Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 99 (2008) 1362–1382

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

The centred parametrization for the multivariate skew-normal distribution

Reinaldo B. Arellano-Valle[a], Adelchi Azzalini[b,*]

[a] Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile
[b] Dipartimento di Scienze Statistiche, Università di Padova, Italy

Available online 12 February 2008

Abstract

For statistical inference connected to the scalar skew-normal distribution, it is known that the so-called centred parametrization provides a more convenient parametrization than the one commonly employed for writing the density function. We extend the definition of the centred parametrization to the multivariate case, and study the corresponding information matrix.

AMS 2000 subject classifications: 62F12; 62H12; 62E20

Keywords: Index of skewness; Skew-normal distribution; Parametrization; Singular information matrix

1. Background and motivation

1.1. The skew-normal distribution

In recent years, there has been increasing interest in a scheme for the genesis of families of distributions whose main features are collected in the book edited by Genton [13] and in the review paper by Azzalini [5]. The more emblematic and in a sense simplest representative of these distributions is the so-called skew-normal (SN) distribution, whose density function is, in the one-dimensional case,

$$f_1(x;\xi,\omega^2,\alpha)=2\,\omega^{-1}\,\phi\left(\frac{x-\xi}{\omega}\right)\,\Phi\left\{\alpha\left(\frac{x-\xi}{\omega}\right)\right\},\quad (x\in\mathbb{R}),\tag{1}$$

* Corresponding author.
E-mail address: azzalini@stat.unipd.it (A. Azzalini).

0047-259X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.01.020

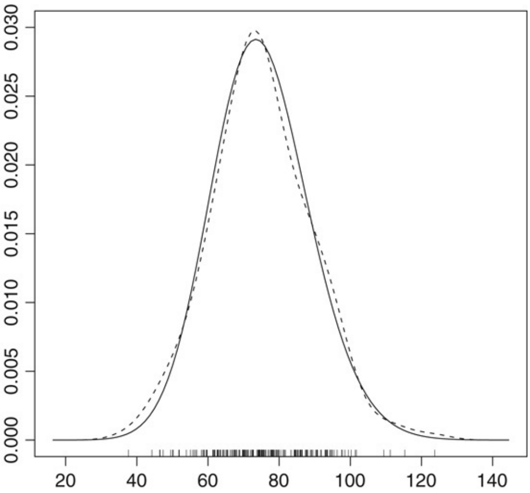

Fig. 1. AIS weight data: the ticks on the horizontal axis correspond to the individual data values, the dashed curve represents the nonparametric estimate of the density function, the continuous curve is the SN density fitted by maximum likelihood.

where ϕ and Φ denote the $N(0, 1)$ density and distribution function, respectively, and ξ, ω and α are location, scale and shape parameters, respectively ($\xi, \alpha \in \mathbb{R}$, $\omega \in \mathbb{R}^+$). The SN family forms a superset of the normal family, which corresponds to the choice $\alpha = 0$; with other values of α, a skewed density is obtained. Various other formal connections between (1) and the normal distribution are recalled in the above-mentioned references.

From the above-quoted general sources, it emerges that substantially more work has been oriented towards the probabilistic aspects of this approach and appreciably less to the statistical ones. Part of the explanation of this fact lies in the striking simplicity of treatment offered by this approach on the mathematical and probabilistic side, while the corresponding statistical work is, by contrast, surprisingly somewhat problematic.

Among these problematic aspects, a peculiar one concerns the behaviour of the likelihood function and other related quantities for a sample from the SN distribution in the neighbourhood of $\alpha = 0$, which is a value of particular relevance since there the SN family reduces to the normal one. We anticipate that the term ‘neighbourhood’ must be interpreted in a fairly ‘wide’ sense, as we shall see shortly.

Consider for illustration the AIS (Australian Institute of Sport) data, repeatedly employed in this stream of literature. These data contain a number of biomedical measurements on 202 Australian athletes, of which we shall use here only the weight (in kg). Fig. 1 reports on the horizontal axis the individual data values, and it displays a nonparametric estimate of the density function obtained by the kernel method, as well as the parametric curve of type (1) selected by maximum likelihood estimation (MLE), such that $(\hat{\xi}, \hat{\omega}^2, \hat{\alpha}) = (64.07, 17.68^2, 1.23)$. The general impression perceived from Fig. 1 is that the SN curve follows quite closely the nonparametric estimate of the density, and both curves indicate a mild departure from normality because of some skewness to the right.

While the fitting process is satisfactory in the sense of an overall indication of adequate parametric fit, there are other aspects which are less pleasant, due to some anomalies of the log-likelihood function. For a random sample $y_1, \ldots, y_n$ from distribution (1), the log-likelihood

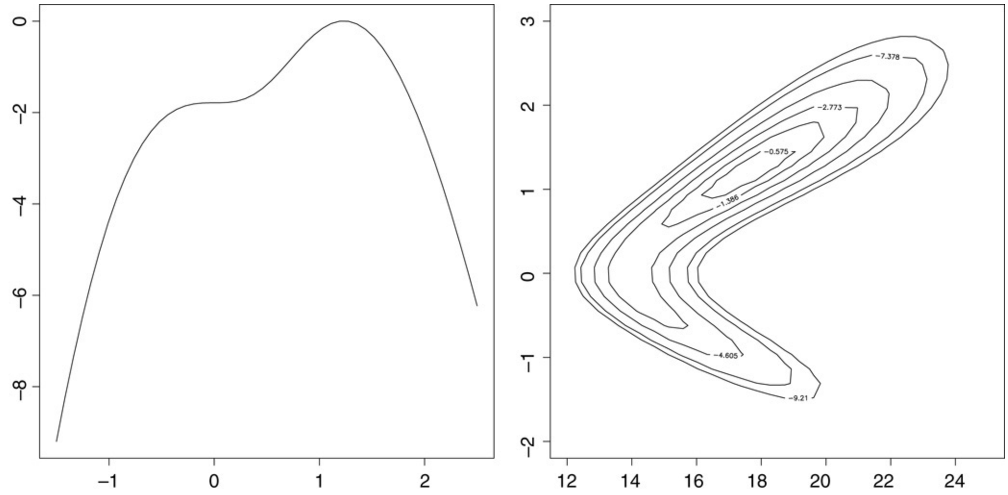

Fig. 2. AIS weight data: profile twice the relative log-likelihood function for α, in the left panel, and for (ξ, ω), in the right panel.

function is

$$\ell_{\mathrm{DP}}(\xi, \omega^2, \alpha) = \text{constant} - \frac{n}{2}\log\omega^2 - \frac{1}{2}\sum_i \left(\frac{y_i - \xi}{\omega}\right)^2 + \sum_i \zeta_0\left\{\alpha\left(\frac{y_i - \xi}{\omega}\right)\right\}, \quad (2)$$

where $\zeta_0(x) = \log\{2\,\Phi(x)\}$; the reason for inserting the subscript DP will be clarified later in the text.

Since there are three parameters involved in (2), we can only display it through the corresponding profile log-likelihood function. For the shape parameter α, this amounts to considering $\ell^*_{\mathrm{DP}}(\alpha) = \ell_{\mathrm{DP}}(\hat{\hat{\xi}}(\alpha), \hat{\omega}^2(\alpha), \alpha)$, where $\hat{\hat{\xi}}(\alpha)$ and $\hat{\omega}^2(\alpha)$ denote the values of ξ and ω^2 which maximize ℓ_{DP} for any given value of α, respectively. To ease readability, the profile log-likelihood has been transformed into a so-called relative profile log-likelihood, by subtracting the maximum value $\ell_{\mathrm{DP}}(\hat{\hat{\xi}}(\alpha), \hat{\omega}^2(\alpha), \hat{\alpha})$; hence the maximum value of the new function is 0. Twice the relative profile log-likelihood for (ω^2, α), or equivalently for (ω, α), is defined in a similar way. The two resulting functions, when constructed for the weight variable of the AIS data, are shown in Fig. 2; the one for (ω, α) is displayed in the form of contour level curves of twice the profile relative log-likelihood.

It is apparent that the two plots in Fig. 2 exhibit a markedly non-quadratic shape of the log-likelihood function. One unusual feature is the stationary point at $\alpha = 0$ in the left-hand side plot. This stationary point is not a peculiar trait of these specific data; it occurs with any sample, as noted by Azzalini [3]. A similar sort of non-quadratic behaviour has been exhibited with other data used by Arnold et al. [2], Azzalini and Capitanio [6, Section 5].

Another unpleasant phenomenon, connected to the presence of this stationary point, is that, at $\alpha = 0$, the expected Fisher information is singular, even if all parameters are identifiable. This fact violates the standard assumptions leading to the asymptotic normal distribution of the MLEs. A situation of this sort falls under the umbrella of the non-standard asymptotic theory studied by Rotnitzky et al. [16], where it is shown that in these circumstances the rate of convergence of the estimates is slower than the usual $O_p(n^{-1/2})$ and that the limiting distribution of the estimates can be bimodal.

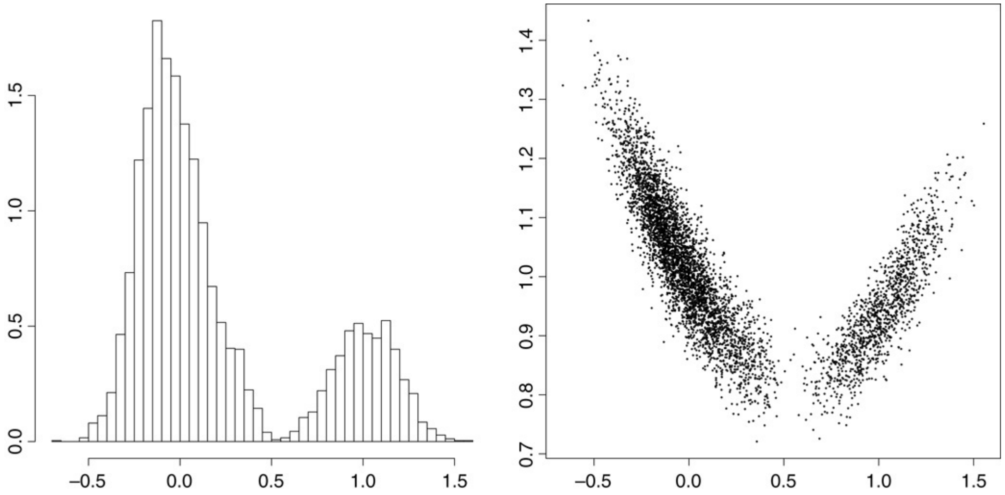

Fig. 3. Estimated distributions of the MLEs when samples of size $n = 200$ are drawn from SN(0, 1, 1); the left panel displays the histogram of $\hat{\xi}$, the right panel displays the scatter plot of $(\hat{\xi}, \hat{\omega})$.

Although for $n \to \infty$ this unusual behaviour of the estimates is limited to the point $\alpha = 0$, in practice for finite sample size this phenomenon propagates even at some distance from the point $\alpha = 0$. Furthermore, because of the slow rate of convergence of the estimates, this situation persists even for quite large sample sizes. To get a direct perception of the problem, we have run a little simulation experiment generating 5000 samples of size $n = 200$ each from SN(0, 1, 1), and for each sample the MLEs $(\hat{\xi}, \hat{\omega}^2, \hat{\alpha})$ have been computed. The sample size $n = 200$ and the shape parameter $\alpha = 1$ were chosen to match approximately those of the AIS weight data. Fig. 3 displays the corresponding empirical distribution of $\hat{\xi}$ and of $(\hat{\xi}, \hat{\omega})$, in the form of an histogram and a scatter plot, respectively.

1.2. An alternative parametrization

It is apparent that standard likelihood-based methods are problematic when they are applied for inference on the parameters (ξ, ω^2, α), at least near $\alpha = 0$. To a large extent, the problems would persist even if one switched to the Bayesian approach, unless a strongly informative prior distribution is adopted, since a diffuse prior would lead to a posterior distribution with a shape not very different, for the AIS weight data, from the one of Fig. 2, hence producing credibility regions with peculiar shapes, especially in the multiparameter case.

The above problems are however entirely due to a parametrization not suitable for estimation, since the parameters themselves are identifiable. This remark is the basis of the proposal of Azzalini [3] of an alternative parametrization for $Y \sim \mathrm{SN}(\xi, \omega^2, \alpha)$, which is defined as follows. Start from the identity

$$Y = \xi + \omega Z = \mu + \sigma Z_0, \tag{3}$$

where $Z = (Y - \xi)/\omega$ is the ‘normalized’ variable with distribution $\mathrm{SN}(0, 1, \alpha)$, and $Z_0 = \sigma_z^{-1}(Z - \mu_z)$ is its standardized version, having set

$$\mu_z = \mathbb{E}\{Z\} = b\,\delta, \qquad \sigma_z^2 = \mathrm{var}\{Z\} = 1 - b^2\,\delta^2$$

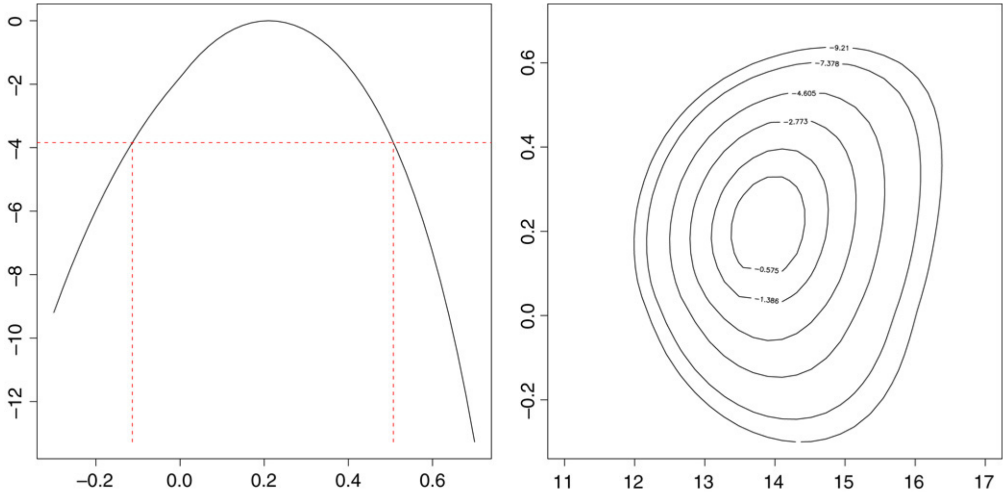

Fig. 4. AIS weight data: profile twice the relative log-likelihood function for γ_1, in the left panel, and for (σ, γ_1), in the right panel.

and $b = \sqrt{2/\pi}$, $\delta = \alpha/\sqrt{1+\alpha^2}$; here μ and σ are defined implicitly so that (3) is satisfied. The alternative parametrization is then formed by $(\mu, \sigma^2, \gamma_1)$ whose explicit expressions are

$$\begin{aligned} \mu &= \mathbb{E}\{Y\} = \xi + \omega\mu_z \\ \sigma^2 &= \text{var}\{Y\} = \omega^2\,(1-\mu_z^2) \\ \gamma_1 &= \frac{\mathbb{E}\{(Y-\mathbb{E}\{Y\})^3\}}{\text{var}\{Y\}^{3/2}} = \frac{4-\pi}{2}\,\frac{\mu_z^3}{(1-\mu_z^2)^{3/2}}, \end{aligned} \tag{4}$$

where γ_1 denotes the Pearson's index of skewness. For later use, we note that the inverse of transformation (4) is

$$\mu_z = \frac{c}{\sqrt{1+c^2}}, \quad c = \left(\frac{2\gamma_1}{4-\pi}\right)^{1/3}. \tag{5}$$

We then denote $(\mu, \sigma^2, \gamma_1)$ as the 'centred parameters' (CP) since they are built via the centred variable Z_0, while the earlier parameters (ξ, ω^2, α) are called 'direct' (DP) because they can be read directly from (1).

The adoption of the CP in place of the DP is highly beneficial for ironing out the peculiar features described earlier. Fig. 4 displays plots which are analogous to those of Fig. 2 but the roles of scale parameter and shape parameter are now played by σ and γ_1, respectively. Clearly the new plots exhibit a definitely more regular behaviour, much closer to quadratic functions, and without a stationary point at $\gamma_1 = 0$. The maximum likelihood estimates of CP are $(75.0, 13.9^2, 0.211)$.

If each of the 5000 estimates $(\hat{\xi}, \hat{\omega}^2, \hat{\alpha})$ of the simulation experiment described in Section 1.1 is transformed to the new parameter estimates $(\hat{\mu}, \hat{\sigma}^2, \hat{\gamma}_1)$, then the empirical distribution of the estimates $\hat{\mu}$ of new location parameter μ is as shown in the left panel of Fig. 5, while that of $(\hat{\mu}, \hat{\sigma})$ is in the right panel of the same figure. Notice that DP $= (0, 1, 1)$ corresponds to CP $= (0.564, 0.826^2, 0.137)$. Clearly these empirical distributions are much closer to normality than those in Fig. 2. In fact, it can be shown that the singularity of the expected Fisher information matrix when the skewness parameter is null does not occur any longer, and

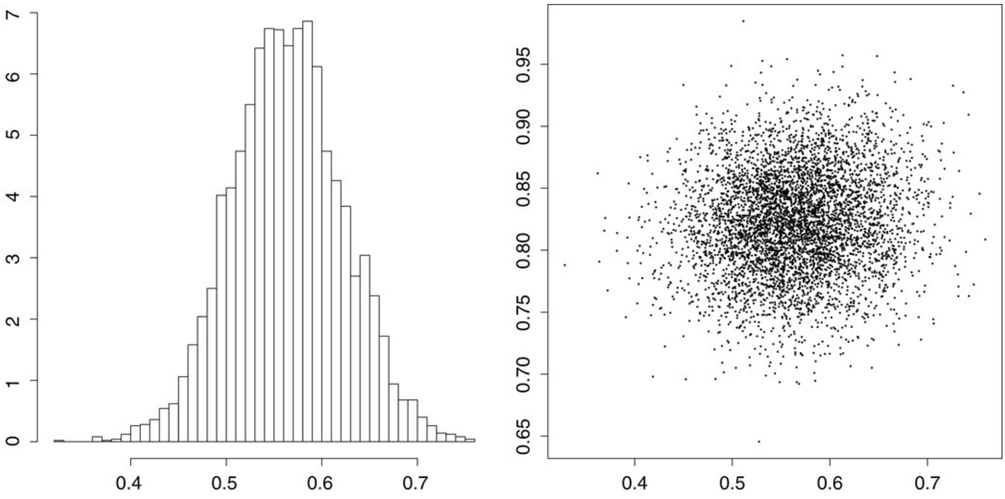

Fig. 5. Estimated distributions of the MLE when samples of size $n = 200$ are drawn from SN(0, 1, 1) but adopting the centred parametrization; the left panel displays the histogram of $\hat{\mu}$, the right panel displays the scatter plot of $(\hat{\mu}, \hat{\sigma})$.

the limiting distribution for the CP estimates is multivariate normal with asymptotic variance $\mathrm{diag}(\sigma^2, 2\,\sigma^4, 6)$; see [3] and the careful derivation of Chiogna [11].

Since the estimation problem is now re-cast in a formulation where standard asymptotics holds, usual likelihood-based methods can then be employed. For instance, a 95% confidence interval for γ_1 is identified in the left panel of Fig. 4 by finding the values where the curve intersects the level -1.96^2, leading to the confidence interval $(-0.113, 0.507)$. If this interval is mapped on the α axis, it corresponds to the interval $(-0.920, 2.20)$. This other interval can be obtained directly from the left panel of Fig. 2, by intersecting the curve with the level -1.96^2. Notice however that the validity of this second procedure rests on the validity of the confidence interval for the CP plus the equivariance property of the likelihood function when we map γ_1 on the α scale; it would not be justified solely on the basis of the properties of the likelihood function in the DP formulation.

The adoption of the CP formulation turns out to be a satisfactory solution of the earlier problems with the DP. In addition, there is a distinct advantage of the CP over the DP as regards interpretability of the individual components. The CP formulation has however been considered only for the scalar SN distribution (1). Given the increasing set of applications of the multivariate SN distribution, it is useful to develop the CP formulation also in the d-dimensional case, a problem which does not appear to have been considered in the literature. This will be the target of the rest of this paper.

The multivariate SN distribution has been discussed by Azzalini and Dalla Valle [8], and subsequently by Azzalini and Capitanio [6] who have introduced a modified parametrization, which is in fact more clearly linked to the scalar version (1), although technically equivalent. Under the latter formulation, the d-dimensional SN density function is

$$f_d(x; \xi, \Omega, \alpha) = 2\,\phi_d(x - \xi; \Omega)\ \Phi\{\alpha^\top \omega^{-1}(x - \xi)\}, \quad (x \in \mathbb{R}^d), \tag{6}$$

where $\phi_d(x; \Omega)$ denotes the $N_d(0, \Omega)$ density function for a $d \times d$ positive definite symmetric matrix Ω, ξ is a vector location parameter, α is a vector shape parameter ($\xi, \alpha \in \mathbb{R}^d$), and ω is a diagonal matrix formed by the standard deviations of Ω.

2. The centred parametrization in the multivariate case

2.1. Definition of CP and plan of work

To define the CP formulation in the multivariate case, we introduce some expressions following the scheme of Azzalini and Capitanio [6]. Define

$$\bar{\Omega} = \omega^{-1}\Omega\omega^{-1}, \qquad \delta = (1+\alpha^\top\bar{\Omega}\alpha)^{-1/2}\bar{\Omega}\alpha$$

and the 'normalized' variable $Z = \omega^{-1}(Y-\xi) \sim \mathrm{SN}_d(0, \bar{\Omega}, \alpha)$, such that

$$\mathbb{E}\{Z\} = b\,\delta = \mu_z, \qquad \mathrm{var}\{Z\} = \bar{\Omega} - \mu_z\mu_z^\top = \bar{\Omega} - b^2\,\delta\delta^\top = \Sigma_z \tag{7}$$

and its standardized version is $Z_0 = \sigma_z^{-1}(Z-\mu_z)$, where $\sigma_z = \mathrm{diag}(\sigma_{z,1}, \ldots, \sigma_{z,d})$ whose terms are the standard deviations of Σ_z such that the j-th term is $\sigma_{z,j} = (1-b^2\delta_{z,j}^2)^{1/2}$, in analogy to the scalar case ($j = 1, \ldots, d$). The underlying idea here is to use again the representation (3) but with ingredients which are vectors and diagonal matrices. The CP is now given by (μ, Σ, γ_1), where

$$\mu = \mathbb{E}\{Y\} = \xi + \omega\,\mu_z, \qquad \Sigma = \mathrm{var}\{Y\} = \Omega - \omega\mu_z\mu_z^\top\omega = \omega\Sigma_z\omega \tag{8}$$

and γ_1 is the d-dimensional vector obtained by applying (4) on each separate component of μ_z.

Clearly, when $d = 1$, this definition of CP reduces to the one defined in Section 1.2. Furthermore, it is easy to show that there is a one-to-one correspondence between DP $= (\xi, \Omega, \alpha)$ and CP $= (\mu, \Sigma, \gamma_1)$. The only slightly peculiar aspect is that, while the three ingredients of DP are variation independent and they can be chosen freely, apart from the constraint $\Omega > 0$, the same fact is not true for CP. Under the condition that a certain choice of (μ, Σ, γ_1) belongs to the admissible CP set, then μ_z can be computed from γ_1 using (5) componentwise; then δ and σ_z are immediately obtained. After building the diagonal matrix σ with the square roots of the diagonal of Σ, the terms ξ, ω and Ω are computed from

$$\xi = \mu - \sigma\sigma_z^{-1}\mu_z, \qquad \omega = \sigma\,\sigma_z^{-1}, \qquad \Omega = \Sigma + \omega\mu_z\mu_z^\top\omega$$

and the last component of DP is

$$\alpha = \frac{1}{\sqrt{1-\delta^\top\bar{\Omega}^{-1}\delta}}\bar{\Omega}^{-1}\delta$$

using (5) of [6].

Hence the CP provides a legitimate parametrization of the SN multivariate family. Under the above assumption that a given choice of (μ, Σ, γ_1) is admissible and it corresponds to some point (ξ, Ω, α) of DP, the log-likelihood function for a random sample $y_1, \ldots, y_n$ is given by evaluating (2) at the corresponding point in the DP space, that is

$$\ell_{\mathrm{CP}}(\mu, \Sigma, \gamma_1) = \ell_{\mathrm{DP}}(\xi, \Omega, \alpha).$$

Our main target is to obtain some associated quantities, especially the Fisher information matrix for the CP, in the expected and the observed form. To this end, we shall proceed in various steps, introducing additional parametrizations; these are not of intrinsic interest, and represent merely a technical means to an end. In this process, we shall also obtain similar quantities for the DP.

2.2. The information matrix for the working parametrization and for DP

We shall build the computation of the information matrix first for a working parametrization $\theta = (\xi, \Omega, \eta)$, where $\eta = \omega^{-1}\alpha$. Since Ω is symmetric, the actual parametrization discards the duplicated elements of Ω, but we shall delay the explicit treatment of this fact until later. This parametrization is not really of direct statistical use as it lacks a simple meaningful interpretation, for the reasons explained by Arellano-Valle and Azzalini [1], but it proves to be a useful technical device. Azzalini and Capitanio [6, Section 6.1] have used this parametrization to develop an efficient algorithm for finding the MLEs.

If y denotes a single observation from $Y \sim \mathrm{SN}_d(\xi, \Omega, \alpha)$, then the contribution from y to the log-likelihood function for the working parametrization is

$$\ell(\theta) = \text{constant} - \frac{1}{2}\log|\Omega| - \frac{1}{2}\mathrm{tr}\{\Omega^{-1} S_0\} + \zeta_0(\eta^\top y_0), \tag{9}$$

where $y_0 = y - \xi$ and $S_0 = y_0 y_0^\top$. Starting from this expression, the overall scheme of the work to be developed is as follows: (i) the information matrix for θ is obtained; (ii) the Jacobian matrix of the transformation from θ to DP is obtained, leading to the information matrix for DP; (iii) similarly, the Jacobian matrix of another transformation which maps θ into CP is computed, leading to the information matrix for CP; this second transformation involves in fact two steps. Finally, some properties of the ensuing information matrices are obtained.

We shall make use of methods in matrix differential calculus, as developed by Magnus and Neudecker [14], from which we shall recall some results as their need occurs. To write the first differential of $\ell(\theta)$, take into account that $\mathrm{d}\log|\Omega| = \mathrm{tr}\{\Omega^{-1}(\mathrm{d}\Omega)\} = -\mathrm{tr}\{(\mathrm{d}\Omega^{-1})\Omega\}$, $\mathrm{d}y_0 = -\mathrm{d}\xi$, $\mathrm{d}S_0 = -[(\mathrm{d}\xi)y_0^\top + y_0(\mathrm{d}\xi)^\top]$, leading to

$$\begin{aligned}
\mathrm{d}\ell(\theta) &= -\frac{1}{2}\mathrm{d}\log|\Omega| - \frac{1}{2}\mathrm{tr}\{(\mathrm{d}\Omega^{-1})S_0 + \Omega^{-1}(\mathrm{d}S_0)\} + \mathrm{d}\zeta_0(\eta^\top y_0)\\
&= \frac{1}{2}\mathrm{tr}\{(\mathrm{d}\Omega^{-1})\Omega\} - \frac{1}{2}\mathrm{tr}\{(\mathrm{d}\Omega^{-1})S_0\} + (\mathrm{d}\xi)^\top\Omega^{-1}y_0 + \zeta_1(\eta^\top y_0)[(\mathrm{d}\eta)^\top y_0 - \eta^\top\mathrm{d}\xi]\\
&= \frac{1}{2}\mathrm{tr}\{(\mathrm{d}\Omega^{-1})[\Omega - S_0]\} + (\mathrm{d}\xi)^\top\Omega^{-1}y_0 + \zeta_1(\eta^\top y_0)[(\mathrm{d}\eta)^\top y_0 - \eta^\top\mathrm{d}\xi],
\end{aligned}$$

where ζ_1 denotes the derivative of ζ_0. Recalling that $\mathrm{d}^2X = \mathrm{d}(\mathrm{d}X) = 0$, the second differential of $\ell(\theta)$ is obtained as follows:

$$\begin{aligned}
\mathrm{d}^2\ell(\theta) &= \frac{1}{2}\mathrm{tr}\{(\mathrm{d}^2\Omega^{-1})[\Omega - S_0] + (\mathrm{d}\Omega^{-1})[\mathrm{d}\Omega - \mathrm{d}S_0]\} + (\mathrm{d}\xi)^\top(\mathrm{d}\Omega^{-1})y_0\\
&\quad - (\mathrm{d}\xi)^\top\Omega^{-1}\mathrm{d}\xi + \zeta_2(\eta^\top y_0)[(\mathrm{d}\eta)^\top y_0 - \eta^\top\mathrm{d}\xi][(\mathrm{d}\eta)^\top y_0 - \eta^\top\mathrm{d}\xi]\\
&\quad - 2\zeta_1(\eta^\top y_0)(\mathrm{d}\eta)^\top\mathrm{d}\xi\\
&= \frac{1}{2}\mathrm{tr}\{(\mathrm{d}^2\Omega^{-1})[\Omega - S_0] + (\mathrm{d}\Omega^{-1})\mathrm{d}\Omega\} + 2(\mathrm{d}\xi)^\top(\mathrm{d}\Omega^{-1})y_0\\
&\quad - (\mathrm{d}\xi)^\top\Omega^{-1}\mathrm{d}\xi + \zeta_2(\eta^\top y_0)[(\mathrm{d}\eta)^\top S_0\mathrm{d}\eta - 2(\mathrm{d}\eta)^\top y_0\eta^\top\mathrm{d}\xi\\
&\quad + (\mathrm{d}\xi)^\top\eta\eta^\top\mathrm{d}\xi] - 2\zeta_1(\eta^\top y_0)(\mathrm{d}\eta)^\top\mathrm{d}\xi\\
&= \frac{1}{2}\mathrm{tr}\{(\mathrm{d}^2\Omega^{-1})[\Omega - S_0] + (\mathrm{d}\Omega^{-1})\mathrm{d}\Omega\} + 2(\mathrm{d}\xi)^\top(\mathrm{d}\Omega^{-1})y_0\\
&\quad - (\mathrm{d}\xi)^\top[\Omega^{-1} - \zeta_2(\eta^\top y_0)\eta\eta^\top]\mathrm{d}\xi\\
&\quad + \zeta_2(\eta^\top y_0)(\mathrm{d}\eta)^\top S_0\mathrm{d}\eta - 2(\mathrm{d}\eta)^\top[\zeta_1(\eta^\top y_0)I_d + \zeta_2(\eta^\top y_0)y_0\eta^\top]\mathrm{d}\xi,
\end{aligned}$$

where ζ_2 denotes the second derivative of ζ_0. Notice that $\zeta_2(x) = -\zeta_1(x)\{x + \zeta_1(x)\}$.

For computing the expected value of the above expression, we shall make use of results based on the following lemma, whose proof is given in the Appendix, along with some corollaries.

Lemma 1. *Let $Y_0 \sim \mathrm{SN}_d(0, \Omega, \alpha)$ and let $\eta = \omega^{-1}\alpha$. Then, provided these expectations exist,*

$$\mathbb{E}\{[\zeta_1(\eta^\top Y_0)]^k g(Y_0)\} = c_k\, \mathbb{E}\{g(W_k)/[\Phi(\eta^\top W_k)]^{k-1}\}, \quad (k = 1, 2, \ldots),$$

where

$$c_k = \frac{2}{(2\pi)^{k/2}\sqrt{1 + k\eta^\top \Omega\eta}} = \frac{2(b/2)^k}{\sqrt{1 + k\eta^\top \Omega\eta}}$$

and $W_k \sim N_d(0, \Omega_k)$ where

$$\Omega_k = (\Omega^{-1} + k\eta\eta^\top)^{-1} = \Omega - k(1 + k\eta^\top \Omega\eta)^{-1}\Omega\eta\eta^\top \Omega.$$

Using this lemma, and the quantities defined therein, we can write

$$\begin{aligned}
&\mathbb{E}\{\zeta_1(\eta^\top Y_0)\} = c_1,\\
&\mathbb{E}\{\zeta_2(\eta^\top Y_0)\} = -c_2 a_0,\\
&\mathbb{E}\{\zeta_2(\eta^\top Y_0) Y_0 \eta^\top\} = -c_1 \Omega_1 \eta\eta^\top - c_2 a_1 \eta^\top,\\
&\mathbb{E}\{\zeta_2(\eta^\top Y_0) S_0\} = -c_2 A_2,
\end{aligned}$$

where

$$\begin{aligned}
a_0 &= c_2^{-1}\mathbb{E}\{[\zeta_1(\eta^\top Y_0)]^2\} = \mathbb{E}\{1/\Phi(\eta^\top W_2)\},\\
a_1 &= c_2^{-1}\mathbb{E}\{[\zeta_1(\eta^\top Y_0)]^2 Y_0\} = \mathbb{E}\{W_2/\Phi(\eta^\top W_2)\},\\
A_2 &= c_2^{-1}\mathbb{E}\{[\zeta_1(\eta^\top Y_0)]^2 Y_0 Y_0^\top\} = \mathbb{E}\{W_2\, W_2^\top/\Phi(\eta^\top W_2)\}
\end{aligned} \tag{10}$$

which must be evaluated numerically. Brute-force d-dimensional numerical integration in the evaluation of (10) can be very slow and unstable; an efficient computational scheme will be described in Section 2.4. We then obtain

$$\begin{aligned}
-\mathbb{E}\{\mathrm{d}^2\ell(\theta)\} &= -\frac{1}{2}\mathrm{tr}\{(\mathrm{d}\Omega^{-1})\mathrm{d}\Omega\} - 2c_1(\mathrm{d}\xi)^\top(\mathrm{d}\Omega^{-1})\Omega\eta + (\mathrm{d}\xi)^\top[\Omega^{-1} + c_2\, a_0\eta\eta^\top]\mathrm{d}\xi\\
&\quad + c_2(\mathrm{d}\eta)^\top A_2 \mathrm{d}\eta + 2\,(\mathrm{d}\eta)^\top[c_1(I_d - \Omega_1\eta\eta^\top) - c_2 a_1\eta^\top]\mathrm{d}\xi\\
&= \frac{1}{2}\mathrm{tr}\{\Omega^{-1}(\mathrm{d}\Omega)\Omega^{-1}\mathrm{d}\Omega\} + 2c_1\mathrm{tr}\{\eta(\mathrm{d}\xi)^\top\Omega^{-1}(\mathrm{d}\Omega)\} + (\mathrm{d}\xi)^\top[\Omega^{-1}\\
&\quad + c_2\, a_0\eta\eta^\top]\mathrm{d}\xi + c_2(\mathrm{d}\eta)^\top A_2\mathrm{d}\eta + 2\,(\mathrm{d}\eta)^\top[c_1(I_d + \Omega\eta\eta^\top)^{-1}\\
&\quad - c_2 a_1\eta^\top]\mathrm{d}\xi
\end{aligned}$$

where we have also used $\mathbb{E}\{Y_0\} = c_1\Omega\eta$, $\mathbb{E}\{S_0\} = \Omega$ and $\mathrm{d}\Omega^{-1} = -\Omega^{-1}(\mathrm{d}\Omega)\Omega^{-1}$.

To incorporate the symmetry of Ω in $\mathrm{d}\Omega$, we must introduce some additional notation. For a $d \times m$ matrix M denote by $\mathrm{vec}(M)$ the dm-dimensional vector formed by stacking the columns of M. If $d = m$ and $M = M^\top$, let $v(M)$ be the $[d(d+1)/2] \times 1$ vector obtained by stacking the lower triangle of M. There exists a $d^2 \times [d(d+1)/2]$ duplication matrix D such that $\mathrm{vec}(M) = Dv(M)$, and $v(M) = D^{+}\mathrm{vec}(M)$, where $D^{+} = (D^\top D)^{-1}D^\top$. Taking into account the symmetry of Ω, the actual parameter set is $\theta = (\xi, v(\Omega), \eta)$, for which we can write

$$\begin{aligned}
-\mathbb{E}\{\mathrm{d}^2\ell(\theta)\} &= \frac{1}{2}\mathrm{vec}(\mathrm{d}\Omega)^\top(\Omega^{-1}\otimes\Omega^{-1})\mathrm{vec}(\mathrm{d}\Omega) + 2c_1\mathrm{vec}(\mathrm{d}\Omega)^\top(\eta\otimes\Omega^{-1})\mathrm{d}\xi \\
&\quad + (\mathrm{d}\xi)^\top[\Omega^{-1} + c_2a_0\eta\eta^\top]\mathrm{d}\xi + c_2(\mathrm{d}\eta)^\top A_2\mathrm{d}\eta \\
&\quad + 2\,(\mathrm{d}\eta)^\top[c_1(I_d + \Omega\eta\eta^\top)^{-1} - c_2a_1\eta^\top]\mathrm{d}\xi \\
&= \frac{1}{2}(\mathrm{d}v(\Omega))^\top D^\top(\Omega^{-1}\otimes\Omega^{-1})D(\mathrm{d}v(\Omega)) + 2c_1(\mathrm{d}v(\Omega))^\top D^\top(\eta\otimes\Omega^{-1})\mathrm{d}\xi \\
&\quad + (\mathrm{d}\xi)^\top[\Omega^{-1} + c_2a_0\eta\eta^\top]\mathrm{d}\xi + c_2(\mathrm{d}\eta)^\top A_2\mathrm{d}\eta \\
&\quad +2\,(\mathrm{d}\eta)^\top\,[c_1(I_d + \Omega\eta\eta^\top)^{-1} - c_2a_1\eta^\top]\mathrm{d}\xi, \qquad (11)
\end{aligned}$$

where $\otimes$ denotes the Kronecker product and we have used

$$\begin{aligned}
\mathrm{tr}\{(\mathrm{d}\Omega)\Omega^{-1}(\mathrm{d}\Omega)\Omega^{-1}\} &= \mathrm{vec}(\mathrm{d}\Omega)^\top(\Omega^{-1}\otimes\Omega^{-1})\mathrm{vec}(\mathrm{d}\Omega) \\
&= (\mathrm{d}v(\Omega))^\top D^\top(\Omega^{-1}\otimes\Omega^{-1})D(\mathrm{d}v(\Omega)), \\
(\mathrm{d}\xi)^\top\Omega^{-1}(\mathrm{d}\Omega)\eta &= \mathrm{tr}\{\Omega^{-1}(\mathrm{d}\Omega)\eta(\mathrm{d}\xi)^\top\} \\
&= (\mathrm{d}\xi)^\top(\eta^\top\otimes\Omega^{-1})\mathrm{vec}(\mathrm{d}\Omega) \\
&= (\mathrm{d}\xi)^\top(\eta^\top\otimes\Omega^{-1})D\mathrm{d}v(\Omega)
\end{aligned}$$

which in turn depend on some results given by Magnus and Neudecker [14], specifically on p. 30–31, 173 and 189.

From (11), by following an argument similar to the one used by Magnus and Neudecker [14, p. 317–8] for the normal case, we obtain the expected information

$$\mathscr{I}(\theta) = \begin{pmatrix} \Omega^{-1} + c_2a_0\eta\eta^\top & c_1(\eta^\top\otimes\Omega^{-1})D & A_1 \\ c_1D^\top(\eta\otimes\Omega^{-1}) & \frac{1}{2}D^\top(\Omega^{-1}\otimes\Omega^{-1})D & 0 \\ A_1^\top & 0 & c_2\,A_2 \end{pmatrix}, \qquad (12)$$

where $A_1 = c_1(I_d + \eta\eta^\top\Omega)^{-1} - c_2\eta a_1^\top = c_1[I_d - (1+\eta^\top\Omega\eta)^{-1}\eta\eta^\top\Omega] - c_2\eta\, a_1^\top$.

If $e_i = (0, \dots, 0, 1, 0, \dots, 0)^\top$ where the non-zero component is in the i-th position and $E_{ii} = e_ie_i^\top$ $(i = 1, \dots, d)$, then

$$\eta = \omega^{-1}\alpha = \sum_{i=1}^{d}(e_i^\top\Omega e_i)^{-1/2}(e_i^\top\alpha)e_i$$

and the Jacobian matrix $D_{\mathrm{DP}}\theta$ of the transformation from θ to DP is given by

$$\begin{aligned}
D_{\mathrm{DP}}\theta &= \begin{pmatrix} D_\xi\xi & D_{v(\Omega)}\xi & D_\alpha\xi \\ D_\xi v(\Omega) & D_{v(\Omega)}v(\Omega) & D_\alpha v(\Omega) \\ D_\xi\eta & D_{v(\Omega)}\eta & D_\alpha\eta \end{pmatrix} \\
&= \begin{pmatrix} I_d & 0 & 0 \\ 0 & I_{d(d+1)/2} & 0 \\ 0 & -\frac{1}{2}\sum_{i=1}^{d}(e_i^\top\Omega e_i)^{-3/2}(\alpha^\top E_{ii}\otimes E_{ii})D & \omega^{-1} \end{pmatrix}, \qquad (13)
\end{aligned}$$

where the notation $D_\nu\psi$ denotes the matrix of partial derivatives $(\partial\psi/\partial\nu^\top)$. Hence the information matrix for DP can be written as

$$\mathscr{I}(\mathrm{DP}) = (D_{\mathrm{DP}}\theta)^\top \mathscr{I}(\theta)\, D_{\mathrm{DP}}\theta. \tag{14}$$

2.3. The information matrix for CP

The next stage of our construction is to convert the information matrix for θ into the one for CP. This mapping is actually accomplished in two steps, introducing an additional parameter set, which again has no intrinsic interest. Let $Y_0 = Y - \xi \sim \mathrm{SN}_d(0, \Omega, \alpha)$ and consider the parameter set $\psi = (\mu, \Sigma, \mu_0)$ where

$$\mu_0 = \mathbb{E}\{Y_0\} = c_1 \Omega\eta = \frac{b}{\sqrt{1+\eta^\top\Omega\eta}}\Omega\eta \tag{15}$$

such that, after some algebra, the inverse transformation from ψ to θ turns out to be

$$\xi = \mu - \mu_0, \qquad \Omega = \Sigma + \mu_0\mu_0^\top, \qquad \eta = q_1\Sigma^{-1}\mu_0,$$

where

$$q_1 = \frac{1}{c_1(1+\beta_0^2)}, \qquad c_1 = \sqrt{\frac{b^2-(1-b^2)\beta_0^2}{1+\beta_0^2}}, \qquad \beta_0^2 = \mu_0^\top\Sigma^{-1}\mu_0. \tag{16}$$

Computation of the Jacobian matrix of the reparametrization starts from the first-order partial differentials

$$\begin{aligned}
&\mathrm{d}_\mu\xi = \mathrm{d}\mu, \qquad \mathrm{d}_\Sigma\xi = 0, \qquad \mathrm{d}_{\mu_0}\xi = -I_d,\\
&\mathrm{d}_\mu\Omega = 0, \qquad \mathrm{d}_\Sigma\Omega = \mathrm{d}\Sigma, \qquad \mathrm{d}_{\mu_0}\Omega = (\mathrm{d}\mu_0)\mu_0^\top + \mu_0(\mathrm{d}\mu_0)^\top,\\
&\mathrm{d}_\mu\eta = 0, \qquad \mathrm{d}_\Sigma\eta = q_1'\Sigma^{-1}\mu_0\,[-\mu_0^\top\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}\mu_0] + q_1\,[-\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}\mu_0],\\
&\mathrm{d}_{\mu_0}\eta = q_1'\Sigma^{-1}\mu_0[2(\mathrm{d}\mu_0)^\top\Sigma^{-1}\mu_0] + q_1[\Sigma^{-1}\mathrm{d}\mu_0],
\end{aligned}$$

where

$$q_1' = \frac{\mathrm{d}q_1}{\mathrm{d}\beta_0^2} = -\frac{1}{2}q_1^2(2c_1 - q_1),$$

implying, from the identification theorem given of Magnus and Neudecker [14, p. 87], that

$$D_\psi\theta = \begin{pmatrix} D_\mu\xi & D_{v(\Sigma)}\xi & D_{\mu_0}\xi \\ D_\mu v(\Omega) & D_{v(\Sigma)}v(\Omega) & D_{\mu_0}v(\Omega) \\ D_\mu\eta & D_{v(\Sigma)}\eta & D_{\mu_0}\eta \end{pmatrix} = \begin{pmatrix} I_d & 0 & -I_d \\ 0 & I_{d(d+1)/2} & D_{23} \\ 0 & D_{32} & D_{33} \end{pmatrix},$$

where

$$\begin{aligned}
&D_{23} = D^+(I_d\otimes\mu_0 + \mu_0\otimes I_d),\\
&D_{32} = -\{\mu_0^\top\Sigma^{-1}\}\otimes\left\{q_1\Sigma^{-1} - q_1q_2\Sigma^{-1}\mu_0\mu_0^\top\Sigma^{-1}\right\}D,\\
&D_{33} = q_1\Sigma^{-1} - 2q_1q_2\Sigma^{-1}\mu_0\mu_0^\top\Sigma^{-1}
\end{aligned}$$

and $q_2 = \frac{1}{2}q_1(2c_1 - q_1)$.

The final mapping from ψ to CP works via the componentwise transformation

$$\gamma_{1,j} = \left(\frac{4-\pi}{2}\right)\left(\frac{\mu_{0j}}{\sigma_j}\right)^3, \quad (j = 1, \ldots, d),$$

for the individual terms of the vector γ_1, and the inverse transformation is

$$\mu_0 = \sigma\,\mathrm{diag}(b_0\,\gamma_1)^{1/3}\,1_d,$$

where $\mathrm{diag}(u)$ denotes the diagonal matrix whose diagonal terms are those of the vector u, and $b_0 = 2/(4-\pi)$. Since

$$\sigma = \sum_{i=1}^{d}(e_i^\top \Sigma e_i)^{1/2}\,E_{ii}, \qquad \mathrm{diag}(b_0\,\gamma_1)^{1/3} = \sum_{i=1}^{d}(b_0\,e_i^\top \gamma_1)^{1/3}\,E_{ii}$$

we have

$$\mathrm{d}_\Sigma\mu_0 = (\mathrm{d}_\Sigma\sigma)\mathrm{diag}(b_0\gamma_1)^{1/3}\,1_d = \frac{1}{2}\left\{\sum_{i=1}^{d}(e_i^\top \Sigma e_i)^{-1/2}\,E_{ii}(d\Sigma)\,E_{ii}\,\mathrm{diag}(b_0\gamma_1)^{1/3}1_d\right\},$$

$$\mathrm{d}_{\gamma_1}\mu_0 = \frac{b_0}{3}\sigma\left\{\sum_{i=1}^{d}(b_0 e_i^\top\gamma_1)^{-2/3}(b_0 e_i^\top \mathrm{d}\gamma_1)\,E_{ii}\right\} = \frac{b_0}{3}\sigma\,\mathrm{diag}(b_0\gamma_1)^{-2/3}\mathrm{d}\gamma_1.$$

Hence, the Jacobian matrix is

$$D_{\mathrm{CP}}\psi = \begin{pmatrix} I_d & 0 & 0 \\ 0 & I_{d(d+1)/2} & 0 \\ 0 & \tilde{D}_{32} & \tilde{D}_{33} \end{pmatrix},$$

where, on setting $\bar{\mu}_0 = \sigma^{-1}\mu_0/\beta_0$,

$$\begin{aligned}\tilde{D}_{32} &= D_{v(\Sigma)}\mu_0 = \frac{1}{2}\sum_{i=1}^{d}(e_i^\top \Sigma e_i)^{-1/2}[1_d^\top\,\mathrm{diag}(b_0\gamma_1)^{1/3}E_{ii}\otimes E_{ii}]D \\ &= \frac{\beta_0}{2}\sum_{i=1}^{d}(e_i^\top \Sigma e_i)^{-1/2}[\bar{\mu}_0^\top E_{ii}\otimes E_{ii}]\,D,\end{aligned}$$

$$\tilde{D}_{33} = D_{\gamma_1}\mu_0 = \frac{b_0}{3}\sigma\,\mathrm{diag}(b_0\gamma_1)^{-2/3} = \frac{b_0}{3\beta_0^2}\sigma\,\mathrm{diag}(\bar{\mu}_0)^{-2}.$$

Finally, the information matrix for CP is given by

$$\mathscr{I}(\mathrm{CP}) = D_{\mathrm{CP}}\psi^\top\,D_\psi\theta^\top\,\mathscr{I}(\theta)D_\psi\theta D_{\mathrm{CP}}\psi \tag{17}$$

which, after lengthy algebra, can be expressed in terms of the components of $\mathscr{I}(\theta)$. Specifically, if I_{rs} denotes the (r,s)-th block component of $\mathscr{I}(\theta)$, as partitioned in (12), and K_{rs} is the corresponding block of $\mathscr{I}(\mathrm{CP})$, then

$$\begin{aligned}K_{11} &= I_{11}, \\ K_{12} &= -I_{11}\tilde{D}_{32} + I_{12}(I_{d(d+1)/2} + D_{23}\tilde{D}_{32}) + I_{13}(D_{32} + D_{33}\tilde{D}_{32}) \\ K_{13} &= -I_{11}\tilde{D}_{33} + I_{12}D_{23}\tilde{D}_{33} + I_{13}D_{33}\tilde{D}_{33},\end{aligned}$$

$$\begin{aligned}
K_{22} &= -\tilde{D}_{32}^\top K_{12} + (I_{d(d+1)/2} + D_{23}\tilde{D}_{32})^\top[-I_{12}^\top \tilde{D}_{32} + I_{22}(I_{d(d+1)/2} + D_{23}\tilde{D}_{32})] \\
&\quad + (D_{32} + D_{33}\tilde{D}_{32})^\top[-I_{13}^\top \tilde{D}_{32} + I_{33}(D_{32} + D_{33}\tilde{D}_{32})], \\
K_{23} &= -\tilde{D}_{32}^\top K_{13} + (I_{d(d+1)/2} + D_{23}\tilde{D}_{32})^\top[-I_{12}^\top \tilde{D}_{33} + I_{22} D_{23}\tilde{D}_{33}] \\
&\quad + (D_{32} + D_{33}\tilde{D}_{32})^\top[-I_{13}^\top \tilde{D}_{33} + I_{33} D_{33}\tilde{D}_{33}], \\
K_{33} &= -\tilde{D}_{33}^\top K_{13} + (D_{23}\tilde{D}_{33})^\top[-I_{12}^\top \tilde{D}_{33} + I_{22} D_{23}\tilde{D}_{33}] \\
&\quad + (D_{33}\tilde{D}_{33})^\top[-I_{13}^\top \tilde{D}_{33} + I_{33} D_{33}\tilde{D}_{33}].
\end{aligned}$$

2.4. Limiting behaviour when asymmetries vanish

The aim of this section is to obtain the limiting form of $\mathscr{I}(\mathrm{CP})$ when $\gamma_1 \to 0$, which corresponds to the $\mathrm{N}_d(\mu, \Sigma)$ distribution; μ and Σ are regarded as fixed. The non-asymmetric case leads to a singular information matrix for DP, while we want to show that for CP the limiting information matrix is positive definite, extending the results known for $d = 1$.

We shall start working with the intermediate parametrization ψ, and examine its behaviour as $\mu_0 \to 0$, which corresponds to $\gamma_1 \to 0$. Since we are concerned with a limit in d components, this can be problematic to study, in general. Luckily, it turns out that all relevant quantities can be written as functions of β_0, the positive square root of β_0^2 defined in (16). Since $\beta_0 < \varepsilon$ for any positive ε implies that the vector μ_0 is within a suitable neighbourhood of the origin, we can equivalently consider the limit as $\beta_0 \to 0$. The fact that the expressions below involve also the term $\bar{\mu}_0$ does not invalidate the above argument, since $\bar{\mu}_0^\top \Sigma^{-1}\bar{\mu}_0 = 1$.

Consider the information matrix $\mathscr{I}(\psi) = D_\psi\theta^\top \mathscr{I}(\theta) D_\psi\theta$ and denote by J_{rs} its (rs)-th block when the matrix is partitioned similarly to (12). For the subsequent development, we make use of various properties, some of which are non-standard, of the Kronecker product and of the duplication matrix D, namely

$$\begin{aligned}
&(DD^+)^\top = DD^+, \qquad (DD^+)^2 = DD^+, \qquad DD^+D = D, \\
&DD^+(A \otimes A) = (A \otimes A)D, \\
&DD^+(w \otimes A)D = \frac{1}{2}(A \otimes w + w \otimes A), \\
&D^\top(A \otimes w - w \otimes A) = D^\top\{DD^+(A \otimes w - w \otimes A)\} = 0,
\end{aligned}$$

where A and w have dimension $d \times d$ and $d \times 1$, respectively. After some straightforward but extensive algebraic computations, one arrives at

$$\begin{aligned}
J_{11} &= \Sigma^{-1} + (c_2 q_1^2 a_0 - c_1 q_1)\beta_0^2 \sigma^{-1}\bar{\Lambda}_0\sigma^{-1}, \\
J_{12} &= q_1^2\beta_0^2(\bar{\mu}_0^\top \overline{\Sigma}^{-1}\sigma^{-1}) \\
&\quad \otimes \left\{\sigma^{-1}\left[\left(\frac{c_1^3 q_1\beta_0}{2b^2}\bar{\mu}_0 + c_2\bar{a}_1\right)^\top \overline{\Sigma}^{-1} - c_2 q_2\beta_0^2(\bar{a}_1^\top \overline{\Sigma}^{-1}\bar{\mu}_0)\bar{\Lambda}_0\right]\sigma^{-1}\right\} D, \\
J_{13} &= -q_1^2\beta_0\sigma^{-1}\left[c_2\overline{\Sigma}^{-1}\bar{\mu}_0\,(a_0\beta_0\bar{\mu}_0 + \bar{a}_1)^\top\,\overline{\Sigma}^{-1}\right. \\
&\quad \left. - \left(\frac{c_1^3 q_1\beta_0}{b^2} + 2c_2 q_2\bar{a}_1^\top \overline{\Sigma}^{-1}\bar{\mu}_0\right)\beta_0^2\bar{\Lambda}_0\right]\sigma^{-1},
\end{aligned}$$

$$
\begin{aligned}
J_{22} &= \frac{1}{2} D^\top (\Sigma + \beta_0^2 \sigma \bar\mu_0 \bar\mu_0^\top \sigma) \otimes (\Sigma + \beta_0^2 \sigma \bar\mu_0 \bar\mu_0^\top \sigma)^{-1} D \\
&\quad + c_2 q_1^2 \beta_0^2 D^\top \left\{ (\sigma^{-1} \bar\Lambda_0 \sigma^{-1}) \otimes \left[\sigma^{-1} (\overline{\Sigma}^{-1} - q_2 \beta_0^2 \bar\Lambda_0) \bar A_2 (\overline{\Sigma}^{-1} - q_2 \beta_0^2 \bar\Lambda_0) \sigma^{-1} \right] \right\} D, \\
J_{23} &= -q_1 \beta_0 D^\top \left(\sigma^{-1} \overline{\Sigma}^{-1} \bar\mu_0 \right) \otimes \left\{ \sigma^{-1} \left(\overline{\Sigma}^{-1} - q_2 \beta_0^2 \bar\Lambda_0 \right) \left[\left(c_2 q_1 \bar A_2 \overline{\Sigma}^{-1} - c_1 I_d \right) \right. \right. \\
&\qquad \left. \left. + c_2 q_1 \beta_0 \bar a_1 \bar\mu_0^\top \overline{\Sigma}^{-1} + q_1 \beta_0^2 \left(\frac{c_1^2}{b^2} \overline{\Sigma} - 2 c_2 q_2 \bar A_2 \right) \bar\Lambda_0 \right] \sigma^{-1} \right\}, \\
J_{33} &= q_1^2 \beta_0 \sigma^{-1} \\
&\quad \times \left[c_2 \overline{\Sigma}^{-1} \bar\mu_0 \left(a_0 \beta_0 \bar\mu_0 + \bar a_1 \right)^\top \overline{\Sigma}^{-1} - \left(\frac{c_1^3 q_1 \beta_0}{b^2} + 2 c_2 q_2 \bar a_1^\top \overline{\Sigma}^{-1} \bar\mu_0 \right) \beta_0^2 \bar\Lambda_0 \right] \sigma^{-1} \\
&\quad + q_1 \sigma^{-1} \left(\overline{\Sigma}^{-1} - 2 q_2 \beta_0^2 \bar\Lambda_0 \right) \\
&\quad \times \left[\left(c_2 q_1 \bar A_2 \overline{\Sigma}^{-1} - c_1 I_d \right) + c_2 q_1 \beta_0 \bar a_1 \bar\mu_0^\top \overline{\Sigma}^{-1} + q_1 \beta_0^2 \left(\frac{c_1^2}{b^2} \overline{\Sigma} - 2 c_2 q_2 \bar A_2 \right) \bar\Lambda_0 \right] \sigma^{-1},
\end{aligned}
$$

where

$$
\begin{aligned}
&\bar a_1 = \sigma^{-1} a_1, \qquad \bar A_2 = \sigma^{-1} A_2 \sigma^{-1}, \\
&\overline{\Sigma} = \sigma^{-1} \Sigma \sigma^{-1}, \qquad \bar\Lambda_0 = \overline{\Sigma}^{-1} \bar\mu_0 \bar\mu_0^\top \overline{\Sigma}^{-1}
\end{aligned}
\tag{18}
$$

such that $\bar\mu_0^\top \overline{\Sigma}^{-1} \bar\mu_0 = 1$. To avoid increasing notational complexity, we do not make explicit that c_1, c_2, q_1, q_2 are to be regarded as functions of β_0^2, as indicated by (16) and that q_2 is defined as a function of the other three quantities.

To obtain an expansion of (10), or equivalently of the scaled versions (18), in terms of β_0^2, consider

$$
\begin{pmatrix} \eta^\top W_2 \\ W_2 \end{pmatrix} \sim N_{1+d} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \dfrac{\tilde\eta^2}{1 + 2\tilde\eta^2} & \dfrac{1}{1 + 2\tilde\eta^2} \eta^\top \Omega \\ \dfrac{1}{1 + 2\tilde\eta^2} \Omega \eta & \Omega - \dfrac{2}{1 + 2\tilde\eta^2} \Omega \eta \eta^\top \Omega \end{pmatrix} \right),
$$

where $\tilde\eta^2 = \eta^\top \Omega \eta$, leading to

$$
(W_2 | \eta^\top W_2 = u) \sim N_d(\mu_c u, \Omega_c),
$$

where $\mu_c = \tilde\eta^{-2} \Omega \eta$, $\Omega_c = \Omega - \tilde\eta^{-2} \Omega \eta \eta^\top \Omega$. Hence, we can rewrite (10) as

$$
a_0 = \mathbb{E} \left\{ \frac{1}{\Phi(U)} \right\}, \quad a_1 = \mathbb{E} \left\{ \frac{U}{\Phi(U)} \mu_c \right\}, \quad A_2 = \mathbb{E} \left\{ \frac{1}{\Phi(U)} (U^2 \mu_c \mu_c^\top + \Omega_c) \right\}, \tag{19}
$$

where

$$
U = \eta^\top W_2 \sim N(0, \bar\alpha^2), \qquad \bar\alpha^2 = \frac{\tilde\eta^2}{1 + \tilde\eta^2} = \frac{\beta_0^2}{b^2 + (1 + b^2) \beta_0^2},
$$

which provides a far more convenient representation, for two reasons. One is that (19) requires only three one-dimensional numerical integrations instead of $(2 + 3d + d^2)/2$ integrations in d

dimensions, with dramatic improvement in the speed and accuracy of computation. The other use of the above argument is to form the basis of the argument leading to the following result, whose proof is given in the Appendix.

Lemma 2. *As $\beta_0 \to 0$, we have*

$$\begin{aligned} a_0 &= 2\left\{1+\beta_0^2-\frac{2(1-b^2)}{b^2}\beta_0^4\right\}+O(\beta_0^6),\\ \bar{a}_1 &= -2\left\{\beta_0+\frac{3b^2-2}{b^2}\beta_0^3\right\}\bar{\mu}_0+O(\beta_0^5)\\ \bar{A}_2 &= 2\left\{\overline{\Sigma}+\beta_0^2\left[\overline{\Sigma}+\frac{3b^2-2}{b^2}\bar{\mu}_0\bar{\mu}_0^\top\right]\right.\\ &\qquad \left.+\beta_0^4\left[\frac{2b^2-2}{b^2}\overline{\Sigma}+\frac{13b^4-12b^2+2}{b^4}\bar{\mu}_0\bar{\mu}_0^\top\right]\right\}+O(\beta_0^6). \end{aligned}$$

After substitution of these expansions into J_{rs} and some additional algebraic reduction, we obtain the following result for $\mathscr{I}(\psi)$ near $\beta_0 = 0$.

Lemma 3. *As $\beta_0 \to 0$, we have*

$$\begin{aligned} &\mathscr{I}(\psi)\\ &= \begin{pmatrix} \Sigma^{-1}+O(\beta_0^2) & O(\beta_0^3) & O(\beta_0^5)\\ O(\beta_0^3) & \frac{1}{2}D^\top\left(\Sigma^{-1}\otimes\Sigma^{-1}\right)D+O(\beta_0^2) & O(\beta_0^5)\\ O(\beta_0^5) & O(\beta_0^5) & \frac{(2b^2-1)^2}{2b^4}\beta_0^4\{\overline{\Sigma}^{-1}+2\,\bar{\Lambda}_0\}+O(\beta_0^6) \end{pmatrix}. \end{aligned}$$

By inserting these expansions in $\mathscr{I}(\mathrm{CP}) = K = D_{\mathrm{CP}}\psi^\top\mathscr{I}(\psi)D_{\mathrm{CP}}\psi$, we finally obtain that, when $\beta_0 \to 0$,

$$\mathscr{I}(\mathrm{CP}) = \begin{pmatrix} \Sigma^{-1}+O(\beta_0^2) & O(\beta_0^3) & O(\beta_0^3)\\ O(\beta_0^3) & \frac{1}{2}D^\top(\Sigma\otimes\Sigma)^{-1}D+O(\beta_0^2) & O(\beta_0^3)\\ O(\beta_0^3) & O(\beta_0^3) & \frac{1}{18}\{\overline{\Sigma}^{-1}+2\,\overline{\Sigma}^{-1}\gamma_0\gamma_0^\top\overline{\Sigma}^{-1}\}+O(\beta_0^2) \end{pmatrix}, \tag{20}$$

where $\gamma_0 = \lim_{\beta_0\to 0}\bar{\mu}_0$, and is such that $\gamma_0^\top\overline{\Sigma}^{-1}\gamma_0 = 1$. Clearly, as $\beta_0 \to 0$, (20) converges to a block diagonal matrix, whose first two blocks coincide with those of the corresponding normal case; see for instance Magnus and Neudecker [14, p. 317–8]. If $d = 1$, we have identically $\overline{\Sigma} = 1$, and hence $\gamma_0^2 = 1$, implying that $\lim_{\beta_0\to 0} K_{33} = 1/6$, in agreement with earlier results.

3. Linear regression and observed information

Consider the linear regression setting where y_i is sampled from $Y_i \sim \mathrm{SN}_d(\xi_i, \Omega, \alpha)$ where $\xi_i = \beta^\top x_i$ for a p-dimensional vector x_i of covariates and some $p \times d$ matrix β of regression parameters, for $i = 1, \ldots, n$ and with independence among the Y_i's. Denote by X the $n \times p$ matrix whose i-th row is $x_i^\top$, by Y the $n \times d$ matrix whose i-th row is $y_i^\top$, and let $Y_0 = Y - X\beta$,

$\bar{S}_0 = \frac{1}{n} Y_0^\top Y_0$. The log-likelihood function for $\theta = (\mathrm{vec}(\beta), v(\Omega), \eta)$ is

$$\ell(\theta) = \text{constant} - \frac{1}{2} n \log|\Omega| - \frac{1}{2} n \,\mathrm{tr}\left\{\Omega^{-1}\bar{S}_0\right\} + 1_n^\top \zeta_0(Y_0\eta), \tag{21}$$

where we adopt the convention that the notation $\zeta_m(u)$ for some vector u denotes the vector formed by applying the function ζ_m to each element of u. After some algebra, we obtain the differential

$$\begin{aligned} \mathrm{d}\ell(\theta) = {} & \frac{1}{2} n \,\mathrm{tr}\left\{(\mathrm{d}\Omega^{-1})(\Omega - \bar{S}_0)\right\} + \mathrm{tr}\left\{\Omega^{-1}(\mathrm{d}\beta)^\top X^\top Y_0\right\} - \eta^\top (\mathrm{d}\beta)^\top X^\top \zeta_1(Y_0\eta) \\ & + (\mathrm{d}\eta)^\top Y_0^\top \zeta_1(Y_0\eta) \end{aligned}$$

which leads to the likelihood equations

$$\begin{aligned} & (\Omega^{-1} \otimes X^\top)\mathrm{vec}(Y_0) - (\eta \otimes X^\top)\zeta_1(Y_0\eta) = 0, \\ & D^\top(\Omega^{-1} \otimes \Omega^{-1})\mathrm{vec}(\bar{S}_0 - \Omega) = 0, \\ & Y_0^\top \zeta_1(Y_0\eta) = 0. \end{aligned}$$

If the solution of the second equation, $\hat{\Omega}(\beta) = \bar{S}_0$, is substituted in (21), we obtain the profile log-likelihood function

$$\ell^*(\beta, \eta) = \text{constant} - \frac{1}{2} n \log|\bar{S}_0| - \frac{1}{2} nd + 1_n^\top \zeta_0(Y_0\eta)$$

similarly to Azzalini and Capitanio [6]. Further algebraic work gives the second differential

$$\begin{aligned} \mathrm{d}^2\ell(\theta) = {} & -\frac{1}{2} n \,\mathrm{tr}\left\{\Omega^{-1}(2\bar{S}_0 - \Omega)\Omega^{-1}(\mathrm{d}\Omega)\Omega^{-1}(\mathrm{d}\Omega)\right\} - 2\mathrm{tr}\left\{\Omega^{-1}(\mathrm{d}\Omega)\Omega^{-1}(\mathrm{d}\beta)^\top X^\top Y_0\right\} \\ & - \mathrm{tr}\left\{X^\top X(\mathrm{d}\beta)\Omega^{-1}(\mathrm{d}\beta)^\top + X^\top Z_2 X(\mathrm{d}\beta)\eta\eta^\top(\mathrm{d}\beta)^\top\right\} \\ & + 2\,\mathrm{tr}\left\{-X^\top \zeta_1(Y_0\eta)(\mathrm{d}\eta)^\top(\mathrm{d}\beta)^\top + \eta(\mathrm{d}\eta)^\top Y_0^\top Z_2 X(\mathrm{d}\beta)\right\} - (\mathrm{d}\eta)^\top Y_0^\top Z_2 Y_0(\mathrm{d}\eta), \end{aligned}$$

where $Z_2 = \mathrm{diag}(-\zeta_2(Y_0\eta)) > 0$, yielding the Hessian matrix which is the negative of

$$\mathscr{J}(\theta) = \begin{pmatrix} \Omega^{-1} \otimes (X^\top X) + (\eta\eta^\top) \otimes (X^\top Z_2 X) & [\Omega^{-1} \otimes (X^\top Y_0 \Omega^{-1})]D & I_d \otimes u - \eta \otimes U \\ D^\top[\Omega^{-1} \otimes (\Omega^{-1} Y_0^\top X)] & \frac{1}{2} n D^\top(\Omega^{-1} \otimes V)D & 0 \\ I_d \otimes u^\top - \eta^\top \otimes U^\top & 0 & Y_0^\top Z_2 Y_0 \end{pmatrix},$$

where $u = X^\top \zeta_1(Y_0\eta)$, $U = X^\top Z_2 Y_0$ and $V = \Omega^{-1}(2\bar{S}_0 - \Omega)\Omega^{-1}$. Note that at $\hat{\theta}$, the MLE of θ, we have $\hat{\Omega} = \bar{S}_0(\hat{\beta})$, and so $\hat{V} = \hat{\Omega}^{-1}$. Clearly $\mathscr{J}(\hat{\theta})$ gives the observed information matrix. Moreover, computation of $\mathbb{E}\{\mathscr{J}(\theta)\}$ gives the expected information

$$\mathscr{I}_n(\theta) = \begin{pmatrix} (\Omega^{-1} + c_2 a_0 \eta\eta^\top) \otimes (X^\top X) & c_1[\Omega^{-1} \otimes (X^\top 1_n \eta^\top)]D & A_1 \otimes (X^\top 1_n) \\ c_1 D^\top[\Omega^{-1} \otimes (\eta 1_n^\top X)] & \frac{1}{2} n D^\top(\Omega^{-1} \otimes \Omega^{-1})D & 0 \\ A_1^\top \otimes (1_n^\top X) & 0 & n\, c_2\, A_2 \end{pmatrix},$$

where we have used that $\mathbb{E}\{Y_0\} = c_1 1_n \eta^\top \Omega$. If $p = 1$ and $X = 1_n$, then $\beta = \xi^\top$, $X\beta = 1_n \xi^\top$, and the above matrix reduces to (12) multiplied by n.

Under the usual setting that the first column of X is 1_n, the transformation from the DP to the CP only changes the intercept term, as explained by Azzalini and Capitanio [6] for the case

$d = 1$; the same fact holds in the transformation from the θ to the CP parameters. In our setting, if $\beta_1^\top$ denotes the first row of β, the corresponding component of CP is $\beta_1^\top + c_1\eta^\top\Omega$, where the second term represents the common value of each row of $\mathbb{E}\{Y_0\}$, while the other components of β remain unchanged.

To compute the CP expected information matrix in the present context, we still use (17), except that $\mathscr{I}(\theta)$ is replaced by $\mathscr{I}_n(\theta)$ above and the Jacobian matrices now are

$$\begin{aligned} D_\psi\theta &= \begin{pmatrix} I_d & 0 & 0 & -I_d \\ 0 & I_{(p-1)d} & 0 & 0 \\ 0 & 0 & I_{d(d+1)/2} & D_{23} \\ 0 & 0 & D_{32} & D_{33} \end{pmatrix}, \\ D_{\mathrm{CP}}\psi &= \begin{pmatrix} I_d & 0 & 0 & 0 \\ 0 & I_{(p-1)d} & 0 & 0 \\ 0 & 0 & I_{d(d+1)/2} & 0 \\ 0 & 0 & \tilde{D}_{32} & \tilde{D}_{33} \end{pmatrix} \end{aligned} \tag{22}$$

while use of (17) with $\mathscr{J}(\hat{\theta})$ in place of $\mathscr{I}(\theta)$ gives the CP observed information matrix. Finally, the DP information matrices, expected and observed, can be obtained by using an expression of type (14), provided the middle term is replaced by $\mathscr{I}_n(\theta)$ or by $\mathscr{J}(\theta)$ and the first block row/column of (13) is modified following the pattern of $D_{\mathrm{CP}}\psi$ in (22).

4. Final remarks

The results provided allow us to adopt the CP for routine use with the multivariate SN distribution. This is not to say that the DP must be dismissed, since it has proved to be a very convenient parametrization for the development of much distribution theory, as documented in the references mentioned earlier and others quoted therein. Our view is that DP is more suitable for probabilistic work, while CP is more suitable for statistical work, for two reasons: (i) a more regular behaviour of the log-likelihood function and related quantities, (ii) simpler interpretation of the parameters.

The superiority of CP over DP in terms of interpretability was clear already for the case $d = 1$, and this superiority is even more apparent in the multivariate case. Consider specifically the skewness parameter α versus γ_1. The latter is a more familiar quantity than α for the statistician when $d = 1$, but this fact could be overcome with practice; however, a more radical problem exists in the multivariate case, since α cannot be interpreted componentwise to assess the skewness of the corresponding marginal distribution. This fact is illustrated by the following two sets of parameters, $(\Omega, \alpha^{(1)})$ and $(\Omega, \alpha^{(2)})$, where

$$\Omega = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 2 & 4 \\ 3 & 4 & 9 \end{pmatrix}, \qquad \alpha^{(1)} = \begin{pmatrix} 5 \\ -3 \\ 4 \end{pmatrix}, \qquad \alpha^{(2)} = \begin{pmatrix} 5 \\ -3 \\ -4 \end{pmatrix}$$

whose corresponding indices of marginal skewness, rounded to two decimal digits, are $\gamma_1^{(1)} = (0.85, 0.04, 0.16)^\top$ and $\gamma_1^{(2)} = (0.00, -0.21, -0.07)^\top$, respectively; clearly, consideration of an individual component of α does not provide information on the corresponding component of γ_1, in fact not even on its sign. In addition, μ is a far more familiar index of location compared to ξ, and so is Σ with respect to Ω as for dispersion.

Throughout the preceding development, a fundamental role is played by the quantity $\beta_0^2 = \mu_0^\top \Sigma^{-1}\mu_0 = \mu_z^\top \overline{\Sigma}^{-1}\mu_z$. Not only does this quantity appear explicitly in (20) as the one which regulates the limiting behaviour of the information matrix when all asymmetries vanish, but also it influences the non-limiting behaviour of the information matrix, since various other key quantities are actually functions of β_0; see for instance the definition of c_1 and q_1 in (16) and of q_2 which is in turn a function of the other two. Additional quantities regulated by β_0 are the integrals (19), via $\bar{\alpha}^2$. These remarks reinforce earlier ones on the important role of β_0 as a summary index of asymmetry, as indicated by Azzalini and Capitanio [6], both to regulate the Mardia indices of skewness and kurtosis and, via the one-to-one transformation α^* of β_0^2, to regulate the parameter of the canonical representation of their Proposition 4. Further evidence on the important role of α^*, and hence of β_0, is provided by Azzalini [4].

Pewsey [15] has shown that the problem of singularity of the information matrix at $\alpha = 0$ occurs with other distributions similar to (1) where the skewing factor is formed by using another symmetric distribution function G in place of Φ, at least in the univariate case. So far, these alternative forms have not been adopted a great deal in practical work, but this might change. In such a case, it would make sense to consider the CP approach in place of DP even with other forms of “skew-normal” distribution.

Because of its greater relevance and of its close connection with the SN distribution, a special mention is due for the so-called “extended skew-normal distribution”, studied by Azzalini [3] and Arnold et al. [2] in the univariate case, and by Arnold and Beaver [2] and Capitanio et al. [10] in the multivariate case. This distribution involves the same parameters of the ordinary skew-normal distribution plus an additional scalar term representing the mean value of an unobservable normal random variable, τ say. The corresponding asymptotic theory of the MLEs has not been developed, as far as we know, not even in the scalar case. It is quite natural to conjecture that the CP formulation discussed earlier, complemented with the extra parameter τ, could represent a plausible extension of the CP to the new setting. An exploration of this direction would however require substantial work, beyond the aims of the present paper.

A very different picture emerges if one considers a “skewed” form of another density, in place of the normal one. Special attention has been paid to the skew-t distribution and the skew-exponential power distribution, whose statistical aspects have been studied by Azzalini and Capitanio [7], DiCiccio and Monti [12] and Azzalini and Genton [9]. DiCiccio and Monti [12] obtain the information matrix for the skew-exponential power distribution, which is free from the singularity problem at $\alpha = 0$, except of course when the tail parameter corresponds to the normal distribution. Although strictly speaking a similar formal proof has not been achieved by Azzalini and Genton [9] for the skew-t distribution, they provide convincing evidence that the singularity problem does not arise also in this case. These facts indicate that there is no need to consider the CP for distributions of this type to avoid the singularity problem, although the issue of interpretation of parameters persists.

The arguments put forward by Azzalini and Genton [9] in support of the widespread use of the skew-t distribution motivate the introduction of a suitable form of centred parametrization, whose components should be aimed towards easier interpretability than that of its direct parameters. In this setting, the non-existence of low-order moments for low degrees of freedom requires however an entirely different treatment, which will be developed elsewhere.

Acknowledgments

We would like to thank Antonella Capitanio, Marc Genton and two referees for their comments which have led to improvement of the presentation. The work of the first author was supported by grant Fondecyt No. 1040865. The work of the second author was supported by MIUR, Italy, under grant PRIN 2006132978.

Appendix

A.1. Proof of Lemma 1 and some corollaries

Since

$$\begin{aligned}\mathbb{E}\{[\zeta_1(\eta^\top Y_0)]^k g(Y_0)\} &= \int_{\mathbb{R}^d} [\zeta_1(\eta^\top y_0)]^k g(y_0) 2\phi_d(y_0;\Omega)\,\Phi(\eta^\top y_0)\,\mathrm{d}y_0\\ &= 2\int_{\mathbb{R}^d} [1/\Phi(\eta^\top y_0)]^{k-1} g(y_0)[\phi(\eta^\top y_0)]^k \phi_d(y_0;\Omega)\,\mathrm{d}y_0,\end{aligned}$$

the proof follows by noting that

$$[\phi(\eta^\top y_0)]^k\phi_d(y_0;\Omega) = (1/2\pi)^{k/2}(|\Omega_k|/|\Omega|)^{1/2}\phi_d(y_0;\Omega_k),$$

where $\Omega_k = (\Omega^{-1} + k\eta\eta^\top)^{-1}$ implying that $|\Omega_k|/|\Omega| = 1/(1 + k\eta^\top\Omega\eta)$.

As by-product of Lemma 1, we have the following results, where a_0, a_1, A_2 are as in (10).

1. Taking $k = 1$ and $g(y_0) = 1$, we obtain $\mathbb{E}\{\zeta_1(\eta^\top Y_0)\} = c_1 = b/\sqrt{1 + \eta^\top\Omega\eta}$.
2. Taking $k = 1$ and $g(\cdot)$ to be an odd function, i.e. $g(-z) = -g(z)$, we have

$$\mathbb{E}\{\zeta_1(\eta^\top Y_0)g(Y_0)\} = c_1\mathbb{E}\{g(W_1)\} = 0,$$

implying, in particular, that $\mathbb{E}\{\zeta_1(\eta^\top Y_0)Y_0\} = c_1\mathbb{E}\{W_1\} = 0$.

3. Using that $\zeta_2(u) = -u\,\zeta_1(u) - [\zeta_1(u)]^2$, and taking $k = 1$ and $k = 2$ for the first and second terms on the right-hand side expression below, respectively, we have

$$\begin{aligned}\mathbb{E}\{\zeta_2(\eta^\top Y_0)\} &= -\mathbb{E}\{\zeta_1(\eta^\top Y_0)\,\eta^\top Y_0\} - \mathbb{E}\{[\zeta_1(\eta^\top Y_0)]^2\}\\ &= -c_1\mathbb{E}\{\eta^\top W_1\} - c_2\mathbb{E}\{1/\Phi(\eta^\top W_2)\}\\ &= -c_2\mathbb{E}\{1/\Phi(\eta^\top W_2)\}\\ &= -c_2 a_0\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\{\zeta_2(\eta^\top Y_0)g(Y_0)\} &= -\mathbb{E}\{\zeta_1(\eta^\top Y_0)(\eta^\top Y_0)g(Y_0)\} - \mathbb{E}\{[\zeta_1(\eta^\top Y_0)]^2 g(Y_0)\}\\ &= -c_1\mathbb{E}\{(\eta^\top W_1)g(W_1)\} - c_2\mathbb{E}\{g(W_2)/\Phi(\eta^\top W_2)\}.\end{aligned}$$

For specific choices of $g(\cdot)$, this result leads to

$$\begin{aligned}\mathbb{E}\{\zeta_2(\eta^\top Y_0)Y_0\} &= -c_1\mathbb{E}\{W_1W_1^\top\eta\} - c_2\mathbb{E}\{[1/\Phi(\eta^\top W_2)]W_2\}\\ &= -c_1\Omega_1\eta - c_2 a_1\\ &= -c_1(1 + \eta^\top\Omega\eta)^{-1}\Omega\eta - c_2 a_1,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\{\zeta_2(\eta^\top Y_0)Y_0Y_0^\top\} &= -c_1\mathbb{E}\{(\eta^\top W_1)W_1W_1^\top\} - c_2\mathbb{E}\{W_2W_2^\top/\Phi(\eta^\top W_2)\}\\ &= -c_2\mathbb{E}\{W_2W_2^\top/\Phi(\eta^\top W_2)\}\\ &= -c_2 A_2.\end{aligned}$$

A.2. Proof of Lemma 2

The conditional argument leading to (19) can be adapted to obtain a Taylor series expansion for (10), or equivalently for (18), since these appear in the terms J_{rs}. Let

$$g(x) = \frac{1}{\Phi(x)} = 2\,\exp\{-\zeta_0(x)\} = \sum_{k=0}^{\infty} \frac{g^{(k)}(0)}{k!} x^k,$$

whose derivatives can be obtained recursively from

$$g^{(k)}(x) = -\sum_{j=0}^{k-1} \binom{k-1}{j} g^{(j)}(x)\zeta_{k-j}(x).$$

From results in [6] (full version of the paper), we have $\zeta_m(0) = \kappa_m$ $(m = 1, 2, \ldots)$, where the κ_m's are the cumulants of a random variable $V \sim \sqrt{\chi_1^2}$. Therefore we obtain

$$g^{(k)}(0) = -\sum_{j=0}^{k-1} \binom{k-1}{j} g^{(j)}(0)\kappa_{k-j}, \quad (k = 1, 2, \ldots).$$

Defining $\bar{W}_2 = \sigma^{-1} W_2$ and $\eta^\top W_2 = q_1\beta_0 \left(\bar{\mu}_0^\top \overline{\Sigma}^{-1} \bar{W}_2\right)$, we can write

$$\bar{A}_2 = \mathbb{E}\{g(\eta^\top W_2)\bar{W}_2\bar{W}_2^\top\} = \sum_{k=0}^{\infty} \frac{g^{(k)}(0)}{k!} \mathbb{E}\{(\eta^\top W_2)^k \bar{W}_2\bar{W}_2^\top\},$$

and the expectations can be computed using a conditioning argument essentially as the one leading to (19), leading to

$$\begin{aligned}
\mathbb{E}\{(\eta^\top W_2)^k \bar{W}_2\bar{W}_2^\top\} &= (q_1\beta_0)^k\, \mathbb{E}\{(\bar{\mu}_0^\top \overline{\Sigma}^{-1} \bar{W}_2)^k \mathbb{E}\{\bar{W}_2\bar{W}_2^\top | \bar{\mu}_0^\top \overline{\Sigma}^{-1} \bar{W}_2\}\} \\
&= (q_1\beta_0)^k \left[\mathbb{E}\{(\bar{\mu}_0^\top \overline{\Sigma}^{-1} \bar{W}_2)^k\}(\overline{\Sigma} - \bar{\mu}_0\bar{\mu}_0^\top) + \mathbb{E}\{(\bar{\mu}_0^\top \overline{\Sigma}^{-1} \bar{W}_2)^{k+2}\}\bar{\mu}_0\bar{\mu}_0^\top\right] \\
&= \begin{cases} \bar{\alpha}^k \left\{\upsilon_k \overline{\Sigma} - (\upsilon_k - \upsilon_{k+2}\omega_0^2)\bar{\mu}_0\bar{\mu}_0^\top\right\}, & \text{for } k \text{ even,} \\ 0 & \text{for } k \text{ odd,} \end{cases}
\end{aligned}$$

where $\omega_0 = 2\,c_2\sqrt{1 + \beta_0^2}/b^2$ and $\upsilon_{2r} = \mathbb{E}\{X^{2r}\} = (2r)!/(2^r\, r!)$ if $X \sim N(0, 1)$. By arguing in a similar way for a_0 and $\bar{a}_1$, we obtain

$$a_0 = \sum_{m=0}^{\infty} \frac{g^{(2m)}(0)\bar{\alpha}^{2m}}{2^m m!},$$

$$\bar{a}_1 = \left\{\sum_{m=1}^{\infty} \frac{g^{(2m-1)}(0)\bar{\alpha}^{2m-1}}{2^{m-1}(m-1)!}\right\} \omega_0\bar{\mu}_0,$$

$$\bar{A}_2 = \sum_{m=0}^{\infty} \frac{g^{(2m)}(0)\bar{\alpha}^{2m}}{(2m)!} \left\{\frac{(2m)!}{2^m m!}\overline{\Sigma} - \left(\frac{(2m)!}{2^m m!} - \frac{[2(m+1)]!}{2^{m+1}(m+1)!}\omega_0^2\right)\bar{\mu}_0\bar{\mu}_0^\top\right\},$$

whose leading terms provide the expansion near $\bar{\alpha} = 0$, namely

$$a_0 = 2 + 2b^2\bar{\alpha}^2 + 2b^2(3b^2 - 1)\bar{\alpha}^4 + O(\bar{\alpha}^6),$$

$$\begin{aligned}
\bar{a}_1 &= -\left\{2b\bar{\alpha} + b(6b^2-1)\bar{\alpha}^3\right\}\omega_0\bar{\mu}_0 + O(\bar{\alpha}^5),\\
\bar{A}_2 &= 2\left\{\overline{\Sigma} - (1-\omega_0^2)\bar{\mu}_0\bar{\mu}_0^\top\right\} + 2b^2\bar{\alpha}^2\left\{\overline{\Sigma} - (1-3\omega_0^2)\bar{\mu}_0\bar{\mu}_0^\top\right\},\\
&\quad + 2b^2(3b^2-1)\bar{\alpha}^4\left\{\overline{\Sigma} - (1-5\omega_0^2)\bar{\mu}_0\bar{\mu}_0^\top\right\} + O(\bar{\alpha}^6).
\end{aligned}$$

After conversion of this expansion based on $\bar{\alpha}$ into one based on β_0, via use of appropriate derivatives, and consideration of the expansions

$$\begin{aligned}
\bar{\alpha} &= \frac{\beta_0}{b} - \frac{1-b^2}{2b^3}\beta_0^3 + O(\beta_0^5),\\
\omega_0 &= 1 + \frac{b^2-2}{2b^2}\beta_0^2 + \frac{3(4+4b^2-b^4)}{b^4}\beta_0^4 + O(\beta_0^6),
\end{aligned}$$

we reach the conclusion.

References

[1] R.B. Arellano-Valle, A. Azzalini, On the unification of families of skew-normal distributions, Scand. J. Statist. 33 (2006) 561–574.

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

[3] A. Azzalini, A class of distributions which includes the normal ones, Scand. J. Statist. 12 (1985) 171–178.

[4] A. Azzalini, A note on regions of given probability of the skew-normal distribution, Metron LIX (2001) 27–34.

[5] A. Azzalini, The skew-normal distribution and related multivariate families (with discussion), Scand. J. Statist. 32 (2005) 159–188 (CR: 189–200).

[6] A. Azzalini, A. Capitanio, Statistical applications of the multivariate skew-normal distributions, J. R. Statist. Soc. B 61 (1999) 579–602.

[7] A. Azzalini, A. Capitanio, Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t* distribution, J. R. Statist. Soc. B 65 (2003) 367–389.

[8] A. Azzalini, A. Dalla Valle, The multivariate skew-normal distribution, Biometrika 83 (1996) 715–726.

[9] A. Azzalini, M.G. Genton, Robust likelihood methods based on the skew-*t* and related distributions, Internat. Statist. Rev. 76 (2008) 106–129.

[10] A. Capitanio, A. Azzalini, E. Stanghellini, Graphical models for skew-normal variates, Scand. J. Statist. 30 (2003) 129–144.

[11] M. Chiogna, A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution, Statist. Methods Appl. 14 (2005) 331–341.

[12] T. DiCiccio, A.C. Monti, Inferential aspects of the skew exponential power distribution, J. Amer. Statist. Assoc. 99 (2004) 439–450.

[13] M.G. Genton (Ed.), Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality, Chapman & Hall/CRC, Boca Raton, 2004.

[14] J.R. Magnus, H. Neudecker, Matrix Differential Calculus, J. Wiley & Sons, New York, 1988.

[15] A. Pewsey, Some observations on a simple means of generating skew distributions, in: N. Balakrishnan, E. Castillo, J.M. Sarabia (Eds.), Advances in Distribution Theory, Order Statistics, and Inference, in: Statistics for Industry and Technology, Birkhäuser, Boston, MA, 2006, pp. 75–84.

[16] A. Rotnitzky, D.R. Cox, M. Bottai, J. Robins, Likelihood-based inference with singular information matrix, Bernoulli 6 (2000) 243–284.