The Annals of Statistics
2004, Vol. 32, No. 1, 219–232

ESTIMATION OF PROPORTIONAL COVARIANCES IN THE PRESENCE OF CERTAIN LINEAR RESTRICTIONS

BY SØREN TOLVER JENSEN AND JESPER MADSEN

University of Copenhagen and Statens Serum Institut

Proportionality of covariance matrices of n independent p-dimensional normal distributions with the same type of linear restrictions of the inverse covariances is considered. Conditions for existence and uniqueness of the maximum likelihood estimator are obtained through the development of general results for scale-invariant natural exponential families.

1. Introduction. Let $X_1, \ldots, X_n$ be independent and p-dimensional normally distributed random variables such that $\mathrm{E}(X_i) = 0$ and $\mathrm{V}(X_i) = \lambda_i \Sigma$, $i = 1, \ldots, n$, where $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$ and where Σ is positive definite. This model is known as the *model of proportional covariances*, and the problem of maximum likelihood estimation of the unknown parameters $(\lambda_1, \ldots, \lambda_n)$ and Σ has been treated by several authors; see, for example, Flury (1986), Eriksen (1987) and Jensen and Johansen (1987). (To obtain a one-to-one parametrization of the model, a constraint like $\prod_{i=1}^n \lambda_i = 1$ should be imposed.)

In Jensen and Johansen (1987), it was proved that when the likelihood function is maximized over Σ then the negative logarithm of the profile likelihood function is strictly convex with probability 1 as a function of $(\log(\lambda_1), \ldots, \log(\lambda_n))$ if $n > p$.

In this paper, we consider an extension of the model as we further shall assume (i) that $\theta = \Sigma^{-1} \in M$, where M is a subspace of S_p ($\equiv$ the vector space of all $p \times p$ symmetric matrices) and (ii) that $(\log(\lambda_1), \ldots, \log(\lambda_n)) \in U$, where U is a subspace of $\mathbb{R}^n$. We present a general necessary and sufficient condition for convexity of the profile likelihood function (Theorem 2.1), and we prove that in the special case where M is a *Jordan algebra*, that is, $I \in M$ and $AB + BA \in M$ for all $A, B \in M$, the condition is fulfilled for all observations for which the profile likelihood function exists (Theorem 2.2, Corollary 2.2, Theorem 3.1). Moreover, in this case, we prove that the profile likelihood function in fact (a) is strictly convex with probability 1 and that, in this case, the maximum likelihood estimator exists; (b) is constant with probability 1; or (c) does not exist for any observations (Theorem 3.2). In Section 4, an algorithm to find the maximum likelihood estimator is discussed, and, finally, in Section 5, we conjecture that the

Received November 2001; revised March 2003.

AMS 2000 subject classifications. Primary 62H12, 62H15; secondary 62H10, 62H20, 62F10, 17C50, 52A20.

Key words and phrases. Proportional covariances, maximum likelihood estimation, profile likelihood function, natural exponential families, convexity, Jordan algebra.

only possible situations where a unique solution to the likelihood equations [for the unknown parameters $(\lambda_1, \ldots, \lambda_n)$ and θ] exists with probability 1 are those where M is a Jordan algebra.

For practical applications, note that the assumption of zero means is not essential. In fact, we can assume that $\mu_i = \mathrm{E}(X_i) \in L$, $i = 1, \ldots, n$, where L is a subspace of $\mathbb{R}^p$ such that

$$\forall \theta \in M : \theta L = L. \tag{1.1}$$

[This can be justified as follows. The condition (1.1) implies that the orthogonal complement to L w.r.t. the inner product on $\mathbb{R}^p$ given by θ is independent of θ. For $(\log(\lambda_1), \ldots, \log(\lambda_n)) \in U$ and $\theta \in M$ fixed, the maximum likelihood estimator for $(\mu_1, \ldots, \mu_n)$ is then given as $\widehat{\mu}_i = PX_i$, $i = 1, \ldots, n$, where P is the orthogonal projection matrix of $\mathbb{R}^p$ onto L (w.r.t. all $\theta \in M$). The profile likelihood function obtained by substitution of $(\mu_1, \ldots, \mu_n)$ by $(\widehat{\mu}_1, \ldots, \widehat{\mu}_n)$ is then proportional to the likelihood function based on the residuals $(Y_1, \ldots, Y_n) = ((I - P)X_1, \ldots, (I - P)X_n)$. Here we have $\mathrm{E}(Y_i) = 0$ and $\mathrm{V}(Y_i) = \lambda_i \Gamma$, where $\Gamma = (I - P)\Sigma(I - P)$, and since linear restrictions on $\theta = \Sigma^{-1}$ imply linear restrictions on Γ^{-1}, we are thus faced with an estimation problem similar to the one stated above.]

Applications of models with the type of restrictions (i) are, for example, situations with observation vectors $X_1, \ldots, X_n$ from independent, balanced orthogonal designs of the same type. The restrictions imposed by random factors of a balanced orthogonal design are linear in the inverse covariance (except for the constraints of nonnegative variance components); see Section 3. In this case, the hypothesis of proportional covariances is equivalent to the additional assumption that correlations between observations are equal across the n designs but the variances ($\equiv$ the sum of all variance components) can vary freely.

Hypotheses of this type occur in problems related to quantitative genetics [see Lynch and Walsh (1998)], where the simplest example is to consider measurements of some quantitative trait in two independent populations of animal offsprings in a so-called *paternal half-sib design*: a number (say m) of males are sequentially mated to a number (say r) of females (i.e., no two males are mated with the same female), and one representative from each litter of offsprings is selected. Offsprings thus occur in clusters of half-sibs of size r.

We therefore have the two observation vectors $X_1 = (X_{1jk} \mid j = 1, \ldots, m, k = 1, \ldots, r)$ and $X_2 = (X_{2jk} \mid j = 1, \ldots, m, k = 1, \ldots, r)$, where X_{ijk} is the value of the quantitative trait of the offspring in the ith population, $i = 1, 2$, that stems from the kth mating of the jth male to a female. The statistical assumption is that

$$X_{ijk} = \mu_i + Y_{ij} + \varepsilon_{ijk},$$

where $\{Y_{ij} \sim \mathrm{N}(0, \sigma^2_{ig})\}$, $\{\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2_{ie})\}$ are independent variables and $\mu_i \in \mathbb{R}$, $\sigma^2_{ie}, \sigma^2_{ig} > 0$ are the unknown parameters.

The components σ_{ig}^2 and σ_{ie}^2 are in genetics theory interpreted as the parts of the total variation in the trait across the ith population, $i = 1, 2$, due to genetic and environmental causes, respectively, and the ratio $\sigma_{ig}^2/(\sigma_{ig}^2 + \sigma_{ie}^2)$ ($\equiv$ the intraclass correlation between half-sibs) is called the *heritability*. The hypothesis of proportionality between the covariances of the two observation vectors X_1 and X_2 then corresponds to the hypothesis of equal heritabilities in the two populations. [The "genetic" component of variance σ_{ig}^2 can by the theory of genetics be split into two parts, the *additive genetic variance* σ_{ia}^2 and the *dominance genetic variance* σ_{id}^2, and sometimes the heritability is defined as $\sigma_{ia}^2/(\sigma_{ig}^2 + \sigma_{ie}^2)$, $i = 1, 2$; see Lynch and Walsh (1998) or the *Encyclopedia of Biostatistics*, pages 1905 and 1906. In this case, the hypothesis of equal heritabilities does not coincide with that of proportional covariances.]

Applications of models with the type of restrictions (ii) could be the (common) situation where U is a regression subspace given by vectors of covariates $z_1, \ldots, z_n \in \mathbb{R}^r$ in which case

$$\mathrm{V}(X_i) = \exp(\beta_i' z_i)\Sigma,$$

where $\beta_i \in \mathbb{R}^r$, $i = 1, \ldots, n$. In the one-dimensional case ($p = 1$), this type of relationship among variances has previously been studied; see Cook and Weisberg (1983) and Aitkin (1987).

2. Convexity of the profile likelihood function. In this section, a necessary and sufficient condition for convexity of the profile likelihood function is derived. Since the condition does not involve the subspace U [which determines the restrictions on $(\log(\lambda_1), \ldots, \log(\lambda_n))$], we shall in the following assume that $U = \mathbb{R}^n$ (eventually see Remark 2.2 for further details).

For computational convenience, we shall formulate the model in the context of natural exponential families [cf., e.g., Morris (1982), Casalis (1996) or Pace and Salvan (1997)]. Therefore, in the following let Y denote a random vector taking values in $\mathbb{R}^k$, $k \in \mathbb{N}$, and assume that the distribution of Y belongs to a natural exponential family (NEF) $(P_\theta \mid \theta \in \Theta)$ of order k; that is, there exists a σ-finite measure ν on $\mathbb{R}^k$ such that, for $y \in \mathbb{R}^k$,

$$\frac{dP_\theta}{d\nu}(y) = \frac{1}{L(\theta)} e^{\langle \theta, y \rangle},$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^k$, that is, $\langle \theta, y \rangle = \theta' y$, θ is the canonical (natural) parameter, $L(\theta)$ is the norming constant ($\equiv$ the Laplace transform of ν) and Θ is the canonical (natural) parameter space ($\equiv \{\theta \in \mathbb{R}^k \mid L(\theta) < \infty\}$). We shall assume that the NEF is regular; that is, Θ is an open nonempty set.

For $\theta \in \Theta$, we denote by $\tau(\theta)$ the mean value parameter, that is, $\tau(\theta) = \mathrm{E}_\theta(Y)$, and by $\mathrm{V}(\theta)$ the variance parameter, that is, $\mathrm{V}(\theta) = \mathrm{V}_\theta(Y)$. We denote by C the

convex support of ν. We can, without loss of generality, assume that the family is *minimal* [cf. Barndorff-Nielsen (1978), page 112], that is, $\text{int}(C) \neq \varnothing$. In this case, τ is one-to-one, $\tau(\Theta) = \text{int}(C)$ and $V(\theta)$ is nonsingular for all $\theta \in \Theta$. We shall denote the inner product on $\mathbb{R}^k$ associated with $V(\theta)^{-1}$ by $\langle \cdot, \cdot \rangle_{V(\theta)^{-1}}$, that is, $\langle y_1, y_2 \rangle_{V(\theta)^{-1}} = \langle y_1, V(\theta)^{-1} y_2 \rangle = y_1' V(\theta)^{-1} y_2$ for $y_1, y_2 \in \mathbb{R}^k$.

In the following, we shall assume that the NEF is scale invariant; that is, if Y has the distribution P_θ, then, for all $\lambda > 0$, λY has the distribution $P_{\theta/\lambda}$ ($\theta \in \Theta$).

First, we note that, due to the scale invariance, C is a cone, that is, $\lambda y \in C$ for $\lambda > 0$ and $y \in C$. This can be seen as follows: in general (for exponential families), we have that the convex support of P_θ equals C for all $\theta \in \Theta$. If y has the distribution P_θ, the convex support of the distribution of λy is λC. On the other hand, due to the scale invariance, the distribution of λy is $P_{\theta/\lambda}$, implying that the convex support is C.

Furthermore, for scale-invariant NEF's we have the following lemma.

LEMMA 2.1. *For* $\theta \in \Theta$,

$$\tau(\theta) = -V(\theta)\theta.$$

PROOF. It follows from the definition that

$$\tau\left(\frac{\theta}{\lambda}\right) = \lambda \tau(\theta), \tag{2.1}$$

and by differentiation of (2.1) w.r.t. λ we get

$$-\frac{1}{\lambda^2} V\left(\frac{\theta}{\lambda}\right)\theta = \tau(\theta) \tag{2.2}$$

for $\lambda > 0$. The lemma now follows by evaluating (2.2) at $\lambda = 1$. □

Next, we consider independent observations $y_1, \ldots, y_n$ such that y_i has the distribution P_{θ_0/λ_i}, $i = 1, \ldots, n$, where $\theta_0 \in \Theta$ is unknown and $\lambda_1, \ldots, \lambda_n$ are unknown positive scalars. To obtain a one-to-one parametrization, we assume that $\prod_{i=1}^n \lambda_i = 1$. Since $y_i \in C$ with probability 1, $i = 1, \ldots, n$, we shall assume that this is the case. The negative logarithm of the likelihood function becomes

$$l(\theta_0, \lambda_1, \ldots, \lambda_n) = \sum_{i=1}^n \log\left(L\left(\frac{\theta_0}{\lambda_i}\right)\right) - \theta_0'\left(\sum_{i=1}^n \frac{y_i}{\lambda_i}\right),$$

$\theta_0 \in \Theta$, $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$ and $\prod_{i=1}^n \lambda_i = 1$. The following lemma gives a necessary and sufficient condition for the existence of the profile likelihood function.

LEMMA 2.2. *For $(\lambda_1, \ldots, \lambda_n)$ fixed, l has a unique minimum $\widehat{\theta}_0$ if and only if $(1/n)\sum_{i=1}^n y_i \in \operatorname{int}(C)$. In this case, θ_0 is given by the likelihood equation*

$$\tau(\widehat{\theta}_0) = \frac{1}{n}\sum_{i=1}^n \frac{y_i}{\lambda_i}. \tag{2.3}$$

PROOF. It is well known [cf. Barndorff-Nielsen (1978), page 151] that, for $(\lambda_1, \ldots, \lambda_n)$ fixed, l has a unique minimum $\widehat{\theta}_0$ if and only if $(1/n)\sum_{i=1}^n y_i/\lambda_i \in \operatorname{int}(C)$ and that, in this case, $\widehat{\theta}_0$ is given by the likelihood equation (2.3). We shall show that, if $(1/n)\sum_{i=1}^n y_i \in \operatorname{int}(C)$, then $(1/n)\sum_{i=1}^n y_i/\lambda_i \in \operatorname{int}(C)$ for all $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$.

We can write

$$\frac{1}{n}\sum_{i=1}^n \frac{y_i}{\lambda_i} = \alpha\frac{1}{n}\sum_{i=1}^n y_i + (1-\alpha)\frac{1}{n}\sum_{i=1}^n \mu_i y_i,$$

where $\alpha \in]0, 1[$ is chosen such that $\alpha < \min(1/\lambda_1, \ldots, 1/\lambda_n)$ and $\mu_i = 1/(1-\alpha) \times (1/\lambda_i - \alpha)$, $i = 1, \ldots, n$. Since $y_i \in C$, $i \in 1, \ldots, n$, and C is a convex cone we must have $(1/n)\sum_{i=1}^n \mu_i y_i \in C$. The lemma now follows from the standard theory for convex sets [cf. Rockafellar (1970), Theorem 6.1] since, by assumption, $(1/n)\sum_{i=1}^n y_i \in \operatorname{int}(C)$. □

The following result then gives a necessary and sufficient condition for convexity of the profile likelihood function. We consider an arbitrary vector of observations $(y_1, \ldots, y_n)$ such that the profile likelihood function exists, that is, $(1/n)\sum_{i=1}^n y_i \in \operatorname{int}(C)$ (cf. Lemma 2.2).

THEOREM 2.1. *The profile likelihood function $\widehat{l}(\lambda_1, \ldots, \lambda_n) = l(\widehat{\theta}_0, \lambda_1, \ldots, \lambda_n)$ is convex as a function of $(\log(\lambda_1), \ldots, \log(\lambda_n))$ if and only if*

$$\sum_{m=1}^n \sum_{j\,:\,m<j} (c_m - c_j)^2 \left\langle \frac{y_m}{\lambda_m}, \frac{y_j}{\lambda_j} \right\rangle_{V(\widehat{\theta}_0)^{-1}} \geq 0 \tag{2.4}$$

for all $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ and all $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$, where $\prod_{i=1}^n \lambda_i = 1$.

PROOF. The profile likelihood function $\widehat{l}$ is given by

$$\widehat{l}(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^n \log\left(L\left(\frac{\widehat{\theta}_0}{\lambda_i}\right)\right) - \widehat{\theta}_0'\left(\sum_{i=1}^n \frac{y_i}{\lambda_i}\right).$$

By differentiation,

$$\begin{aligned}\frac{\partial \widehat{l}}{\partial \lambda_j} &= \sum_{i=1}^n \tau\left(\frac{\widehat{\theta}_0}{\lambda_i}\right)' \frac{\partial \widehat{\theta}_0}{\partial \lambda_j}\frac{1}{\lambda_i} - \tau\left(\frac{\widehat{\theta}_0}{\lambda_j}\right)'\frac{\widehat{\theta}_0}{\lambda_j^2} - \frac{\partial \widehat{\theta}_0'}{\partial \lambda_j}\left(\sum_{i=1}^n \frac{y_i}{\lambda_i}\right) + \widehat{\theta}_0'\frac{y_j}{\lambda_j^2} \\ &= \widehat{\theta}_0'\left(\frac{y_j}{\lambda_j^2} - \tau(\widehat{\theta}_0)\frac{1}{\lambda_j}\right),\end{aligned}$$

where (2.1) and (2.3) are used, and by setting $\beta_j = \log(\lambda_j)$, we get

$$\begin{aligned}\frac{\partial \widehat{l}}{\partial \beta_j} &= \widehat{\theta}_0'\Big(\frac{y_j}{\lambda_j^2} - \tau(\widehat{\theta}_0)\frac{1}{\lambda_j}\Big)\lambda_j \\ &= \widehat{\theta}_0'\Big(\frac{y_j}{\lambda_j} - \tau(\widehat{\theta}_0)\Big),\end{aligned} \tag{2.5}$$

$j = 1, \ldots, n$. Differentiation of (2.3) yields

$$\frac{\partial \tau(\widehat{\theta}_0)}{\partial \lambda_j} = -\frac{1}{n}\frac{y_j}{\lambda_j^2}, \tag{2.6}$$

but, on the other hand, by successive differentiation,

$$\frac{\partial \tau(\widehat{\theta}_0)}{\partial \lambda_j} = \mathrm{V}(\widehat{\theta}_0)\frac{\partial \widehat{\theta}_0}{\partial \lambda_j}, \tag{2.7}$$

and thus combining (2.6) and (2.7) yields

$$\mathrm{V}(\widehat{\theta}_0)\frac{\partial \widehat{\theta}_0}{\partial \lambda_j} = -\frac{1}{n}\frac{y_j}{\lambda_j^2}, \tag{2.8}$$

$j = 1, \ldots, n$. Then, by differentiation of (2.5),

$$\begin{aligned}\frac{\partial^2 \widehat{l}}{\partial \lambda_j\, \partial \beta_j} &= \frac{\partial \widehat{\theta}_0'}{\partial \lambda_j}\Big(\frac{y_j}{\lambda_j} - \tau(\widehat{\theta}_0)\Big) - \widehat{\theta}_0'\Big(\frac{y_j}{\lambda_j^2} + \frac{\partial \tau(\widehat{\theta}_0)}{\partial \lambda_j}\Big) \\ &= \tau(\widehat{\theta}_0)'\mathrm{V}(\widehat{\theta}_0)^{-1}\frac{y_j}{\lambda_j^2} - \frac{1}{n}\frac{y_j'}{\lambda_j^2}\mathrm{V}(\widehat{\theta}_0)^{-1}\frac{y_j}{\lambda_j},\end{aligned}$$

where (2.6), (2.8) and Lemma 2.1 have been used. Hence,

$$\begin{aligned}\frac{\partial^2 \widehat{l}}{\partial \beta_j^2} &= \frac{\partial^2 \widehat{l}}{\partial \lambda_j\, \partial \beta_j}\lambda_j \\ &= \tau(\widehat{\theta}_0)'\mathrm{V}(\widehat{\theta}_0)^{-1}\frac{y_j}{\lambda_j} - \frac{1}{n}\frac{y_j'}{\lambda_j}\mathrm{V}(\widehat{\theta}_0)^{-1}\frac{y_j}{\lambda_j},\end{aligned} \tag{2.9}$$

$j = 1, \ldots, n$. Analogously, we get

$$\frac{\partial^2 \widehat{l}}{\partial \beta_m\, \partial \beta_j} = -\frac{1}{n}\frac{y_m'}{\lambda_m}\mathrm{V}(\widehat{\theta}_0)^{-1}\frac{y_j}{\lambda_j} \tag{2.10}$$

for $m \neq j$, $m, j = 1, \ldots, n$.

To prove convexity of the profile likelihood function, we shall show that the matrix $\mathrm{D}^2\widehat{l}$ determined by (2.9) and (2.10) is positive semidefinite for all $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$. For notational convenience we let $z_j = y_j/\lambda_j$,

$j = 1, \ldots, n$. Note that $\tau(\widehat{\theta}_0) = (1/n)\sum_{i=1}^n z_i$ by (2.3). For $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$, we then have

$$\begin{aligned} c'(\mathrm{D}^2\widehat{l})c &= \sum_{m=1}^n \sum_{j=1}^n c_m \frac{\partial^2 \widehat{l}}{\partial\beta_m\,\partial\beta_j} c_j \\ &= \sum_{m=1}^n c_m^2 z_m' \mathrm{V}(\widehat{\theta}_0)^{-1}\left(\frac{1}{n}\sum_{i=1}^n z_i\right) - \frac{1}{n}\sum_{m=1}^n c_m^2 z_m' \mathrm{V}(\widehat{\theta}_0)^{-1} z_m \\ &\quad - \frac{1}{n}\sum_{m=1}^n \sum_{j\,:\,j\neq m} c_m c_j z_m' \mathrm{V}(\widehat{\theta}_0)^{-1} z_j \\ &= \frac{1}{n}\sum_{m=1}^n \sum_{j=1}^n c_m^2 z_m' \mathrm{V}(\widehat{\theta}_0)^{-1} z_j - \frac{1}{n}\sum_{m=1}^n \sum_{j=1}^n c_m c_j z_m' \mathrm{V}(\widehat{\theta}_0)^{-1} z_j \\ &= \frac{1}{n}\sum_{m=1}^n \sum_{j\,:\,j\neq m} (c_m^2 - c_m c_j) z_m' \mathrm{V}(\widehat{\theta}_0)^{-1} z_j \\ &= \frac{1}{n}\sum_{m=1}^n \sum_{j\,:\,m<j} (c_m - c_j)^2 z_m' \mathrm{V}(\widehat{\theta}_0)^{-1} z_j, \end{aligned}$$

and, hence, the theorem follows. □

COROLLARY 2.1. *The profile likelihood function* $\widehat{l}(\lambda_1, \ldots, \lambda_n) = l(\widehat{\theta}_0, \lambda_1, \ldots, \lambda_n)$ *is convex as a function of* $(\log(\lambda_1), \ldots, \log(\lambda_n))$ *if*

$$\forall \theta \in \Theta\ \forall m < j : \langle y_m, y_j \rangle_{\mathrm{V}(\theta)^{-1}} \geq 0. \tag{2.11}$$

PROOF. If (2.11) is fulfilled, then clearly (2.4) is fulfilled for all $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ and all $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$. □

REMARK 2.1. If the inequality in (2.4) is strict for all $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ where not all coordinates are equal and all $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$, where $\prod_{i=1}^n \lambda_i = 1$, then the profile likelihood function is strictly convex.

In the following, S_p denotes the vector space of all symmetric $p \times p$ matrices, P_p the cone of all positive-definite $p \times p$ matrices and PS_p the cone of all positive-semidefinite $p \times p$ matrices ($p \in \mathbb{N}$). In our specific case study, we consider, as the underlying exponential family model, the normal distributions $\mathrm{N}(0, \Sigma)$ on $\mathbb{R}^p$, with the restriction $\Sigma^{-1} \in M_+ \equiv M \cap P_p$, where M denotes a subspace of S_p such that $M_+ \neq \varnothing$. For $\Sigma \in P_p$, we denote the inner product on S_p associated with the Kronecker product $\Sigma \otimes \Sigma$ by $\langle \cdot, \cdot \rangle_\Sigma$, that is, $\langle A, B \rangle_\Sigma = \mathrm{tr}(A\Sigma B\Sigma)$, $A, B \in S_p$, and we denote by Q_Σ the orthogonal projection onto M w.r.t. $\langle \cdot, \cdot \rangle_\Sigma$. We let $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_I$ ($\equiv$ the standard inner product on S_p) and $Q = Q_I$.

The density of $\mathrm{N}(0, \Sigma)$ w.r.t. the Lebesgue measure on $\mathbb{R}^p$ can be written as

$$\frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}}\exp\left\{\left\langle -\frac{1}{2}\Sigma^{-1}, Q(xx')\right\rangle\right\},$$

where $x \in \mathbb{R}^p$. As the NEF in the above context, we thus consider instead the family of (generalized Wishart) distributions of $y = Q(xx')$, where x is distributed according to $\mathrm{N}(0, \Sigma)$ and $\Sigma^{-1} \in M_+$. (In the simple case where $M = S_p$, $y = xx'$ is the Wishart distribution with one degree of freedom.) We have thus that the canonical parameter is $\theta = -\frac{1}{2}\Sigma^{-1}$, the canonical parameter set is $\Theta = -M_+$, the mean value parameter is

$$\tau(\theta) = \mathrm{E}_\theta\big(Q(xx')\big) = Q\big(\mathrm{E}_\theta(xx')\big) = Q(\Sigma) = -\tfrac{1}{2}Q(\theta^{-1}),$$

and the variance function is (by differentiation of τ) given by

$$\mathrm{V}(\theta)(B) = \mathrm{D}\tau(\theta)(B) = \tfrac{1}{2}Q(\theta^{-1}B\theta^{-1}) \tag{2.12}$$

for $B \in M$, $\theta \in \Theta$. Furthermore, $\mathrm{int}(C) = \tau(\Theta) = Q(M_+^{-1})$. Note that since τ is a mapping from $-M_+$ to $Q(M_+^{-1}) \subseteq M$, $\mathrm{V}(\theta)$ defines a linear mapping of M onto itself ($\theta \in \Theta$).

First, we give an equivalent algebraic condition for the case where (2.11) is fulfilled for all $(y_1, \ldots, y_n) \equiv (Q(x_1x_1'), \ldots, Q(x_nx_n'))$, $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$.

THEOREM 2.2. *For $\theta = -\frac{1}{2}\Sigma^{-1} \in \Theta$, the conditions*

$$Q_\Sigma(PS_p) \subseteq PS_p, \tag{2.13}$$

$$\forall\, x_1, x_2 \in \mathbb{R}^p : \langle Q(x_1x_1'), Q(x_2x_2')\rangle_{\mathrm{V}(\theta)^{-1}} \geq 0 \tag{2.14}$$

are equivalent.

PROOF. Let $\theta = -\frac{1}{2}\Sigma^{-1} \in \Theta$. For $x_2 \in \mathbb{R}^p$, we define $A(x_2) = \mathrm{V}(\theta)^{-1} \times Q(x_2x_2')$, that is, $A(x_2) \in M$ and $Q(x_2x_2') = \frac{1}{2}Q(\theta^{-1}A(x_2)\theta^{-1})$, and (2.14) is thus that $\mathrm{tr}(Q(x_1x_1')A(x_2)) = \mathrm{tr}(x_1x_1'A(x_2)) = x_1'A(x_2)x_1 \geq 0$ for all $x_1, x_2 \in \mathbb{R}^p$ or, equivalently, that $A(x_2)$ is positive semidefinite for all $x_2 \in \mathbb{R}^p$. But for $x_2 \in \mathbb{R}^p$ we have

$$\begin{aligned}\mathrm{tr}\big(\Sigma A(x_2)\Sigma B\big) &= \mathrm{tr}\big(\tfrac{1}{4}Q\big(\theta^{-1}A(x_2)\theta^{-1}\big)B\big)\\ &= \mathrm{tr}\big(\tfrac{1}{2}Q(x_2x_2')B\big) = \mathrm{tr}\big(\tfrac{1}{2}x_2x_2'B\big) = \mathrm{tr}\big(\Sigma(2\theta x_2x_2'\theta)\Sigma B\big)\end{aligned}$$

for all $B \in M$, which implies that $A(x_2) = Q_\Sigma(2\theta x_2x_2'\theta) = Q_\Sigma(2(\theta x_2)(\theta x_2)')$. From this expression, it then follows that $A(x_2)$ is positive semidefinite for all $x_2 \in \mathbb{R}^p$ if and only if $Q_\Sigma(PS_p) \subseteq PS_p$. □

For an arbitrary vector of observations $(x_1, \ldots, x_n)$ for which the profile likelihood function exists, that is, $Q((1/n)\sum_{i=1}^n x_ix_i') \in Q(M_+^{-1})$ (cf. Lemma 2.2), we then obtain the following corollary directly from Theorems 2.1 and 2.2.

COROLLARY 2.2. *If* $Q_\Sigma(PS_p) \subseteq PS_p$ *for all* $\Sigma^{-1} \in M_+$, *then the profile likelihood function is convex as a function of* $(\log(\lambda_1), \ldots, \log(\lambda_n))$.

REMARK 2.2. Due to the fact that the restriction to a linear subspace of a convex function is convex, all results about convexity of the profile likelihood function for the investigated model hold as well for the case where $(\log(\lambda_1), \ldots, \log(\lambda_n))$ is restricted to vary in some subspace U of $\mathbb{R}^n$ (we still impose the restriction $\prod_{i=1}^n \lambda_i = 1$ in order to obtain a one-to-one parametrization).

3. The Jordan algebra case. In this section, we consider the case where M is a *Jordan algebra*, that is, $I \in M$ and $AB + BA \in M$ for all $A, B \in M$. This general type of covariance hypothesis was introduced by Jensen (1988), who gave a complete characterization of the corresponding normal models. It turns out that the models correspond to products of i.i.d. normal models where the observations have a covariance matrix with real ($\equiv$ the unrestricted case), complex or quaternion structure, or a parametrization given by means of the Clifford algebra. Products of normal models with covariances given by any of the first three cases coincide with the so-called group symmetry models [for a brief introduction, see Andersson and Madsen (1998), Appendix A].

The condition that M is a Jordan algebra can equivalently be formulated as $M_+^{-1} = M_+$. Note that, in particular, when M corresponds to restrictions given by random factors of a balanced orthogonal design, M is a Jordan algebra. This follows from the fact that M in this case can be parametrized as $M = \{\alpha_1 Q_1 + \cdots + \alpha_m Q_m \mid (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m\}$, where $Q_1, \ldots, Q_m$ are the orthogonal projection matrices of a partition of $\mathbb{R}^p$, that is, $Q_i^2 = Q_i$ and $Q_i Q_j = 0$ for $i \neq j$, $i, j = 1, \ldots, m$ [cf. Tjur (1984)].

First, we prove that the assumption in Corollary 2.2 is fulfilled.

THEOREM 3.1. *If* M *is a Jordan algebra, then* $Q_\Sigma(PS_p) \subseteq PS_p$ *for all* $\Sigma^{-1} \in M_+$.

PROOF. For $A \in S_p$ and $\Sigma \in P_p$, we have, in general, that $\langle A, B\rangle_\Sigma = \langle Q_\Sigma(A), B\rangle_\Sigma$ for all $B \in M$. Now let $\Sigma^{-1} \in M_+$. Since M is a Jordan algebra, $\Sigma \in M_+$, and, furthermore,

$$\Sigma B \Sigma = \tfrac{1}{2}\big((\Sigma B + B\Sigma)\Sigma + \Sigma(B\Sigma + \Sigma B) - (B\Sigma^2 + \Sigma^2 B)\big),$$

that is, $\Sigma B \Sigma \in M$ for $B \in M$. Thus, on the other hand, for $A \in S_p$,

$$\langle A, B\rangle_\Sigma = \langle A, \Sigma B \Sigma\rangle = \langle Q(A), \Sigma B \Sigma\rangle = \langle Q(A), B\rangle_\Sigma$$

for all $B \in M$, implying that $Q_\Sigma(A) = Q(A)$. Hence, $Q_\Sigma = Q$ for all $\Sigma^{-1} \in M_+$. It therefore suffices to show that $Q(xx') \in PS_p$ for all $x \in \mathbb{R}^p$. But $Q(xx') \in C$ and since

$$\operatorname{int}(C) = \tfrac{1}{2} Q(M_+^{-1}) = \tfrac{1}{2} Q(M_+) = \tfrac{1}{2} M_+ = M_+ \subseteq P_p,$$

we must have $Q(xx') \in PS_p$, $x \in \mathbb{R}^p$. $\square$

Thus, in particular, it follows from Corollary 2.2 that in the case where M is a Jordan algebra, the profile likelihood function [for observations $(x_1, \ldots, x_n)$ for which it exists] is convex as a function of $(\log(\lambda_1), \ldots, \log(\lambda_n))$. We shall now prove that, in fact, the profile likelihood function (a) is strictly convex with probability 1, (b) is constant with probability 1 or (c) does not exist for any observations.

THEOREM 3.2. *If M is a Jordan algebra, then either the profile likelihood function exists with probability* 1 *or it never exists for any observation* $(x_1, \ldots, x_n)$. *In the case where the profile likelihood function exists, it is either strictly convex as a function of* $(\log(\lambda_1), \ldots, \log(\lambda_n))$ *with probability* 1 *or it is constant for all observations. Furthermore, the maximum likelihood estimator exists with probability* 1 *if the profile likelihood function is strictly convex.*

PROOF. First note that, when M is a Jordan algebra, the variance function (2.12) becomes

$$\mathrm{V}(\theta)(B) = \tfrac{1}{2}\theta^{-1}B\theta^{-1}$$

since $\theta^{-1}B\theta^{-1} \in M$, and thus

$$\mathrm{V}(\theta)^{-1}(A) = 2\theta A\theta$$

for $A, B \in M$, $\theta \in \Theta$.

From Jensen (1988), it follows that the model corresponds to a product of i.i.d. normal models where the observations have a covariance matrix with real ($\equiv$ the unrestricted case), complex or quaternion structure, or a parametrization given by means of the Clifford algebra. The profile likelihood function therefore becomes a sum of the profile likelihoods for the factors in the product. In the real case, the profile likelihood function does not exist for $n < p$ (this follows directly from Lemma 2.2). For $n = p$, it is constant, since in this case [by (2.5)]

$$\begin{aligned}\frac{\partial \hat{l}}{\partial \beta_j} &= -\frac{p}{2}\operatorname{tr}\left(\left(\sum_{i=1}^{p}\frac{x_i x_i'}{\lambda_i}\right)^{-1}\left(\frac{x_j x_j'}{\lambda_j} - \frac{1}{p}\sum_{i=1}^{p}\frac{x_i x_i'}{\lambda_i}\right)\right)\\ &= -\frac{p}{2}\left(\frac{x_j'}{\lambda_j}\left(\sum_{i=1}^{p}\frac{x_i x_i'}{\lambda_i}\right)^{-1}\frac{x_j}{\lambda_j} - 1\right) = 0,\end{aligned}$$

where the last equation is due to the fact that, in general,

$$z_j'\left(\sum_{i=1}^{p} z_i z_i'\right)^{-1} z_j = 1$$

for $z_1, \ldots, z_p \in \mathbb{R}^p$, $j = 1, \ldots, p$.

For $n > p$, we let Ω denote the set of observations $(x_1, \ldots, x_n)$ for which any subfamily $(x_{i_1}, \ldots, x_{i_r})$ of size $r \leq p$ is linearly independent. Clearly, an observation $(x_1, \ldots, x_n)$ belongs to this set with probability 1, and we shall show that the profile likelihood function for such an observation is strictly convex. Thus, assume that, for $(x_1, \ldots, x_n) \in \Omega$, the profile likelihood function is *not* strictly convex. Then, by Remark 2.1, there exists $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ where not all coordinates are equal and $(\lambda_1, \ldots, \lambda_n) \in]0, \infty[^n$ where $\prod_{i=1}^n \lambda_i = 1$, such that

$$
\begin{aligned}
0 &= \sum_{m=1}^{n} \sum_{j\,:\,m<j} (c_m - c_j)^2 \langle z_m z_m', z_j z_j' \rangle_{\mathrm{V}(\widehat{\theta})^{-1}} \\
&= \frac{1}{2n} \sum_{m=1}^{n} \sum_{j\,:\,m<j} (c_m - c_j)^2 \operatorname{tr}\left(z_m z_m' \left(\sum_{i=1}^{n} z_i z_i' \right)^{-1} z_j z_j' \left(\sum_{i=1}^{n} z_i z_i' \right)^{-1} \right) \\
&= \frac{1}{2n} \sum_{m=1}^{n} \sum_{j\,:\,m<j} (c_m - c_j)^2 \left(z_m' \left(\sum_{i=1}^{n} z_i z_i' \right)^{-1} z_j \right)^2,
\end{aligned}
$$

where $z_i = x_i / \sqrt{\lambda_i}$, $i = 1, \ldots, n$. This equation implies that

$$
z_m' \left(\sum_{i=1}^{n} z_i z_i' \right)^{-1} z_j = 0,
$$

whenever $c_m \neq c_j$, that is, $z_m \perp z_j$ w.r.t. the inner product on $\mathbb{R}^p$ given by

$$
\left(\sum_{i=1}^{n} z_i z_i' \right)^{-1}.
$$

Let $I_1, \ldots, I_r$ be the partitioning of $I = \{1, \ldots, n\}$ in nonempty sets such that i, j belong to the same set I_l if and only if $c_i = c_j$, $i, j = 1, \ldots, n$, $l = 1, \ldots, r$. The corresponding families $(z_{i_1})_{i_1 \in I_1}, \ldots, (z_{i_r})_{i_r \in I_r}$ are thus linearly independent. Since $n > p$, there exists an l, $l = 1, \ldots, r$, such that the vectors $(z_i)_{i \in I_l}$ are linearly dependent, and we must have $|I_l| > p$ since any subset of $\{z_1, \ldots, z_n\}$ of size less than or equal to p is linearly independent. For the same reason, we must have $\dim(\operatorname{span}\{z_i : i \in I_l\}) = p$, contradicting the fact that $r \geq 2$.

In the complex (quaternion) case, we consider observations $x_1, \ldots, x_n \in \mathbb{R}^{2p}$ ($\mathbb{R}^{4p}$) that have a one-to-one correspondence to observations $y_1, \ldots, y_n \in \mathbb{C}^p$ ($\mathbb{H}^p$) [see Andersson (1975) or Andersson, Brøns and Jensen (1983)]. In this case, the symmetric $2p \times 2p$ ($4p \times 4p$) matrix $Q(x_i x_i')$ corresponds to the Hermitian $p \times p$ matrix $y_i y_i'$, $i = 1, \ldots, n$, and thus the above considerations for the real case apply in this situation too, since these only use basic results about finite-dimensional vector spaces (over arbitrary fields). In the case of the Clifford algebra, it follows from Jensen [(1988), Theorem 4] that $Q(x_1 x_1'), \ldots, Q(x_n x_n')$ are

all positive definite with probability 1, except for four cases where $(\dim(M), p) = (3, 2), (4, 4), (6, 8)$ or $(10, 16)$. Thus, in this situation, the inequality in (2.4) is strict for all $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ where not all coordinates are equal, and hence the profile likelihood function is strictly convex for all $n \geq 2$. The four exceptional cases can all be handled as the two-dimensional complex normal distribution, and it is seen that the profile likelihood function is constant for $n = 2$ and strictly convex for $n \geq 3$.

By a similar inspection of the above cases, it can be seen that the maximum likelihood estimator exists with probability 1 if the profile likelihood function is strictly convex. □

4. Calculation of the maximum likelihood estimator. For the model given at the beginning of the Introduction, the likelihood equations are

$$\lambda_i = \operatorname{tr}(\Sigma^{-1} X_i X_i^*)/C, \tag{4.1}$$

where $i = 1, \ldots, n$ and $C = \prod_{i=1}^n \operatorname{tr}(\Sigma^{-1} X_i X_i^*)$, and

$$Q\left(\sum_{i=1}^n X_i X_i^*/\lambda_i\right) = n\,Q(\Sigma). \tag{4.2}$$

For a specific model, the problem is to solve (4.2) in Σ. By Lemma 2.1, the solution is unique if it exists.

The equation is also given in Anderson (1969, 1970) and can, in general, be solved by the Newton–Raphson iterative process; see Anderson (1970), (4.7)–(4.9).

In the Jordan algebra case, one has the linear equation

$$\Sigma = Q\left(\sum_{i=1}^n X_i X_i^*/\lambda_i\right)\Big/ n.$$

For a decomposable covariance selection model, an explicit solution is given by Lauritzen [(1996), Proposition 5.9], and for a general covariance selection model, the equation can be solved by an IPS algorithm; compare Lauritzen (1996), Theorem 5.4.

The iterative process given in Eriksen (1987), Theorem 3.2, can then be generalized by successively solving (4.1) and (4.2). If the maximum likelihood estimator exists and there is a unique solution to the likelihood equations, the algorithm converges to the maximum likelihood estimator. If the likelihood function has several local maxima, the algorithm may converge to any one of them depending on the starting value.

For a decomposable covariance selection model, however, there might be several solutions to the likelihood equations (4.1) and (4.2).

5. Further investigations. In this section, we present two additional conjectures, which, unfortunately, we have not been able to prove, except in the case $\dim(M) = 2$. The first is purely algebraic and concerns the converse of Theorem 3.1.

CONJECTURE 1. *If $Q_\Sigma(PS_p) \subseteq PS_p$ for all $\Sigma^{-1} \in M_+$, then M is a Jordan algebra.*

Note that, for $\Sigma \in P_p$, the condition $Q_\Sigma(PS_p) \subseteq PS_p$ can be formulated alternatively that the cone M is self-dual [see Faraut and Korányi (1994)] w.r.t. the inner product $\langle \cdot, \cdot \rangle_\Sigma$ on S_p.

The second conjecture states that the only cases where a unique solution to the likelihood equations for the unknown parameters $(\lambda_1, \ldots, \lambda_n)$ and θ exists with probability 1 are those where M is a Jordan algebra. Thus, in particular (if the conjecture holds), we cannot have existence and uniqueness of the ML estimator with probability 1, except in the case where M is a Jordan algebra. In this situation, the profile likelihood function is then strictly convex as a function of $(\log(\lambda_1), \ldots, \log(\lambda_n))$ (cf. Theorem 3.2).

CONJECTURE 2. *If there exists an $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$, there exists a unique solution to the likelihood equations for the unknown parameters $(\lambda_1, \ldots, \lambda_n)$ and θ with probability* 1, *then M is a Jordan algebra.*

Acknowledgments. Thanks to the referees for useful comments. Moreover, the authors thank Ib Skovgaard and Per Brockhoff for discussion of possible applications of the models.

REFERENCES

AITKIN, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Appl. Statist.* **36** 332–339.

ANDERSON, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Proc. Second International Symposium on Multivariate Analysis* (P. R. Krishnaiah, ed.) 55–66. Academic Press, New York.

ANDERSON, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith, eds.) 1–24. Univ. North Carolina Press, Chapel Hill.

ANDERSSON, S. (1975). Invariant normal models. *Ann. Statist.* **3** 132–154.

ANDERSSON, S. A., BRØNS, H. K. and JENSEN, S. T. (1983). Distribution of eigenvalues in multivariate statistical analysis. *Ann. Statist.* **11** 392–415.

ANDERSSON, S. A. and MADSEN, J. (1998). Symmetry and lattice conditional independence in a multivariate normal distribution. *Ann. Statist.* **26** 525–572.

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

CASALIS, M. (1996). The $2d + 4$ simple quadratic natural exponential families on $\mathbb{R}^d$. *Ann. Statist.* **24** 1828–1854.

COOK, R. D. and WEISBERG, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70** 1–10.

ERIKSEN, P. S. (1987). Proportionality of covariance matrices. *Ann. Statist.* **15** 732–748.

FARAUT, J. and KORÁNYI, A. (1994). *Analysis on Symmetric Cones*. Clarendon, Oxford.

FLURY, B. K. (1986). Proportionality of k covariance matrices. *Statist. Probab. Lett.* **4** 29–33.

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

JENSEN, S. T. and JOHANSEN, S. (1987). Estimation of proportional covariances. *Statist. Probab. Lett.* **6** 83–85.

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford Univ. Press, New York.

LYNCH, M. and WALSH, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

MORRIS, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65–80.

PACE, L. and SALVAN, A. (1997). *Principles of Statistical Inference*. World Scientific, Singapore.

ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.

TJUR, T. (1984). Analysis of variance models in orthogonal designs (with discussion). *Internat. Statist. Rev.* **52** 33–81.

DEPARTMENT OF STATISTICS
AND OPERATIONS RESEARCH
UNIVERSITY OF COPENHAGEN
UNIVERSITETSPARKEN 5
DK-2100 COPENHAGEN Ø
DENMARK
E-MAIL: soerent@math.ku.dk

BIOSTATISTICS UNIT
STATENS SERUM INSTITUT
ARTILLERIVEJ 5
DK-2300 COPENHAGEN S
DENMARK
E-MAIL: jpm@ssi.dk