal{G}$.

The pairwise τ–$\mathcal{G}$ Markov property is used in Cox & Wermuth (1993) to interpret covariance graphs (see Fig. 1b). The global concept provides the strongest formulation, i.e. it implies the local and pairwise property. In general, however, the three Markov properties above are not equivalent. In the following, we prove their equivalence for distributions which fulfill

$$A \perp B \quad \text{and} \quad A \perp C \Rightarrow A \perp (B \cup C) \tag{2}$$

for disjoint subsets A, B and C of V. Note that one may consider (2) as a dual version to (1).

It is directly verified that property (2) holds in the normal distribution. However, we did not find other simple distributions for which (2) is valid. In particular, it is not fulfilled for

the multinomial distribution unless restrictions are imposed on the parameter space. For the three-dimensional case, these restrictions are given by a vanishing log-linear three factor interaction parameter, as shown by Darroch (1962, 1974). In higher dimensions, property (2) holds at least approximately in binary quadratic exponential families, as discussed in Cox & Wermuth (1994). In general, however, the restrictions on the parameter space are complicated in higher dimensions and therefore it does not seem appropriate to dualize graphical models for contingency tables.

For distributions fulfilling (2), that means in particular for the normal distribution, we can formulate the following result.

Proposition 2.2
Let $\mathcal{P}$ be a family of distributions that fulfill (2). *The following statements are then equivalent: (i) $\mathcal{P}$ fulfills the pairwise τ–$\mathcal{G}$ Markov property; (ii) $\mathcal{P}$ fulfills the local τ–$\mathcal{G}$ Markov property; (iii) $\mathcal{P}$ fulfills the global τ–$\mathcal{G}$ Markov property.*

Proof. The proof is very similar to the results derived in Frydenberg (1990a). In detail, (iii) $\Rightarrow$ (ii) follows by definition and (ii) $\Rightarrow$ (i) due to the implication $A \perp (B \cup C) \Rightarrow A \perp B$ (see Dawid, 1979). Hence, the only thing left to show is (i) $\Rightarrow$ (iii).

Let B and C be separated by $D = V \backslash \{A, B, C\}$. Now, let $A_1 \subset A$ denote the set of all vertices in A which are not adjacent to B. Defining $A_2 := A \backslash A_1$ (which may be the empty set) gives A_2 being not adjacent to C. This is easily seen by using the following argument. Let $\alpha \in A_2$ and assume that α is adjacent to C. Then, since $\alpha \notin A_1$, i.e. α is adjacent to B, we have a path between B and C passing through α. This contradicts the assumption that B and C are separated by D. The same argument shows that A_1 is not adjacent to A_2.

Now, for all $\beta \in B$ and $\gamma \in C$, we have $\beta \perp \gamma$, and by iteration over the elements of B and C respectively, using (2), we get (a) $B \perp C$. Similarly, we get (b) $B \perp A_1$, (c) $C \perp A_2$ and (d) $A_1 \perp A_2$. Again, applying (2), we obtain with (a) and (b) $B \perp (C \cup A_1)$ and with (c) and (d) $A_2 \perp (C \cup A_1)$, which together yield $(B \cup A_2) \perp (C \cup A_1)$. This implies by standard arguments, given in Dawid (1979),

$$(B \cup A_2) \perp (C \cup A_1) \Rightarrow B \perp (C \cup A_1) \mid A_2 \Rightarrow B \perp C \mid A. \qquad \square$$

The remaining part of the paper deals with the normal distribution which fulfills (1) and (2). Hence, following the equivalence statements of propositions 2.1 and 2.2, we do not need to distinguish between the Markov properties introduced above. Therefore, we simply use the notation θ–$\mathcal{G}$ Markov property and τ–$\mathcal{G}$ Markov property below.

3. Graphical Gaussian models

3.1. Definition

We consider families of normal distributions with conditional or marginal independence restrictions respectively. In both cases, a graph $\mathcal{G}$ represents the independence relations. In particular, we define the following classes of models.

We denote with $\mathcal{N}_0(\mathcal{G})$ the family of multivariate normal distributions, over the set V of variables, which fulfill the θ–$\mathcal{G}$ Markov property, i.e. $\mathcal{N}_0(\mathcal{G})$ is the class of graphical Gaussian models discussed in Speed & Kiiveri (1986). We use the notation $\mathcal{N}_\tau(G)$ for the family of multivariate normal distributions over the variables in set V which fulfill the τ–$\mathcal{G}$ Markov property. Completely analogous definitions are obtained for subgraphs, which are denoted by $\mathcal{N}_0(\mathcal{G}_A)$ and $\mathcal{N}_\tau(\mathcal{G}_A)$ respectively. Families of marginalized distributions are

denoted by $\mathcal{N}_0(\mathcal{G})_A := \{f_A : f \in \mathcal{N}_0(\mathcal{G})\}$, whilst a notation with superscript defines the family of conditional normal distributions, conditioned on the variables in set $V \backslash A$, i.e. $\mathcal{N}_\tau(\mathcal{G})^A := \{f_{A \mid (V \backslash A)} : f \in \mathcal{N}_\tau(\mathcal{G})\}$. In the following we derive results concerning the equivalence of some of the models above. In this respect, the notation $\mathcal{N}_0(\mathcal{G})^A$ and $\mathcal{N}_\tau(\mathcal{G})_A$ respectively is redundant, since we have $\mathcal{N}_0(\mathcal{G})^A = \mathcal{N}_0(\mathcal{G}_A)$ and $\mathcal{N}_\tau(\mathcal{G})_A = \mathcal{N}_\tau(\mathcal{G}_A)$.

3.2. Equivalent models

In order to formulate equivalence statements for families of distributions, it is necessary to introduce the concept of Markov perfect families. Since we deal with two different Markov properties, it seems natural that two different definitions of Markov perfectness are needed here.

Definition 3.1
A family $\mathcal{P}$ of distributions is called *θ-Markov perfect*, if $B \perp C \mid A$, for every distribution in $\mathcal{P}$, implies that B and C are separated by A in $\mathcal{G}$. We call $\mathcal{P}$ *τ-Markov perfect*, if $B \perp C \mid A$, for every distribution in $\mathcal{P}$, implies that B and C are separated by $D = V \backslash \{A, B, C\}$ in $\mathcal{G}$.

The definition of θ-Markov perfect families of distributions is a concept introduced by Frydenberg (1990a). The above definition allows formulation of the following result.

Theorem 3.1
The family $\mathcal{N}_0(\mathcal{G})$ is θ-Markov perfect, whereas the family $\mathcal{N}_\tau(\mathcal{G})$ is τ-Markov perfect.

The proof of the first statement is given in Frydenberg (1990a). We obtain a proof for the second statement in a similar manner (see also Geiger & Pearl, 1993).

Proof. Let $\mathcal{G}$ be a graph and assume $B \perp C \mid A$ for every distribution in $\mathcal{N}_\tau(\mathcal{G})$. The object is to prove that $D = V \backslash \{A, B, C\}$ separates B and C. Equivalently, we can show that whenever B and C are not separated by D, we find a distribution in $\mathcal{N}_\tau(\mathcal{G})$ with $B \not\perp C \mid A$.

Let B and C be connected by a path not passing a vertex in D, that is we have vertices $\beta = \alpha_0 \in B$ and $\gamma = \alpha_n \in C$ being connected by the path $(\alpha_1, \ldots, \alpha_{n-1})$ with elements $\alpha_i \in A$ for $0 < i < n$ (in the simplest case there is an edge between β and γ). Let f^0 denote the normal density with covariance matrix Σ defined by

$$\Sigma_{ij} = \begin{cases} 1 & \text{for } i = j \\ -\frac{1}{2} & \text{for } i = \alpha_l \text{ and } j = \alpha_{l+1} \\ & \text{or } i = \alpha_{l+1} \text{ and } j = \alpha_l \text{ for } 0 \leqslant l \leqslant n-1 \\ 0 & \text{otherwise,} \end{cases}$$

where it is easily checked that $f^0 \in \mathcal{N}_\tau(\mathcal{G})$.

Let $\Sigma_{\alpha,\alpha}$ denote the marginal covariance matrix of $X_\alpha = (X_{\alpha_0}, \ldots, X_{\alpha_n})$. This has the specified structure given above. Some elementary matrix algebra shows that the $(1, n)$ element of $(\Sigma_{\alpha,\alpha})^{-1}$, that is $(\Sigma_{\beta\gamma \mid \{\alpha_1, \ldots, \alpha_{n-1}\}})^{-1}$, is not equal to zero. Hence, β and γ are dependent for given A in the distribution $f^0 \in \mathcal{N}_\tau(\mathcal{G})$, which completes the proof. □

Theorem 3.2 permits us to derive the first equality statement. It is related to Jensen's (1988; lem. 1) models which are linear in the covariance matrix and in its inverse. The content of the theorem is also found in Wermuth *et al.* (1995; p. 26).

Theorem 3.2
We have $\mathcal{N}_0(\mathcal{G}) = \mathcal{N}_\tau(\mathcal{G})$ if and only if $\mathcal{G}$ is either complete or consists of unconnected complete subgraphs.

Proof. Let $\mathcal{N}_0(\mathcal{G}) = \mathcal{N}_\tau(\mathcal{G})$ and assume $B \perp C \mid A$ for all distributions in $\mathcal{N}_0(\mathcal{G})$ and $\mathcal{N}_\tau(\mathcal{G})$ respectively (if no such independence relation exists, we easily find that $\mathcal{G}$ is complete, which proves the result). Since both families are Markov perfect in the above particular definitions, we obtain that B and C are separated by A in $\mathcal{G}$, due to the θ-Markov perfectness of $\mathcal{N}_0(\mathcal{G})$, and they are separated by $D = V\backslash\{A, B, C\}$ in $\mathcal{G}$, due to the τ-Markov perfectness of $\mathcal{N}_\tau(\mathcal{G})$. This gives a contradiction unless B and C belong to two unconnected subgraphs. Finally, it is easily seen that the resulting unconnected subgraphs of $\mathcal{G}$ are complete.

If, on the other hand, $\mathcal{G}$ is either complete or consists of unconnected complete subgraphs, the verification of equality $\mathcal{N}_0(\mathcal{G}) = \mathcal{N}_\tau(\mathcal{G})$ is direct due to a factorization of the densities in $\mathcal{N}_0(\mathcal{G})$ and $\mathcal{N}_\tau(\mathcal{G})$ respectively. □

Frydenberg (1990a) utilizes the definition of a simplical collection in a graph $\mathcal{G}$ (see definition given in section 1 above) to prove equality statements in $\mathcal{N}_0(\mathcal{G})$. We formulate his results in a dual version for family $\mathcal{N}_\tau(\mathcal{G})$ here.

Theorem 3.3
We have $\mathcal{N}_0(\mathcal{G})_A = \mathcal{N}_0(\mathcal{G}_A)$ if and only if $V\backslash A$ is a simplicial collection in $\mathcal{G}$. Similarly, we have $\mathcal{N}_\tau(\mathcal{G})^A = \mathcal{N}_\tau(\mathcal{G}_A)$ if and only if $V\backslash A$ is a simplicial collection in $\mathcal{G}$.

The proof of the first statement is found in Porteous (1985) (see also Frydenberg, 1990a). We give a proof of the second statement here.

Proof. Assume $\mathcal{N}_\tau(\mathcal{G})^A = \mathcal{N}_\tau(\mathcal{G}_A)$ and $A_1 \perp A_2 \mid C$ for all distributions in $\mathcal{N}_\tau(\mathcal{G}_A)$, where A_1, A_2 and C are disjoint subsets of A (if no such independence condition exists, we have $\mathcal{G}_A$ as a complete subgraph and nothing remains to be proved). Since $\mathcal{N}_\tau(G_A)$ is τ-Markov perfect, the independence relation implies that A_1 and A_2 are separated by $D = A\backslash\{A_1, A_2, C\}$ in $\mathcal{G}_A$. By assumption, every distribution in $\mathcal{N}_\tau(\mathcal{G}_A)$ is element of $\mathcal{N}_\tau(\mathcal{G})^A$. This gives $A_1 \perp A_2 \mid (C \cup (V\backslash A))$ for all distributions in $\mathcal{N}_\tau(\mathcal{G})^A$ and hence in $\mathcal{N}_\tau(\mathcal{G})$. Again due to the fact that $\mathcal{N}_\tau(\mathcal{G})$ is τ-Markov perfect, we obtain that A_1 and A_2 are separated by $D = V\backslash\{A_1, A_2, C, V\backslash A\}$ in $\mathcal{G}$. The result of Asmussen & Edwards (1983), given previously, shows that $V\backslash A$ is a simplicial collection in $\mathcal{G}$.

Let, on the other hand, $V\backslash A$ be a simplicial collection in $\mathcal{G}$ and assume that A_1 and A_2 are separated by $D = A\backslash\{A_1, A_2, C\}$ in $\mathcal{G}_A$, where A_1, A_2 and C are disjoint subsets of A (if no such separation exists, $\mathcal{G}_A$ is complete and the proof is direct). We have $A_1 \perp A_2 \mid C$ for every distribution in $\mathcal{N}_\tau(\mathcal{G}_A)$. Applying the result of Asmussen & Edwards (1983), we get that A_1 and A_2 are separated by D in $\mathcal{G}$, which gives $A_1 \perp A_2 \mid (C \cup (V\backslash A))$ for all distributions in $\mathcal{N}_\tau(\mathcal{G})$ and hence in $\mathcal{N}_\tau(\mathcal{G})^A$. Now, the variables in set $V\backslash A$ are considered to be fixed in the family $\mathcal{N}_\tau(\mathcal{G})^A$. Hence, we have $A_1 \perp A_2 \mid C$ for all distributions in $\mathcal{N}_\tau(\mathcal{G})^A$, which yields $\mathcal{N}_\tau(\mathcal{G})^A \subset \mathcal{N}_\tau(\mathcal{G}_A)$.

On the other hand, it is easy to verify that $\mathcal{N}_\tau(\mathcal{G}_A) \subset \mathcal{N}_\tau(\mathcal{G})^A$ holds generally. Take an arbitrary distribution $f_A \in \mathcal{N}_\tau(\mathcal{G}_A)$ and denote with $f^0 \in \mathcal{N}_\tau(\mathcal{G})$ a distribution with complete independence among the variables in set V. Defining $\tilde{f} = f_A f^0_{V\backslash A}$ obviously gives an element of $\mathcal{N}_\tau(\mathcal{G})$ and conditioning $\tilde{f}$ on the variables in set $V\backslash A$ gives $\tilde{f}_{A \mid V\backslash A} = f_A \in \mathcal{N}_\tau(\mathcal{G})^A$.

Combining these arguments gives $\mathcal{N}_\tau(\mathcal{G})^A = \mathcal{N}_\tau(\mathcal{G}_A)$, which completes the proof. □

3.3. Parallel foliations

The definition of simplicial collections in a graph $\mathcal{G}$ yields further equivalence statements in $\mathcal{N}_0(\mathcal{G})$ and $\mathcal{N}_\tau(\mathcal{G})$ which appear in the context of exponential families. Frydenberg (1990a) discusses relations in $\mathcal{N}_0(\mathcal{G})$ between the graph theoretic approach and the concept of a cut in exponential families (Barndorff-Nielsen, 1978), equivalently defined as θ-parallel foliation in Barndorff-Nielsen & Blæsild (1983). As dual version to a cut in exponential families, Barndorff-Nielsen & Blæsild define a τ-parallel foliation. Explicit definitions are given below. Before we proceed to show equivalences in $\mathcal{N}_\tau(\mathcal{G})$ between τ-parallel foliations and the graphical concept, some basic notations for exponential families are required.

The family of distributions $\mathcal{P}$ of a random vector X is called an *exponential family* if the density function of a distribution $P \in \mathcal{P}$ (for a suitable measure μ) can be written in the form

$$\frac{dP}{d\mu} = f(x) = \exp(\theta' t(x) - \kappa(\theta)), \tag{3}$$

with $\theta \in \Theta \subset \mathfrak{R}^m$, where θ denotes the natural parameter and Θ the parameter space, $t(\,) \in \mathfrak{R}^m$ is a measurable function, m denotes the dimension of the family and $\kappa(\,)$ the log Laplace transform defined by $\kappa(\theta) = \ln(\int \exp(\theta' t(x))\, d\mu)$. The vectors θ and t are assumed to be affine independent. For a random sample $X^1, \ldots, X^N$, we have $\bar{t} = N^{-1}(t(X^1) + \cdots + t(X^N))$ as sufficient statistic, distributed with natural parameter $N\theta$ and log Laplace transform $\kappa_N(N\theta) = N\kappa(\theta)$.

The family $\mathcal{P}$ is called *regular*, if Θ is open with $\Theta = \{\theta : \int \exp(\theta' t(x))\, d\mu < \infty\}$. The expectation parameter τ results by the invertible function

$$\tau(\theta) = \frac{\partial \kappa(\theta)}{\partial \theta} = E_\theta(t(x)),$$

so that the expectation parameter space $\mathcal{T}$ equals $\tau(\text{int}\,\Theta)$, where int Θ is the interior of Θ. Partitions of θ are denoted by (θ_1, θ_2), and similarly we denote partitions of τ. The corresponding parameter spaces are denoted by Θ_1, for instance, where $\Theta_1 := \{\theta_1 :$ there exist a θ_2 such that $(\theta_1, \theta_2) \in \Theta\}$. This notation now allows us to give the following definition (see Barndorff-Nielsen & Bæsild, 1983).

Definition 3.2 (θ-parallel foliation)
Given an exponential family $\mathcal{P}$ with open and connected parameter space Θ, we partition θ into (θ_1, θ_2) and similarly τ into (τ_1, τ_2). We say θ_1 forms a *θ-parallel foliation* if there exist functions $\varphi(\,)$ and $\chi(\,)$ such that for all $\tau_1 \in \mathcal{T}_1$ and $\theta_2 \in \Theta_2$
(i) $\theta_1 = \varphi(\tau_1) + \chi(\theta_2)$, and
(ii) τ_1 and θ_2 are variation independent, this means that $(\varphi(\tau_1) + \chi(\theta_2), \theta_2) \in \Theta$.

The property of variation independence is important. It means, in particular, that the range of possible parameter values of τ_1 does not depend on θ_2. Note that in regular exponential families condition (ii) above is always valid (Barndorff-Nielsen, 1978, th. 8.4).

The corresponding dual concept, relating to the expectation parameter space $\mathcal{T}$, is introduced by Barndorff-Nielsen & Blæsild (1983) as τ-parallel foliation.

Definition 3.3 (τ-parallel foliation)
Given an exponential family $\mathcal{P}$ with open and connected parameter space Θ, we partition θ into (θ_1, θ_2) and similarly τ into (τ_1, τ_2). We say τ_1 forms a *τ-parallel foliation* if there exist

functions $\varphi^*(\,)$ and $\chi^*(\,)$ such that for all $\tau_1 \in \mathcal{T}_1$ and $\theta_2 \in \Theta_2$
(i) $\tau_1 = \varphi^*(\theta_1) + \chi^*(\tau_2)$, and
(ii) θ_1 and τ_2 are variation independent, this means that $(\varphi^*(\theta_1) + \chi^*(\tau_2), \tau_2) \in \mathcal{T}$.

In the following, we illustrate the existence of parallel foliations in the normal distribution. Let $X^1, \ldots, X^N$ be a random sample of a normal distribution with covariance matrix Σ. We have the log likelihood function, neglecting an additive term,

$$l(K, S) = N\{-\operatorname{tr}(KS) + \ln|K|\}, \tag{4}$$

where K denotes the inverse of the covariance matrix, i.e. $K = \Sigma^{-1}$, and S denotes the sample covariance matrix, i.e. $S = N^{-1}\Sigma_n(X^n - \bar{X})(X^n - \bar{X})^{\mathrm{T}}$, with $\bar{X} = N^{-1}\Sigma_n X^n$. Both, K and S, are assumed to be positive definite. Hence, K is the natural parameter θ and Σ is the expectation parameter τ of the underlying exponential family. We partition X into (X_A, X_B) and similarly we partition Σ into $\tau_1 = \Sigma_{AA}$ and $\tau_2 = (\Sigma_{BA}, \Sigma_{BB})$, and K into $\theta_1 = K_{AA}$ and $\theta_2 = (K_{BA}, K_{BB})$. The equalities (see for instance Wermuth, 1992)

$$K_{AA} = \underbrace{\Sigma_{AA}^{-1}}_{=:\varphi(\Sigma_{AA})} + \underbrace{K_{AB}K_{BB}^{-1}K_{BA}}_{=:\chi(K_{BA}, K_{BB})} \tag{5}$$

$$\Sigma_{AA} = \underbrace{K_{AA}^{-1}}_{=:\varphi^*(K_{AA})} + \underbrace{\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}}_{=:\chi^*(\Sigma_{BA}, \Sigma_{BB})} \tag{6}$$

show that K_{AA} forms a θ-parallel foliation, that is a cut in the exponential family if and only if Σ_{AA} and (K_{BA}, K_{BB}) are variation independent, whilst Σ_{AA} forms a τ-parallel foliation if and only if K_{AA} and $(\Sigma_{BA}, \Sigma_{BB})$ are variation independent.

The following theorem shows the relation between parallel foliations and the graphical approach. The first part of the theorem is due to Frydenberg (1990a), the second is its dual version.

Theorem 3.4
We have $\mathcal{N}_0(\mathcal{G}_A) = \mathcal{N}_0(\mathcal{G})_A$ if and only if K_{AA} forms a θ-parallel foliation. Similarly, we have $\mathcal{N}_\tau(\mathcal{G}_A) = \mathcal{N}_\tau(\mathcal{G})^A$ if and only if Σ_{AA} forms a τ-parallel foliation.

Proof. We give the proof of the second statement here. We have to show the variation independence of K_{AA} and $(\Sigma_{BA}, \Sigma_{BB})$. This is fulfilled if every zero restriction on Σ can be expressed as restriction either on K_{AA} or on $(\Sigma_{BA}, \Sigma_{BB})$. Hence, we have to show that every zero restriction on Σ_{AA} implies a restriction on K_{AA} but not on $(\Sigma_{BA}, \Sigma_{BB})$.

Let A_1, A_2 and C be disjoint subsets of A, with $D = A\backslash\{A_1, A_2, C\}$ separating A_1 and A_2 in G_A (if no such sets exist we have $\mathcal{G}_A$ complete and the proof is direct). With theorem 3.3 we get that $V\backslash A$ is a simplicial collection in $\mathcal{G}$. Then, utilizing the result of Asmussen & Edwards (1983) shows that $A_1 \perp A_2 \mid C$ implies $A_1 \perp A_2 \mid (C \cup (V\backslash A))$. The first independence condition is a restriction on Σ_{AA} whereas the second is a restriction $K_{AA} = (\Sigma_{AA\mid(V\backslash A)})^{-1}$. This directly yields the variation independence of K_{AA} and $(\Sigma_{BA}, \Sigma_{BB})$.

Let on the other hand K_{AA} form a τ-parallel foliation, that is K_{AA} and $(\Sigma_{BA}, \Sigma_{BB})$ are variation independent. We show $\mathcal{N}_\tau(\mathcal{G})^A \subset \mathcal{N}_\tau(\mathcal{G}_A)$. Let $f^A \in \mathcal{N}_\tau(\mathcal{G})^A$, where f^A is parameterized with variance matrix K_{AA}^{-1}, and let f denote the distribution with parameters K_{AA} and $(\Sigma_{BA}, \Sigma_{BB})$. Assume that $\Sigma_{BA} = 0$ and $\Sigma_{BB} = I$, with 0 and I denoting the zero matrix and the identity matrix respectively, with matching dimensions. Due to equality (6) and the assumed variation independence, we find $f \in \mathcal{N}_\tau(\mathcal{G})$ if and only if $f^A \in \mathcal{N}_\tau(\mathcal{G}_A)$. This gives $\mathcal{N}_\tau(\mathcal{G})^A \subset \mathcal{N}_\tau(\mathcal{G}_A)$ which together with $\mathcal{N}_\tau(\mathcal{G}_A) \subset \mathcal{N}_\tau(\mathcal{G})^A$, as shown previously, completes the proof. □

4. Duality of hypotheses

The object of the remaining part of the paper is parameter estimation in the family $\mathcal{N}_\tau(\mathcal{G})$. If $\mathcal{G}$ is not complete, i.e. marginal independence constraints are imposed on Σ, the maximum likelihood estimator for Σ is in general not available in an analytical form. Exactly specified exceptions are found in Pearl & Wermuth (1994). Iterative procedures for determining the maximum likelihood estimator are given in the context of linear structural equation models (Jöreskog, 1973).

We provide the dual maximum likelihood estimation (Efron, 1978; Barndorff-Nielsen, 1978; Brown, 1986; Christensen, 1989) as alternative estimation approach here. A brief presentation is given below.

Let $K(\theta_f, \theta_g) = E(\ln(f/g))$ denote the Kullback Leibler information, where the expectation is taken with respect to density f, and θ_f and θ_g are the parameters of the exponential family densities f and g respectively. We denote by $\acute{\theta}$ the observed natural parameter, i.e. $\acute{\theta} = \theta(\bar{t})$, where $\theta(\,)$ is the functional inverse of $\tau(\,)$. It can be shown that the maximum likelihood estimation corresponds to the minimization problem $K(\acute{\theta}; \hat{\theta}) = \inf\{K(\acute{\theta}; \theta); \theta \in \Theta\}$.

Since the Kullback Leibler information is not symmetrical in its arguments, we get a different minimization problem if the observed and unknown parameters are exchanged. This yields the dual maximum likelihood estimator $\check{\theta}$ defined by $K(\check{\theta}; \acute{\theta}) = \inf\{K(\theta; \acute{\theta}); \theta \in \Theta\}$. In particular, we get the dual log likelihood function

$$l^*(\tau, \acute{\theta}) = N\{\tau\acute{\theta} - \varrho(\tau)\}, \tag{7}$$

where $\varrho(\tau) := \theta(\tau)\tau - \kappa(\theta(\tau))$ is known as log Legendre transform of $\kappa(\,)$. Taking the first derivative of (7) gives the dual likelihood equation

$$\frac{\partial l^*}{\partial \tau} = N\{\acute{\theta} - \theta(\check{\tau})\} = 0,$$

which is solved by the dual maximum likelihood estimator $\check{\tau}$. Note that it is easily checked that the maximum likelihood estimator and the dual maximum likelihood estimator coincide in regular exponential families (see for instance Brown, 1986). Asymptotic properties of $\check{\tau}$ are discussed in Christensen (1989). In particular, she shows that $\check{\tau}$ is asymptotically unbiased and normally distributed.

Coming back to the normal distribution with likelihood function (4), we get the dual likelihood function, when neglecting additive terms,

$$l^*(\Sigma, \acute{K}) = N\{-\operatorname{tr}(\Sigma\acute{K}) + \ln|\Sigma|\}, \tag{8}$$

which is intended to be maximized with respect to Σ.

Before we proceed, it seems necessary to discuss the determination of $\acute{K}$. The direct choice is to set $\acute{K} = S^{-1}$ as observed natural parameter. This is the only correct definition if the family has full dimension, i.e. if no zero restrictions on K exist. However, if the family considered has reduced dimension, we find that S^{-1} does not lie in the natural parameter space of the underlying regular exponential family. This means for model $\mathcal{N}_\tau(\mathcal{G})$ that there exists a graph $\tilde{\mathcal{G}}$ with $\mathcal{N}_\tau(\mathcal{G}) \subset \mathcal{N}_0(\tilde{\mathcal{G}}) \subset \mathcal{N}_0(\mathcal{G}^c)$, where $\mathcal{G}^c$ denotes the complete graph, and $\tilde{\mathcal{G}}$ is the smallest graph having more edges than $\mathcal{G}$ such that the above relation holds. Theorem 3.2 shows that $\tilde{\mathcal{G}}$ is either complete, i.e. $\tilde{\mathcal{G}} = \mathcal{G}^c$, or consists of unconnected complete subgraphs $\tilde{\mathcal{G}}_{A_1}, \ldots, \tilde{\mathcal{G}}_{A_m}$, say. In this sense, the family $\mathcal{N}_0(\tilde{\mathcal{G}})$ represents the smallest regular family in which $\mathcal{N}_\tau(\mathcal{G})$ is embedded. This suggests to set $\acute{K}$ as the observed

natural parameter in $\mathcal{N}_0(\tilde{\mathcal{G}})$. We define (see Dempster, 1972)

$$\acute{K}_{A_iA_j} = \begin{cases} 0 & \text{for } i \neq j \\ S^{-1}_{A_iA_i} & \text{for } i = j. \end{cases}$$

It is important to remark a relevant difference between the dual likelihood and the likelihood approach. In the case of a connected but not necessarily complete graph $\mathcal{G}$, the family $\mathcal{N}_0(\mathcal{G}_A)$ yields a reduction of the dimension of the sufficient statistic, i.e. the elements of S corresponding to the missing edges in $\mathcal{G}$ can be neglected (Speed & Kiiveri, 1986). A similar property does not hold for the dual likelihood estimation concept in $\mathcal{N}_\tau(\mathcal{G})$. This means, although some elements of $\acute{K}$ can be neglected in (8), all elements of the observed sample covariance matrix are required to obtain $\acute{K}$, since $\acute{K} = S^{-1}$ here. Hence, the dual likelihood approach does not lead to a reduction of the dimension of the sufficient statistic. This is a drawback particularly if the marginal independencies in $\mathcal{N}_\tau(\mathcal{G})$ correspond to a recursive factorization of the density, as discussed by Pearl & Wermuth (1994).

It is conspicuous that (4) and (8) have the same structure, which motivates the following consideration. Let Z^i, $i = 1, \ldots, N$, be a random sample from a normal distribution with covariance matrix $\Sigma^{\{Z\}} := \Sigma^{-1}$ and sample covariance matrix $S^{\{Z\}} := \acute{K}$, where Σ is the covariance matrix of Y and $\acute{K}$ is the observed natural parameter of Y. The log likelihood function of $\Sigma^{\{Z\}}$ equals

$$l(K^{\{Z\}}, S^{\{Z\}}) = N\{-\operatorname{tr}(S^{\{Z\}}K^{\{Z\}}) + \log|K^{\{Z\}}|\} \tag{9}$$

which obviously equals (8). Now, the τ–$\mathcal{G}$ Markov property concerning Σ corresponds to a θ–$\mathcal{G}$ Markov property concerning $K^{\{Z\}} = (\Sigma^{\{Z\}})^{-1} = \Sigma$. For the latter case, however, estimation results available in graphical models for conditional independencies can be applied in order to maximize (8) and (9) respectively. We find the dual likelihood equations as dual versions of the likelihood equations for (9) (Dempster, 1972), i.e.

$$\check{K}_{\alpha\beta} = \acute{K}_{\alpha\beta} \quad \text{whenever } \alpha \text{ and } \beta \text{ are adjacent in } \mathcal{G} \text{ or } \alpha = \beta,$$

$$\check{\Sigma}_{\alpha\beta} = 0 \quad \text{whenever } \alpha \text{ and } \beta \text{ are not adjacent in } \mathcal{G}.$$

The following main result is now easily verified.

Theorem 4.1
Let S be a positive definite sample covariance matrix of a random sample from $\mathcal{N}_\tau(\mathcal{G})$ with $\acute{K}$ as corresponding observed natural parameter. Let $S^{\{Z\}} = \acute{K}$ be the sample covariance matrix of a random sample from $\mathcal{N}_0(\mathcal{G})$. Then, the dual maximum likelihood estimator for Σ in $\mathcal{N}_\tau(\mathcal{G})$ equals the maximum likelihood estimator for $K^{\{Z\}} = (\Sigma^{\{Z\}})^{-1}$ in $\mathcal{N}_0\mathcal{G}$), i.e. we get

$$\check{\Sigma} = (\hat{\Sigma}^{\{Z\}})^{-1}.$$

As a direct consequence of this theorem, we can transfer the following result of Frydenberg & Lauritzen (1989).

Corollary 4.1
The dual maximum likelihood estimator $\check{\Sigma}$ in $\mathcal{N}_\tau(\mathcal{G})$ can be obtained analytically (for sample size N sufficiently large) if $\mathcal{G}$ is decomposable.

5. Conclusion and example

Corollary 4.1 shows an interesting benefit of the dual maximum likelihood estimation in the family $\mathcal{N}_\tau(\mathcal{G})$. The dual estimation problem is solved analytically if $\mathcal{G}$ is decomposable, which

Table 1. *(a) Observed and (b) dual maximum likelihood estimator for marginal correlations (lower half), partial correlations given the remaining variables (upper half), means and standard deviation for n = 39 patients*

(a) Variable	Y	X	V	W	(b) Variable	Y	X	V	W
Y	1	−0.431	−0.407	−0.262	Y	1	−0.431	−0.407	−0.223
X	−0.334	1	−0.111	−0.517	X	−0.351	1	−0.175	−0.517
V	−0.404	**0.042**	1	−0.28	V	−0.373	**0**	1	−0.091
W	**−0.071**	−0.460	**0.060**	1	W	**0**	−0.479	**0**	1
Mean	17.49	12.57	3.71	10.40	Mean	17.49	12.57	3.71	10.40
SD	2.07	7.86	92.0	5.72	SD	20.04	7.92	91.56	5.73

is in contrast to the maximum likelihood estimator for Σ. The following example demonstrates this point.

The data listed in Table 1a were collected by Kohlmann *et al.* (1991) on 39 diabetic patients, where the variables stand for: Y = glucose control, X = score for the knowledge about the illness, V = duration of illness and W = score, measuring the attitude of the patients to the illness. Cox & Wermuth (1993) presented the data as an example for the independence hypotheses $Y \perp W$, $X \perp V$ and $V \perp W$, i.e. as an example for the decomposable covariance graph given in Fig. 1b.

The dual maximum likelihood estimator, given in Table 1b, is obtained analytically by using the results presented above. Christensen (1989) shows that an obviously constructed dual likelihood ratio statistic is asymptotically chi-squared distributed. Utilizing this result, we get 5.82 as the value of this statistic at 3 degrees of freedom, when testing the fit of the model $Y \perp W$, $X \perp V$ and $V \perp W$.

Note that the dual maximum likelihood estimator for arbitrary graphs $\mathcal{G}$, as well as the dual likelihood ratio statistic, can be calculated with standard software packages for graphical models for conditional independence. We used MIM (Edwards, 1991) in our example.

Acknowledgements

The author is grateful to Nanny Wermuth for enlightening conversation and drawing his attention to the topic discussed. He is also very indebted to two referees and an associate editor for working through earlier versions of this paper and correcting a mistake in the proof of proposition 2.2.

References

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

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

Barndorff-Nielsen, O. & Blæsild, P. (1983). Exponential models with affine dual foliations. *Ann. Statist.* **11,** 753–769.

Brown, L. D. (1986). Fundamentals of statistical exponential families. *IMS Lecture Notes*, IMS, Hayward, CA.

Christensen, S. (1989). Statistical properties if I-projections within exponential families. *Scand. J. Statist.* **16,** 307–318.

Cox, D. R. & Wermuth, N. (1993). Linear dependences represented by chain graphs (with discussions). *Statist. Sci.* **8,** 204–218, 247–277.

Cox, D. R. & Wermuth, N. (1994). A note on the binary quadratic exponential distribution. *Biometrika* **81,** 403–408.

Darroch, J. N. (1962). Interactions in multi-factor contingency tables. *J. Roy. Statist. Soc. Ser. B* **24,** 257–263.

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

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

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

Edwards, D. E. (1991). *A Guide to MIM Version 2.0.* HyperGraph Software, DK 4000 Roskilde.

Efron, B. (1978). The geometry of exponential families. *Ann. Statist.* **6,** 362–376.

Frydenberg, M. (1990a). Marginalisation and collapsibility in graphical interaction models. *Ann. Statist.* **18,** 790–805.

Frydenberg, M. (1990b). The chain graph Markov property. *Scand. J. Statist.* **17,** 333–353.

Frydenberg, M. & Lauritzen, S. L. (1989). Decomposition of maximum likelihood estimator in mixed graphical interaction models. *Biometrika* **76,** 539–555.

Geiger, D. & Pearl, J. (1993). Logical and algorithmic properties of conditional independence and graphical models. *Ann. Statist.* **21,** 2001–2021.

Jensen, S. T. (1988). Covariance hypotheses which are linear in the covariance and the inverse covariance. *Ann. Statist.* **16,** 302–322.

Jöreskog, K. G. (1973). A general method for estimating linear structural equation systems. In *Structural equation models in the social science* (eds A. S. Goldberger & O. D. Duncan), 85–112. Seminar Press, New York.

Kohlmann, C. W., Krohne H. W., Küstner, E., Schrezenmeir, J., Walther, U. & Eyer, J. (1991). Der IPC-Diabetes-Fragebogen. *Diagnostica* **37,** 252–270.

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

Pearl, J. & Paz, T. (1986). Graphoids. A graph-based logic for reasoning about relevancy relations. *Proceedings 7th European Conference on Artificial Intelligence*, Brighton, UK, 1986.

Pearl, J. & Wermuth, N. (1994). When can association graphs admit a causal explanation? In *Selecting models and data, artificial intelligence and statistics* (eds P. Cheeseman & W. Oldford), 205–214. Springer-Verlag, New York.

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

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

Wermuth, N. (1992). On block-recursive regression equations. *Braz. J. Probab. Statist.* **6,** 1–56.

Wermuth, N. & Lauritzen, S. L. (1990). On substantive research hypothesis, conditional independence graphs and graphical chain models. *J. Roy. Statist. Soc. Ser. B* **52,** 21–51.

Wermuth, N. Cox, D. R. & Pearl, J. (1995). Explanations for multivariate structures derived from multivariate recursive regression. Submitted to *Bernoulli.*

Received December 1993, in final form March 1995

Göran Kauermann, Technische Universität Berlin, Institut für Quantitative Methoden, Franklinstrasse 28/29, 10587 Berlin, Germany.