STATISTICAL APPLICATIONS OF THE MULTIVARIATE SKEW-NORMAL DISTRIBUTION

A. Azzalini
Department of Statistical Sciences
University of Padua, Italy
e-mail: azzalini@mailhost.stat.unipd.it

A. Capitanio
Department of Statistical Sciences
University of Bologna, Italy
e-mail: capitani@stat.unibo.it

February 1998
(revision of December 1998, with amendment of September 2001)

This is the full-length version of the paper with the same title which appears in: J. Roy. Statist. Soc., series B, vol.61, no. 3

Summary

Azzalini & Dalla Valle (1996) have recently discussed the multivariate skew-normal distribution which extends the class of normal distributions by the addition of a shape parameter. The first part of the present paper examines further probabilistic properties of the distribution, with special emphasis on aspects of statistical relevance. Inferential and other statistical issues are discussed in the following part, with applications to some multivariate statistics problems, illustrated by numerical examples. Finally, a further extension is described which introduces a skewing factor of an elliptical density.

1 INTRODUCTION

There is a general tendency in the statistical literature towards more flexible methods, to represent features of the data as adequately as possible and reduce unrealistic assumptions. For the treatment of continuous multivariate observations within a parametric approach, one aspect which has been little affected by the above process is the overwhelming role played by the assumption of normality which underlies most methods for multivariate analysis. A major reason for this state of affairs is certainly the unrivaled mathematical tractability of the multivariate normal distribution, in particular its simplicity when dealing with fundamental operations like linear combinations, marginalization and conditioning, and indeed its closure under these operations.

From a practical viewpoint, the most commonly adopted approach is transformation of the variables to achieve multivariate normality, and in a number of cases this works satisfactorily. There are however also problems: (i) the transformations are usually on each component separately, and achievement of joint normality is only hoped for; (ii) the transformed variables are more difficult to deal with as for interpretation, especially when each variable is transformed using a different function; (iii) when multivariate homoscedasticity is required, this often requires a different transformation from the one for normality.

Alternatively, there exist several other parametric classes of multivariate distributions to choose from, although the choice is not as wide as in univariate case; many of them are reviewed by Johnson & Kotz (1972). A special mention is due to the hyperbolic distribution and its generalized version, which form a very flexible and mathematically fairly tractable parametric class; see Barndorff-Nielsen & Blæsild (1983) for a summary account, and Blæsild (1981) for a detailed treatment of the bivariate case and a numerical example.

As for extensions of distribution theory of classical statistical methods, the direction which seems to have been explored more systematically in this context is the extension of distribution theory of traditional sample statistics to the case of elliptical distribution of the underlying population; elliptical distributions represent a natural extension of the concept of symmetry to the multivariate setting. The main results in this area are summarized by Fang, Kotz & Ng (1990); see also Muirhead (1982, chapters 1 and 8).

Except for data transformation, however, no alternative method to the multivariate normal distribution has been adopted for regular use in applied work, within the framework considered here of a parametric approach to handle continuous multivariate data.

The present paper examines a different direction of the above broad problem, namely the possibility to extend some of the classical methods to the class of multivariate skew-normal distributions which has recently been discussed by Azzalini & Dalla Valle (1996). This distribution represents a mathematically tractable extension of the multivariate normal density with the addition of a parameter to regulate skewness.

We aim at demostrating that this distribution achieves a reasonable flexibility in real data fitting, while it maintains a number of convenient formal properties of the normal one. In particular, associated distribution theory of linear and quadratic forms remains largely valid.

More specifically, the targets of the paper are as follows: (a) to extend the analysis of the probabilistic aspects of the multivariate skew-normal distribution, especially when they reproduce or resemble similar properties of the normal distribution; (b) to examine the potential applications of this distribution in statistics, with special emphasis on multivariate analysis. Correspondingly, after a summary of known results about the distribution,

sections 3, 4 and 5 deal with distribution of linear and quadratic forms of skew-normal variates, and other probabilistic aspects; sections 6 and 7 deal with issues of more direct statistical relevance, with some numerical examples for illustration. In addition, section 8 sketches an additional level of generalization by introducing a skew variant of elliptical densities.

2 THE MULTIVARIATE SKEW-NORMAL DISTRIBUTION

We first recall the definition and a few key properties of the distribution, as given by Azzalini & Dalla Valle (1996) except for re-arrangement of the results. A k-dimensional random variable Z is said to have a multivariate skew-normal distribution if it is continuous with density function

$$2\phi_k(z;\Omega)\,\Phi(\alpha^\top z), \qquad (z \in \mathbb{R}^k), \tag{1}$$

where $\phi_k(z;\Omega)$ is the k-dimensional normal density with zero mean and correlation matrix Ω, $\Phi(\cdot)$ is the $N(0,1)$ distribution function, and α is a k-dimensional vector. For simplicity, Ω is assumed to be of full rank.

When $\alpha = 0$, (1) reduces to the $N_k(0,\Omega)$ density. We then refer to α as a 'shape parameter', in a broad sense, although the actual shape is regulated in a more complex way, as it will emerge in the course of the paper.

The above density does not allow location and scale parameters. Clearly, these are essential in practical statistical work, but we defer their introduction until later, to keep notation simple as long as possible.

The matrix Ω and the vector α appearing in (1) were defined in Azzalini & Dalla Valle (1996) as functions of other quantities, namely another correlation matrix Ψ and a vector $\lambda \in \mathbb{R}^k$; hence a member of the parametric family was identified by the pair (λ, Ψ). It is in fact possible to identify the member of the family directly by the pair (α, Ω); i.e. this pair provides an equivalent parametrization of the class of densities. The proof of this fact is of purely algebraic nature, and it is given in an appendix, together with some related results. For the purposes of the present paper, this parametrization appears preferable and we shall adopt the notation

$$Z \sim \mathrm{SN}_k(\Omega, \alpha)$$

to indicate that Z has density function (1).

The cumulant generating function is

$$K(t) = \log M(t) = \tfrac{1}{2} t^\top \Omega t + \log\{2\,\Phi(\delta^\top t)\} \tag{2}$$

where

$$\delta = \frac{1}{(1+\alpha^\top\Omega\alpha)^{1/2}}\Omega\alpha. \tag{3}$$

Hence the mean vector and the variance matrix are

$$\mu_z = \mathbb{E}\{Z\} = (2/\pi)^{1/2}\,\delta, \qquad \mathrm{var}\{Z\} = \Omega - \mu_z\,\mu_z^\top. \tag{4}$$

The following result provides a stochastic representation of Z, useful for computer generation of random numbers and for theoretical purposes.

Proposition 1 *Suppose that*

$$\begin{pmatrix} X_0 \\ X \end{pmatrix} \sim N_{k+1}(0, \Omega^*), \quad \Omega^* = \begin{pmatrix} 1 & \delta^\top \\ \delta & \Omega \end{pmatrix}$$

where X_0 is a scalar component and Ω^* is a correlation matrix. Then

$$Z = \begin{cases} X & \text{if } X_0 > 0 \\ -X & \text{otherwise} \end{cases}$$

is $\mathrm{SN}_k(\Omega, \alpha)$ *where*

$$\alpha = \frac{1}{(1 - \delta^\top \Omega^{-1} \delta)^{1/2}} \Omega^{-1} \delta. \tag{5}$$

Also, we shall make repeated use of the Sherman–Morrison–Woodbury formula for matrix inversion, which states

$$(A + UBV)^{-1} = A^{-1} - A^{-1}UB(B + BVA^{-1}UB)^{-1}BVA^{-1} \tag{6}$$

for any conformable matrices, provided the inverses involved exist; see for instance Rao (1973, exercise 2.9, p. 33).

3 LINEAR AND QUADRATIC FORMS

A key feature of the multivariate normal distribution is its simplicity to handle linear and quadratics forms. We now explore the behaviour of the skew-normal distribution in these cases.

3.1 MARGINAL DISTRIBUTIONS

It is implicit in the genesis of the multivariate skew-normal variate, as described by Azzalini & Dalla Valle (1996), that the marginal distribution of a subset of the components of Z is still a skew-normal variate. In the marginalization operation, the (λ, Ψ) parametrization works in a very simple manner, since one only needs to extract the relevant components of λ and Ψ. With the (Ω, α) parametrization, specific formulae must be developed.

Proposition 2 *Suppose that $Z \sim \mathrm{SN}_k(\Omega, \alpha)$ and Z is partitioned as $Z^\top = (Z_1^\top, Z_2^\top)$ of dimensions h and $k - h$, respectively; denote by*

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}, \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

the corresponding partitions of Ω and α. Then the marginal distribution of Z_1 is $\mathrm{SN}_h(\Omega_{11}, \bar{\alpha}_1)$, where

$$\bar{\alpha}_1 = \frac{\alpha_1 + \Omega_{11}^{-1}\Omega_{12}\alpha_2}{(1 + \alpha_2^\top \Omega_{22\cdot 1}\alpha_2)^{1/2}}, \qquad \Omega_{22\cdot 1} = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}.$$

The proof follows from straightforward integration, with the aid of Proposition 4 of Azzalini & Dalla Valle (1996).

3.2 LINEAR TRANSFORMS

Proposition 3 *If $Z \sim \mathrm{SN}_k(\Omega, \alpha)$, and A is a non-singular $k \times k$ matrix such that $A^\top \Omega A$ is a correlation matrix, then*

$$A^\top Z \sim \mathrm{SN}_k(A^\top \Omega A, A^{-1}\alpha).$$

The proof follows from standard rule of transformation of random variables. The above condition that $A^\top \Omega A$ is a correlation matrix is there for the sake of simplicity of exposition, and it can be removed; see section 5.

Proposition 4 *For a variable $Z \sim \mathrm{SN}_k(\Omega, \alpha)$, there exists a linear transform $Z^* = A^* Z$ such that $Z^* \sim \mathrm{SN}_k(I_k, \alpha^*)$ where at most one component of α^* is not zero.*

Proof. By using the factorization $\Omega = C^\top C$, we first transform Z into a variable $Y = (C^\top)^{-1} Z$ such that $Y \sim \mathrm{SN}_k(I_k, C\alpha)$. Now consider an orthogonal matrix P with one column on the same direction of $C\alpha$, and define $Z^* = P^\top Y$ which fulfills the conditions.

The above result essentially defines a sort of 'canonical form' whose components are mutually independent, with a single component 'absorbing' all asymmetry of the multivariate distribution. This linear transformation plays a role similar to the one which converts a multivariate normal variable into a spherical form. Further, notice that the component transformations of A^* are invertible; hence it is possible to span the whole class $\mathrm{SN}_k(\Omega, \alpha)$ starting from Z^* and applying suitable linear transformations. The density of Z^* is of the form

$$2 \prod_{i=1}^{k} \phi(u_i)\, \Phi(\alpha_m^* u_m)$$

where

$$\alpha_m^* = \left(\alpha^\top \Omega \alpha\right)^{1/2} \tag{7}$$

is the only non-zero component of α^*.

For the rest of this section, we examine conditions for independence among blocks of components of a linear transform $Y = A^\top Z$. Before stating the main conclusion, we need the following intermediate result.

Proposition 5 *Let $Z \sim \mathrm{SN}_k(\Omega, \alpha)$ and A is as in Proposition 3, and consider the linear transform*

$$Y = A^\top Z = \begin{pmatrix} Y_1 \\ \vdots \\ Y_h \end{pmatrix} = \begin{pmatrix} A_1^\top \\ \vdots \\ A_h^\top \end{pmatrix} Z \tag{8}$$

where the matrices $A_1, \ldots, A_h$ have $m_1, \ldots, m_h$ columns, respectively. Then

$$Y_i \sim \mathrm{SN}_{m_i}(\Omega_{Y_i}, \alpha_{Y_i})$$

where

$$\Omega_{Y_i} = A_i^\top \Omega A_i, \qquad \alpha_{Y_i} = \frac{(A_i^\top \Omega A_i)^{-1} A_i^\top \Omega \alpha}{\left(1 + \alpha^\top (\Omega - \Omega A_i (A_i^\top \Omega A_i)^{-1} A_i^\top \Omega)\alpha\right)^{1/2}}$$

Proof. Without a loss of generality, we consider the case $h = 2$ and $i = 1$. Write $A = (A_1, A_2)$ and denote its inverse by

$$A^{-1} = \begin{pmatrix} A_1^{(-1)} \\ A_2^{(-1)} \end{pmatrix}$$

where the number of columns of the blocks of A matches the number of rows of the blocks of A^{-1}. Since $AA^{-1} = I_k$, then the identity $A_1 A_1^{(-1)} + A_2 A_2^{(-1)} = I_k$ holds. On partitioning $A^\top \Omega A$ in an obvious way, and

$$A^{-1}\alpha = \begin{pmatrix} A_1^{(-1)}\alpha \\ A_2^{(-1)}\alpha \end{pmatrix},$$

the result follows after some algebra by applying Proposition 2 to the parameters of $A^\top Z$, taking into account the above identity.

We now turn to examine the issue of independence among blocks of a linear transform $A^\top Z$ where A satisfies the condition of Proposition 3. To establish independence among the Y_i's, a key role is played by the $\Phi(\cdot)$ component in (1). Since $\Phi(u+v)$ cannot be factorized as the product $\Phi(u)\,\Phi(v)$, it follows that at most one of the Y_i can be a 'proper' skew-normal variate, while the others must have the skewness parameter equal to 0, hence be regular normal variates, if mutual independence holds.

Proposition 6 *If $Z \sim \mathrm{SN}_k(\Omega, \alpha)$, and $A^\top \Omega A$ is a positive definite correlation matrix, then the variables $(Y_1, \ldots, Y_h)$ defined by (8) are independent if and only if the following conditions hold simultaneously:*

(a) $A_i^\top \Omega A_j = 0$ *for* $i \neq j$,

(b) $A_i^\top \Omega \alpha \neq 0$ *for at most one* i.

Proof. Prove sufficiency first. By Proposition 3 and condition (a), the joint distribution of Y is $\mathrm{SN}_k(\Omega_Y, \alpha_Y)$ where

$$\begin{aligned} \Omega_Y &= \mathrm{diag}(A_1^\top \Omega A_1, \ldots, A_h^\top \Omega A_h), \\ \alpha_Y &= (A^\top \Omega A)^{-1} A^\top \Omega \alpha = \begin{pmatrix} (A_1^\top \Omega A_1)^{-1} A_1^\top \Omega \alpha \\ \vdots \\ (A_h^\top \Omega A_h)^{-1} A_h^\top \Omega \alpha \end{pmatrix}. \end{aligned}$$

If condition (b) is satisfied too, only one of the blocks of α_Y is not zero. Hence the joint density can be factorized in obvious manner.

To prove necessity, note that if independence holds the density of Y can be factorized as the product of the densities of the Y_i's, given by Proposition 5. Since the function Φ cannot be factorized, only one block of α_Y can be not zero, and Ω_Y must be a block-diagonal matrix. These requirements can be met only if conditions (a) and (b) are satisfied.

Notice that the parameters of the Y_i's are equal to the corresponding blocks of (Ω_Y, α_Y) only if independence holds.

3.3 QUADRATIC FORMS

One appealing feature of the one-dimensional skew-normal distribution is that the square of a random variate of this kind is a χ^2_1. This property carries on in the multivariate case since $Z^\top \Omega^{-1} Z \sim \chi^2_k$, irrespectively of α. These facts are special cases of the more general results presented below.

Proposition 7 *If $Z \sim \mathrm{SN}_k(\Omega, \alpha)$, and B is a symmetric positive semi-definite $k \times k$ matrix of rank p such that $B\Omega B = B$, then $Z^\top B Z \sim \chi^2_p$.*

Proof. Consider first the case of a random variable $Y \sim \mathrm{SN}_p(I_p, \alpha)$. Since $Y^\top Y = Y^\top A A^\top Y$ for any orthogonal matrix A, hence in particular it holds for a matrix having a column on the same direction of α, i.e. we are considering the canonical form associated to Y. It then follows that $Y^\top Y \sim \chi^2_p$ independently of α.

In the general case, let us write $B = MM^\top$ where M is a full-rank $k \times p$ matrix ($p \leq k$), and notice that $M^\top \Omega M = I_p$ is equivalent to $B\Omega B = B$; to see this, it is sufficient to left-multiply each side of the latter equality by$(M^\top M)^{-1} M^\top$ and right-multiply by its transpose. Then $Z^\top B Z = Y^\top Y$ where $Y = M^\top Z \sim SN_p(I_p, \alpha_Y)$ for some suitable vector α_Y. Therefore the statement holds because $Y^\top Y \sim \chi^2_p$.

Corollary 8 *If $Z \sim \mathrm{SN}_k(\Omega, \alpha)$, and C is a full-rank $k \times p$ matrix ($p \leq k$), then*

$$Z^\top C (C^\top \Omega C)^{-1} C^\top Z \sim \chi^2_p.$$

Proposition 9 *If $Z \sim \mathrm{SN}_k(\Omega, \alpha)$, and B_i is a symmetric positive semi-definite $k \times k$ matrix of rank p_i $(i = 1, 2, \ldots, h)$ such that*

(a) $B_i \Omega B_j = 0$ for $i \neq j$,

(b) $\alpha^\top \Omega B_i \Omega \alpha \neq 0$ for at most one i,

then the quadratic forms $Z^\top B_i Z$ $(i = 1, 2, \ldots, h)$ are mutually independent.

Proof. Similarly to the proof of Proposition 7, write $B_i = M_i M_i^\top$ where M_i has rank p_i. Clearly the quadratic forms $Z^\top B_i Z$ are mutually independent if this is true for the linear forms $M_i^\top Z$. It is easy to see that $M_i^\top \Omega M_j = 0$ is equivalent to $B_i^\top \Omega B_j = 0$ for $i \neq j$; similarly $M_i^\top \Omega \alpha \neq 0$ is equivalent to $\alpha^\top \Omega B_i \Omega \alpha \neq 0$. This completes the proof.

Proposition 10 (Fisher–Cochran) *If $Z \sim \mathrm{SN}_k(I_k, \alpha)$ and $B_1, \ldots, B_h$ are symmetric $k \times k$ matrices of rank $p_1, \ldots, p_h$, respectively, such that $\sum B_i = I_k$ and $B_i \alpha \neq 0$ for at most one choice of i, then the quadratic forms $Z^\top B_i Z$ are independent $\chi^2_{p_i}$ if and only if $\sum p_i = k$.*

Proof. The proof follows the steps of the usual one of Fisher–Cochran theorem, as given for instance by Rao (1973, p. 185 ff.), taking into account Proposition 9 for independence of the quadratic forms, and Proposition 7 as for their marginal distributions.

It would be possible to develop this section via a different approach, on the basis of Proposition 1. For most of the results, this route would offer a simple treatment, but for some others it would be quite cumbersome, especially for the results about independence of components.

4 CUMULANTS AND INDICES

To study higher order cumulants besides those given in Section 2, we need some preliminary results about the cumulants of the half-normal distribution, i.e. the distribution of $V = |U|$, where $U \sim N(0,1)$. Its cumulant generating function is

$$K^V(t) = \tfrac{1}{2}t^2 + \zeta_0(t)$$

where

$$\zeta_0(x) = \log(2\Phi(x)).$$

For later use, define

$$\zeta_m(x) = \frac{\mathrm{d}^m}{\mathrm{d}x^m}\zeta_0(x) \qquad (m = 1, 2, \ldots).$$

Clearly, $\zeta_1(x) = \phi(x)/\Phi(x)$; the subsequent derivatives can be expressed as functions of the lower order derivatives, e.g.

$$\begin{aligned}
\zeta_2(x) &= -\zeta_1(x)\{x + \zeta_1(x)\},\\
\zeta_3(x) &= -\zeta_2(x)\{x + \zeta_1(x)\} - \zeta_1(x)\{1 + \zeta_2(x)\},\\
\zeta_4(x) &= -\zeta_3(x)\{x + 2\zeta_1(x)\} - 2\zeta_2(x)\{1 + \zeta_2(x)\},
\end{aligned}$$

hence as functions of $\zeta_1(x)$. Computation of ζ_m at $x = 0$ gives the corresponding cumulant κ_m^V. Unfortunately, it is not clear how to obtain a closed or recursive formula for the $\zeta_m(x)$'s.

An alternative route for computing κ_m^V is as follows: since $V \sim (\chi_1^2)^{1/2}$ then

$$\mathbb{E}\{V^m\} = \frac{2^{m/2}}{\sqrt{\pi}}\Gamma\left(\frac{m+1}{2}\right)$$

which admits the recurrence formula

$$\mathbb{E}\{V^m\} = (m-1)\mathbb{E}\left\{V^{m-2}\right\}, \qquad (m \geq 2).$$

Hence the cumulant κ_m^V can be obtained from the set $\mathbb{E}\{V^r\}$, $r = 1, \ldots, m$, using well-known results; see e.g. Table 2.1.2 of David, Kendall & Burton (1966) for expressions connecting cumulants to moments up to order 8. In particular, we obtain for V that

$$\kappa_3^V = (2/\pi)^{1/2}(4/\pi - 1), \quad \kappa_4^V = 4(2 - 6/\pi)/\pi.$$

Returning to cumulant generating function (2), its first two derivatives are

$$\frac{\mathrm{d}K(t)}{\mathrm{d}t} = \Omega t + \zeta_1(x)\delta, \quad \frac{\mathrm{d}^2K(t)}{\mathrm{d}t\ \mathrm{d}t^\top} = \Omega + \zeta_2(x)\ \delta\delta^\top$$

where $x = \delta^\top t$, and its evaluation at $t = 0$ confirms (4). Higher order cumulants are obtained from

$$\frac{\mathrm{d}^m K(t)}{\mathrm{d}t_i\ \mathrm{d}t_j \cdots \mathrm{d}t_r} = \zeta_m(x)\ \delta_i\delta_j \cdots \delta_r$$

which needs to be evaluated only at $x = 0$ where

$$\zeta_m(x)|_{x=0} = \kappa_m^V$$

which can has been obtained as described above.

One use of these expressions is to obtain summary indicators for the SN_k distribution. The most popular ones are those introduced by Mardia (1970, 1974) to measure multivariate skewness and kurtosis. In our case, the index of skewness takes the form

$$\begin{aligned}\gamma_{1,k} &= \beta_{1,k} = (\kappa_3^V)^2 \sum_{rst} \sum_{r's't'} \delta_r \delta_s \delta_t \delta_{r'} \delta_{s'} \delta_{t'} \sigma^{rr'} \sigma^{ss'} \sigma^{tt'} \\ &= \left(\frac{4-\pi}{2}\right)^2 \left(\mu_z^\top \Sigma^{-1} \mu_z\right)^3\end{aligned}$$

where $\Sigma = \Omega - \mu_z \mu_z^\top = (\sigma_{rs})$ with inverse $\Sigma^{-1} = (\sigma^{rs})$. Similarly, the index of kurtosis is

$$\begin{aligned}\gamma_{2,k} &= \beta_{2,k} - k(k+2) = \kappa_4^V \sum_{rstu} \delta_r \delta_s \delta_t \delta_u \sigma^{rs} \sigma^{tu} \\ &= 2(\pi - 3) \left(\mu_z^\top \Sigma^{-1} \mu_z\right)^2.\end{aligned}$$

There exists an alternative multivariate index of skewness discussed in the literature; see e.g. McCullagh (1987, p.40). However this differs from $\gamma_{1,k}$ only by a different way of matching the indices of the cumulants, but this has no effect in the present case because of the special pattern of the cumulants of order higher than 2. Hence, in our case the two indices of skewness coincide.

Using (6), one can re-write

$$\mu_z^\top \Sigma^{-1} \mu_z = \frac{\mu_z^\top \Omega^{-1} \mu_z}{1 - \mu_z^\top \Omega^{-1} \mu_z}$$

which allows easier examination of the range of $\mu_z^\top \Sigma^{-1} \mu_z$, by considering the range of $\delta^\top \Omega^{-1} \delta$. On using (3), we write

$$\delta^\top \Omega^{-1} \delta = \frac{\alpha^\top \Omega \alpha}{1 + \alpha^\top \Omega \alpha} = \frac{a}{1+a}$$

where a is the square of α_m^*, defined by (7). Since a spans $[0, \infty)$, then

$$\mu_z^\top \Sigma^{-1} \mu_z = \frac{2a}{\pi + (\pi - 2)a} \in [0, 2/(\pi - 2))$$

and the approximate maximal values for $\gamma_{1,k}$ and $\gamma_{2,k}$ are 0.9905, and 0.869, respectively, in agreement with the univariate case. Since both $\gamma_{1,k}$ and $\gamma_{2,k}$ depend of (Ω, α) only via α_m^*, this reinforces the role of the latter as the summary quantity of the distribution shape.

5 SOME EXTENSIONS

5.1 LOCATION AND SCALE PARAMETERS

For the subsequent development of the paper, we need to introduce location and scale parameters, which have been omitted in the expression (1) of the density of Z. Write then

$$Y = \xi + \omega Z \tag{9}$$

where

$$\xi = (\xi_1, \ldots, \xi_k)^\top, \qquad \omega = \mathrm{diag}(\omega_1, \ldots, \omega_k)$$

are location and scale parameters, respectively; the components of ω are assumed to be positive. The density function of Y is

$$2\,\phi_k(y-\xi;\Omega)\,\Phi\{\alpha^\top\omega^{-1}(y-\xi)\} \tag{10}$$

where

$$\Omega = \omega\Omega_z\omega$$

is a covariance matrix and, from now on, Ω_z replaces the symbol Ω used in the previous sections. Hence, for instance, (3) must now be read with Ω replaced by Ω_z. We shall use the notation

$$Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha)$$

to indicate that Y has density function (10). In the sequel, we shall also use the notation $\sqrt{A}$ to denote the diagonal matrix of the square root of the diagonal elements of a positive definite matrix A; hence, for instance, $\omega = \sqrt{\Omega}$.

Earlier results on linear and quadratic forms for Z carry on for Y, apart for some slight complication in the notation. For instance, for a linear transform $A^\top Y$ where A is a $k \times h$ matrix, a simple extension of Proposition 5 gives

$$X = A^\top Y \sim \mathrm{SN}_h(\xi_X, \Omega_X, \alpha_X) \tag{11}$$

where

$$\xi_X = A^\top\xi, \qquad \Omega_X = A^\top\Omega A, \qquad \alpha_x = \frac{\omega_X\Omega_X^{-1}B^\top\alpha}{\left(1+\alpha^\top(\Omega_z - B\Omega_X^{-1}B^\top)\alpha\right)^{1/2}}$$

and

$$\omega_X = \sqrt{\Omega_X}, \qquad B = \omega^{-1}\Omega A.$$

Similar extensions could be given for other results of Section 3. For later reference, we write the new form of the cumulant generating function

$$K(t) = t^\top\xi + \tfrac{1}{2}t^\top\Omega t + \log\{2\,\Phi(\delta^\top\omega t)\}. \tag{12}$$

5.2 Conditional distributions

Suppose that Y has density function (10), and it is partitioned in two components, Y_1 and Y_2, of dimensions h and $k-h$, respectively, with a corresponding partition for ξ, Ω and α. To examine the distribution of Y_2 conditionally on $Y_1 = y_1$, write

$$\xi_2^c = \xi_2 + \Omega_{21}\Omega_{11}^{-1}(y_1-\xi_1), \quad \Omega_{22\cdot1} = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}, \quad \bar\alpha_1 = \frac{\alpha_1 + \omega_1\Omega_{11}^{-1}\Omega_{12}\omega_2^{-1}\alpha_2}{(1+\alpha_2^\top\bar\Omega_{22\cdot1}\alpha_2)^{1/2}},$$

where

$$\omega_1 = \sqrt{\Omega_{11}}, \qquad \omega_2 = \sqrt{\Omega_{22}}, \qquad \bar\Omega_{22\cdot1} = \omega_2^{-1}\Omega_{22\cdot1}\omega_2^{-1}.$$

Here ξ_2^c and $\Omega_{22\cdot1}$ are given by the usual formulae for the conditional mean and variance of a normal variable, and $\bar\alpha_1$ is the shape parameter of the marginal distribution of Y_1. After

some straightforward computation, it follows that the cumulant generating function of the conditional distribution is

$$K_c(t) = t^\top \xi_2^c + \tfrac{1}{2} t^\top \Omega_{22\cdot 1} t + \log \Phi(x_0 + \tilde{\delta}_2^\top \omega_2 t) - \log \Phi(x_0)$$

where

$$x_0 = \bar{\alpha}_1^\top \omega_1^{-1}(y_1 - \xi_1)$$

and $\tilde{\delta}_2$ is computed similarly to (3), with Ω and α replaced by $\bar{\Omega}_{22\cdot 1}$ and α_2, respectively. This gives immediately

$$\mathbb{E}\{Y_2|y_1\} = \xi_2^c + \zeta_1(x_0)\tau, \qquad \mathrm{var}\{Y_2|y_1\} = \Omega_{22\cdot 1} + \zeta_2(x_0)\tau\tau^\top \tag{13}$$

where $\tau = \omega_2 \tilde{\delta}_2$; higher order cumulants of order m are of the form

$$\zeta_m(x_0)\, \underbrace{\tau_r \tau_s \cdots \tau_u}_{m \text{ terms}}, \qquad (m > 2),$$

where τ_r denotes the r-th component of τ.

Clearly, $K_c(t)$ is of form (12). This special case occurs only if $x_0 = 0$; this condition is essentially equivalent to $\bar{\alpha}_1 = 0$, i.e. Y_1 is marginally normal.

The expression of the conditional density in the general case is easily written down, namely

$$\phi_{k-h}(y_2 - \xi_2^c; \Omega_{22\cdot 1})\, \Phi\{\alpha_2^\top \omega_2^{-1}(y_2 - \xi_2^c) + x_0'\}/\Phi(x_0) \tag{14}$$

where $x_0' = \left(1 + \alpha_2^\top \bar{\Omega}_{22.1}\, \alpha_2\right)^{1/2} x_0$. In the case $k - h = 1$, this distribution has been discussed by several people, including Chou & Owen (1984), Azzalini (1985), Cartinhour (1990) and Arnold *et al.* (1993). From (14), it is easy to see that conditions for independence among components are the same of the unconditional case, with $\Omega_{22\cdot 1}$ and α_2 replacing Ω and α, confirming again the usefulness of the adopted parametrization.

The shape of (14) depends on a number of ingredients; however, for most cases, the plot of this density function displays a remarkable similarity with the one of the skew-normal density. This similarity suggests the approximation of the conditional density by a skew-normal density which matches cumulants up to the third order.

The resulting equations allow explicit solution, except for extreme situations when the exact conditional density has an index of skewness outside the range of the skew-normal one; these unfeasible cases are very remote. In the overwhelming majority of cases, the equations can be solved, and the approximate density is close to the exact one. Figure 1 shows the contour levels of the two densities for two combinations of parameter values when $k - h = 2$; the left panel shows one of the worst cases which have been observed, while the right panel displays a much better, and also more frequently observed, situation.

Besides the generally small numerical discrepancy between the approximate and the exact density, the following two properties hold.

- ◇ *Independence is respected.* If two components of Y_2 are independent conditionally on $Y_1 = y_1$ with respect to the exact conditional density, so they are with respect to the approximate one, and *vice versa.*

- ◇ *Interchange of marginalization and conditioning.* Integrating out some components of Y_2 after conditioning produces the same result of integration followed by conditioning. This fact is obvious when using the exact density; it still holds for the approximate one.

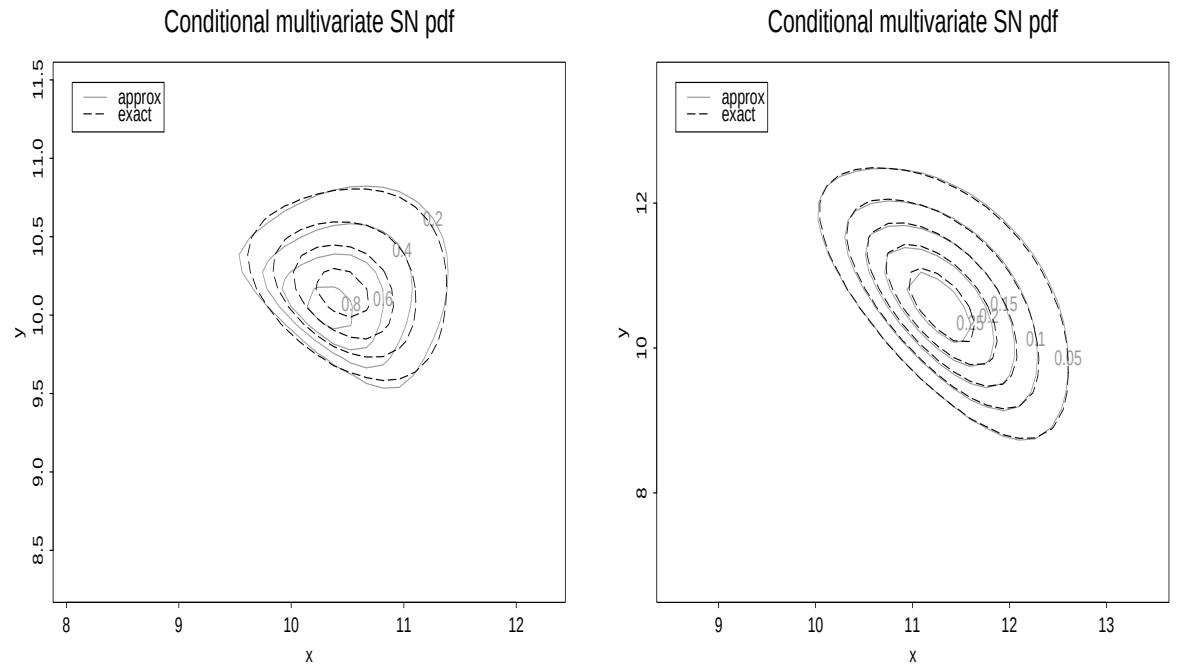

Figure 1: *Contour levels of the exact (dashed lines) and approximate (continuous lines) conditional density of a multivariate skew-normal variable, plotted for two sets of values of the parameters and of the conditioning variable*

To prove the first statement, denote by (a, b) a partition of set of indices composing Y_2. Conditional independence of Y_a and Y_b implies that $\Omega_{22\cdot1}$ is block diagonal and that one of the two components, Y_a say, has no skewness; hence $\tilde{\delta}_a = 0$ and $\tau_a = 0$. Therefore all off-diagonal blocks composing the variance in (13) are 0, and the same structure must hold in the matching quantity of the approximating distribution. The converse statement can be proved similarly. To prove the second statement, simply notice that the approximation preserves exact cumulants up to the third order, which uniquely identify a member of the SN family; hence also the cumulants of the marginal distribution are preserved up to the same order.

The degree of accuracy of the approximation jointly with the above two properties support routine use of the approximate conditional density in place of the exact one. In this sense, we can say that the skew-normal class of density is closed with respect to the conditioning operation.

6 STATISTICAL ISSUES IN THE SCALAR CASE

6.1 DIRECT PARAMETERS

Starting from this section, we switch attention to inferential aspects, and other issues of more direct statistical relevance, initially by considering univariate distributions.

Some of the issues discussed in this subsection have a close connection with the problem considered by Copas & Li (1997) and the sociological literature on Heckman's model referenced there; see also Aigner *et al.* (1977) and the literature of stochastic frontier mod-

els.

In the univariate case, write $Y \sim \mathrm{SN}(\xi, \omega^2, \alpha)$, dropping the subscript k for simplicity. If a random sample $y = (y_1, \ldots, y_n)^\top$ is available, the loglikelihood function for the direct parameters $DP = (\xi, \omega, \alpha)$ is

$$\ell(DP) = -n \log \omega - \tfrac{1}{2} z^\top z + \sum_i \zeta_0(\alpha z_i) \tag{15}$$

where $z = \omega^{-1}(y - \xi 1_n)$ and z_i denotes its i-th component; here 1_n is the $n \times 1$ vector of all ones. We shall denote by $\hat{\alpha}$ the maximum likelihood estimate (MLE) of α, and similarly for the other parameters. The likelihood equations are immediately written down, namely

$$\begin{aligned} &\sum z_i - \alpha \sum p_{1i} = 0, \\ &\sum z_i^2 - \alpha \sum p_{1i} z_i - n = 0, \\ &\sum p_{1i} z_i = 0 \end{aligned}$$

where $p_{1i} = \zeta_1(\alpha z_i)$. There are however two sort of problems with this parametrization. Firstly, there is always an inflection point at $\alpha = 0$ of the profile loglikelihood. Correspondingly, at $\alpha = 0$, the expected Fisher information becomes singular. This phenomenon is a special case of the problem studied in greater generality by Rotnitzky *et al.* (1999).

In addition, the likelihood function itself can be problematic; its shape can be far from quadratic even when α is not near 0. This aspect is clearly illustrated by the plots given by Arnold *et al.* (1993) who have analysed a dataset of size 87, later referred to as the Otis data; see also Figure 2, which refers to the same data.

For evaluation of the MLE, gradient-based methods have been considered, but better results were obtained using the EM algorithm, with the introduction of a fictitious unobserved variable which is essentially $|X_0|$ of Proposition 1. This method works satisfactorily, at least when the initial values are chosen by the method of moments. As typical for the EM algorithm, reliability rather than speed is its best feature. Methods for accelerating the EM algorithm are available; see for instance Meng & van Dyk (1997) and references therein. However, we prefer to expand in greater detail the discussion of another approach, for the reasons explained in the next subsection.

6.2 Centred parameters

To avoid the singularity problem of the information matrix at $\alpha = 0$, Azzalini (1985) has reparameterized the problem by writing

$$Y = \mu + \sigma Z^\circ,$$

where

$$Z^\circ = (Z - \mu_z)/\sigma_z, \quad \sigma_z = \left(1 - \mu_z^2\right)^{1/2},$$

and considering the centred parameters $CP = (\mu, \sigma, \gamma_1)$ instead of the DP parameters. Here γ_1 is the usual univariate index of skewness, which is equal to the square root of the multivariate index of skewness of Section 4, taken with the same sign of α. Clearly, there is the correspondence

$$\xi = \mu - \sigma \sigma_z^{-1} \mu_z, \qquad \omega = \sigma \sigma_z^{-1}.$$

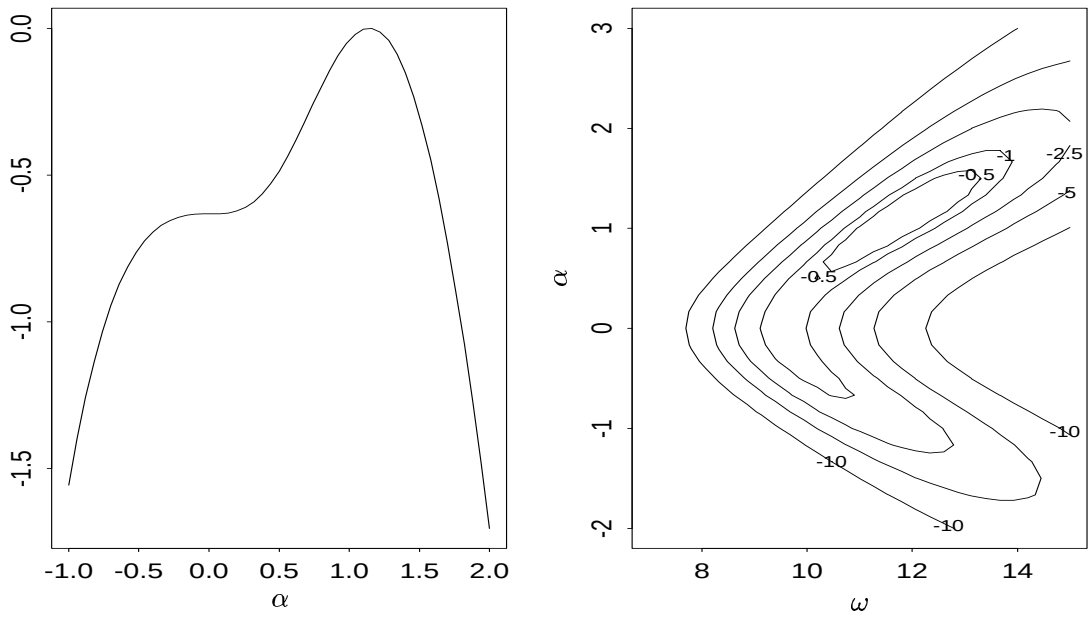

Figure 2: *Twice relative profile loglikelihood of α (left) and contour levels of the similar function of (ω, α) (right) for the Otis data, when the direct parametrization is used*

In the case of a regression problem, write $\mathbb{E}\{Y_i\} = x_i^\top \beta$, where x_i is a vector of p covariates and β is vector parameter. The corresponding loglikelihood is then

$$\ell(CP) = n \log(\sigma_z/\sigma) - \tfrac{1}{2} z^\top z + \sum \zeta_0(\alpha z_i)$$

where

$$z_i = \mu_z + \sigma_z \sigma^{-1}(y_i - x_i^\top \beta) = \mu_z + \sigma_z r_i, \qquad z = (z_1, \ldots, z_n)^\top.$$

In case we wanted to reformulate the regression problem in terms of direct parameters, then only the first component must be adjusted, namely

$$\beta_1^{DP} = \beta_1^{CP} - \sigma \mu_z / \sigma_z$$

in a self-explanatory notation.

The gradient and the Hessian matrix of the loglikelihood in the CP parametrization are more involved than with the DP parametrization, and we confine the details in an appendix. The effects of the reparametrization are however beneficial in various respects and worth the algebraic complications, for the following reasons.

- ◇ The reparametrization removes the singularity of the information matrix at $\alpha = 0$. This fact was examined numerically by Azzalini (1985), and checked by detailed analytic computations by Chiogna (1997).

- ◇ Although not orthogonal, the components of CP are less correlated than those of DP, especially μ and the γ_1. This fact can be checked numerically with the aid of the expressions given in an appendix.

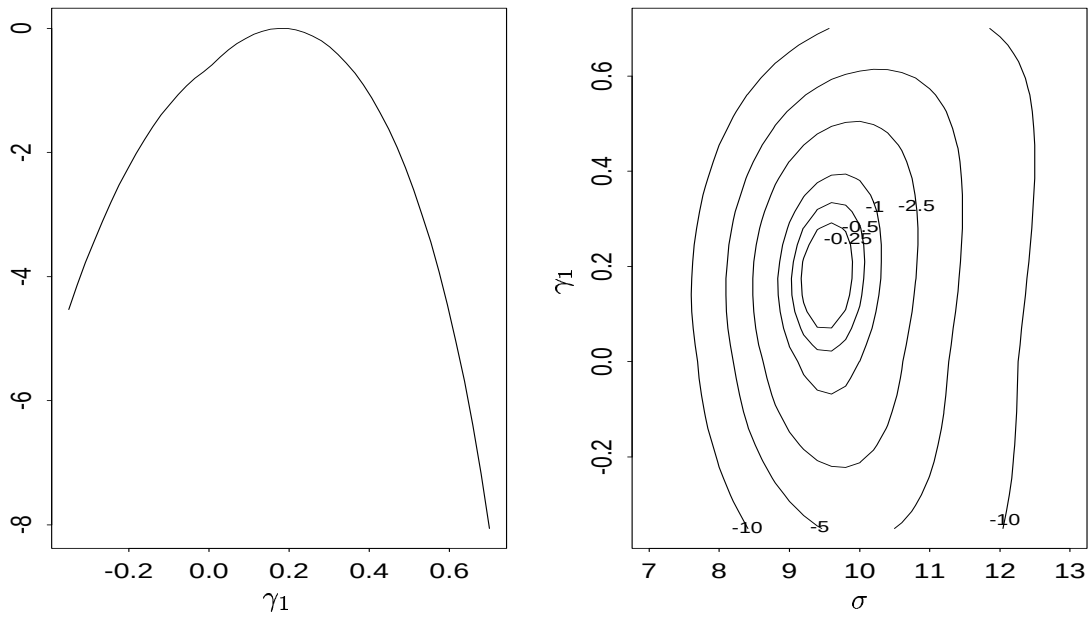

Figure 3: *Twice relative profile loglikelihood of γ_1 (left) and contour level of the similar function of (σ, γ_1) (right) for the Otis data, when the centred parametrization is used*

- ◇ The likelihood shape is generally much improved. This is illustrated by Figure 3, which refers to the same data of Figure 2; the left panel refers to twice the relative profile loglikelihood for the new shape parameter γ_1, and the right panel refers to the pair (σ, γ_1). There is a distinct improvement over the earlier figure, in various respects:
  - ◦ the inflection point at $\alpha = 0$ of the first panel of Figure 2 has been removed, with only a mild change of slope at $\gamma_1 = 0$ left;
  - ◦ the overall shape of the profile loglikelihood has changed into one appreciably closer to a quadratic shape;
  - ◦ near the MLE point, the axes of the approximating ellipsis are now more nearly alligned to the orthogonal axes than before.
- ◇ Simulation work, whose details are not reported here, showed that the marginal distribution of $\hat{\xi}$ can be bimodal when n and $|\alpha|$ are small or moderate; for instance it happens with $n = 50$, sampling from $\mathrm{SN}(0,\, 1,\, 1)$. Such an unusual distribution of the MLE is in qualitative agreement with the findings of Rotnitzky *et al.* (1999). Again, this unpleasant feature disappeared with the CP parametrization, in the sense that the distribution of the new location parameter $\hat{\mu}$ exhibited a perfectly regular behaviour.

The advantages of CP over DP are not only on the theoretical side but also practical, since the more regular shape of the loglikelihood leads to faster convergence of the numerical maximization procedures when computing the MLE.

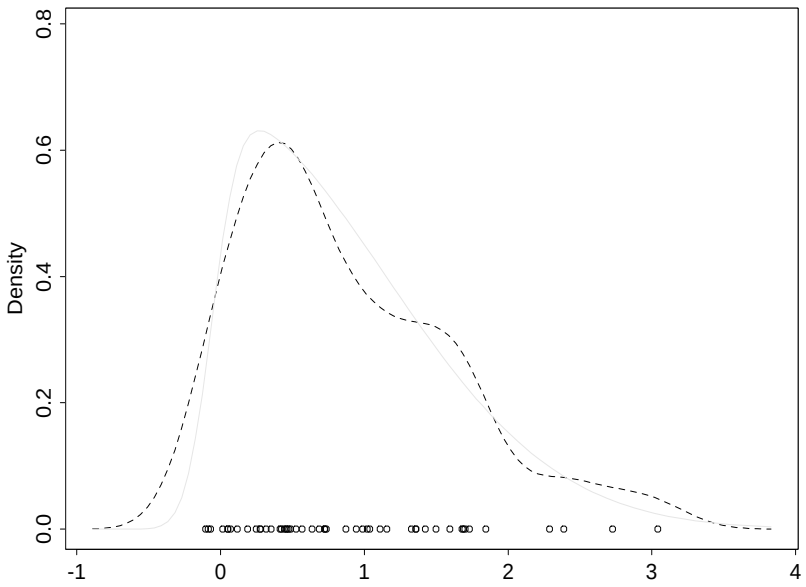

Figure 4: *Simulated data points (small circles) leading to* $\hat{\alpha} = \infty$*, with nonparametric density estimate (dashed curve) and parametric curve with* $\alpha = 8.14$ *(continuous curve)*

For numerical computation of the MLE, we have obtained satisfactory results by adopting the following scheme: (i) choose initial values by the method of moments; (ii) optionally, improve these estimates by a few EM iterations; (iii) obtain the MLE either by Newton–Raphson or by quasi-Newton methods. Only in a few cases, the third stage did not converge; full EM iteration was then used, and this always led to convegence.

The set of S-Plus routines developed for these computations, as well as those related to the problems discussed later, will be made freely available on the WorldWideWeb.

6.3 Anomalies of MLE

Notwithstanding what is stated near the end of the previous subsection, there are still cases where the likelihood shape and the MLE are problematic. We are not referring here to difficulties with numerical maximization, but to the intrinsic properties of the likelihood function, not removable by change of parametrization.

An illustration is provided by Figure 4; here 50 data points, sampled from $\mathrm{SN}(0,1,5)$, are plotted on the horizontal axis, together with a nonparametric estimate of the density (dashed curve) and another (continuous) curve representing a skew-normal density. This parametric curve has $\alpha = 8.14$ but it is not the one of the MLE, however: the MLE has $\alpha = \infty$, which corresponds to the half-normal density.

This divergence of $\hat{\alpha}$ (or equivalently $\hat{\gamma}_1 \to 0.99527$, its maximal value) looks rather surprising, since apparently there is nothing pathological in the data pattern of Figure 4; the sample index of skewness is 0.9022, which is inside the feasible region of γ_1. Similar situations occur with a non-negligible frequency when n is small to moderate, but they disappear when n increases.

The source of this sort of anomaly is easy to understand in the one-parameter case with ξ and ω known; $\xi = 0$, $\omega = 1$, say. If all sample values have the same sign, the final term of (15) increases with $\pm\alpha$, depending the sign of the data but irrespective of their actual values, as it has been remarked by Liseo (1990). For instance, if 25 data are sampled from $\mathrm{SN}(0, 1, 5)$, the probability that they are all positive is about 0.20.

When all three DP parameters are being estimated, the explanation of this fact is not so clear, but it is conceivable that a similar mechanism is in action.

In cases of this sort, the behaviour of the MLE appears qualitatively unsatisfactory, and an alternative estimation method is called for. Tackling this problem is beyond the scope of the present paper, however. As a temporary solution we adopted the following simple strategy: when the maximum occurs on the frontier, re-start the maximization procedure and stop it when it reaches a loglikelihood value not significantly lower than the maximum. This was the criterion used for choosing the parametric curve plotted in Figure 4; in this case the difference from the maximum of the loglikelihood is 2.39, far below the 95% significant point of a $\chi^2_3/2$ distribution.

The above proposal leaves some degree of arbitrariness, since it does not say exactly how much below the maximum to stay. In practice the choice is not so dramatic, because the boundary effect involves only α, and when this is large, $\alpha > 20$ say, the actual shape of the density varies very slowly. Moreover, in the numerical cases which have been examined, the loglikelihood function was very flat only along the α-axis, while it was far more curved with along the location and scale parameters which were then little affected by the specific choice of α, within quite wide limits.

7 APPLICATIONS TO MULTIVARIATE ANALYSIS

7.1 FITTING MULTIVARIATE DISTRIBUTIONS

In the case of independent observations $(y_1, \ldots, y_n)$ sampled from $\mathrm{SN}_k(\xi_i, \Omega, \alpha)$ for $i = 1, \ldots, n$, the loglikelihood is

$$\ell = -\tfrac{1}{2} n \log|\Omega| - \tfrac{1}{2} n \,\mathrm{tr}(\Omega^{-1} V) + \sum_i \zeta_0\{\alpha^\top \omega^{-1}(y_i - \xi_i)\} \tag{16}$$

where

$$V = n^{-1} \sum_i (y_i - \xi_i)(y_i - \xi_i)^\top.$$

The location parameters have been considered to be different having in mind a regression context where ξ_i is related to p explanatory variables x_i via

$$\xi_i^\top = x_i \beta, \qquad (i = 1, \ldots, n),$$

for some $p \times k$ matrix β of parameters.

It would be ideal to reproduce in this setting the centred parametrization introduced in the scalar case. This approach poses difficulties, and we follow a different direction to obtain the MLE. Once the estimates have been computed, they could be converted componentwise to the centred parameters.

The letters y, X, ξ will denote the matrices of size $n \times k$, $n \times p$ and $n \times k$ containing the y_i's, the x_i's, and the ξ_i's, respectively. Also, a notation of type $\zeta_m(z)$ represents the vector obtained by applying the function $\zeta_m(\cdot)$ to each element of the vector z.

Regarding $\eta = \omega^{-1}\alpha$ as a parameter in replacement of α separates the parameters in (16) in the following sense: for fixed β and η, maximization of ℓ with respect Ω is equivalent to maximizing the analogous function for normal variates for fixed β, which has the well known solution

$$\hat{\Omega}(\beta) = V(\beta) = n^{-1}u^\top u$$

where $u = (y - X\beta)$. Replacing this expression in ℓ gives the profile loglikelihood

$$\ell^*(\beta, \eta) = -\tfrac{1}{2}n \log|V(\beta)| - \tfrac{1}{2}nk + 1_n^\top \zeta_0(u\eta)$$

with substantial reduction of dimensionality of the maximization problem. Numerical maximization of ℓ^* is required; this process can be speeded up substantially if the partial derivatives

$$\begin{aligned} \frac{\partial \ell^*}{\partial \beta} &= X^\top u\, V(\beta)^{-1} - X^\top \zeta_1(u\eta)\eta^\top, \\ \frac{\partial \ell^*}{\partial \eta} &= u^\top \zeta_1(u\eta), \end{aligned}$$

are supplied to a quasi-Newton algorithm. Upon convergence, numerical differentiation of the gradient leads to approximate standard errors for β and η, hence for α after multiplication by ω.

The above computational scheme has been used satisfactorily in numerical work with non-trivial dimensions of the arrays X, y, β. A very simple illustration is provided by Figure 5 which refers to a subset of the AIS (Australian Institute of Sport) data examined by Cook & Weisberg (1994), which contains various biomedical measurements on a group of Australian athletes; we then have $k = 4$, $p = 1$, $n = 202$. Figure 5 displays the scatter plot of each pair of the four variables considered superimposed with the contour lines of the marginal density obtained by marginalization of the fitted SN_4 density.

Visual inspection of Figure 5 indicates a satisfactory fit of the density to the data. However, to obtain a somewhat more comprehensive graphical display, consider the Mahalanobis distances

$$d_i = (y_i - \xi)^\top \Omega^{-1}(y_i - \xi), \qquad (i = 1, \ldots, n), \tag{17}$$

which are sampled from a χ^2_k if the fitted model is appropriate, by using Proposition 7. In practice, estimates must be replaced to the exact parameter values in (17). The above d_i's must be sorted and plotted versus the χ^2_k percentage points. Equivalently, the cumulative χ^2_k probabilities can be plotted against their nominal values $1/n, 2/n, \ldots, 1$; the points should lie on the bisection line of the quadrant. This diagnostic method is a natural analogue of a well-know diagnostics used in normal theory context (Healy, 1968).

Figure 6 diplays the second variant of this plot for the AIS data, in its right-hand side panel; the left-hand side panel shows the similar traditional plot under assumption of normality. Comparison of the two plots indicates a substantial improvement of the skew-normal fit over the normal one.

A similar conclusion is achieved by considering a parametric test for normality which is provided by the likelihood ratio test for the null hypothesis $\alpha = 0$, that is

$$2\{\ell(\hat{\xi}, \hat{\Omega}, \hat{\alpha}) - \ell(\hat{\mu}, \hat{\Sigma}, 0)\}$$

where $(\hat{\mu}, \hat{\Sigma})$ denote the MLE of (ξ, Ω) under the assumption of normality. The observed value of the test statistics in the above example is over 103, and the associated value of the χ^2_4 distribution function does not even need to be computed.

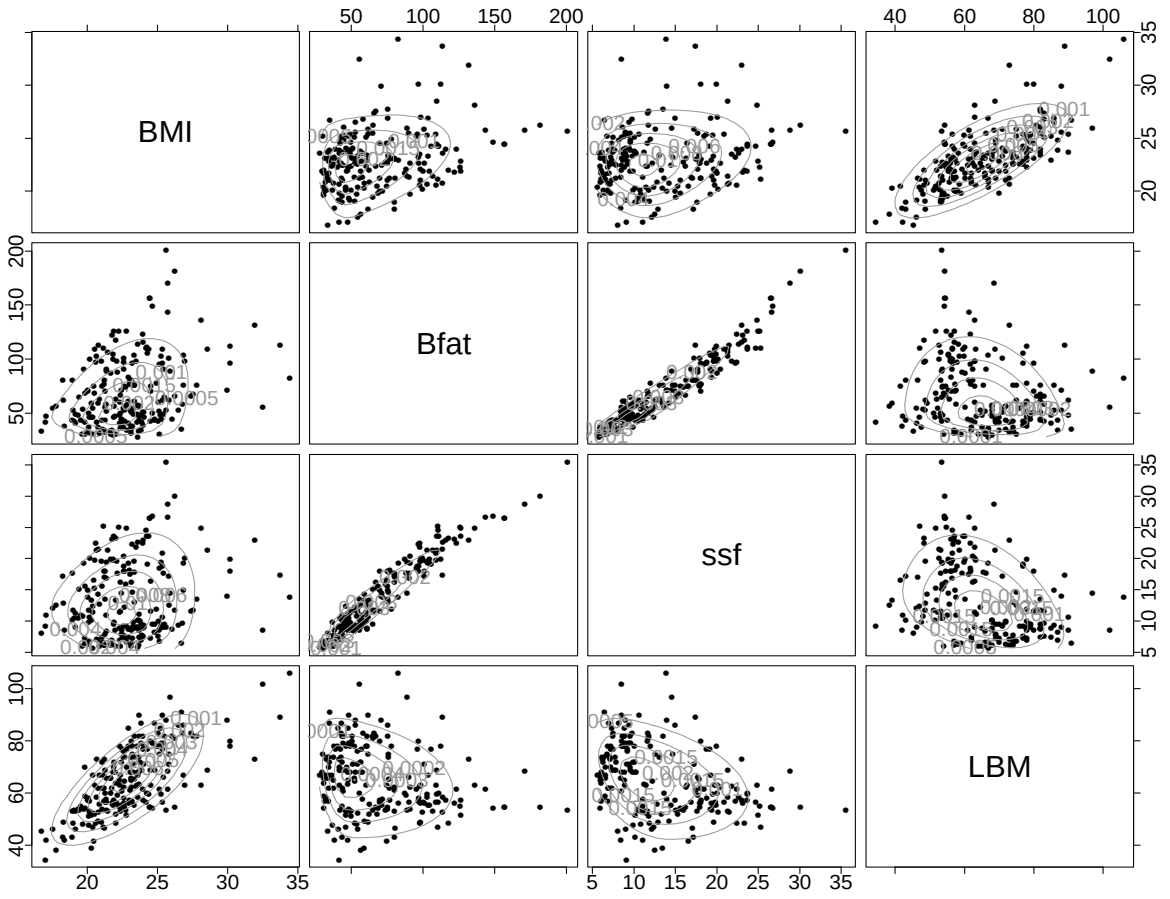

Figure 5: *Scatterplots of some pairs of the AIS variables and contour levels of the fitted distribution*

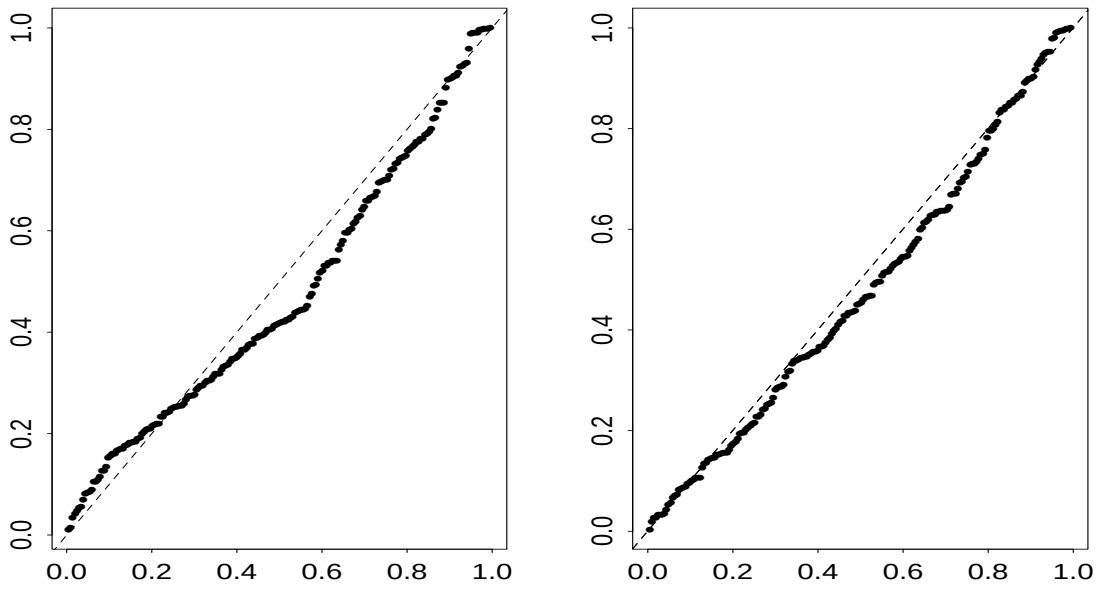

Figure 6: *Healy's plot when either a normal distribution (left panel) or a skew-normal distribution (right panel) is fitted to the AIS data*

7.2 DISCRIMINANT ANALYSIS

The results of Section 3, once reinterpreted in the more general setting introduced in Section 5, provide tools to examine the behaviour of many classical multivariate techniques, when based on linear transforms of the data, in the more general context of SN variables. For the present discussion, however, we shall restrict ourselves to a rather simple problem of discrimination between two populations, under the traditional hypothesis that they differ only in the location parameters.

If $Y_i \sim \mathrm{SN}_k(\xi_i, \Omega, \alpha)$ denote the random variables associated to the two populations $(i = 1, 2)$, then the likelihood-based discrimination rule allocates a new unit with observed vector y to population 1 if

$$(\xi_1 - \xi_2)^\top \Omega^{-1}(y - \tfrac{1}{2}(\xi_1 + \xi_2)) + \zeta_0(w_1) - \zeta_0(w_2) + \log(\pi_1/\pi_2) > 0 \tag{18}$$

where $w_i = w_i(y) = \alpha^\top \omega^{-1}(y - \xi_i)$ and π_i is the prior probability of the i-th population $(i = 1, 2)$.

Nonlinearity of the left-hand side of the above inequality prevents explicit solution. However, some properties can be obtained; one is that the likelihood-based discriminant function is a linear function of y when either of the following conditions holds:

$$(\xi_1 - \xi_2)^\top \omega^{-1}\alpha = 0, \tag{19}$$

$$\omega^{-1}\alpha = c\,\Omega^{-1}(\xi_1 - \xi_2) \tag{20}$$

where c ia non-zero scalar constant. The proof is omitted.

The natural alternative to (18) is the Fisher linear discriminant functions, whose commonly used expression is

$$(\mu_1 - \mu_2)^\top \Sigma^{-1} \left(y - \tfrac{1}{2}(\mu_1 + \mu_2)\right) + \log(\pi_1/\pi_2) > 0,$$

using a self-explanatory notation; in the present case, this can be re-written as

$$(\xi_1 - \xi_2)^\top (\Omega - \omega\mu_z\mu_z^\top\omega)^{-1} \left(y - \tfrac{1}{2}(\xi_1 + \xi_2 + 2\omega\mu_z)\right) + \log(\pi_1/\pi_2) > 0. \tag{21}$$

Proposition 11 *When condition (19) holds, the discriminant rules (18) and (21) coincide.*

Proof. First, notice that (19) implies that $w_1(y) = w_2(y)$ in (18). Next, use (6) to invert $(\Omega - \omega\mu_z\mu_z^\top\omega)$ in (21), leading to

$$(\xi_1 - \xi_2)^\top \Omega^{-1}(y - \tfrac{1}{2}(\xi_1 + \xi_2) - \omega\mu_z) > 0.$$

Then, on using (19) again and noticing that the vectors $\Omega^{-1}\omega\mu_z$ and $\omega^{-1}\alpha$ have the same direction, one obtains the result.

In the general case, (18) and (21) can only be compared numerically. The various cases considered differ for the relative positions of the locations parameters, while the other parameters have been kept fixed; specifically, we have set $k = 2$, $\pi_1 = \pi_2$, $\omega = I_2$, Ω equal to the correlation matrix with off-diagonal elements equal to 0.4, $\alpha = (3,3)^\top$, and $\|\xi_1 - \xi_2\|^2 = 1$. This choice of the parameters, such that α is an eigenvector of Ω, has been made for the sake of simplicity, in the following sense. It turns out that the quantities regulating the basic behaviour of the classification rules are the angle θ_1 between the vectors $\omega^{-1}\alpha$ and $\xi_1 - \xi_2$, and the angle θ_2 between $\omega^{-1}\alpha$ and $\Omega^{-1}(\xi_1 - \xi_2)$. The above choice of α and Ω makes it easier to choose values of $\xi_1 - \xi_2$ fulfilling conditions (19) and (20), i.e. such that $\cos\theta_1 = 0$ and $\cos\theta_2 = 1$.

Figure 7 shows the relevant entities for a few cases. Each panel of the figure displays the contour levels of the two population densities with superimposed the separation lines of the two discriminant rules. The bottom-right panel corresponds to a case satisfying (19) and only one discrimination line is then visible; the top-right panel corresponds to fulfilling (20) and the two discriminant lines are parallel.

Table 1 contains summary values of the numerical work, in particular misclassification probabilities, for a larger number of cases. For the Fisher rule, classification probabilities can be computed exactly with the aid of (11); for (18), the corresponding probabilities have been evaluated by simulation methods, using 100000 replicates for each case. The main qualitative conclusions from these figures are as follows: (a) the total misclassification probability is lower for the likelihood-based rule than for the Fisher linear discriminant, as expected from known results (Rao, 1947); (b) the Fisher rule is however not much worse than the other one, and its two components are more balanced that the analogous ones of the likelihood-based rule, which could be considered as an advantage on its own; (c) for some values of θ_1 and θ_2, the fraction of cases which are classified differently by the two rules is not negligible; hence the choice of the method can be relevant even if the probabilities of misclassification are similar.

For numerical illustration, we have applied the two discriminant rules to the same subset of the AIS data used in subsection 7.1. The individuals were divided by sex, obtaining two groups of 102 male and 100 female athletes, respectively, and prior probabilities were

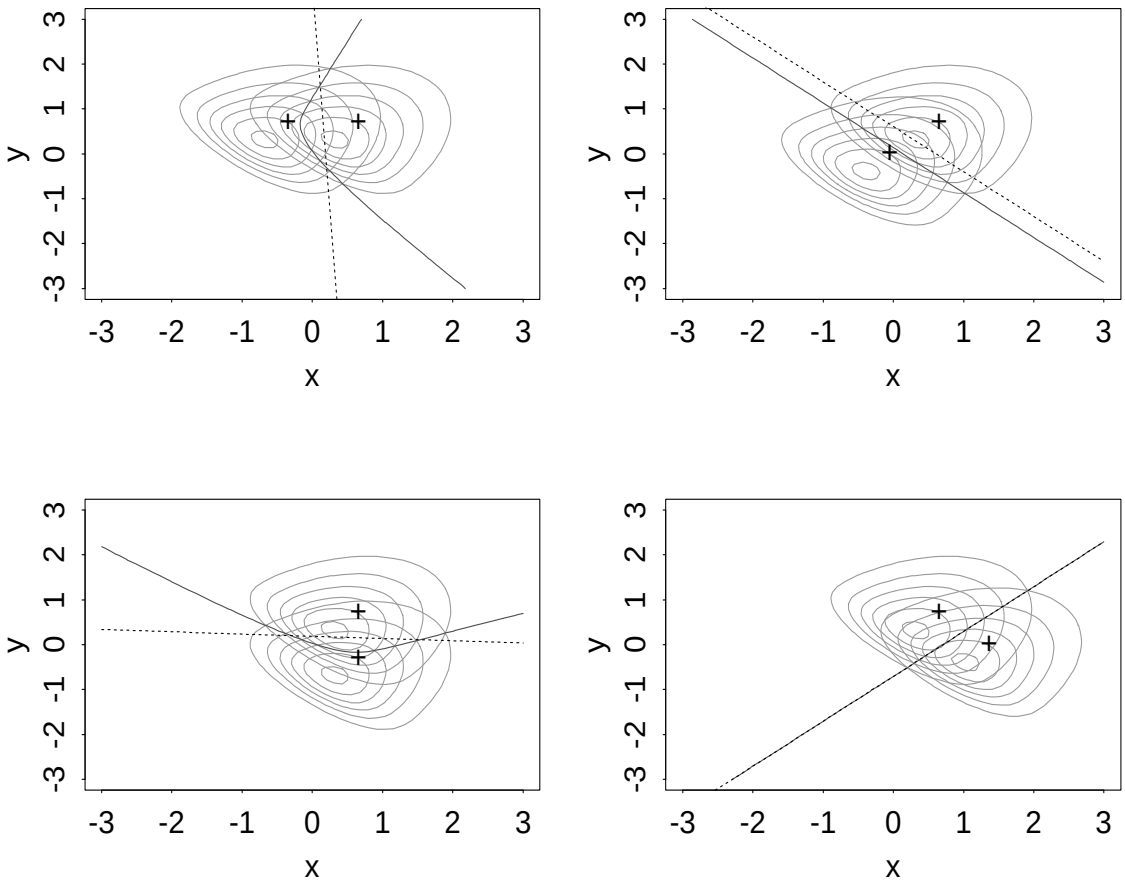

Figure 7: *Contour plots of four pairs of* SN_2 *variables, with likelihood discriminant function (continuous line) and Fisher linear discriminant function (dashed line)*

p_{1L}	p_{1F}	p_{2L}	p_{2F}	p^*	$\cos\theta_1$	$\cos\theta_2$
0.35	0.23	0.10	0.28	0.84	1.000	1.000
0.35	0.23	0.11	0.28	0.85	0.907	0.981
0.34	0.23	0.13	0.27	0.87	0.719	0.924
0.31	0.23	0.16	0.26	0.89	0.530	0.831
0.29	0.24	0.19	0.26	0.91	0.394	0.707
0.27	0.25	0.21	0.26	0.92	0.275	0.556
0.26	0.26	0.24	0.26	0.94	0.175	0.383
0.26	0.26	0.25	0.26	0.96	0.085	0.195
0.26	0.26	0.26	0.26	1.00	0.000	0.000
0.25	0.26	0.26	0.26	0.96	-0.085	-0.195
0.24	0.26	0.26	0.26	0.94	-0.175	-0.383
0.21	0.26	0.27	0.25	0.92	-0.275	-0.556
0.19	0.26	0.29	0.24	0.91	-0.394	-0.707
0.16	0.26	0.31	0.23	0.89	-0.530	-0.831
0.13	0.27	0.33	0.23	0.87	-0.719	-0.924
0.10	0.28	0.35	0.23	0.85	-0.907	-0.981
0.10	0.28	0.35	0.23	0.84	-1.000	-1.000

Table 1: *Classification probabilities of likelihood-based and Fisher linear discriminant rules. The entries are:* p_{1L}*, misclassification error probability using likelihood based rule, when sampling from population 1;* p_{1F}*, misclassification error probability using Fisher linear discriminant function, when sampling from population 1;* p_{2L} *and* p_{2F} *are similar quantities in the case of sampling from population 2;* p^**, probability that the discriminant rules coincide;* θ_1 *and* θ_2 *are angles associated to the relative position of the location parameters, as described in the text*

	Actual groups			
Allocated groups	G_1	G_2	G_3	G_4
G_1	55, 55	5, 5	2, 2	0, 0
G_2	2, 2	36, 37	2, 4	0, 0
G_3	0, 0	0, 0	22, 20	10, 11
G_4	0, 0	3, 2	14, 14	67, 66
Total	57	44	40	77

Table 2: *Discrimination of the four groups of the hepatic data; the data indicate the number of individuals classified by likelihood rule (first entry) and by the Fisher discriminat function (second entry)*

set equal to the observed frequencies. In this case $\theta_1 = 1.54041$ radians, a situation not so far from the one associated with (19), i.e coincidence of the two discriminant functions. In fact the total number of misclassified subjects differs only for one unit: more precisely Fisher rule fails in three units, while the likelihood-based one fails in two. Further numerical work has been done using data reported by Albert & Harris (1987, chapter 5), fairly often used for illustration in the context of discriminant methods. An overall sample of 218 individuals affected by liver problems are divided into four groups, corresponding to severity of their status: acute viral hepatitis (group G_1, 57 patients), persistent chronic hepatitis (G_2, 44 patients), aggressive chronic hepatitis (G_3, 40 patients), and post-necrotic cirrhosis (G_4, 77 patients). Albert & Harris (1987) construct a discrimination rule based on data on three of four available liver enzymes: aspartate aminotransferase (AST), alanine aminotransferase (ALT) and glutamate dehydrogenase (GLDH); the data have been logarithmically transformed because of extreme skewness in the original variables. To ease comparison, we employed the same variables and applied the same data transformation.

Goodness-of-fit and graphical diagnostics, along the lines of subsection 7.1, confirm the adequacy of the skew-normal distribution in modeling this set of variables. Prior probabilities were set equal to the observed frequencies, i.e. $\pi_1 = 0.26$, $\pi_2 = 0.20$, $\pi_3 = 0.18$ and $\pi_4 = 0.35$. The summary results, shown in Table 2, indicate a slight improvement using the SN distribution instead of the normal one, in the sense that 3 data points which were incorrectly classified by the Fisher rule are now correctly classified, and only one is moved in the reverse direction.

7.3 Regression and graphical models

Graphical models are currently a much studied research topic. This subsection examines some related issues when the assumption of normal distribution of the variable is replaced by (1). We adopt Cox & Wermuth (1996) as a reference text for background material.

In the construction of a graphical model of normal variables, a key ingredient is the concentration matrix, i.e. the inverse of the covariance matrix, possibly scaled to obtain unit diagonal elements. When the (i, j)-th entry of the concentration matrix is 0, this indicates that the two corresponding components, Y_i and Y_j say, are independent conditionally on all the others. The associated concentration graph has then no edge between Y_i and Y_j.

The results of sections 3 and 5 enable us to transfer the above scheme in the context of skew-normality; consider in particular Proposition 6 and expression (14). Hence, two

components, Y_i and Y_j say, of $Y \sim \mathrm{SN}_k(\xi, \Omega, \alpha)$ are independent conditionally on the others if the (i, j)-th entry of Ω^{-1} is zero and at most one of α_i and α_j is different from zero. Hence Ω^{-1} plays a role analogous to the concentration matrix in normal theory context, but also α must be considered now.

Building a graphical model from real data involves to follow essentially the strategy presented by Cox & Wermuth (1996) for the normal case. The main difference is in the distinction between regression and conditioning, which are essentially coincident in the normal case but not here.

Since it seems best to illustrate the actual construction of a graphical model in a specific example, we consider the data analysed by Cox & Wermuth (1996, chapter 6), concerning 68 patients with fewer than 25 years of diabetes. This dataset is of rather small sample size for an adequate fitting of a multivariate SN distribution, but it has been adopted here because it is a 'standard' one in this context. For each patient, eight variables are recorded; of these, glucose control (Y) and knowledge about illness (X) are the primary response and the intermediate response variables, respectively; the special role of these two variables drives the subsequent analysis. Of the other variables, W, A and B are explanatory variables regarded as given, with A and B binary; Z, U and V are other stochastic variables. See the above reference for a full description of the variables and some background information.

A preliminary analysis, using the methods described at the end of subsection 7.1, shows the presence of a significant skewness in the distribution of some of the variables; this is largely due to the X component but not only to this one. Therefore, we introduce a multivariate regression model of type

$$(Y, X, Z, U, V) \sim \mathrm{SN}_5(\xi, \Omega, \alpha)$$

where ξ is a linear combination of $(1, W, A, B)$, and Ω and α are constant across individuals. Fit of the above model, using the algorithm described in section 7.1, led to a boundary solution, in the sense that the components of $\hat{\alpha}$ diverged. Adopting the simple method described in section 6.3 to handle these cases, a set of parameters has been chosen inside the parameter space having a loglikelihood value about 7.7 units lower than the maximum, which is a very minor loss in consideration of the large number of parameters being estimated.

Table 3 gives the partial correlation matrix, $\hat{\Omega}^*$, which is $\hat{\Omega}^{-1}$ after scaling to obtain unit diagonal entries and changing signs of the off-diagonal entries, and the shape parameters with standard errors and t-ratios.

Because of the different role played by the variables in the present problem, the most relevant entries of Table 3 are those of the first two rows of Ω^*. Joint inspection of both components of Table 3 indicates conditional independence of (Y, Z), (Y, U) and (Y, V), while there is conditional dependence between (X, Z) and between (Y, X). Moreover the results concerning the regression component suggest dropping B from the model.

Additional numerical work not reported here has been carried out to examine the sensitivity of the results to the choice of the point where the MLE iteration sequence was stopped. The overall conclusions are as follows: the regression coefficients and their observed significances are stable over a wide range of stopping points; the individual components of $\hat{\alpha}$ are not so stable, but the overall significance of the test for normality described at the end of Section 7.1 remains well below 1%. The instability of the components of $\hat{\alpha}$ is not

$$\hat{\Omega}^* = \begin{array}{c} \\ Y \\ X \\ Z \\ U \\ V \end{array} \begin{array}{c} \begin{array}{ccccc} Y & X & Z & U & V \end{array} \\ \begin{pmatrix} 1 & -0.49 & 0.09 & -0.16 & 0.06 \\ -0.49 & 1 & -0.38 & -0.04 & 0.17 \\ 0.09 & -0.38 & 1 & 0.42 & -0.25 \\ -0.16 & -0.04 & 0.42 & 1 & -0.07 \\ -0.06 & 0.17 & -0.25 & -0.07 & 1.00 \end{pmatrix} \end{array}$$

	Y	X	Z	U	V
$\hat{\alpha}$	1.53	-32.89	-3.49	-1.16	-2.41
std.error	6.4	11.68	2.89	7.27	2.70
t-ratio	0.24	-2.81	-1.21	-0.16	-0.89

Table 3: *Matrix $\hat{\Omega}^*$, $\hat{\alpha}$ and other quantities associated to the regression analysis of (Y, X, Z, U, V) on $(1, W, A, B)$ for the glucose data*

$$\hat{\Omega}^* = \begin{array}{c} \\ Y \\ X \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} Y & X & Z \end{array} \\ \begin{pmatrix} 1.00 & -0.50 & 0.00 \\ -0.50 & 1.00 & -0.52 \\ 0.00 & -0.52 & 1.00 \end{pmatrix} \end{array}$$

	Y	X	Z
$\hat{\alpha}$	2.50	-21.42	-1.43
std. error	1.23	5.15	1.52
t ratio	2.04	-4.16	-0.94

Table 4: *Matrix $\hat{\Omega}^*$, $\hat{\alpha}$ and other quantities associated to the regression analysis of (Y, X, Z) on $(1, W, A)$ for the glucose data*

surprising considering that the sample size, $n = 68$, is small in this context, as discussed in Section 6.3.

Reduction of the model, dropping components because of the non-significant coefficients or because of their irrelevance to the variables of interest, leads to consideration of the triplet (Y, X, Z) with explanatory variables $(1, W, A)$. The new matrix $\hat{\Omega}^*$ and the vector $\hat{\alpha}$ are as reported in Table 4.

The final graphical model has an edge between (X, Y) and between (X, Z) to represent conditional dependence, for fixed values of (A, W) as indicated by $\hat{\Omega}^*$ and $\hat{\alpha}$ in Table 4; background information can be used to choose a direction on these arcs. Additional arcs are added from the fixed variables to the stochastic ones with the aid of the estimates and related t-ratios obtained from the last regression analysis, namely directed arcs between (A, Y), and (W, Z).

The pictorial representation of the graphical model is similar to the regression graph of Figure 6.4 of Cox & Wermuth (1996, p. 141), except for the arcs for they added on the basis of univariate regressions. Clearly, the building procedures and the associated interpretations are a bit different, and the two types of arcs (arising from conditional dependence and from regression) should be kept graphically distinct in our case.

We stress again that the above discussion intended to illustrate the use of the conditional independence techniques with the aid of a well-known dataset, not to produce a full data analysis. Moreover, the estimation method presented in Section 7.1 must be used with caution with small samples like this one.

8 AN EXTENSION TO ELLIPTICAL DENSITIES

The univariate skew-normal distribution was obtained by applying a skewing factor to the standard normal density, but the same method is applicable to any symmetric density, as stated in Lemma 1 of Azzalini (1985). This lemma can be extended to the k-dimensional case where the notion of symmetric density is replaced by the notion of elliptical density. The following lemma is a direct generalization of Lemma 1 of Azzalini (1985), of which it also follows the same line of argument in the proof.

Lemma 12 *Denote by X a continuous random variable with density function G' symmetric about 0 and by $Y = (Y_1, \ldots, Y_k)^\top$ a continuous random variable with density function f, such that X and Y are independent. Suppose that the real-valued transform $W(Y)$ has symmetric density about 0. Then*

$$\tilde{f}(y) = 2\, f(y) G(W(y)) \tag{22}$$

is a k-dimensional density function.

Proof. Since $X - W(Y)$ is symmetric about 0, then

$$\tfrac{1}{2} = \mathbb{P}\{X \le W(Y)\} = \mathbb{E}_Y\{\mathbb{P}\{X \le W(Y)|Y\}\} = \int_{\mathbb{R}^k} G(W(y))\, f(y)\, dy.$$

Corollary 13 *Suppose that X and Y satisfy the conditions of the above lemma, and in addition that Y has elliptical density centred at the origin; if*

$$W(Y) = a_1 Y_1 + \cdots + a_k Y_k = a^\top Y \tag{23}$$

then (22) is a k-dimensional density function for any choice of a.

Proof. The statement follows by noticing that $a^\top Y$ has 1-dimensional elliptical distribution, i.e. its density is symmetric about 0. See Theorem 2.16 of Fang, Kotz & Ng (1990) for the distribution of a linear transform of elliptical variables.

Clearly, (22) with $W(y)$ of type (23) includes the SN_k density for suitable choice of f, G and any choice of a.

In principle, Lemma 12 can be applied also to non-elliptical densities. For instance, if $Y \sim \mathrm{SN}_k$ and a is chosen suitably, according to Proposition 3, the density of W can be made normal, hence symmetric. There is however a major difference: in this case, the property holds for specific choices of a depending on the given choice of f, while with the elliptical densities it holds for all a's.

Implicit in the proof of the lemma there is an acceptance–rejection idea, hence a conditioning argument, similar to the one of Azzalini (1986), leading to the following method for random number generation. If X and Y are as in above lemma, and

$$Z = \begin{cases} Y & \text{if } X < W(Y), \\ -Y & \text{if } X > W(Y), \end{cases}$$

then the density function of Z is (22). In fact, its density at point z is

$$f(z)G(W(z)) + f(-z)\{1 - G(W(-z))\}$$

which is equal to $2\, f(z)\, G(W(z))$ if $f(z) = f(-z)$, a condition fulfilled e.g. by elliptical densities centred at 0.

9 FURTHER WORK

Various issues related to the SN family have been discussed, but many others remain pending. Broadly speaking, these fall in two categories: open questions and further applications.

Among the open questions, the anomalous behaviour of MLE in cases described in section 6.3 is worth exploration even *per se*. In the multivariate case, construction of more accurate standard errors would be welcome. A more radical solution would be the introduction of the centred parametrization which has not been carried on from the univariate to the multivariate case.

Besides applications to numerically more substantial applied problems than those discussed here, it is worth exploring the relevance of the distribution in other areas of multivariate statistics, in addition to those touched in section 7. A natural aspect to consider is the behaviour of other linear statistical methods outside normality, not only discriminant analysis. Another relevant use could be in connection with sampling affected by bias selection; this has been discusses by Copas & Li (1997) and references quoted therein, in the case of a scalar response variable. The skew-normal distribution offers the framework for a multivariate treatment of the same problem, by consideration of its genesis via conditioning.

The generalization to skew-elliptical densities has been left completely unexplored. An adequate treatment of the connected distributional and statistical issues requires the space of an entire paper. Hence, this direction has not been explored here, but a brief mention seemed to be appropriated, partly because of its close connection with the SN distribution.

ACKNOWLEDGMENTS

In the development of this paper, we much benefited from helpful and stimulating discussions with several colleagues. Specifically, we are most grateful to John Aitchison for suggesting the reparametrization adopted in subsection 7.1, to Ann Mitchell for introducing us to elliptical densities, to Paul Ruud for discussions about the EM algorithm, to Monica Chiogna and David Cox for additional general discussions, and to Samuel Kotz for his constant encouragement. Additional extensive comments from the referees and the editor have led to much better presentation of the material. We also thank W. Q. Meeker for kindly providing the Otis data, and A. Albert for the liver data and associated informations.

A substantial part of this work has been developed while the first author was at Nuffield College, Oxford, within the Jemolo Fellowship scheme; the generous hospitality of the College is gratefully acknowledged. Additional support has been provided by the 'Ministero per l'Università e per la Ricerca Scientifica e Tecnologica' and by 'Consiglio Nazionale delle Ricerche', Italy (grant no. 97.01331.CT10).

APPENDICES

TWO EQUIVALENT PARAMETRIZATIONS

We want to show that the (Ω, α) parametrization adopted in this paper is equivalent to the (λ, Ψ) parametrization of Azzalini & Dalla Valle (1996).

The matrix Ω and the vector α appearing in (1) were defined in Azzalini & Dalla Valle (1996) in terms of a correlation matrix Ψ and a vector $\lambda = (\lambda_1, \ldots, \lambda_k)^\top$; specifically, they

defined

$$\Delta = \text{diag}\left(\left(1+\lambda_1^2\right)^{-1/2},\ldots,\left(1+\lambda_k^2\right)^{-1/2}\right), \tag{24}$$

$$\Omega = \Delta(\Psi+\lambda\lambda^\top)\Delta, \tag{25}$$

$$\alpha = \left(1+\lambda^\top\Psi^{-1}\lambda\right)^{-1/2}\Delta^{-1}\Psi^{-1}\lambda. \tag{26}$$

Also, they defined $\delta=(\delta_1,\ldots,\delta_k)^\top$ where $\delta_j=\lambda_j\left(1+\lambda_j^2\right)^{-1/2}$ for $j=1,\ldots,k$.

With some algebraic work, it can be shown that (25) and (26) are invertible, obtaining

$$\Psi = \Delta^{-1}(\Omega-\delta\delta^\top)\Delta^{-1} \tag{27}$$

and (3), which then gives λ using $\lambda_j=\delta_j\left(1-\delta_j^2\right)^{-1/2}$. As a by-product, (5) is obtained.

Clearly, for any choice of the (λ,Ψ) pair, we obtain a feasible (Ω,α) pair; hence, we must only show the following.

Proposition 14 *For any choice of the correlation matrix Ω and of the vector $\alpha\in\mathbb{R}^k$, (1) is a density of SN_k type.*

Proof. Given α and Ω, compute δ using (3). This vector must satisfy condition $\Omega-\delta\delta^\top\geq 0$, required by (27); hence we must check that

$$\Omega-(1+\alpha^\top\Omega\alpha)^{-1}\Omega\alpha\alpha^\top\Omega\geq 0.$$

By using (6), the left-hand side can be seen to be equal to $(\Omega^{-1}+\alpha\alpha^\top)^{-1}$ which is positive definite. Moreover, fulfillment of $\Omega-\delta\delta^\top\geq 0$ implies that all components of δ are less than 1 in absolute value. Algebraic equivalence of (24)–(26) and (3), (27) completes the proof.

GRADIENT AND HESSIAN OF THE CENTRED PARAMETERS

The partial derivatives of $\ell(CP)$ defined in Section 6.2 with respect to (β,σ,λ) are

$$\begin{aligned}
\frac{\partial\ell}{\partial\beta} &= (\sigma_z/\sigma)^2X^\top\{y-X\beta-\sigma\sigma_z^{-1}(\lambda p_1-\mu_z 1_n)\},\\
\frac{\partial\ell}{\partial\sigma} &= -n/\sigma+\sigma_z(y-X\beta)^\top(z-p_1\lambda)/\sigma^2,\\
\frac{\partial\ell}{\partial\lambda} &= \frac{n}{\sigma_z}\sigma_z'-z^\top z'+p_1^\top(z+\lambda z')
\end{aligned}$$

where z' denotes the derivative with respect to λ, and

$$\begin{aligned}
&p_1=\zeta_1(\lambda z), \qquad z'=\mu_z'+\sigma^{-1}(y-X\beta)\sigma_z'=\mu_z'+r\sigma_z',\\
&\mu_z'=\frac{(2/\pi)^{1/2}}{(1+\lambda^2)^{3/2}}, \qquad \sigma_z'=-\frac{\mu_z}{\sigma_z}\mu_z'.
\end{aligned}$$

To obtain the partial derivatives with respect to γ_1, use

$$\frac{\partial\ell}{\partial\gamma_1}=\frac{\partial\ell}{\partial\lambda}\Big/\frac{\text{d}\gamma_1}{\text{d}\lambda}, \qquad \frac{\text{d}\gamma_1}{\text{d}\lambda}=\frac{3(4-\pi)}{2}\,\frac{\mu_z^2(\mu_z'\sigma_z-\mu_z\sigma_z')}{\sigma_z^4}.$$

or equivalently

$$\frac{\partial \ell}{\partial \gamma_1} = \frac{\partial \ell}{\partial \lambda}\,\frac{\mathrm{d}\lambda}{\mathrm{d}\gamma_1}, \qquad \frac{\mathrm{d}\lambda}{\mathrm{d}\gamma_1} = \frac{2}{3(4-\pi)}\left(\frac{1}{T\,R^2} + \frac{1-2/\pi}{T^3}\right)$$

where

$$R = \frac{\mu_z}{\sigma_z} = \left(\frac{2\gamma_1}{4-\pi}\right)^{1/3} \qquad T = \left(2/\pi - (1-2/\pi)R^2\right)^{1/2}.$$

The above derivatives lead immediately to the likelihood equations for $CP = (\beta, \sigma, \gamma_1)$. We need second derivatives for numerical efficient computations, and for computing the observed information matrix. The entries of the Hessian matrix for (β, σ, λ) are given by

$$\begin{aligned}
-\frac{\partial^2 \ell}{\partial \beta \partial \beta^\top} &= (\sigma_z/\sigma)^2 X^\top (I_n - \lambda^2 P_2) X, \\
-\frac{\partial^2 \ell}{\partial \beta \partial \sigma} &= (\sigma_z/\sigma^2) X^\top (z - \lambda p_1 + (I_n - \lambda^2 P_2)(z - \mu_z 1_n)), \\
-\frac{\partial^2 \ell}{\partial \beta \partial \lambda} &= \sigma^{-1} X^\top \{\sigma_z'(-2r\sigma_z + \lambda p_1 - \mu_z 1_n) + \sigma_z(p_1 + \lambda P_2 \tilde{z} - \mu_z' 1_n)\}, \\
-\frac{\partial^2 \ell}{\partial \sigma^2} &= \sigma^{-2}\{-n + 2\sigma_z r^\top (z - \lambda p_1) + \sigma_z^2 r^\top (I_n - \lambda^2 P_2) r\}, \\
-\frac{\partial^2 \ell}{\partial \sigma \partial \lambda} &= -\sigma^{-1} r^\top \left(\sigma_z'(z - \lambda p_1) + \sigma_z(z' - p_1 - \lambda P_2 \tilde{z})\right), \\
-\frac{\partial^2 \ell}{\partial \lambda^2} &= n\frac{(\sigma_z')^2 - \sigma_z \sigma_z''}{\sigma_z^2} + (z')^\top z' + z^\top z'' - \tilde{z}^\top P_2 \tilde{z} - p_1^\top (2z' + \lambda z'')
\end{aligned}$$

where

$$\begin{aligned}
&r = \sigma^{-1}(y - X\beta), \qquad \tilde{z} = z + \lambda z', \\
&p_1 = \zeta_1(\lambda z), \qquad P_2 = \mathrm{diag}(p_2) = \mathrm{diag}(\zeta_2(\lambda z)), \\
&z'' = \frac{\mathrm{d}z'}{\mathrm{d}\lambda} = \mu_z'' + \sigma_z'' \sigma^{-1}(y - X\beta), \\
&\mu_z'' = \frac{\mathrm{d}\mu_z'}{\mathrm{d}\lambda} = -\frac{3\mu_z}{(1+\lambda^2)^2}, \qquad \sigma_z'' = \frac{\mathrm{d}\sigma_z'}{\mathrm{d}\lambda} = -\left(\frac{\mu_z'(\mu_z'\sigma_z - \mu_z\sigma_z')}{\sigma_z^2} + \frac{\mu_z \mu_z''}{\sigma_z}\right).
\end{aligned}$$

Again, to obtain the Hessian matrix with respect to γ_1 instead of λ, the last row and last column of the above matrix must be multiplied by $\mathrm{d}\lambda/\,\mathrm{d}\gamma_1$, except the bottom right element which is computed as

$$-\frac{\partial^2 \ell}{\partial \gamma_1^2} = -\frac{\partial^2 \ell}{\partial \lambda^2}\left(\frac{\mathrm{d}\lambda}{\mathrm{d}\gamma_1}\right)^2 - \frac{\partial \ell}{\partial \lambda}\left(\frac{\mathrm{d}^2\lambda}{\mathrm{d}\gamma_1^2}\right).$$

The final term of this expression is given by

$$\frac{\mathrm{d}^2\lambda}{\mathrm{d}\gamma_1^2} = -\frac{2}{3(4-\pi)}\left(\frac{T'}{(T\,R)^2} + \frac{2R'}{T\,R^3} + \frac{3(1-2/\pi)T'}{T^4}\right)$$

where

$$R' = \frac{\mathrm{d}R}{\mathrm{d}\gamma_1} = \frac{2}{3\,R^2(4-\pi)}, \qquad T' = \frac{\mathrm{d}T}{\mathrm{d}\gamma_1} = -(1-2/\pi)\frac{R\,R'}{T}.$$

For practical numerical work, the above quantities suffices. If the expected Fisher information matrix I_{CP} is needed, this is given by

$$I_{CP} = D^\top I_{DP} D$$

where I_{DP} is the information matrix for the DP parameters, given by Azzalini (1985) in the case $X = 1_n$, and

$$D = \left(\frac{\partial (DP)_i}{\partial (CP)_j}\right) = \begin{pmatrix} 1 & -\dfrac{\mu_z}{\sigma_z} & \dfrac{\partial \xi}{\partial \gamma_1} \\ 0 & \dfrac{1}{\sigma_z} & \dfrac{\partial \omega}{\partial \gamma_1} \\ 0 & 0 & \dfrac{\partial \lambda}{\partial \gamma_1} \end{pmatrix}$$

where

$$\frac{\partial \xi}{\partial \gamma_1} = -\frac{\sigma \mu_z}{3\sigma_z \gamma_1}, \qquad \frac{\partial \omega}{\partial \gamma_1} = -\frac{\sigma \sigma_z'}{\sigma_z^2} \frac{\mathrm{d}\lambda}{\mathrm{d}\gamma_1}.$$

REFERENCES

Aigner, D. J., Lovell, C. A. K. & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function model. *J. Econometrics* **12**, 21–37.

Albert, A. & Harris, E. K. (1987). *Multivariate Interpretation of Clinical Laboratory Data*. Dekker, New York and Basel.

Arnold, B.C., Beaver, R.J., Groeneveld, R.A. & Meeker, W.Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* **58**, 471-478.

Azzalini, A. (1985). A class of distribution which includes the normal ones. *Scand. J. Statist.* **12**, 171–8.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* **46**, 199–208.

Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–26.

Barndorff-Nielsen, O. & Blæsild, P. (1983). Hyperbolic distributions. In: *Encyclopedia of Statistical Sciences* (ed. N.L.Johnson, S.Kotz & C.B.Read), vol. 3, 700–707. Wiley, New York.

Blæsild, P. (1981). The two-dimensional hyperbolic distribution and related distributions, with an application to Johansen's bean data. *Biometrika*, **68**, 251–63.

Chiogna, M. (1997). Notes on estimation problems with scalar skew-normal distributions. Technical report 1997.15, Department of Statistical Sciences, University of Padua.

Cartinhour, J. (1990). One dimensional marginal density function of a truncated multivariate Normal density function. *Comm. Statist., Theory and Methods* **19**, 197–203.

Chou, Y.-M. & Owen, D. B. (1984). An approximation to the percentiles of a variable of the bivariate normal distribution when the other variable is truncated, with applications. *Comm. Statist., Theory and Methods*, **13**, 2535–47.

Cook, R. D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

Copas, J. B. & Li, H. G. (1997). Inference for non-random samples (with discussion). *J. Roy. Statist. Soc. B*, **59**, 55–95.

Cox, D, R, & Wermuth, N. (1996). *Multivariate dependencies: models, analysis and interpretation*. Chapman & Hall, London.

David, F. N., Kendall, M. G. & Barton, D. E. (1966) *Symmetric functions and allied tables*. Cambridge University Press.

Fang, K.-T., Kotz, S. & Ng, K. (1990). *Symmetric multivariate and related distributions*. Chapman & Hall, London.

Healy, M. J. R. (1968). Multivariate normal plotting. *Appl. Statist.* **17**, 157–161.

Johnson, N. L. & Kotz, S. (1972). *Distributions in statistics: continuous multivariate distributions*. Wiley, New York

Liseo, B. (1990). The skew-normal class of densities: Inferential aspects from a Bayesian viewpoint (in Italian). *Statistica*, **50**, 59–70.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā* **B 36**, 115-28.

McCullagh, P. (1987). *Tensor methods in statistics*. Chapman & Hall, London.

Meng, X.-L. & van Dyk, D. (1997). The EM-algorithms — an old folk-song sung to a fast new tune (with discussion). *J.Roy. Statist. Soc. B* **59**, 511–67.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. Wiley, New York.

Rao, C. R. (1947). The problem of classification and distance between two populations. *Nature*, **159**, 30–31.

Rao, C. R. (1973). *Linear statistical inference*, 2nd edition. Wiley, New York.

Rotnitzky, A, Cox, D. R. Bottai, M. & Robins, J. (1999). Likelihood-based inference with singular information matrix. To appear.