The skew-normal distribution and related multivariate families

Adelchi Azzalini
Dipartimento di Scienze Statistiche
Università di Padova, Italia
azzalini@stat.unipd.it

2nd November 2004
amended 1st March 2005

This paper has appeared in
Scandinavian Journal of Statistics
Volume 32, pages 159–188, June 2005
doi:10.1111/j.1467-9469.2005.00426.x

Abstract

This paper provides an introductory overview of a portion of distribution theory which is currently under intense development. The starting point of this topic has been the so-called skew-normal distribution, but the connected area is becoming increasingly broad, and its branches include now many extensions, such as the skew-elliptical families, and some forms of semi-parametric formulations, extending the relevance of the field much beyond the original theme of 'skewness'. The final part of the paper illustrates connections with various areas of application, including selective sampling, models for compositional data, robust methods, some problems in econometrics, non-linear time series, especially in connection with financial data, and more.

Some key-words: skew-normal distribution, skew-elliptical distribution, skew-t distribution, heavy tails, selective sampling, stochastic frontier models, flexible parametric family, hidden truncation model, graphical models.

1 Background and motivation

The development of parametric families and the study of their properties has ever been a persistent theme of the statistical literature, although of course not constantly with the same intensity. In the last few years, there has been another spurt of activity in this area, with several results appearing in rapid succession, and at increasing rate of generation. A substantial part of this recent literature is broadly related to the skew-normal (SN) distribution, which represents a superset of the normal family. This paper intends to provide an introductory overview of the literature concerned with the SN family of distributions and with other families for continuous variables connected to the SN class in the mechanism of genesis. The discussion refers to both the univariate and the multivariate context, and it covers the main probabilistic properties, issues in statistical inference, and some of the applied work based on these distributions.

Given all work which has been devoted in the past to the development of so many parametric families of distributions, and the variety of already available families, it is natural to ask why to put some additional effort in this direction, and what radically new and appealing can be found in the realm of families of distributions. The following points attempt to answer these questions; a few of them anticipate some aspects which will be clarified in the subsequent development.

- ◇ The normal family is the limiting or the boundary case for very many parametric families of probability densities, but it seldom is an interior point of the family. On the contrary, in practical statistical work, the normal family is quite naturally perceived as the 'central' form of a range of densities.

- ◇ It is desirable to have available families of distributions which retain at least partly the mathematical tractability and some nice formal properties of the standard parametric classes, such as the normal family.

- ◇ While in the univariate case one has available a very large set of alternatives for data modelling, the options are far less numerous in the multivariate case. The choice is even more restricted if one inserts the requirement of the previous remark.

- ◇ A parametric family should possibly have associated at least one reasonably simple mechanism of genesis of variates. These stochastic representation provide support for use of the family of distributions in practical data modelling, and they can be used for generation of pseudo-random numbers. In addition, the availability of some forms of stochastic representations can often allow simple derivation of formal properties of the distribution.

- ◇ Ideally, one would like to have available families where a few parameters regulate distributions with high flexibility of shape and of their main characteristics: skewness, kurtosis and, in the multivariate case, dependence structure.

- ◇ A step beyond the standard concept of parametric class is achieved if the degree of flexibility of the family of distributions is selectable freely, by suitably increasing the number of parameters. Such a construction builds a bridge between the parametric and the nonparametric context, under appropriate conditions on the achievable degree of approximation to an arbitrary target distribution.

It is clearly unfeasible to discuss here in any detail why the various existing proposals do not fulfill the above desiderata in some way or another. Only to illustrate the point, consider the question of constructing a bivariate distribution with normal marginals, and assume that we tackle the problem using the standard tool represented by the Farlie–Gumbel–Morgenstern formula, which in the bivariate case takes the form

$$F(x_1, x_2) = F_1(x_1)\, F_2(x_2)\, [1 + \alpha\{1 - F_1(x_1)\}\, \{1 - F_2(x_2)\}]$$

where F_1 and F_2 denote the marginal distributions and α is a parameter varying in $(-1, 1)$. If we set $F_1(x) = F_2(x) = \Phi(x)$, where $\Phi(\cdot)$ denotes the $N(0, 1)$ distribution function, then we get an expression which extends the notion of normal distribution from the univariate to the bivariate case, but it does so in a purely formal manner, not linked to the intrinsic properties of the components F_1 and F_2. A specific unsatisfactory aspect is that, when $F_1(x) = F_2(x) = \Phi(x)$, the above expression does not produce the bivariate normal distribution, except that $\alpha = 0$ corresponds to an extremely special case of it. In addition, the range of achievable correlations is limited to the interval $(-\frac{1}{3}, \frac{1}{3})$ only.

Of course, we are not implying that the approach to be described in the rest of the paper meets, simultaneously and to the full extent, all desiderata described above. It does however makes significant progress in a number of ways, and it presents a number of interesting and novel features.

An aspect to be stressed right away is that this stream of literature is not concerned with skewness only. The addition of an extra parameter to allow for skewness in the normal distribution was the starting point, but current advances of the literature allow much more than this simple operation. Given any symmetric density function, possibly multivariate, it is now possible to modify it up to an extreme extent, and still retain some features of the original symmetric density. Therefore, in the title of various recent papers, the key-prefix 'skew' appears more as identifier of the approach, rather than as a descriptor of the technical content.

2 The univariate skew-normal distribution

2.1 A useful lemma

The following lemma, and its extension to the multivariate case, are central to our development. Since its proof is remarkably simple, yet instructive, it is worth presenting it.

Lemma 1 *If f_0 is a one-dimensional probability density function symmetric about 0, and G is a one-dimensional distribution function such that G' exists and is a density symmetric about 0, then*

$$f(z) = 2\, f_0(z)\, G\{w(z)\}\,, \qquad (-\infty < z < \infty), \tag{1}$$

is a density function for any odd function $w(\cdot)$.

Proof. If $Y \sim f_0$ and $X \sim G'$ are independent random variables, then

$$\tfrac{1}{2} \;=\; \mathbb{P}\{X - w(Y) \le 0\} = \mathbb{E}_Y\{\mathbb{P}\{X - w(Y) \le 0|Y\}\} = \int_{\mathbb{R}} G\{w(z)\}\, f_0(z)\, \mathrm{d}z$$

on noticing that $w(Y)$ and $X - w(Y)$ also have symmetric distribution about 0. QED

This lemma allows us to manipulate a symmetric 'basis' density f_0 through a 'perturbation' function $G\{w(x)\}$ to get a new legitimate density f. Since there is much freedom for the choice of the ingredients G and w, then the set of distributions which can be obtained starting from a given 'basis' f_0 is vast. The set of 'perturbed' densities always include the 'basis' density, since $w(x) \equiv 0$ gives $f_0 = f$. Some simple but interesting results connected to Lemma 1 are as follows.

Stochastic representation If $X \sim G'$ and $Y \sim f_0$ are independent variables, then

$$Z = \begin{cases} Y & \text{if } X < w(Y), \\ -Y & \text{otherwise}, \end{cases} \tag{2}$$

has density function (1). This expression provides a simple mean for random number generation.

Perturbation invariance If $Y \sim f_o$ and $Z \sim f$, then $|Y| \stackrel{d}{=} |Z|$, where the notation $\stackrel{d}{=}$ denotes equality in distribution. Among other properties, the result implies that all even moments of Y and Z are the same.

Lemma 1 appears in Azzalini (1985), in the special case $w(y) = \alpha y$, where α is a constant; some implications are given in Azzalini (1986). The more general version reported above has not appeared until much later, in the multivariate setting to be discussed in Section 3.1.

2.2 Definition and some properties

On using the Lemma 1 with $f_0 = \phi$ and $G = \Phi$, the density function and the distribution function of a $N(0,1)$ variate, respectively, and $w(x) = \alpha\, x$ where $\alpha \in \mathbb{R}$, we get the density

$$\phi(z; \alpha) = 2\, \phi(z)\, \Phi(\alpha\, z), \qquad (-\infty < z < \infty), \tag{3}$$

which is called skew-normal distribution with shape parameter α, denoted by $\mathrm{SN}(\alpha)$. If $Z \sim \mathrm{SN}(\alpha)$ and $Y = \xi + \omega\, Z$, where $\xi \in \mathbb{R}$, $\omega \in \mathbb{R}^+$, then we shall write $Y \sim \mathrm{SN}(\xi, \omega^2, \alpha)$.

The following properties of the density (3) hold. For most of them the proof is immediate; for property (d), recall the perturbation invariance property mentioned in the previous subsection.

(a) If $\alpha = 0$, we obtain the $N(0,1)$ density.

(b) If $Z \sim SN(\alpha)$, then $-Z \sim SN(-\alpha)$.

(c) As $\alpha \to \infty$, (3) converges pointwise to the half-normal density, namely $2\phi(z)$ for $z \geq 0$.

(d) If $Z \sim SN(\alpha)$, then $Z^2 \sim \chi^2_1$.

(e) For fixed α, density (3) is strongly unimodal, i. e. $\log f(z; \alpha)$ is a concave function of z.

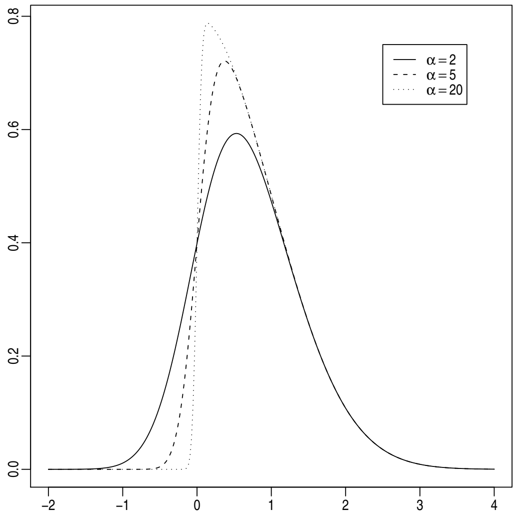

Figure 1: *Density function* $\mathrm{SN}(\alpha)$ *for a few values of* α

(f) The corresponding distribution function is given by

$$\Phi(z;\alpha) = \Phi(z) - 2\,T(z,\alpha)$$

where $T(z,\alpha)$ is the function studied by Owen (1956), and it satisfies the relationship

$$\Phi(z;-\alpha) = 1 - \Phi(-z;\alpha).$$

(g) If $U \sim N(0,1)$ is independent of $Z \sim \mathrm{SN}(\alpha)$, then

$$\frac{a\,U + b\,Z}{\sqrt{a^2+b^2}} \sim \mathrm{SN}\left(\frac{b\,\alpha}{\sqrt{a^2(1+\alpha^2)+b^2}}\right) \tag{4}$$

for any $a, b \in \mathbb{R}$.

The shape of density (3) is shown in Figure 1 for a few values of the parameter. Only positive values of α are considered, since for negative values the density is mirrored on the opposite side of the vertical axis, by property (b).

The moment generating function of $\mathrm{SN}(\xi,\omega^2,\alpha)$ is given by

$$M(t) = \mathbb{E}\{e^{tY}\} = 2\exp(\xi t + \omega^2 t^2/2)\,\Phi(\delta\omega t) \tag{5}$$

where $\delta = \alpha/\sqrt{1+\alpha^2} \in (-1,1)$. The result is immediate taking into account the following statement, given for instance by Ellison (1964) and Zacks (1981, pp. 53–54).

Lemma 2 *If* $U \sim N(0,1)$ *and* $a, b \in \mathbb{R}$, *then*

$$\mathbb{E}\{\Phi(a+b\,U)\} = \Phi\{a/\sqrt{1+b^2}\}. \tag{6}$$

From (5), if follows that

$$\mathbb{E}\{Y\} = \xi + \omega\,\mu_z,$$

$$
\begin{aligned}
\mathrm{var}\{Y\} &= \omega^2(1-\mu_z^2),\\
\gamma_1 &= \frac{4-\pi}{2}\,\frac{\mu_z^3}{(1-\mu_z^2)^{3/2}},\\
\gamma_2 &= 2(\pi-3)\,\frac{\mu_z^4}{(1-\mu_z^2)^2},
\end{aligned}
$$

where $\mu_z = \sqrt{2/\pi}\,\delta$, and γ_1 and γ_2 denote the standardised third and fourth order cumulants, respectively. The range of γ_1 is approximately $(-0.9953, 0.9953)$. It can also be shown that

$$
\mathbb{E}\{Z^r\} = \begin{cases} 1\times 3\times\cdots\times(r-1) & \text{if } r \text{ is even,}\\ \dfrac{\sqrt{2}\,(2k+1)!\,\alpha}{\sqrt{\pi}(1+\alpha^2)^{k+1/2}\,2^k} \displaystyle\sum_{m=0}^{k} \frac{m!(2\alpha)^{2m}}{(2m+1)!\,(k-m)!} & \text{if } r = 2k+1 \text{ and } k = 0,1,\ldots. \end{cases}
$$

Another use of Lemma 2 is to show that

$$
\phi(z)\,\Phi(\tau\,\sqrt{1+\alpha^2}+\alpha z)/\Phi(\tau) \tag{7}
$$

is also a proper density function for each choice of $\tau \in \mathbb{R}$. We shall refer to (7) as the extended skew-normal distribution, since it reduces to (3) when $\tau = 0$. This extended version shares a few of the above-listed formal properties of (3), possible after suitable modification, but not for instance the chi-square property (d).

The above presentation of the SN distribution is based largely on work of Azzalini (1985), with additional results of Henze (1986), Azzalini (1986) and Chiogna (1998). The extended version (7) has been considered briefly in Azzalini (1985) and more extensively in Arnold et al. (1993).

2.3 Early appearances and additional stochastic representations

In addition to (2), there are various other stochastic constructions leading to a variable with distribution of type (3) or (7). These other representations have appeared in earlier papers considering manipulations of normal random variates of various forms.

Conditional inspection and selective sampling Motivated by a practical problem in educational testing, Birnbaum (1950) has considered a problem whose essential aspects are as follows. Denote by U_1 the score obtained by a given subject in an attitudinal or educational test, where possibly U_1 is obtained as a linear combination of several such tests, and denote by U_0 the score obtained by the same subject in the admission examination. Assume that, after suitable scaling, (U_0, U_1) is distributed as a bivariate normal random variable with unit marginals and correlation ρ. Since individuals are examined in the subsequent tests conditionally on the fact that the admission score exceeds a certain threshold τ', the distribution of interest is the one of $Z = (U_1|U_0 > \tau')$, and this turns out to be of type (7) with $\alpha = \rho/\sqrt{1-\rho^2}$ and $\tau = -\tau'$. There is no loss of generality in assuming that the marginal distributions of U_0 and U_1 have the same location parameters since any difference can be absorbed in τ. Departures from zero location and unit scale parameter are handled by the transformation $Y = \xi + \omega Z$.

This scheme is in turn connected to the question of biased and selective sampling; see Section 6.

Selecting the largest/smallest value If (U_0, U_1) is as in the previous paragraph, consider the distribution of $\max(U_0, U_1)$ and of $\min(U_0, U_1)$. Roberts (1966) has examined this problem in the context of twin studies, when U_0 and U_1 represent suitably standardised measurements taken on a pair of twins. Given the type of individuals being measured, it makes sense to assume equal distribution of the two components. The distribution of $\max(U_0, U_1)$ is $\mathrm{SN}(\sqrt{(1-\rho)/(1+\rho)})$; for $\min(U_0, U_1)$ the sign of the shape parameter is reversed. Roberts (1966) also obtained the chi-square property (d). Recently, similar results have been re-obtained independently by Loperfido (2002).

Convolution of normal and truncated-normal Weinstein (1964) initiated a discussion in *Technometrics* about the cumulative distribution function of the sum of two independent normal variables, V_0 and V_1, say, when V_0 is truncated by constraining it to exceed a certain threshold. The ensuing discussion, summarised by Nelson (1964), lead to an expression for computing the required probability, which is in essence the distribution function of (7).

Although formulated quite differently, a closely related construction has been considered by O'Hagan & Leonard (1976), working in a Bayesian context. Here θ denotes the mean value of a Normal population for which prior considerations suggest that $\theta \geq 0$ but we are not completely convinced about this. This uncertainty is handled by a two-stage construction of the prior distribution for θ, assuming that $\theta|\mu \sim N(\mu, \sigma^2)$ and that μ has distribution of type $N(\mu_0, \sigma_0^2)$ truncated below 0. The resulting distribution for θ corresponds again to the sum of a normal and a truncated normal variable.

In the case when the threshold value of the variable V_0 coincides with $\mathbb{E}\{V_0\}$, the above-discussed sum is equivalent to the form $a\,|V_0| + b\,V_1$, for some real values a and b, and $|V_0|$ is distributed as an half-normal variable. There is no real loss of generality is considering the special version

$$Z = \delta\,|V_0| + \sqrt{1-\delta^2}\,V_1 \tag{8}$$

where V_0 and V_1 are independent $N(0,1)$ variables, and $\delta \in (-1, 1)$. The distribution of Z is $\mathrm{SN}(\delta/\sqrt{1-\delta^2})$.

This Z has the structure of the random term appearing in a stream of econometric literature dealing with stochastic frontier analysis; a leading paper is Aigner et al. (1977). In this context, the response variable is given by the output produced by some economic unit of a given type, and a regression model is built to express the relationship between the response variable and a set of covariates which represent the input factors employed to obtain the corresponding output. The key difference from ordinary regression models is that here the stochastic component is the sum of two ingredients: one is a standard error term centred at 0 and the other is an essentially negative quantity which represents the inefficiency of a production unit, producing an output level somewhat below the curve of technical efficiency. In the common case when the purely random term is normal, such as V_1 in (8), and the inefficiency is assumed to be of type $\delta\,|V_0|$ with $\delta < 0$, we are effectively considering a regression model with error term of type SN.

A further related case is the nonlinear autoregressive stochastic process studied by Anděl et al. (1984) satisfying a relationship essentially of type

$$Z_t = \delta|Z_{t-1}| + \sqrt{1-\delta^2}\,\varepsilon_t \qquad (t = \ldots, -1, 0, 1, \ldots)$$

where the $\{\varepsilon_t\}$'s form a sequence of independent variables $N(0,1)$. The integral equation for the stationary distribution of the process $\{Z_t\}$ has a solution of type (3).

As a general remark, there is a difference in the focus of interest of the earlier papers just discussed and the more recent work summarised in the previous sections. The main target of these older papers appears to be completed in the study of the properties of certain transformations of normal variates, while the more recent papers regard (3) and (7) as new parametric families, representing an extension of independent interest to the normal family, especially for their ability to incorporate skewness in the data modelling process.

Although mathematically implicit in what has already been presented, the following example is useful to illustrate the stochastic mechanism underlying the SN distribution. Consider a bivariate normal variable

$$\begin{pmatrix} H \\ W \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$$

where the variables H and W may for instance represent 'height' and 'weight', respectively, of some people. The distribution of W conditionally on the fact that subjects have 'weight above average' can be intended in the sense $(W|W > \mu_2)$ or in a more informative form $W|U > 0$ where $U = W - \mathbb{E}\{W|H\}$, taking into account the 'height' of a subject to state whether it is 'overweight'. It turns out that $W|U > 0 \sim \mathrm{SN}(\mu_2, \sigma_{22}, \delta/\sqrt{1-\delta^2})$, where $\delta = \sqrt{1-\sigma_{12}^2/(\sigma_{11}\,\sigma_{22})}$.

The left-hand side panel of Figure 2 displays a simulated sample from a certain bivariate normal populations, marking points which are above the regression line $\mathbb{E}\{W|H\}$ differently from those of the points below that line. The right-hand side plot of the same figure shows the marginal distribution of $W \sim N(\mu_2, \sigma_{22})$, and compares it with the theoretical distribution $\mathrm{SN}(\mu_2, \sigma_{22}, \delta/\sqrt{1-\delta^2})$ and the observed histogram of $W|U > 0$.

If the selection condition is modified to $U + \tau > 0$, then we get a distribution of type (7).

2.4 Statistical aspects

While the study of the SN distribution on the probabilistic side leads nicely to a collection of appealing properties which extend in a natural way those of the normal distribution, the connected inferential process has some unusual aspects, especially if we focus on a neighbourhood of point $\alpha = 0$ which corresponds to the normal distribution.

Consider first the case of a simple random sample $y = (y_1, \ldots, y_n)^\top$ from $\mathrm{SN}(\xi, \omega^2, \alpha)$, with corresponding log-likelihood function

$$\ell(\xi, \omega^2, \alpha) = \mathrm{const} - \tfrac{1}{2} n \log \omega^2 + z^\top z + \sum_i \zeta_0(\alpha\, z_i) \tag{9}$$

where

$$z = (z_1, \ldots, z_n)^\top = \omega^{-1}(y - \xi\, 1_n) \qquad \zeta_0(x) = \log\{2\,\Phi(x)\} \tag{10}$$

and 1_n is the n-dimensional vector of all 1's. It is immediate to extend the log-likelihood (9) to regression models; these will however be discussed in § 3.6.

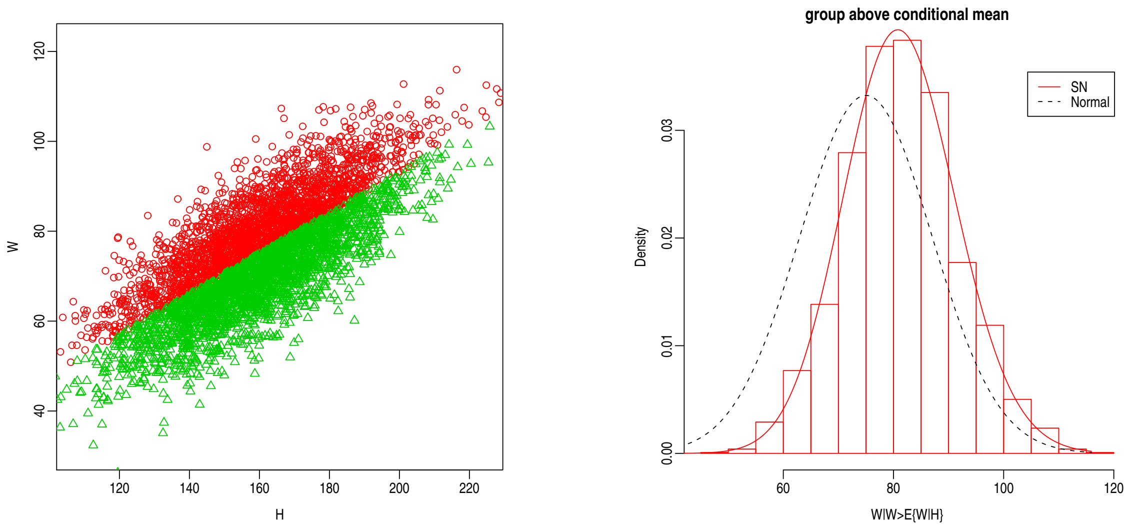

Figure 2: *The left-hand side plot shows simulated data from a bivariate normal population, using different symbols for points which are above the regression line and below that line. The right-hand side plot display the histogram and the theoretical distribution of values of the second component which are above the regression line, and the marginal distribution of the second component*

Besides the common fact that the likelihood equations need to be solved numerically, there are two additional problems. Firstly, there is always an inflection point at $\alpha = 0$ of the profile log-likelihood. Correspondingly, at $\alpha = 0$, the expected Fisher information becomes singular.

To have a direct view of the problem, consider the left-hand side plot of Figure 3 which shows the behaviour of the profile log-likelihood for the (ω, α) based of a sample of $n = 87$ data; the twisted contour levels lines are markedly different from the ideal quadratic shape. As n increases, this unpleasant sort of behaviour vanishes, except at the point $\alpha = 0$. For finite n, however, the effect propagates at some distance away from $\alpha = 0$, and in this context the phrase 'moderate size n' typically refers to a somewhat larger n than in other settings.

To get around these difficulties, consider the reparametrisation from (ξ, ω^2, α) to $(\mu, \sigma^2, \gamma_1)$ obtained on re-writing

$$Y = \xi + \omega\, Z = \mu + \sigma\, Z_0, \qquad Z_0 = (Z - \mu_z)/\sqrt{1 - \mu_z^2}\,. \tag{11}$$

Since Z_0 is a standardised variable, μ and σ denote the mean and the standard deviation of Y, respectively; γ_1 denotes the index of skewness. With this reparametrisation, the log-likelihood function and the Fisher information matrix have a regular behaviour. The right-hand side plot of Figure 3 refers to the same data used for the left-hand side plot, and it demonstrates the clear improvement in the shape of the log-likelihood obtained with the alternative reparametrisation.

These unusual aspects of the likelihood function and of the expected Fisher information matrix have been discussed by Azzalini (1985), Chiogna (1997), Azzalini & Capitanio (1999), Pewsey (2000a). The phenomenon of singularity at $\alpha = 0$ of the expected Fisher information matrix is special case of the problem studied in great generality by Rotnitzky et al. (2000), but motivated by a problem closely related to the present one.

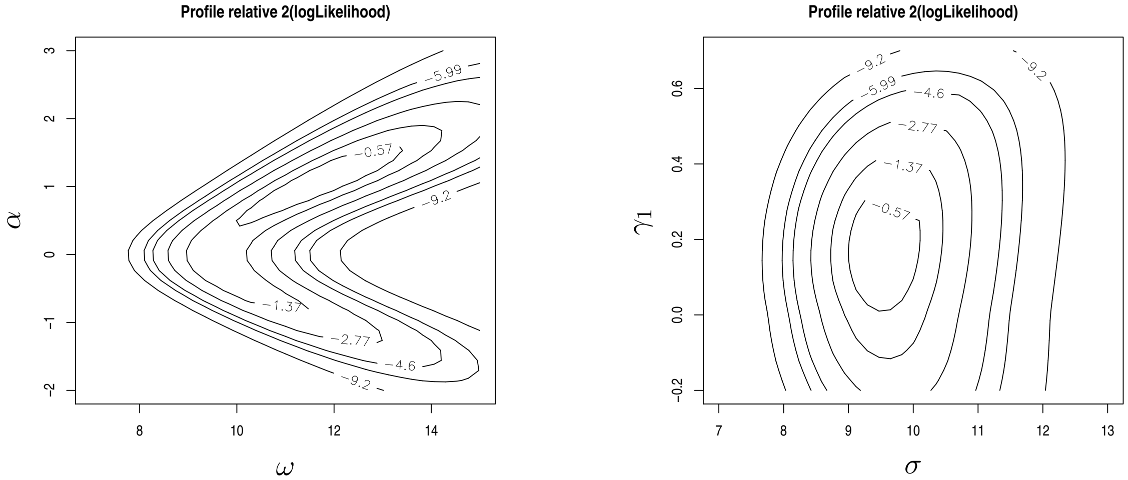

Figure 3: *Profile relative twice log-likelihood function for a sample with $n = 87$, using two different parameterisations. The contour lines correspond to upper percentage points of level 0.25, 0.5, 0.75, 0.90, 0.95, 0.99 of a χ^2_2 distribution*

A second peculiar aspect of the log-likelihood function (9) is that, for small or possibly even moderate value of n, it can be monotonically increasing or decreasing in α; hence the maximum likelihood estimate $\hat{\alpha}$ is $\pm\infty$. The phenomenon is particularly easy to examine in the one-parameter version of the likelihood, when ξ and ω are known to be 0 and 1, say, respectively. In this case, (9) reduces to $\sum_i \zeta_0(\alpha y_i)$ and, if the sample values all have the same sign, then $\hat{\alpha}$ occurs at $\pm\infty$, as noted by Liseo (1990). The root of the problem is in the limited theoretical range $(-0.9953, 0.9953)$ for γ_1 under the SN model; there is however no direct connection between the phenomenon of $\hat{\alpha} = \pm\infty$ and the sample index of skewness g_1 falling outside the theoretical admissible range.

It can be argued that the occurrence of $\hat{\alpha}$ at $\pm\infty$ must be regarded similarly to an estimate 0 or 1 for the parameter of a binomial variable, that is a boundary but not 'unacceptable' value. On the other hand, there is a difference from that case, since these boundary estimates in the binomial case occur when the data give a strong indication in that sense, while for the SN distribution estimates at $\pm\infty$ can occur even for data which appear to be free of any peculiar pattern; see Figure 4 of Azzalini & Capitanio (1999) for an example.

Therefore $\hat{\alpha} = \pm\infty$ is perceived as an anomalous outcome, and various proposals have been put forward to avoid it. In the classical approach, the method proposed by Sartori (2005) is based on a second-order modification of the likelihood equation which never produces boundary estimates. See also Monti (2003) for another proposal, based on minimum chi-square. In the Bayesian approach, Liseo & Loperfido (2004) show that the Jeffreys' prior for α is a proper distribution, a most special situation given that the range of α in unbounded. Hence the posterior distribution for α is a proper distribution too, and use of its mode or median produces finite estimates, which are shown to have good frequentist properties.

For testing parametrically the hypothesis of normality within the SN class, Salvan (1986) has shown that the sample index of skewness g_1 leads to the test locally uniformly most powerful among those which are location and scale invariant to test that the null hypothesis $H_0 : \alpha = 0$ against a one-sided alternative. The procedure is based on a form of invariant likelihood, whose connections with the Bayesian approach are discussed by Liseo (1990). Muliere & Nikitin (2002) and Durio & Nikitin (2003) examine local Bahadur efficiency of various nonparametric tests within the class (3).

Extending the work of O'Hagan & Leonard (1976), Mukhopadhyay & Vidakovic (1995) use (3) and similar other densities produced from Lemma 1 to represent the prior distribution for the expected value of a normal variable.

2.5 Some related distributions

Application of Lemma 1 to other densities f_0 in place of ϕ produces a variety of skew variants of f_0, in fact one for each choice of G. Of the vast set of possible options, a certain number have been considered in the literature, with special attention to cases where $G = \int f_0$. Since the original formulation of Lemma 1 referred to $w(x) = \alpha x$, this condition carries on in many papers until very recent ones.

In this subsection, we shall not discuss distributions for which a multivariate treatment has been developed, since these cases will be presented in a subsequent section, and confine ourselves to those examined only in the univariate case.

Gupta et al. (2002) examine a range of possible options for f_0, which is in turn selected to be normal, uniform, logistic and G is taken to be $\int f_0$; the treatment of the t case is somewhat different and it is in line with the multivariate case to be discussed later. For these distributions, moments and some other characteristics are obtained. The paper of Nadarajah & Kotz (2003) has a similar target in the case when f_0 is kept equal to ϕ and G varies.

A particularly important feature in applications is the possibility to regulate the thickness of the tails of a symmetric distribution via a shape parameter ψ, say. In a location or a regression model where the error terms are assumed to belong to a parametric family of this sort inferential methods can be expected to have good robustness properties, since the adjustable parameter ψ can adapt to the presence of outlying observations.

An important special case of this type is the Subbotin's distribution (Subbotin, 1923) with density $f_0(x) = C_\psi \exp(-|x|^\psi/\psi)$ where $\psi > 0$ and C_ψ is a normalising constant; $\psi = 2$ corresponds to the normal density. This distribution is also known with other names, including exponential power distribution and generalised error distribution. Combination of this f_0 with a skewing factor $2\,G(\alpha x)$ has been considered by Azzalini (1986) in two forms: (i) $G = \int f_0$, and (ii) $G = \Phi$. Inferential aspects for case (ii) have been examined by DiCiccio & Monti (2004), providing a detailed study of asymptotic properties of maximum likelihood estimation. These are very much similar to those discussed earlier for the SN case, with more technical complications due to the additional parameter ψ. A reparametrisation similar to (11) turns out to be useful to investigate the asymptotic properties of maximum likelihood when $\alpha = 0$. DiCiccio & Monti (2004) provide some illustrative applications of the methodology in some regression and time-series problems when robustness issues arise.

Replacing the normal density by a skew-normal one in the wrapped normal distribution on the circle, Pewsey (2000b) has developed a wrapped skew-normal distribution, whose density function is

$$f(\theta) = \frac{2}{\omega} \sum_{r=-\infty}^{\infty} \phi\left(\frac{\theta + 2\pi\, r - \xi}{\omega}\right) \Phi\left(\frac{\alpha(\theta + 2\pi\, r - \xi)}{\omega}\right), \qquad 0 \le \theta < 2\pi,$$

where ξ, ω, α are location, scale and skewness parameters. Various formal properties, inferential aspects and an illustrative example are discussed in the quoted paper.

3 The multivariate skew-normal distribution

3.1 Some general preliminaries

For introducing the multivariate version of the skew-normal distribution, it is useful to consider the following extension of Lemma 1 to the d-dimensional case; the proof is similar to the univariate case. In the multivariate context, the concept of symmetric density is not defined in a unique way, as for $d = 1$. The result below refers to the assumption of the condition $f_0(x) = f_0(-x)$, called 'central symmetry' by some authors; see for instance Serfling (2004).

Lemma 3 *If f_0 is a d-dimensional probability density function such that $f_0(x) = f_0(-x)$ for $x \in \mathbb{R}^d$, G is a one-dimensional differentiable distribution function such that G' is a density symmetric about 0, and w is real-valued function such that $w(-x) = -w(x)$ for all $x \in \mathbb{R}^d$, then*

$$f(z) = 2\, f_0(z)\, G\{w(z)\}\,, \qquad z \in \mathbb{R}^d, \tag{12}$$

is a density function on $\mathbb{R}^d$.

Besides the formal analogy with Lemma 1 for constructing new densities starting from a symmetric density f_0, also some other results of § 2.1 carry on in the multivariate case.

Stochastic representation If $X \sim G'$ and $Y \sim f_0$ are independent variables, Z defined as in (2) has distribution (12).

Perturbation invariance If $Y \sim f_0$ and $Z \sim f$, then

$$t(Y) \overset{d}{=} t(Z) \tag{13}$$

for any real-valued function such that $t(x) = t(-x)$ for all $x \in \mathbb{R}^d$, irrespectively of the choice of G and w.

A statement more general but less operative than Lemma 3 has been given by Azzalini & Capitanio (1999, § 7). The above formulation is in the form presented by Azzalini & Capitanio (2003) and, in a slightly different way, by Genton & Loperfido (2005); despite the discrepancy in publication dates, these two papers have been developed independently and more or less simultaneously. One difference between the two formulations is that Genton & Loperfido (2005) restrict the statement to the case when f_0 is of elliptical type. Another difference is that they replace $G\{w(x)\}$ by a function $\pi(x)$ satisfying the conditions

$$\pi(x) \ge 0, \qquad \pi(x) + \pi(-x) = 1, \tag{14}$$

since the decomposition $\pi(x) = G\{w(x)\}$ is not unique. On the other hand, the actual construction of a function $\pi(x)$ with the prescribed properties is immediate if one selects G and w as indicated.

Formulations more general than (12) can certainly be considered. If U_0 and U_1 are continuous variables of dimension m and d respectively, then Arellano-Valle et al. (2002) put forward the very general formula for the distribution of $Z = (U_1|U_0 > 0)$, where the notation $U_0 > 0$ is intended as a set of component-wise inequalities,

$$f_Z(z) = f_{U_1}(z)\,\frac{\mathbb{P}\{U_0 > 0|U_1 = z\}}{\mathbb{P}\{U_0 > 0\}} \tag{15}$$

in a self-explanatory notation. Notice that here the choice of the threshold 0 is without loss of generality, since no symmetry around 0 is assumed. The problem is that the actual computation of the integrals involved is amenable only in some cases.

For the rest of this section, we shall concentrate on the special case when f_0 in (12) is the multivariate normal density. Application of Lemma 3 to more general cases will be discussed in Sections 4 and 5.

3.2 Definition of skew-normal density and some properties

Consider the case that $f_0(x)$ in (12) is $\phi_d(x;\Omega)$, the density function of a $N_d(0,\Omega)$ variable, where Ω is a positive definite matrix; also, take $G = \Phi$ and w to be a linear function. Allowing for the presence of a d-dimensional location parameter ξ, the density function is

$$f(y) = 2\,\phi_d(y-\xi;\Omega)\,\Phi(\alpha^\top\omega^{-1}(y-\xi)), \qquad (y \in \mathbb{R}^d), \tag{16}$$

where α is shape parameter ($\alpha \in \mathbb{R}^d$) and ω is the diagonal matrix formed by the standard deviations of Ω. If a d-dimensional continuous random variable Y has density (16), we say that its distribution is multivariate skew-normal and write $Y \sim \mathrm{SN}_d(\xi,\Omega,\alpha)$. The reason to have the apparently redundant term ω^{-1} in the argument of Φ is to keep the shape parameter α unaltered when a location and scale transformation of type $Y' = a + bY$ is applied to Y, for some positive definite diagonal matrix b.

Figure 4 shows contour levels of two cases of SN distribution where $d = 2$. To illustrate how the down-weighting factor $\Phi(\alpha_1 z_1+\alpha_2 z_2)$ of the original normal density function operates, some lines with constant value of $\alpha_1 z_1 + \alpha_2 z_2$ are also indicated.

A simple extension of Lemma 2 leads readily to the moment generating function corresponding to (16), that is

$$M(t) = 2\,\exp(\xi^\top t + \tfrac{1}{2}t^\top\Omega t)\,\Phi(\delta^\top\omega t), \qquad t \in \mathbb{R}^d, \tag{17}$$

where

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

and $\bar\Omega = \omega^{-1}\Omega\omega^{-1}$ is the correlation matrix associated to Ω. From $M(t)$, one obtains

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

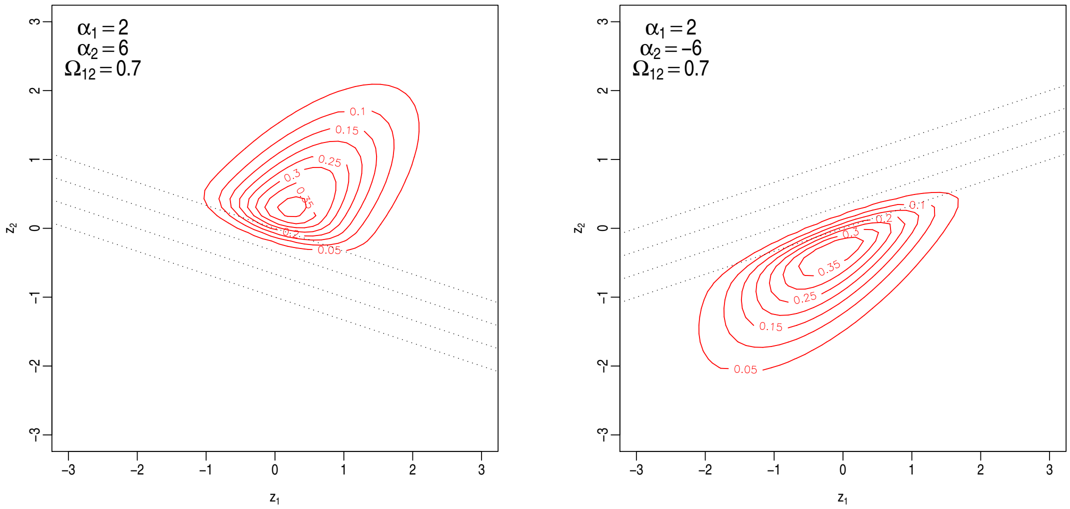

Figure 4: Contour levels of bivariate skew-normal distributions for some choices of the parameters; in both cases the location parameter is $(0,0)$, $\Omega_{11} = \Omega_{22} = 1$ and $\Omega_{12} = 0.7$. The dashed lines are those with $\alpha_1 z_1 + \alpha_2 z_2 = k$ for $k = -6, -4, -2, 0$

where $\mu_z = \sqrt{2/\pi}\,\delta$ is the mean value of the reduced variable $Z = \omega^{-1}(Y - \xi) \sim \mathrm{SN}_d(0, \bar\Omega, \alpha)$. The multivariate indices of skewness and kurtosis are

$$\gamma_{1,d} = \left(\frac{4-\pi}{2}\right)^2 \left(\frac{\mu_z^\top \bar\Omega^{-1}\mu_z}{1-\mu_z^\top \bar\Omega^{-1}\mu_z}\right)^3, \qquad \gamma_{2,d} = 2(\pi-3)\left(\frac{\mu_z^\top \bar\Omega^{-1}\mu_z}{1-\mu_z^\top \bar\Omega^{-1}\mu_z}\right)^2,$$

whose approximate ranges are (0, 0.9905) and (0, 0.869), respectively.

Another direct consequence of (17) is that the sum of a multivariate skew-normal variate and an independent multivariate normal variate is still skew-normal. This property is immediate on noticing that multiplication of (17) and $\exp(\mu t + \frac{1}{2}t^\top \Sigma t)$, say, gives another expression of type (17) with δ replaced by $\psi^{-1}\omega\delta$ where ψ is the diagonal matrix of the standard deviations of $\Omega + \Sigma$. This fact is essentially the multivariate version of property (4).

In analogy with (7), we also introduce an 'extended' form of multivariate skew-normal distribution, with density

$$f(y) = \phi_d(y - \xi; \Omega)\, \Phi(\alpha_0 + \alpha^\top \omega^{-1}(y - \xi))/\Phi(\tau) \tag{18}$$

where $\alpha_0 = \tau\,(1 + \alpha^\top \bar\Omega \alpha)^{1/2}$. In this case, we write $Y \sim \mathrm{SN}_d(\xi, \Omega, \alpha, \tau)$. If $\tau = 0$, then also $\alpha_0 = 0$ and (18) reduces to (16).

The corresponding distribution function can be computed via the distribution function of a $(d+1)$-dimensional normal variate. Specifically, if $Z \sim \mathrm{SN}_d(0, \bar\Omega, \alpha, \tau)$, then

$$\mathbb{P}\{Z \le z\} = \Phi_{d+1}((\tau, z^\top)^\top, \tilde\Omega)/\Phi(\tau)$$

where the notation $Z \le z$ means that each component of Z does not exceed the corresponding component of z, and

$$\tilde\Omega = \begin{pmatrix} 1 & -\delta^\top \\ -\delta & \bar\Omega \end{pmatrix}.$$

Both families (16) and (18) are closed under marginalisation. Specifically, if $Y \sim \mathrm{SN}_d(\xi, \Omega, \alpha, \tau)$ and its parameters are partitioned as follows

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \tag{19}$$

where Y_1 is of dimension h, then

$$Y_1 \sim \mathrm{SN}_h(\xi_1, \Omega_{11}, \alpha_{1(2)}, \tau)$$

where

$$\bar{\Omega}_{22\cdot 1} = \bar{\Omega}_{22} - \bar{\Omega}_{21}\bar{\Omega}_{11}^{-1}\bar{\Omega}_{12}, \qquad \alpha_{1(2)} = \frac{\alpha_1 + \bar{\Omega}_{11}^{-1}\bar{\Omega}_{12}\alpha_2}{\left(1 + \alpha_2^\top \bar{\Omega}_{22\cdot 1}\alpha_2\right)^{1/2}}. \tag{20}$$

The expression of the parameters of the marginal distribution is simply produced by extracting appropriate components of the original parameters, for the ξ and Ω components, but for α the expression is somewhat more involved, and it depends both on α_1 and on α_2. In this sense, it would be simpler to adopt δ as the skewness parameter, since marginalisation would lead to simple extraction of a subset of its components, as it is apparent from (17). The drawback of adopting δ as the shape parameter is that Ω and δ are not variation independent, as we can see from the expression of $\mathrm{var}\{Y\}$, while this property holds for the pair (Ω, α).

Although families (16) and (18) are much similar in many respects, there are important differences, since (18) does not fit in the scheme of Lemma 3. One implication is that the perturbation invariance property (13) holds for (16) but not for (18). One advantage of (18) over (16) is the property of closure under conditioning. Specifically, if Y has a density of type extended skew-normal (18), then $Y_1|Y_2 = y_2$ is still a member of the same family, but the same is not true in general for the sub-family with $\tau = 0$.

Many of the above results have appeared in Azzalini & Dalla Valle (1996) and in Azzalini & Capitanio (1999), where the focus is on the case $\tau = 0$. The extended form (18) has been discussed by Arnold & Beaver (2000a) and by Capitanio et al. (2003). The latter paper develops a theory of graphical models for skew-normal variates, apparently the first case of graphical models for continuous non-Gaussian variables. Explicit expressions of moments of $\mathrm{SN}_d(\xi, \Omega, \alpha)$ up to order four are given by Genton et al. (2001), but not reported here due to their lengthy expression. Azzalini (2001) provides accurate approximations to select regions of given probability with minimum geometric volume. A theory of skew-normal stochastic process related to the theory of communication and signal processing has been developed by Gualtierotti (2004).

3.3 Stochastic representations

The univariate skew-normal distribution can be obtained by several stochastic mechanisms. Some of them carry on to the multivariate case, in addition to the general representation (2).

Representation via conditioning If the d-dimensional random variable U is partitioned in components U_0 and U_1 of size 1 and d, respectively, such that

$$U = \begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim N_{1+d}(0, \Omega^*), \qquad \Omega^* = \begin{pmatrix} 1 & \delta^\top \\ \delta & \bar{\Omega} \end{pmatrix} \tag{21}$$

where Ω^* is a positive definite correlation matrix, then

$$Z = (U_1 | U_0 > 0) \sim N_d(0, \bar{\Omega}, \alpha), \qquad \alpha = \left(1 - \delta^\top \bar{\Omega}^{-1} \delta\right)^{-1/2} \bar{\Omega}^{-1} \delta, \tag{22}$$

The affine transformation $Y = \xi + \omega Z$ leads to density (16).

Representation via convolution Assume that $V_0 \sim N(0,1)$ and $V_1 \sim N_d(0, \Psi)$ are independent variables, where Ψ is a correlation matrix, and let $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_d)$ where $\delta_j \in (-1, 1)$ for all j's. Then

$$Z = \Delta\, 1_d\, |V_0| + (I_d - \Delta^2)^{1/2}\, V_1 \tag{23}$$

has distribution $\mathrm{SN}_d(0, \bar{\Omega}, \alpha)$, with a known relationship between the (Ψ, Δ) and the ($\bar{\Omega}$, α) sets of parameters.

It is reassuring that both representations, (22) and (23), lead to the same family of distributions, since otherwise we would have to choose between two legitimate multivariate extensions of the univariate distribution (3). In a slightly more elaborate version, similar representations exist for the extended form (18).

3.4 Linear and quadratic forms

The class (18) is closed under affine transformations. Specifically, if $Y \sim SN_d(\xi, \Omega, \alpha, \tau)$, $a \in \mathbb{R}^h$ and B is a matrix of size $h \times d$, then

$$W = a + B\,Y \sim \mathrm{SN}_h(\xi_W, \Omega_W, \alpha_W, \tau) \tag{24}$$

where

$$\begin{aligned} \xi_W &= a + B\xi, \\ \Omega_W &= B \Omega B^\top, \\ \alpha_W &= \frac{1}{\left(1 + \alpha^\top (\bar{\Omega} - H \Omega_W^{-1} H^\top) \alpha\right)^{1/2}} \omega_W \Omega_W^{-1} H^\top \alpha, \qquad H = \omega^{-1} \Omega B^\top \end{aligned} \tag{25}$$

and ω_W denotes the diagonal matrix of standard deviations of Ω_W.

Closure of the class under affine transformation resembles a well known fact for normal variates, and the formal rules for transforming the location parameter ξ and the dispersion matrix Ω are the same as the normal case.

There exists a particular full-rank affine transformation of type (24) which produces a sort of ‘canonical form’, in the sense W has $d-1$ components with 0 asymmetry and a single component which ‘absorbs’ all asymmetry of the others. The skewness parameters of this component is a summary index of asymmetry featuring in many other expressions.

The connection with the normal family is even more striking in the light of the following result which is the analogue of a known characterisation of the multivariate normal family.

Theorem 4 *For a d-dimensional random variable Z, denote by $\Omega = \mathbb{E}(Z\,Z^\top)$ the matrix of second-order moments, and suppose that, for all $h \in \mathbb{R}^d$ such that $h^\top \Omega h = 1$, $h^\top Z$ has skew-normal distribution of type (3), with shape parameter depending on h. Then Z has multivariate skew-normal distribution (16).*

Due to the general property of perturbation invariance (13), a vast set of results on quadratic forms of multivariate normal variables carry on for the skew-normal distribution. For instance, if Y has distribution (16), it is immediate to state:

(a) if C is a full-rank $p \times d$ matrix ($p \le d$), then
$$(Y-\xi)^\top C(C^\top \Omega C)^{-1}C^\top (Y-\xi) \sim \chi^2_p\,, \tag{26}$$
for which the case $C = \Omega^{-1}$ is of special interest;

(b) $(Y-\xi)^\top A(Y-\xi)$ and $(Y-\xi)^\top B(Y-\xi)$ are independent if and only if $A\Omega B = 0$.

The above discussion summarises only the most basic properties on linear and quadratic forms of skew-normal variates, as given by Azzalini & Capitanio (1999), Gupta & Huang (2002) and Capitanio et al. (2003). Additional distributional properties can be found in Loperfido (2001) and Genton et al. (2001). In particular, explicit expressions for the moment generating function and the lower order moments of a general quadratic form $Y^\top AY$ are available. Notice that all these properties of quadratic forms refer to (16), and do not carry on to the extended version (18).

3.5 Further extensions: several latent variables

One limitation of the formulation discussed so far is that, when we consider it in the light of the representation via conditioning, only one latent variable U_0 is admitted. However, if one thinks of a form of selective sampling as mentioned in Section 2.3, it is quite natural to allow that two or more constraints of the form $U_0 + \tau > 0$ are operating simultaneously on a set of m latent normal variables with 0 mean and covariance matrix Γ.

The integrals involved by (15) can easily be handled, thanks to well-known properties of the normal distribution, and we arrive at a density function of the form
$$f(x) = \phi_d(x-\xi;\Omega)\,\frac{\Phi_m\{\tau+\Upsilon(x-\xi),\Gamma^*\}}{\Phi_m(\tau,\Gamma)} \tag{27}$$
where $\Phi_m(a, A)$ denotes the distribution function of a $N_m(0, A)$ variate evaluated at point $a \in \mathbb{R}^m$, τ is now an m-dimensional parameter vector, Υ is a matrix $m \times d$, and Γ^* is a suitable function of Γ, Υ, and Ω.

Alternatively to the above mechanism of conditioning, a density of type (27) can be obtained via a form of convolution of normal and truncated normal variables.

Quite recently, several papers have put forward distributions essentially of type (27) although formulated in various different ways. These formulations include proposals by Sahu et al. (2003), Liseo & Loperfido (2003a), González-Farías et al. (2004a), González-Farías et al. (2004b), Arellano-Valle & Genton (2005). The question of their interconnections, and in fact of the essential equivalence among them, is discussed by Arellano-Valle & Azzalini (2006).

Because of the more technical nature of this set of proposals and the fact that the theme is still under development, we do not discuss in detail these extensions. There are important properties to be mentioned, however: (i) an improvement over the basic SN formulation is the closure of the class under sum of independent components; (ii) properties of closure of the class under marginalisation, affine transformations and conditioning to given values of some components still hold.

3.6 Statistical aspects

Suppose that a set of independent d-dimensional observations $y_1, \ldots, y_n$ is available, and the i-th observation is associated with a p-dimensional vector x_i of concomitant variables $(i = 1, \ldots, n)$. These data are arranged in two matrices, y and X, say, of size $n \times d$ and $n \times p$, respectively.

Under the assumption that y_i is sampled from a $\mathrm{SN}_d(\xi_i, \Omega, \alpha)$ distribution $(i = 1, \ldots, n)$, the corresponding log-likelihood function is

$$\ell = \text{constant} - \tfrac{1}{2} n \log|\Omega| - \tfrac{1}{2} \mathrm{tr}\left(\Omega^{-1} D\right) + \sum_i \zeta_0\{\alpha^\top \omega^{-1}(y_i - \xi_i)\} \tag{28}$$

where $D = \sum_i (y_i - \xi_i)(y_i - \xi_i)^\top$.

In many cases, a regression model for the location parameters is introduced, of the form $\xi_i^\top = x_i^\top \beta$ for some matrix of regression parameters β of size $p \times d$, for $i = 1, \ldots, n$. The optimisation process of (28) is much simplified if we introduce $\eta = \omega^{-1}\alpha$, in place of α. This reparametrisation allows explicit maximisation of (28) with respect to Ω, given by

$$\hat{\Omega}(\beta) = n^{-1}(y - X\beta)^\top (y - X\beta),$$

for any fixed β. Hence the profile log-likelihood for (β, η) is

$$\ell^*(\beta, \eta) = \text{constant} - \tfrac{1}{2} n \log|\hat{\Omega}(\beta)| + 1_n^\top \zeta_0\{(y - X\beta)\eta\}$$

where it is intended that the function ζ_0 is computed component-wise. The function ℓ^* must be maximised numerically, but searching a space of substantially smaller dimension than (28). Further improvement in the numerical maximisation is achieved by supplying the algorithm with the expression of the partial derivatives of ℓ^*, which are available in closed form. Numerical differentiation of these derivatives leads to Fisher observed information matrix and standard errors for the estimates.

After the fitting process, a diagnostic tool for the adequacy of the assumed model can be constructed based on property (26) with $C = \Omega^{-1}$, and adapting the Healy's diagnostic plots to the present context; see Healy (1968). If the model is correctly specified, then

$$d_i = (y_i - \xi_i)^\top \Omega^{-1} (y_i - \xi_i), \qquad (i = 1, \ldots, n), \tag{29}$$

is distributed as χ_d^2; in practice one must replace the parameters by their estimates, and the statement is approximate. The plot of the n pairs of points $(q_i, d_{(i)})$, where q_i is quantile of the χ_d^2 distribution at level $i/(n+1)$ and $d_{(i)}$ is the i-th term of the d_i's arranged in increasing order, is expected to lie along the identity line. A variant of the same method operates on the probability scale instead of the quantile, plotting the pairs $(i/(n+1), p_{(i)})$, where $p_{(i)}$ is the distribution function of χ_d^2 evaluated at $d_{(i)}$.

For a more detailed discussion of the above material and illustrative numerical examples, see Azzalini & Capitanio (1999). This paper considers also other aspects of inference, specifically discriminant analysis and graphical regression models. An expanded treatment of graphical models is given by Capitanio et al. (2003).

As an illustration of the working of the methodology, we have used data collected at the Australian Institute of Sport on 202 athletes, and described by Cook & Weisberg (1994).

For this example we consider the pair (LBM,BMI) as the response variable, where the two acronyms denote Lean Body Mass and Body Mass Index, respectively. A linear model has been introduced expressing the response variable as a function of two continuous covariates, Height and Weight, and two factors, Sport (10 levels) and Sex (2 levels), under two alternative distributional assumptions: bivariate normal and bivariate skew-normal. The top-left panel in Figure 5 shows the bivariate residuals, with superimposed the contour levels of the fitted error distribution under assumption of normality, while the bottom-left panel is the corresponding plot under skew-normality, with some visible improvement in the adequacy of the fit. This indication is confirmed by the two plots on the right-hand side, which represent the Healy's plot on the probability scale, under the two distributional assumptions, and it is further supported by the likelihood ratio test for nullity of the parameter α, whose observed value 60.3 is highly significant on the χ_2^2 scale. Even if the improvement provided by the SN distribution over the normal is appreciable, the remaining curvature in the bottom-right plot indicates that there is still room for improvement.

Another area where the SN distribution is conveniently applied is longitudinal data with random effects. A typical formulation of linear type for the response profile of the i-th individual observed at d successive occasions is

$$y_i = \beta^\top x_i + Z_i\, b_i + \varepsilon_i, \qquad (i = 1, \ldots, n),$$

where b_i is a q-dimensional vector of random effects associated to individual i, Z_i denotes a $d \times q$ matrix of additional covariates ($q < d$) and ε_i is a d-dimensional vector of random errors independent of b_i. It is only for simplicity of notation that here we keep d constant across individuals.

The standard assumption on the random components b_i and ε_i is joint normality with 0 mean, and some suitable constraint on the covariance matrices, $\mathrm{var}\{b_i\}$ and $\mathrm{var}\{\varepsilon_i\}$, to ensure identifiability. Often the structure of $\mathrm{var}\{b_i\}$ is assumed to be of diagonal form, and $\mathrm{var}\{\varepsilon_i\}$ is chosen to have a patterned form corresponding to some form of serial correlation structure. A formulation with increased generality is obtained by replacing the assumption of normality by the skew-normality for either b_i or ε_i, retaining the assumption of normality for the other one. Thanks to one of the properties established earlier, the linear combination $Z_i\, b_i + \varepsilon_i$ is multivariate skew-normal. Therefore the log-likelihood structure is again of type (28), after suitable transformation of the parameters, and the procedure described above for its maximisation can be used. Alternatively, Arellano-Valle et al. (2005a) have developed an EM-type algorithm for this case. They also consider the more general case where both components b_i and ε_i are of skew-normal type, leading to a distribution for y_i of type (27) with $m = 2$.

Arellano-Valle et al. (2005b) have examined measurement error models where the usual hypothesis of normality of the stochastic components is replaced by skew-normality, and have shown that the marginal joint distribution of the observables is of type (27) with $m = 3$.

The use of skew-normal distributions to model spatial data has been studied by Kim & Mallick (2004), regarding the d components of the vector y as the observations on a response variable at d geographical locations $x_1, \ldots, x_d$. In this context, the dispersion parameter Ω is conveniently taken to be of the form $\sigma^2\, K(\theta)$ where σ^2 is a scale factor and $K(\theta)$ is scale-free positive-definite matrix which reflects the dependence structure, often via a function of a parameter θ and the distances $\|x_r - x_s\|$ between points. For estimation of the parameters

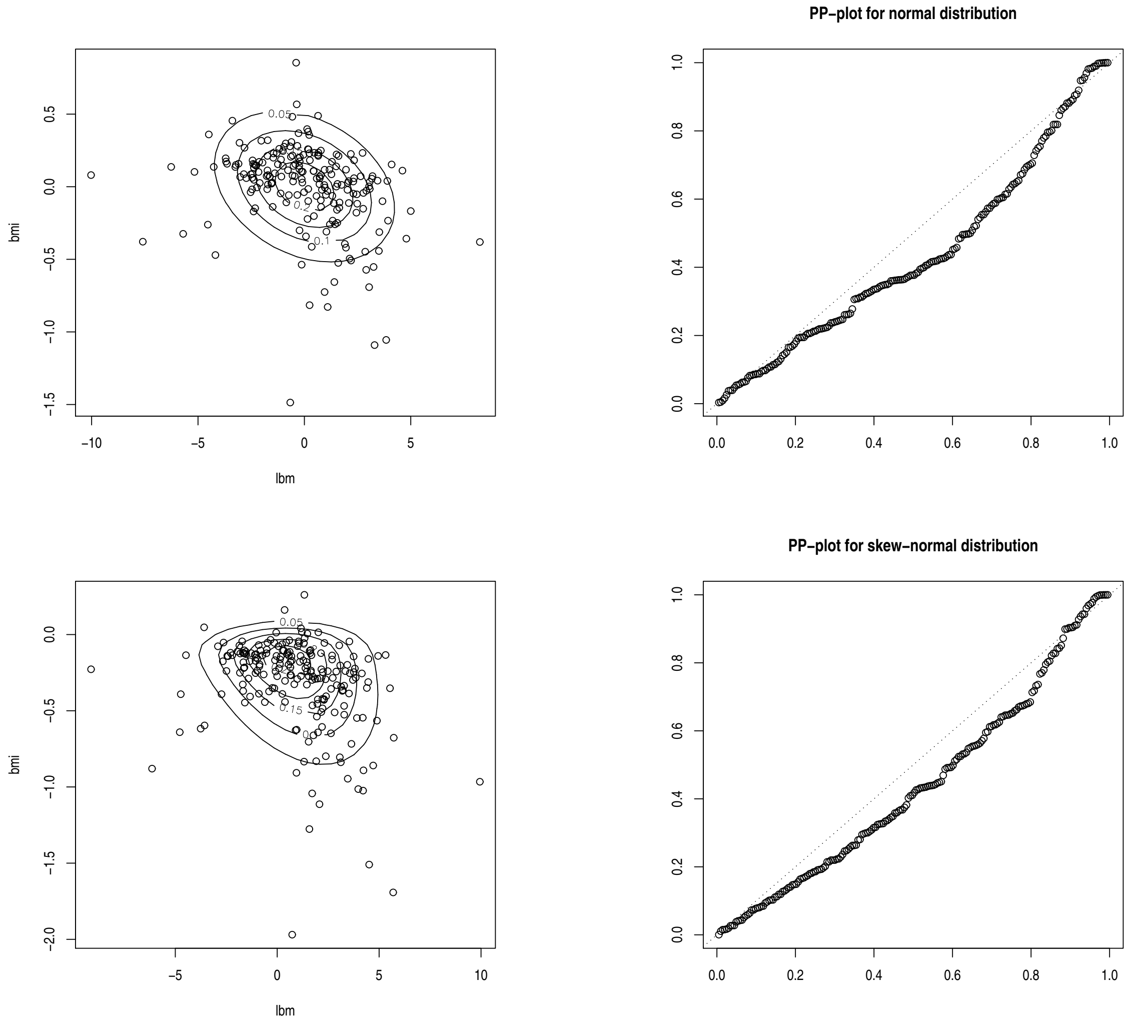

Figure 5: *AIS data: plot of residuals (left-hand side) and Healy's type plot (right-hand side) after fitting a linear model to the bivariate response variable (LBM, BMI) under the assumption of bivariate normal distribution (top row) and skew-normal distribution (bottom row) for the error term*

and subsequent prediction of the response at a new location x_0, Kim & Mallick (2004) frame the problem in the Bayesian framework and develop a MCMC algorithm.

It has been recalled earlier that, in the scalar case, the use of Jeffreys' criterion for uninformative prior of the shape parameter produces a proper distribution (Liseo & Loperfido, 2004). The extension of this property to the multivariate case is studied by Liseo & Loperfido (2003b) and Liseo (2004).

3.7 Other families

Arnold & Beaver (2000a, 2000b, 2000c) have examined skewed versions of various non-normal distributions, in a formulation closely similar to (16) or (18) using different ingredients. In most cases, they consider a constant term α_0 in the skewing factor, leading to a density of type

$$\text{constant} \times f_0(x)\, G(\alpha_0 + \alpha^\top x),$$

similarly to (18) for the SN family, but now with a non-normal f_0 and G. Since this form of skewing function does not satisfies the condition of Lemma 3, the normalising constant must be computed afresh for each choice of f_0 and G.

A favourable case is when f_0 is the product of d replicas of the Cauchy density, $\psi(u) = \{\pi\,(1+u^2)\}^{-1}$, and G is the integral of ψ, namely $\Psi(u) = \frac{1}{2} + \pi^{-1}\arctan(u)$. The normalising constant is then available in explicit form, leading to a density of type

$$f_C(x) = \prod_{j=1}^{d} \psi(x_j)\ \Psi(\alpha_0 + \alpha^\top x)/\Psi\{\alpha_0/(1 + 1_d^\top \alpha)\} \qquad (30)$$

for $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. An appealing aspect is that all marginal and conditional densities of f_C are still members of the same class. Arnold & Beaver (2000b) also provide a numerical example of use of the distribution to data-fitting.

4 Skew-elliptical distributions

4.1 General aspects

So far we have focused on modifications of the normal density function, or possibly of some other density, in a way that essentially introduces a form of skewness in the original density. However, in many practical problems, it is important to be able to regulate a density in a more flexible way.

Moving in this direction, the natural subsequent step is the ability to regulate both skewness and kurtosis of a distribution. In particular, we want to consider distributions with thicker tails than the normal, due to the relevance of this aspect in applications.

This target is not possible without taking the basis function f_0 in (12) to be something different from the normal density. The basic reason is that the modification applied to f_0 by $G\{w(y)\}$ can only make tails thinner than the original f_0. Therefore, we need to replace the normal density by a parametric family with adjustable thickness of the tails.

An extensively studied class of densities of the required type is the family of elliptical distributions. A d-dimensional continuous random variable Y belongs to this class if its density function is constant on ellipsoids, hence it is of the form

$$f_0(y;\xi,\Omega) = \frac{c_d}{|\Omega|^{1/2}}\,\tilde{f}\{(y-\xi)^\top\Omega^{-1}(y-\xi)\}, \qquad y\in\mathbb{R}^d, \tag{31}$$

where $\xi\in\mathbb{R}^d$, Ω is a covariance matrix, $\tilde{f}$ is a suitable function from $\mathbb{R}^+$ to $\mathbb{R}^+$, called the 'density generator', and c_d is a normalising constant; we then write $Y\sim \mathrm{Ell}_d(\xi,\Omega,\tilde{f})$. The set of elliptical densities includes many parametric families, notably the normal one, to which it reduces when $\tilde{f}(x)=\exp(-x/2)$ and $c_d=(2\pi)^{-d/2}$. For a detailed account on elliptical distributions, see Fang et al. (1990).

An initial exploration of the use of (12) with f_0 of type (31) has been done by Azzalini & Capitanio (1999). Proceeding along a different route, Branco & Dey (2001) start from a $(d{+}1)$-dimensional variable $U=(U_0,U_1^\top)^\top$ with location and scale parameters as in (21) but with assumption of normality replaced the one of elliptical form $\mathrm{Ell}(0,\Omega^*,\tilde{f}^{(d+1)})$. Consideration of $Z=(U_1|U_0>0)$ extends the conditioning mechanism (22) from the normal to the elliptical family. The expression of the density of $Y=\xi+\omega\,Z$ involves the marginal distribution of the d-dimensional marginal U_1, whose density generator is

$$\tilde{f}^{(d)}(u) = \frac{2\,\pi^{d/2}}{\Gamma(d/2)}\int_0^\infty \tilde{f}^{(d+1)}(r^2+u)\,r^{d-1}\,\mathrm{d}r, \qquad u>0$$

and the distribution function $F_{\tilde{f}_{q(y)}}$ of $\mathrm{Ell}(0,1,\tilde{f}_{q(y)})$, where

$$\tilde{f}_{q(y)} = \frac{\tilde{f}^{(d+1)}(u+q(y))}{\tilde{f}^{(d)}(q(y))}$$

and $q(y)=(y-\xi)^\top\bar{\Omega}^{-1}(y-\xi)$. The density function of Y is then

$$f_{\tilde{f}^{(d)}}(y)\,F_{\tilde{f}_{q(y)}}(\alpha^\top(y-\xi)) \tag{32}$$

where $f_{\tilde{f}^{(d)}}$ is the density of $\mathrm{Ell}(\xi,\Omega,\tilde{f}^{(d)})$ and α is as in (22).

From the general expression (32), Branco & Dey (2001) formulate 'skew versions' of various multivariate families: logistic, stable, Subbotin (or exponential power), t, Pearson type II, and more. Various formal properties are obtained concerning marginalisation, affine transformations, and quadratic forms, which parallel those of the skew-normal distribution.

Since the Branco & Dey (2001) construction is not set in the form (12) with f_0 of elliptical type as it had been considered by Azzalini & Capitanio (1999), the natural question arises of how the two approaches are related. The problem has been tackled by Azzalini & Capitanio (2003) where, although a general coincidence could not be established, it has been shown that this is however valid at least for various important cases, notably the multivariate Pearson type II and type VII families; the latter family is of special relevance because it includes the t distribution, which will be discussed separately. The essential equivalence of the two above-mentioned formulations is reassuring, as otherwise we would have two separate concepts of skew-elliptical distributions.

For the skew-elliptical family one can reproduce a large set of facts established for the skew-normal distribution: (i) both representations of conditioning type and convolution type hold and lead to the same distribution, (ii) for the case $d = 2$, an analogue of the construction obtained selecting maxima/minima holds, (iii) a sort of canonical transformation exists which transfers all asymmetry into a single component, and so on. See Fang (2003) and Azzalini & Capitanio (2003) for a detailed analysis of these aspects.

The question of equivalence between the genesis via conditioning and via convolution has been examined in a more general framework by Arellano-Valle et al. (2002), showing that the essential requirement is independence of absolute value of U_1 and the signs of the m-dimensional latent variable U_0.

4.2 An interesting case: the skew-t distribution

As already mentioned, it is of special interest from the applied view-point to have at hand families of distributions for which we can regulate both skewness and thickness of the tails. Among the various alternatives sketched above, an appealing option is offered by a skewed version of the t density, since this it is fairly tractable from the algebraic viewpoint and, in its symmetric version, it has already been used for similar purposes.

We shall say that a continuous random variable Y has multivariate skew-t distribution if its density is of type

$$f_T(y) = 2\, t_d(y;\xi,\Omega,\nu)\, T_1\left(\alpha^\top\omega^{-1}(y-\xi)\left(\frac{\nu+d}{Q_y+\nu}\right)^{1/2};\nu+d\right) \tag{33}$$

where ξ, Ω and ω are as in Section 3.2, $Q_y = (y-\xi)^\top\Omega^{-1}(y-\xi)$,

$$t_d(y;\xi,\Omega,\nu) \quad = \quad \frac{\Gamma(\frac{1}{2}(\nu+d))}{|\Omega|^{1/2}\,(\pi\nu)^{d/2}\,\Gamma(\frac{1}{2}\nu)}\,\frac{1}{(1+Q_y/\nu)^{(\nu+d)/2}}$$

is the density function of a d-dimensional t variate with ν degrees of freedom, and $T_1(x;\nu+d)$ denotes the scalar t distribution function with $\nu+d$ degrees of freedom. In this case, we write $Y \sim \mathrm{ST}(\xi,\Omega,\alpha,\nu)$.

From the mathematical viewpoint, (33) seems a convincing formulation since it arises from several different mechanisms:

- ◇ apart from the shift of location from 0 to ξ, it can be obtained from the general formulation of Lemma 3 when $f_0 = t_d$, $G(w) = T_1(w;\nu+d)$ and $w(x) = \alpha^\top\omega\; x/(x^\top\Omega^{-1}x)$;
- ◇ it falls within the formulation (32) when the generator $\tilde{f}^{(d+1)}$ is taken to be the one of the t distribution;
- ◇ it can be obtained via a convolution of type (23) with (V_0, V_1) of elliptical type;
- ◇ it can be generated by the same construction used for the regular multivariate t distribution, namely
$$Y = \xi + \frac{Z}{\sqrt{W/\nu}} \tag{34}$$
where $W \sim \chi^2_\nu$, if Z is an independent variable which is now taken to be $\mathrm{SN}_d(0,\Omega,\alpha)$ in place of the $N_d(0,\Omega)$ distribution used to produce the regular t.

In addition, the family (33) enjoys various appealing formal properties, some of which indicate a strong link with the SN distribution:

- ⋄ if $\nu \to \infty$, (33) converges to the SN density (16), as it is clear from representation (34);
- ⋄ $(Y - \xi)^\top \Omega^{-1}(Y - \xi)/d \sim F(d, \nu)$, again obvious from (34), and this allows to build Healy-type diagnostic plots, similar to those described in Section 3.6 for the SN family, but using the Snedecor's F as reference distribution in place of the χ^2;
- ⋄ the class of densities is closed under affine transformations, with the same rules (25) of the SN case for transformation of the parameters ξ, Ω, α, and ν kept constant;
- ⋄ it allows unlimited range for the indices of skewness and kurtosis for the individual components.

The above statements summarise results of Branco & Dey (2001), Gupta (2003), and Azzalini & Capitanio (2003). Expressions of joint moments of order up to four and some moments of quadratic forms of skew-t variates are given by Kim & Mallick (2003); these results are then used to obtain properties of some sample statistics relevant in time series and spatial statistics. Notice that setting $\nu = 1$ in (33) lends another type of skew-Cauchy density, different from (30).

The data and the linear model used for Figure 5 have been reconsidered under the assumption of skew-t distribution for the error term, and the outcome is summarised in graphical form in Figure 6. The estimated degrees of freedom are 3.45 with standard error 0.73, pointing to a fitted distribution with tails markedly different from the normal or skew-normal ones. Correspondingly, the plot on the left-hand side has contour levels quite concentrated in the centre, with several points falling in the tail area. The Healy's plot using a scaled F as reference distribution for the observed Mahalanobis distances is shown in right-hand side plot, and it indicates that the skew-t distribution provides a satisfactory description of the expected behaviour of the residuals.

Thanks to the possibility of regulating the thickness of the tails via the parameter ν, the symmetric t distribution has been used to produce robust inferential methods; see for instance Lange et al. (1989). It has however been remarked in various occasions that in practical applications outliers do not occur in all directions with equals chance; a detailed study of this sort has been conducted by Hill & Dixon (1982). Adoption of the skew-t in place of the regular t distribution incorporates this fact into the stochastic model and it can be expected to improve inferences. More details and some numerical illustrations are given by Azzalini & Capitanio (2003). Another proposal for robust inference, based of the skew version of the Subbotin distribution, has been mentioned in Section 2.5 (DiCiccio & Monti, 2004).

A different form of multivariate skew-t distribution has been put forward by Sahu et al. (2003), with some illustration of its use for Bayesian inference. Another skew-t distribution, presented by Arellano-Valle & Azzalini (2006), is an analogue of (27) in the sense it uses t components and it reduces to (33) when $m = 1$.

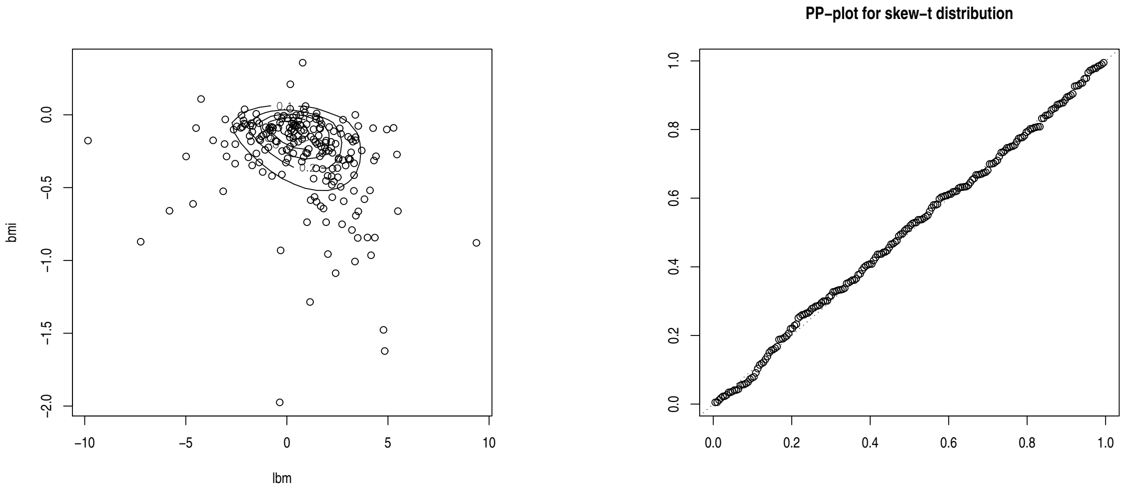

Figure 6: *AIS data: plot of residuals (left-hand side) and Healy's type plot (right-hand side) after fitting a linear model to the bivariate response variable (LBM, BMI) under the assumption of bivariate skew-t distribution for the error term*

5 Flexible parametric classes

So far we have considered modifications of a centrally symmetric density $f_0(x)$ via a perturbation function $2\,G\{w(x)\}$ which has been taken to have a simple parametric structure. In most cases $w(x)$ has been taken to have a linear structure or, in (33), a mildly non-linear and simple form.

Lemma 3 however allows us to make use of much more radical forms of perturbation, and with a very wide range of options in the choice of the perturbation. As a mere graphical illustration of the achievable level of flexibility, consider Figure 7 which shows contour level plots for two perturbations of the bivariate normal density with independent unit marginals, using $G = \Phi$ and

$$w(x) = w(x_1, x_2) = (x_1,\ x_2,\ x_1\,x_2^2,\ x_1^2\,x_2,\ x_1^3,\ x_2^3)\,\alpha$$

is a third-degree polynomial with only non-zero coefficients for odd-degree terms, and $\alpha = (5, -5, -3, 5, -3, 3)^\top$ and $\alpha = (2, 2, 2, -2, -2, 2)^\top$ in the first and the second panel, respectively.

Nothing prevents us from increasing the degree of the polynomial function, as well as choosing different ingredients f_0, G and w. There is then the question of how much flexible are the densities generated in this way, and whether we can approximate well a given density function by this method. This important problem has been examined by Wang et al. (2004) and Ma & Genton (2004), whose results can be summarised as follows. In their formulation, the term $G\{w(x)\}$ in (12) is replaced by the perturbation function $\pi(x)$, satisfying conditions (14).

Consider an arbitrary d-dimensional continuous density function $f(x)$. For any fixed but

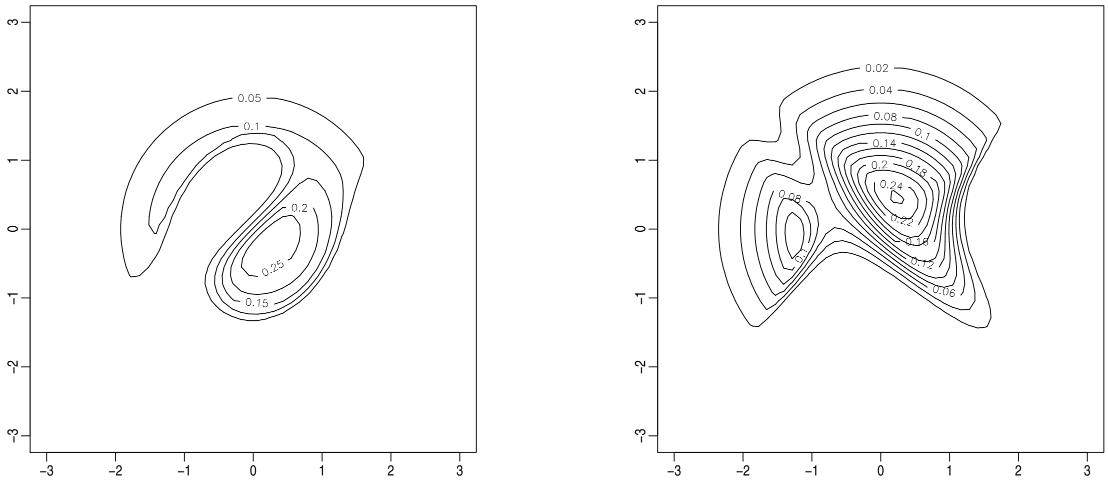

Figure 7: *Bivariate standard normal density perturbed using $G = \Phi$ and $w(x)$ of third-degree polynomial type, for two choices of the coefficients of the polynomial*

arbitrary point $\xi \in \mathbb{R}^d$, there exists a factorisation of the form

$$f(x) = 2\, f_\xi(x - \xi)\, \pi_\xi(x - \xi), \qquad (x \in \mathbb{R}^d),$$

where f_ξ is a density centrally symmetric around 0, and π_ξ is a perturbation function of type (14). If $\xi = 0$, this representation is of type (12).

Under regularity conditions, a representation of type $f(x) = 2\, f_0(x)\, \pi(x)$ can be approximated arbitrarily closely in the L^∞ norm by a member of the set of functions $2\, f_0(x)\, G\{w_K(x)\}$ where

$$w_K(x) = \sum_{j=1}^{K} \alpha_j^{(K)}\, x^{2j-1}, \qquad K = 0, 1, \ldots \tag{35}$$

and G is an arbitrary distribution function such that G' exists and is symmetric about 0.

These results build a bridge between the parametric context considered so far and a semi-parametric or ‘flexible parametric’ one where the term f_0 is taken to be a member of a parametric class and the perturbation function is either modelled with high flexibility via (35) or it is handled in a non-parametric fashion. In the latter option, inferential procedures aim only at the parameters of interest in f_0, neutralising the nuisance component $\pi(x)$.

The use of flexible parametric models in a context of linear mixed models for longitudinal data has been explored by Ma et al. (2004). Methods where $\pi(x)$ is handled non-parametrically have been examined by Loperfido (2004) and by Genton (2004b, § 5.4.3).

6 Some connected fields and applications

There are two purposes of this section, which are however partly overlapping: (i) to highlight the connections with some well-established areas of work; (ii) to mention some applications of the above-described results to practical problems.

Selective sampling In Section 2.3, we have already mentioned the connection with the area of biased and selective sampling. Under normality assumption of the overall population, the effect of this selection it to produce what in our terminology is called a skew-normal distribution for the observable data. This theme has been extensively studied in the quantitative sociology under the heading of 'Heckman model', whose interplay with statistical theory is discussed by Copas & Li (1997).

The literature on Heckman model focus strongly on normality assumption. It is plausible that the above-discussed extensions can produce similar but more flexible and realistic methods. Skew-elliptical distributions with heavy tails, especially the skew-t distribution, can be expected to be useful.

Stochastic frontier models Another link with existing literature mentioned in Section 2.3 points to the econometric literature on stochastic frontier models. Even here the distributional results now available can contribute to improve the stochastic modelling and the derivation of formal properties. Some work in this direction has been done by Tancredi (2002) and Domínguez-Molina et al. (2004).

Compositional data For the analysis of compositional data, a standard device is to transform the $d + 1$ original components belonging to the simplex to d components in $\mathbb{R}^d$ using the additive log-ratio transform, followed by analysis based on methods for normal data; see Aitchison (1986) for an exhaustive discussion of this approach.

The additive log-ratio transformation can be followed by assumption of skew-normality on the transformed data, to improve adequacy in data fitting. This assumption on $\mathbb{R}^d$ induces back a distribution on the simplex which enjoys certain formal desirable properties, which are due to the properties of closure under marginalisation and affine transformation of the skew-normal distribution, inducing some corresponding properties on the simplex. For a discussion of these issues, see Aitchison & Bacon-Shone (1999), Mateu-Figueras (2003), Aitchison et al. (2003).

Financial markets Although the assumption of normality underlying many models for financial market behaviour has often been remarked to be inadequate, still the associated methodology has been retained due to the lack of adequate replacements of the normal assumption. Since the SN distribution reproduces various formal properties of the normal, Adcock (2004) has shown how one can maintain standard formulations of market behaviour while adopting a more realistic distributional assumption.

Another relevant area for financial applications are special time series formulations, such as GARCH or stochastic volatility models. The replacement of normal assumption with skew-normal one has been considered by various authors: Goria (1999), Pietrobon (2003), De Luca & Loperfido (2004), Cappuccio et al. (2004).

Others The above list of application areas is by no means exhaustive. Many others exist which unfortunately we cannot examine here in any detail for space reasons. Luckily, quite a few of them are collected in the recent book edited by Genton (2004a), hence accessible collectively.

The overall message emerging from all these contributions demonstrates the lively activity and the relevant potential of this area of research, both on the theoretical and on the applied side.

Acknowledgements

I am grateful to Antonella Capitanio, Tom DiCiccio, Samuel Kotz, Marc Genton and Anna Clara Monti for various remarks on earlier versions of the paper. This work is based on review talk presented at the 20$^{\text{th}}$ Nordic Conference on Mathematical Statistics, held in Jyväskylä in June 2004; I would like to thank the organisers of the conference for the invitation leading eventually to the compilation of the paper.

References

Adcock, C. (2004). Capital asset pricing in UK stocks under the multivariate skew-normal distribution. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 11, (pp. 191–204). Chapman & Hall/CRC.

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

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Aitchison, J. & Bacon-Shone, J. (1999). Convex linear combinations of compositions. *Biometrika*, 86, 351–364.

Aitchison, J., Mateu-Figueras, G., & Ng, K. W. (2003). Characterization of distributional forms for compositional data and associated distributional tests. *Math. Geol.*, 35(6), 667–680.

Anděl, J., Netuka, I., & Zvára, K. (1984). On threshold autoregressive processes. *Kybernetika*, 20, 89–106. Academia, Praha.

Arellano-Valle, R. B. & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scand. J. Statist.*, 33, to appear.

Arellano-Valle, R. B., Bolfarine, H., & Lachos, V. H. (2005a). Skew-normal linear mixed models. *Journal of Data Science*, 3, to appear.

Arellano-Valle, R. B., del Pino, G., & San Martín, E. (2002). Definition and probabilistic properties of skew-distributions. *Statist. Probab. Lett.*, 58(2), 111–121.

Arellano-Valle, R. B. & Genton, M. G. (2005). On fundamental skew distributions. *J. Multivariate Anal.*, 96, 93–116.

Arellano-Valle, R. B., Ozán, S., Bolfarine, H., & Lachos, V. H. (2005b). Skew-normal measurement error models. *J. Multivariate Anal.*, 96, 265–281.

Arnold, B. C. & Beaver, R. J. (2000a). Hidden truncation models. *Sankhyā, ser. A*, 62(1), 22–35.

Arnold, B. C. & Beaver, R. J. (2000b). The skew-Cauchy distribution. *Statist. Probab. Lett.*, 49, 285–290.

Arnold, B. C. & Beaver, R. J. (2000c). Some skewed multivariate distributions. *Amer. J. of Mathematical and Management Sciences*, 20, 27–38.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A., & Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, 58(3), 471–478.

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

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, XLVI(2), 199–208.

Azzalini, A. (2001). A note on regions of given probability of the skew-normal distribution. *Metron*, LIX(3–4), 27–34.

Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distributions. *J. R. Stat. Soc., ser. B*, 61(3), 579–602.

Azzalini, A. & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *J. R. Stat. Soc., ser. B*, 65(2), 367–389.

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

Birnbaum, Z. W. (1950). Effect of linear truncation on a multinormal population. *Ann. Math. Statist.*, 21, 272–279.

Branco, M. D. & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.*, 79(1), 99–113.

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

Cappuccio, N., Lubian, D., & Raggi, D. (2004). MCMC Bayesian estimation of a skew-GED stochastic volatility model. *Studies in nonlinear dynamics and econometrics*, 8(2). `http://www.bepress.com/snde/vol8/iss2/art6`.

Chiogna, M. (1997). *Notes on estimation problems with scalar skew-normal distributions.* Technical Report 15, University of Padua, Dept Statistical Sciences.

Chiogna, M. (1998). Some results on the scalar skew-normal distribution. *J. Ital. Statist. Soc*, 7, 1–13.

Cook, R. D. & Weisberg, S. (1994). *An introduction to regression graphics.* John Wiley & Sons.

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

De Luca, G. & Loperfido, N. M. R. (2004). A skew-in-mean GARCH model. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 12, (pp. 205–222). Chapman & Hall/CRC.

DiCiccio, T. J. & Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *J. Amer. Statist. Assoc.*, 99(466), 439–450.

Domínguez-Molina, J. A., González-Farías, G., & Ramos-Quiroga, R. (2004). Skew-normality in stochastic frontier analysis. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 13, (pp. 223–242). Chapman & Hall/CRC.

Durio, A. & Nikitin, Y. Y. (2003). Local Bahadur efficiency of some goodness-of-fit tests under skew alternatives. *J. Statist. Plann. Inference*, 115(1), 171–179.

Ellison, B. E. (1964). Two theorems for inferences about the normal distribution with applications in acceptance sampling. *J. Amer. Statist. Assoc.*, 59, 89–95.

Fang, B. Q. (2003). The skew elliptical distributions and their quadratic forms. *J. Multivariate Anal.*, 87(2), 298–314.

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

Genton, M. G., Ed. (2004a). *Skew-elliptical distributions and their applications: a journey beyond normality*. Chapman & Hall/CRC.

Genton, M. G. (2004b). Skew-symmetric and generalized skew-elliptical distributions. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 5, (pp. 81–100). Chapman & Hall/CRC.

Genton, M. G., He, L., & Liu, X. (2001). Moments of skew-normal random vectors and their quadratic forms. *Statist. Probab. Lett.*, 51, 319–325.

Genton, M. G. & Loperfido, N. (2005). Generalized skew-elliptical distributions and their quadratic forms. *Ann. Inst. Statist. Math.*, 57, 389–401.

González-Farías, G., Domínguez-Molina, J. A., & Gupta, A. K. (2004a). Additive properties of skew normal random vectors. *J. Statist. Plann. Inference*, 126, 521–534.

González-Farías, G., Domínguez-Molina, J. A., & Gupta, A. K. (2004b). The closed skew-normal distribution. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 2, (pp. 25–42). Chapman & Hall/CRC.

Goria, S. (1999). Uso di generalizzazioni della distribuzione normale nei modelli per serie storiche. Tesi di laurea, Facoltà di Scienze Statistiche, Università di Padova, Padova, Italia.

Gualtierotti, A. (2004). A family of (skew-normal) stochastic processes that can model some non-Gaussian random signals in dependent Gaussian noise. In *8th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2004)*, volume VI (pp. 88–94). Orlando, Florida.

Gupta, A. K. (2003). Multivariate skew t-distribution. *Statistics*, 37(4), 359–363.

Gupta, A. K., Chang, F. C., & Huang, W.-J. (2002). Some skew-symmetric models. *Random Operators and Stochastic Equations*, 10(2), 133–140.

Gupta, A. K. & Huang, W.-J. (2002). Quadratic forms in skew normal variates. *J. Math. Anal. Appl.*, 273, 558–564.

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

Henze, N. (1986). A probabilistic representation of the 'skew-normal' distribution. *Scand. J. Statist.*, 13, 271–275.

Hill, M. A. & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics*, 38, 377–396.

Kim, H.-M. & Mallick, B. K. (2003). Moments of random vectors with skew t distribution and their quadractic forms. *Statist. Probab. Lett.*, 63(4), 417–423.

Kim, H.-M. & Mallick, B. K. (2004). A Bayesian prediction using the skew Gaussian distribution. *J. Statist. Plann. Inference*, 120(1–2), 85–101.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the t-distribution. *J. Amer. Statist. Assoc.*, 84, 881–896.

Liseo, B. (1990). La classe delle densità normali sghembe: aspetti inferenziali da un punto di vista bayesiano. *Statistica*, L, 59–70.

Liseo, B. (2004). Skew-elliptical distributions in Bayesian inference. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 9, (pp. 153–171). Chapman & Hall/CRC.

Liseo, B. & Loperfido, N. (2003a). A Bayesian interpretation of the multivariate skew-normal distribution. *Statist. Probab. Lett.*, 61, 395–401.

Liseo, B. & Loperfido, N. (2003b). Integrated likelihood inference for the shape parameter of the multivariate skew-normal distribution. In *Proceedings of the 2003 Meeting of the Italian Statistical Society.*

Liseo, B. & Loperfido, N. (2004). Default Bayesian analysis of the skew-normal distribution. *J. Statist. Plann. Inference*, (pp. to appear).

Loperfido, N. (2001). Quadratic forms of skew-normal random vectors. *Statist. Probab. Lett.*, 54, 381–387.

Loperfido, N. (2002). Statistical implications of selectively reported inferential results. *Statist. Probab. Lett.*, 56(1), 13–22.

Loperfido, N. M. R. (2004). Generalized skew-normal distributions. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 4, (pp. 65–80). Chapman & Hall/CRC.

Ma, Y. & Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scand. J. Statist.*, 31, 459–468.

Ma, Y., Genton, M. G., & Davidian, M. (2004). Linear mixed models with flexible generalized skew-elliptical random effects. In M. G. Genton (Ed.), *Skew-elliptical distributions and their applications: a journey beyond normality* chapter 20, (pp. 339–358). Chapman & Hall/CRC.

Mateu-Figueras, G. (2003). *Models de distribució sobre el símplex*. PhD thesis, Universitat Politècnica de Catalunya, Barcelona.

Monti, A. C. (2003). A note on the estimation of the skew normal and the skew exponential power distributions. *Metron*, XLI(2), 205–219.

Mukhopadhyay, S. & Vidakovic, B. (1995). Efficiency of linear bayes rules for a normal mean: skewed prior class. *J. R. Stat. Soc., ser. D*, 44, 389–397.

Muliere, P. & Nikitin, Y. (2002). Scale-invariant test of normality based on Polya's characterization. *Metron*, LX(1–2), 21–33.

Nadarajah, S. & Kotz, S. (2003). Skewed distributions generated by the normal kernel. *Statist. Probab. Lett.*, 65(3), 269–77.

Nelson, L. S. (1964). The sum of values from a normal and a truncated normal distribution. *Technometrics*, 6, 469–471.

O'Hagan, A. & Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63, 201–202.

Owen, D. B. (1956). Tables for computing bivariate normal probabilities. *Ann. Math. Statist.*, 27, 1075–1090.

Pewsey, A. (2000a). Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics*, 27, 859–770.

Pewsey, A. (2000b). The wrapped skew-normal distribution on the circle. *Commun. Statist. – Theory & Methods*, 29(11), 2459–2472.

Pietrobon, E. (2003). Modelli ARCH, GARCH e SV con componenti di errore aventi distribuzione t asimmetrica. Tesi di laurea, Facoltà di Scienze Statistiche, Università di Padova, Padova, Italia.

Roberts, C. (1966). A correlation model useful in the study of twins. *J. Amer. Statist. Assoc.*, 61, 1184–1190.

Rotnitzky, A., Cox, D. R., Bottai, M., & Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, 6(2), 243–284.

Sahu, K., Dey, D. K., & Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canad. J. Statist.*, 31(2), 129–150.

Salvan, A. (1986). Test localmente più potenti tra gli invarianti per la verifica dell'ipotesi di normalità. In *Atti della XXXIII Riunione Scientifica della Società Italiana di Statistica*, volume II (pp. 173–179). Bari: Cacucci.

Sartori, N. (2005). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J. Statist. Plann. Inference*, (pp. in press). doi:10.1016/j.jspi.2005.08.043.

Serfling, R. (2004). Multivariate symmetry and asymmetry. In S. Kotz, N. Balakrishnan, C. B. Read, & B. Vidakovic (Eds.), *Encyclopedia of Statistical Sciences*, volume in press. J. Wiley & Sons, II edition.

Subbotin, M. T. (1923). On the law of frequency of error. *Matematicheskii Sbornik*, 31, 296–301.

Tancredi, A. (2002). *Accounting for heavy tails in stochastic frontier models.* Working paper 16, Dipartimento di Scienze Statistiche, Università di Padova, Padova, Italia.

Wang, J., Boyer, J., & Genton, M. G. (2004). A skew-symmetric representation of multivariate distribution. Statistica Sinica, 14, 1259–1270.

Weinstein, M. A. (1964). The sum of values from a normal and a truncated normal distribution. *Technometrics*, 6, 104–105.

Zacks, S. (1981). *Parametric statistical inference.* Oxford: Pergamon Press.