International Statistical Review (2008), 76, 1, 106–129 doi:10.1111/j.1751-5823.2007.00016.x

Robust Likelihood Methods Based on the Skew-*t* and Related Distributions

Adelchi Azzalini[1] and Marc G. Genton[2,3]

[1] *University of Padua, Italy. E-mail: azzalini@stat.unipd.it*
[2] *University of Geneva, Geneva, Switzerland. E-mail: Marc.Genton@metri.unige.ch*
[3] *Texas A&M University, College Station, TX, USA. E-mail: genton@stat.tamu.edu*

Summary

The robustness problem is tackled by adopting a parametric class of distributions flexible enough to match the behaviour of the observed data. In a variety of practical cases, one reasonable option is to consider distributions which include parameters to regulate their skewness and kurtosis. As a specific representative of this approach, the skew-*t* distribution is explored in more detail and reasons are given to adopt this option as a sensible general-purpose compromise between robustness and simplicity, both of treatment and of interpretation of the outcome. Some theoretical arguments, outcomes of a few simulation experiments and various wide-ranging examples with real data are provided in support of the claim.

Key words: Kurtosis; maximum likelihood; multivariate distributions; profile likelihood; robustness; singular information matrix; skewness.

1 Introduction

1.1 Motivation and General Remarks

One approach to the robust estimation problem is based on the introduction of a parametric class of probability distributions whose tail behaviour can be regulated by one of the component parameters, hence allowing to accommodate the presence of possible outliers by a suitably chosen tail parameter. Two well-known formulations which follow this approach are Box & Tiao (1973) and Lange *et al.* (1989); as a reference distribution for data modelling, they adopt the so-called exponential power distribution (Subbotin, 1923) and the Student's *t*, respectively. The use of these distributions in the data-fitting process goes back however many years; see Box & Tiao (1973) and Lange *et al.* (1989) for references to earlier work.

An important advantage of this sort of formulation with respect to other approaches to robustness is an explicit statement of the probabilistic setting, leading to a clearer interpretation of the results, as compared, for instance, to M-estimators. This aspect becomes especially relevant in the case of asymmetric distribution of the outlying observations. While most robustness studies employ symmetric error distributions with heavy tails but are otherwise regular, there is empirical evidence that, when present in real data, outlying observations display other forms of departures from the normal distribution. An important one is that real data are often asymmetrically

distributed, as reported by Hill & Dixon (1982); see also Stigler (1977). Under asymmetric data distribution, the quantity to which an M-type estimator converges when the sample size diverges is in general not available in an explicit form, since it is given by the solution of an equation involving the expected value of a non-linear transformation of the observed random variable; see, for instance, Huber (1981, pp. 132–133). Only under symmetry of the underlying distribution and of the psi-function employed, can one state immediately that this solution is the centre of symmetry of the distribution.

On the contrary, the maximum likelihood estimates (MLEs) converge to clearly defined quantities, namely the parameters of the specified class of distributions. There is, of course, the issue of adequacy of the fitted parametric class, to ensure that the minimal Kullback–Leibler discrepancy between the actual data distribution and the chosen parametric class is as small as possible, ideally zero. This requirement can be tackled with the adoption of a sufficiently flexible class of distributions.

The last sentence opens immediately the subsequent issue: how much flexible? The formulation to be adopted in the rest of the paper allows selection of any degree of flexibility between a "base" distribution (the normal one, say) and a non-parametric approach. In practice, one will often choose an intermediate option, but choice of the degree of flexibility does not have a precise answer, especially in general terms, without reference to a specific context. However, as a general-purpose strategy, a plausible option is to select a class of distributions which include parameters to regulate skewness and kurtosis, and this is the direction that we will explore in greater detail. For an alternative route, based on a non-parametric form of perturbation of a parametric "base" distribution of elliptical type, see Ma *et al.* (2005) and Ma & Hart (2007).

To develop the above-described program, we work within the general framework of distributions whose probability density function is, up to a location parameter, of type

$$f(x) = 2\, f_0(x)\, G\{w(x)\}, \qquad x \in \mathbb{R}^d, \tag{1}$$

where $f_0(\cdot)$ is a density function in $\mathbb{R}^d$, symmetric in the sense that $f_0(x) = f_0(-x)$ for all $x \in \mathbb{R}^d$, $G(\cdot)$ is a one-dimensional cumulative distribution function whose derivative G' exists and satisfies $G'(x) = G'(-x)$, and $w(\cdot)$ is a real-valued odd function in $\mathbb{R}^d$, hence $w(-x) = -w(x) \in \mathbb{R}$ for all $x \in \mathbb{R}^d$. The term "symmetric" will be used throughout this paper in the sense just indicated. The essence of the proof that (1) produces a proper density function is contained in the following simple argument: if X and Y are independent random variables of dimension 1 and d, with densities G' and f_0, respectively, then $X - w(Y)$ has a symmetric density such that $2\mathbb{P}\{X - w(Y) \le 0\} = 1$; the complete argument is given in Proposition 1 of Azzalini & Capitanio (2003). For an overview of the work developed in connection with (1), we refer the reader to the book edited by Genton (2004) and the review paper of Azzalini (2005); here we will only mention results of direct relevance to the present paper.

1.2 Some Skewed Distributions and Related Inferential Issues

The basic case of (1) is obtained for $d = 1$ on setting $f_0(x) = \phi(x)$ and $G = \Phi$, the $N(0, 1)$ density function and distribution function, respectively, and $w(x)$ of the linear type. With the addition of a location parameter, ξ, and scale parameter, ω, we obtain the univariate skew-normal (SN) distribution studied by Azzalini (1985), that is,

$$f_{\mathrm{SN}}(x; \xi, \omega, \alpha) = 2\omega^{-1}\, \phi(z)\, \Phi(\alpha z), \qquad x \in \mathbb{R}, \tag{2}$$

where $z = \omega^{-1}(x - \xi)$ and $\alpha \in \mathbb{R}$ is a shape parameter; on setting $\alpha = 0$ we return to the normal distribution.

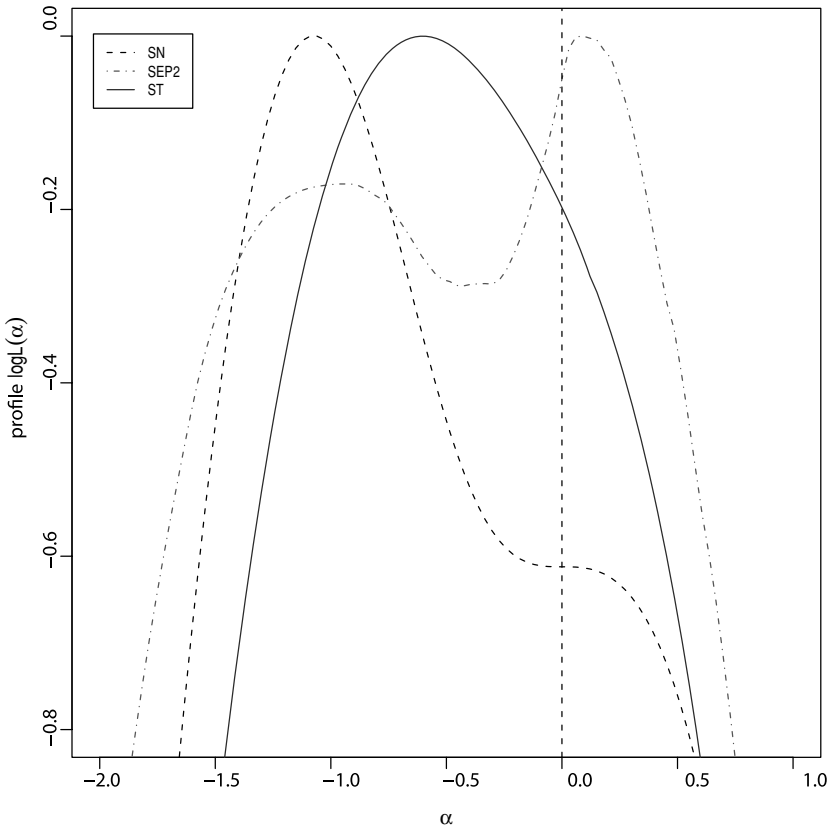

Figure 1. *Profile log-likelihoods for the shape parameter α of the univariate SN, SEP2, and ST distributions fitted to the variable Ht of the AIS data.*

Although, as stated in Section 1.1, we will focus on other forms of (1), notably those where $f_0(x)$ is a density with heavy tails, it is, however, appropriate to digress slightly to recall some problematic aspects connected to inference for (2), as discussed by Azzalini (1985), Azzalini & Capitanio (1999, Section 5), Pewsey (2000) and Chiogna (2005), because of their connection with our subsequent developments.

One of these problematic aspects is that the profile log-likelihood for the shape parameter α of the univariate SN distribution always has a stationary point at $\alpha = 0$, leading to the singularity of the Fisher information matrix of the three parameters in (2). Except for very peculiar cases, this stationary point at $\alpha = 0$ does not correspond to the maximum of the profile log-likelihood function, and it really is a saddle point. This issue is now illustrated on data collected at the Australian Institute of Sport (AIS) on 202 athletes and described by Cook & Weisberg (1994). Figure 1 depicts the profile log-likelihood for the shape parameter α of the univariate SN distribution (dashed curve) fitted to the variable describing the height (*Ht*) of the athletes. The vertical dashed line at $\alpha = 0$ identifies the stationary point. This problem can be circumvented by using a suitable reparametrization, which has been examined by Azzalini (1985) and Chiogna (2005) in the scalar case. Pewsey (2006) has shown that the aforementioned problem is not unique to the univariate SN distribution. It also occurs for other univariate skewed distributions based on a normal kernel ϕ and skewing functions of the form $G(\alpha x)$, if G satisfies the conditions for (1) and in addition $G''(0)$ exists, hence it is 0.

Another peculiar aspect connected to the inference on the parameters of (2) is that the MLE of α can take on boundary values $\pm\infty$, sometimes even for samples not exhibiting extreme skewness. The probability of this event vanishes as the sample size increases, but it is non-negligible for small sample size, especially if $|\alpha|$ is large.

Bearing in mind the discussion of Section 1.1, an appropriate choice for $f_0(x)$ in (1) is the exponential power distribution (Subbotin, 1923). In order to model simultaneously skewness and heavy tails, Azzalini (1986) introduced the univariate skew-exponential power (SEP) distribution, in two variant forms. We consider the type-II version (SEP2) with density

$$f_{\mathrm{SEP2}}(x;\xi,\omega,\alpha,\psi) = 2\,C_\psi\,\omega^{-1}\,\exp\{-|z|^{2\psi}/(2\psi)\}\,\Phi\{\mathrm{sgn}(\alpha z)|\alpha z|^\psi/\sqrt{\psi}\}, \quad x\in\mathbb{R}, \quad (3)$$

where $C_\psi = [2(2\psi)^{1/(2\psi)-1}\Gamma\{1/(2\psi)\}]^{-1}$, $z=\omega^{-1}(x-\xi)$, $\xi\in\mathbb{R}$, $\omega>0$, $\alpha\in\mathbb{R}$, and $\psi>0$. The skewness is regulated by the shape parameter α and the tails of the distribution are controlled by ψ. The normal distribution is obtained when $\alpha=0$ and $\psi=1$. DiCiccio & Monti (2004) have recently investigated likelihood inference for the univariate SEP2 distribution, with only a slight change in the parameterization. In particular, they have proved that in the normal case, corresponding to $\alpha=0$ and $\psi=1$, the Fisher information matrix of the four parameters in (3) is singular. This problem is similar to the one described above for the SN distribution.

Figure 1 depicts the profile log-likelihood for the shape parameter α of the univariate SEP2 distribution (dashed-dotted curve) fitted to the variable *Ht* of the AIS data. The rather unpleasant shape of the SEP2 profile log-likelihood function exhibited here is not a unique case. Although it does not appear in all possible samples, it is not rare at all to encounter a multimodal SEP2 profile log-likelihood, sometimes with an even more irregular shape. Another problem with the multivariate exponential power is the lack of closure of this family under marginalization, which makes it not an appealing model for multivariate data. On the other hand, the exponential power, hence the SEP, enjoys the useful feature of allowing both lighter as well as heavier tails compared to the normal ones, depending on whether $\psi<1$ or $\psi>1$. Although there are arguments in both directions, it seems to us that the disadvantages of the SEP family overtake the advantages, at least if we do not have a specific context in mind, and prefer to move in another direction.

An alternative option for modelling skewness and heavy tails is to consider a skew-*t* distribution, which arises when $f_0(x)$ in (1) is a Student's *t* density function. The specific form which we will consider is the one studied by Branco & Dey (2001, 2002) and in equivalent forms by Azzalini & Capitanio (2003) and Gupta (2003); see also Kim & Mallick (2003) for additional results. Other forms of skew-*t* distribution have been considered by Jones & Faddy (2003), Sahu *et al.* (2003) and Ma & Genton (2004). We follow the notation of Azzalini & Capitanio (2003) for the skew-*t* (ST) distribution whose density in the univariate case takes the form

$$f_{\mathrm{ST}}(x;\xi,\omega,\alpha,\nu) = 2\omega^{-1}\,t(z;\nu)\,T(\alpha z\sqrt{(\nu+1)/(\nu+z^2)};\nu+1), \qquad x\in\mathbb{R}, \qquad (4)$$

where $z=\omega^{-1}(x-\xi)$, t and T denote the univariate standard Student t density function and distribution function, respectively, and $\xi\in\mathbb{R}$, $\omega>0$, $\alpha\in\mathbb{R}$, and ν denote the degrees of freedom. The skewness is regulated by the shape parameter α and the tails of the distribution are controlled by ν. The regular t distribution is recovered by setting $\alpha=0$; the SN distribution is obtained when $\nu\to\infty$ and the normal distribution when both $\alpha=0$ and $\nu\to\infty$.

Figure 1 depicts the profile log-likelihood for the shape parameter α of the univariate ST distribution (solid curve) fitted to the variable *Ht* of the AIS data. This curve appears to be very well behaved with a single maximum and no stationary point at the origin $\alpha=0$. The absence of a stationary point at $\alpha=0$ is not specific to this example. Indeed, it appeared regularly in all numerical examples of Azzalini & Capitanio (2003), and remarked as a welcome feature, but no theoretical support of this empirical fact has been provided. Confirmation of this property is

important to remove or at least alleviate the need for an alternative parameterization, as the one required for the SN case, leading to a simpler inference.

To anticipate the overall content of the paper, this can be schematically summarized as follows. We argue that the univariate ST distribution (4) and its multivariate extension represent a flexible general-purpose class of distributions, suitable to handle adequately a variety of practical cases. In addition, we show that the ST family alleviates some of the inferential problems associated with similar families, specifically the SN and the SEP2 families. In particular, we provide evidence that the Fisher information matrix associated with the univariate ST distribution (4) is never singular at $\alpha = 0$ for finite values of the degrees of freedom ν. Moreover, some numerical work supports the conjecture that the same property carries on in higher dimensions.

A good portion of this paper is largely concerned with maximum likelihood inference for the parameters of the ST distribution, in the univariate and in the multivariate case. In particular, the properties of the profile log-likelihood of the shape parameters are studied in Section 2. The finite-sample performance of the MLEs in the univariate and bivariate case are investigated with simulations in Section 3. Examples of application of the ST model in regression problems, time series, spatio-temporal modelling, and classification, are reported in Section 4. The paper concludes with a discussion in Section 5. Proofs are given in the Appendix.

2 Likelihood Inference

2.1 SN and Other Skewed Variants of the Normal Distribution

As mentioned in the Introduction, some of the distributions of type (1) lead to a profile log-likelihood function with some peculiar behaviour at the point $\alpha = 0$. One aspect of this phenomenon is that the profile log-likelihood function always has a stationary point at $\alpha = 0$, irrespective of the sample observations. A connected unusual feature is that the expected Fisher information is singular at $\alpha = 0$, although the estimation problem is identifiable.

These facts have been noticed by Azzalini (1985) for the basic situation of a simple random sample from (2), and they have been extended by Pewsey (2006) to the case where the term Φ in (2) is replaced by any scalar distribution function G such that G' is a symmetric density. Further extensions in various directions are possible, some of which are discussed in the rest of this section.

One of these extensions refers to the family of flexible skew-symmetric (FSS) distributions introduced by Ma & Genton (2004) with density

$$f_{\mathrm{FSS}}(x; \xi, \omega, \alpha_1, \ldots, \alpha_K) = 2\omega^{-1} f_0(z)\, G\{P_K(z)\}, \qquad x \in \mathbb{R}, \tag{5}$$

where f_0 and G are symmetric univariate density function and distribution function, respectively, $P_K(x) = \alpha_1 x + \alpha_3 x^3 + \cdots + \alpha_K x^K$ is an odd polynomial of degree K (i.e. a polynomial including only terms of odd degree), $z = \omega^{-1}(x - \xi)$, $\xi \in \mathbb{R}$, $\omega > 0$, and $\alpha_1, \alpha_3, \ldots, \alpha_K \in \mathbb{R}$ are shape parameters. The symmetric density function f_0 is obtained when $\alpha_1 = \alpha_3 = \cdots = \alpha_K = 0$. If a random sample $y_1, \ldots, y_n$ is drawn from a random variable with density (5), then the corresponding log-likelihood function is

$$\ell(\xi, \omega, \alpha_1, \ldots, \alpha_K) = \text{constant} - n \log \omega + \sum_{i=1}^{n} \log f_0(z_i) + \sum_{i=1}^{n} \log G\{P_K(z_i)\} \tag{6}$$

where $z_i = \omega^{-1}(y_i - \xi)$ for $i = 1, \ldots, n$.

The following result shows that the FSS distribution has a stationary point in the profile log-likelihood for the shape parameters when the "base" function f_0 is the normal density and the

shape parameters are all equal to zero. It therefore represents an extension of the similar result of Pewsey (2006) from the case where the argument of G is $\alpha\, x$ to the case of an odd polynomial. Its proof is given in the Appendix.

PROPOSITION 1. *Denote by $y_1, \ldots, y_n$ a random sample of size $n \geq 3$ from an FSS distribution with density function given by (5) when $f_0 = \phi$, the standard normal density. Assume G is a continuously differentiable symmetric univariate distribution function. If we denote the sample mean by $\bar{y}$ and the sample variance by s^2, then*

(1) $\xi = \bar{y}, \omega = s, \alpha_1 = \alpha_3 = \cdots = \alpha_K = 0$, is a solution to the score equations for (6);
(2) with the additional assumption that G'' is continuous, the expected Fisher information matrix is singular when $\alpha_1 = \alpha_3 = \cdots = \alpha_K = 0$; moreover, if $K = 1$, then the observed Fisher information matrix is singular when $\alpha_1 = 0$.

Consider now a different setting, concerning the multivariate skew-normal distribution, which is the d-dimensional extension of (2), denoted by $SN_d(\xi, \Omega, \alpha)$. This has been introduced by Azzalini & Dalla Valle (1996), but here we will adopt the alternative although mathematically equivalent parameterization used by Azzalini & Capitanio (1999), so that the density function at $x \in \mathbb{R}^d$ is

$$f_{\mathrm{SN}}(x; \xi, \Omega, \alpha) = 2\, \phi_d(x - \xi; \Omega)\, \Phi\{\alpha^\top \omega^{-1}(x - \xi)\}, \tag{7}$$

where $\phi_d(x; \Omega)$ denotes the $N_d(0, \Omega)$ density function, ω is the diagonal matrix formed by the square root of the diagonal elements of Ω, and ξ and α are now d-dimensional vectors.

In many cases, when n independent observations are considered, the location parameter ξ_i of the i-th individual is related to a set of p covariates x_i via a linear regression model of the form

$$\xi_i = \beta^\top x_i \tag{8}$$

where β is a $p \times d$ matrix of parameters ($i = 1, \ldots, n$). Denote by X the $n \times p$ matrix whose i-th row is $x_i^\top$ and by y the $n \times d$ matrix whose i-th row is $y_i^\top$ if y_i is the observed response vector generated by (7), for $i = 1, \ldots, n$.

In this setting, Azzalini & Capitanio (1999, Section 6.1) have shown that it is convenient to reparameterize the problem using $\eta = \omega^{-1}\alpha$ in place of α. This device separates the parameters in the expression of the log-likelihood $\ell(\beta, \Omega, \eta)$ in the following sense: for fixed β and η, maximization of ℓ with respect to Ω is equivalent to maximizing the analogous function for normal variates for fixed β, which has the well-known solution

$$\hat{\Omega}_\beta = n^{-1} u_\beta^\top u_\beta$$

where $u_\beta = y - X\beta$. On replacing this expression of $\hat{\Omega}_\beta$ in ℓ, we obtain the profile log-likelihood

$$\ell^*(\beta, \eta) = \text{constant} - \frac{1}{2} n \log |\hat{\Omega}_\beta| + 1_n^\top \zeta_0(u_\beta\, \eta) \tag{9}$$

where $\zeta_0(x) = \log\{2\Phi(x)\}$, and it is intended that the notation $\zeta_0(x)$ when x is a vector denotes the vector formed by the component-wise evaluation of the function. Besides decreasing the dimensionality of the numerical maximization problem, $\ell^*(\beta, \eta)$ has the advantage that its partial derivatives are available in explicit form, offering additional numerical efficiency; these are

$$\frac{\partial \ell^*}{\partial \beta} = X^\top u_\beta\, \hat{\Omega}_\beta^{-1} - X^\top \zeta_1(u_\beta\, \eta)\eta^\top, \tag{10}$$

$$\frac{\partial \ell^*}{\partial \eta} = u_\beta^\top \zeta_1(u_\beta\, \eta), \tag{11}$$

where $\zeta_1(x)$ denotes the derivative of $\zeta_0(x)$.

In addition, it is easy to show, from these partial derivatives, that the point with coordinates $(\tilde\beta, 0)$, where $\tilde\beta = (X^\top X)^{-1} X^\top y$, is a stationary point for (9), under the mild condition that 1_n belongs to the space spanned by the columns of X. In fact, the two summands on the right-hand side of (10) are both 0 at $(\tilde\beta, 0)$; nullity of (11) follows immediately on writing $\zeta_1(u_\beta\, 0) = 1_n 2\phi(0)$ and recalling that $u_{\tilde\beta}^\top 1_n = 0$. We therefore have obtained another extension of the result mentioned at the beginning of this section, namely that $\alpha = 0$ is always a stationary point of the profile log-likelihood from a multivariate SN regression model, irrespectively of the observed y.

Furthermore, the adoption of the reparameterization (9) and the fact that $(\tilde\beta, 0)$ is a stationary point for the profile log-likelihood holds more generally for any family similar to (7) but with $\Phi(\cdot)$ replaced by any other distribution function $G(x)$ such that G' exists and is symmetric about 0. Under this new setting, $\zeta_0(x)$ and $\zeta_1(x)$ in (9)–(11) must denote $\log\{2\,G(x)\}$ and its derivative, respectively; for the rest, the above argument holds unchanged.

Since $\eta = 0$ corresponds to $\alpha = 0$ and a stationary point of ℓ^* corresponds to a stationary point of ℓ, we have the conclusion that the log-likelihood function of a multivariate linear regression model with error structure of type (7), possibly with G in place of Φ, has a stationary point at $(\tilde\beta, \hat\Omega_{\tilde\beta}, 0)$ for all samples, under the above-mentioned mild condition on X. We can summarize our conclusion in the following statement.

PROPOSITION 2. *Consider the case of n independent d-dimensional observations, such that the i-th of these is sampled from the density function*

$$f_{\mathrm{SNG}}(x; \xi_i, \Omega, \alpha) = 2\,\phi_d(x - \xi_i; \Omega)\, G\big\{\alpha^\top \omega^{-1}(x - \xi_i)\big\}, \qquad x \in \mathbb{R}^d, \tag{12}$$

where G is a continuous distribution function with symmetric density, ξ_i follows the multivariate regression model (8) which includes an intercept term ($i = 1, \ldots, n$), and the other ingredients are as in (7). Then the likelihood equations for (β, Ω, α) have a stationary point at $\beta = \tilde\beta$, $\Omega = \hat\Omega_{\tilde\beta}$, $\alpha = 0$.

We have therefore seen various cases where distributions of type (1) give rise to a stationary point of the profile log-likelihood function at $\alpha = 0$. Since, however, all these cases refer to distributions where f_0 in (1) is of normal type, a natural question of interest is whether this unusual behaviour of the profile log-likelihood holds also for $f_0 \neq \phi$.

For simplicity of argument, we restrict our discussion to the case $d = 1$ and density functions of type (5), but the same sort of logic applies more generally. If we consider the likelihood equations associated to (6), evaluated at the point $\alpha_1 = \alpha_3 = \cdots = \alpha_K = 0$, the corresponding value of ξ is required to satisfy

$$\sum_{i=1}^{n} h((y_i - \xi)/\omega) = 0 \tag{13}$$

where $h(x) = -f_0'(x)/f_0(x)$; see (23) and (26) in the Appendix for a detailed derivation. In case $f_0 \equiv \phi$, then $h(x) = x$ and clearly $\hat\xi = \bar y$, the arithmetic mean of the observations, is a solution to (13). A relatively lesser known fact is that the converse is true: essentially, if the sample mean is a solution to (13) for all possible samples, then f_0 is the normal density, under some very mild regularity conditions. This property represents a classical characterization theorem of the normal distribution, whose roots go back to Gauss; see Azzalini & Genton (2007) for a recent account.

The property just mentioned implies that other symmetric distributions besides the normal one cannot have a stationary point at $\alpha = 0$ for all possible samples. It is indeed easily possible to construct counter-examples for a given choice of f_0 and specific patterns of the samples. To this end, consider any symmetric density f_0, such that h is an odd function, and select a sample whose values are symmetrically placed around an arbitrary center of symmetry, m say; this requires that a sample value is equal to m if the sample size is odd. Then it is immediate that m coincides with the sample mean and it satisfies (13), irrespective of the value of ω, even if $f_0 \neq \phi$. However, such a sample pattern occurs with probability 0. All of this supports the idea, although it does not provide a formal proof, that the above-mentioned issue of a stationary point of the profile log-likelihood function at $\alpha = 0$ does not arise with distributions of type (1) when $f_0 \neq \phi$.

On another front, a different sort of problem is, in principle, due to the existence of poles in the log-likelihood function. Fernández & Steel (1999) have proved that this phenomenon arises in a regression context where the error term has a symmetric Student's t distribution with unspecified $\nu \in (0, \infty)$. However, these poles are confined to the interval $(0, \nu_0)$ of ν, where ν_0 depends on the design matrix X and y and it typically is very small compared to practically all relevant values; for instance, in the case of a simple random sample with no ties, $\nu_0 = d/(n-1)$. It is plausible that the same mechanism operates also with the ST distribution, but Azzalini & Capitanio (2003) have reported that these poles are in practice hard to locate and, when located, substantially smaller than the already small upper bound ν_0. Specifically, in one case there was a pole at $\nu = 0.06$ when $\nu_0 = 8/13$, and otherwise the maximum of the log-likelihood was at $\hat{\nu} = 1.14$. From a practical viewpoint, such extreme values like $\nu = 0.06$ are not really relevant, because of the completely peculiar behaviour of the corresponding distribution, and there is no real limitation in excluding values of ν which are so close to 0.

2.2 Multivariate ST Distribution

For dealing with multivariate data, we consider the d-dimensional version of the ST distribution (4), denoted by $ST_d(\xi, \Omega, \alpha, \nu)$. Its density function at $x \in \mathbb{R}^d$, in the form adopted by Azzalini & Capitanio (2003), is

$$f_{\mathrm{ST}}(x; \xi, \Omega, \alpha, \nu) = 2\, t_d(x - \xi; \Omega, \nu)\; T\left\{\alpha^\top \omega^{-1}(x - \xi)\left(\frac{\nu + d}{\nu + Q(x)}\right)^{1/2}; \nu + d\right\}, \quad (14)$$

where $Q(x) = (x - \xi)^\top \Omega^{-1}(x - \xi)$ and

$$t_d(x; \Omega, \nu) = \frac{\Gamma((\nu + d)/2)}{|\Omega|^{1/2}(\nu\pi)^{d/2}\,\Gamma(\nu/2)}\left(1 + \frac{Q(x)}{\nu}\right)^{-(\nu+d)/2}$$

denotes the commonly adopted form of d-dimensional Student's t distribution with 0 location, Ω scale matrix and ν degrees of freedom. Notice that, owing to the form of the "base" function $f_0 = t_d$, there is a single parameter ν to regulate the tail thickness of all components of t_d, hence also of f_{ST}.

From (14), it is immediate to write down the log-likelihood function for a regression model of type (8) and ST error terms. For numerical computation of the MLEs, it is advantageous to make use of the expressions of the derivatives of the log-likelihood, available in the full version of the paper of Azzalini & Capitanio (2003).

Because of the algebraic complication it is difficult to examine precisely the formal properties of this log-likelihood function. It is, however, plausible that the property of non-stationary point at $\alpha = 0$ carries on from the univariate to the multivariate ST. This statement is illustrated

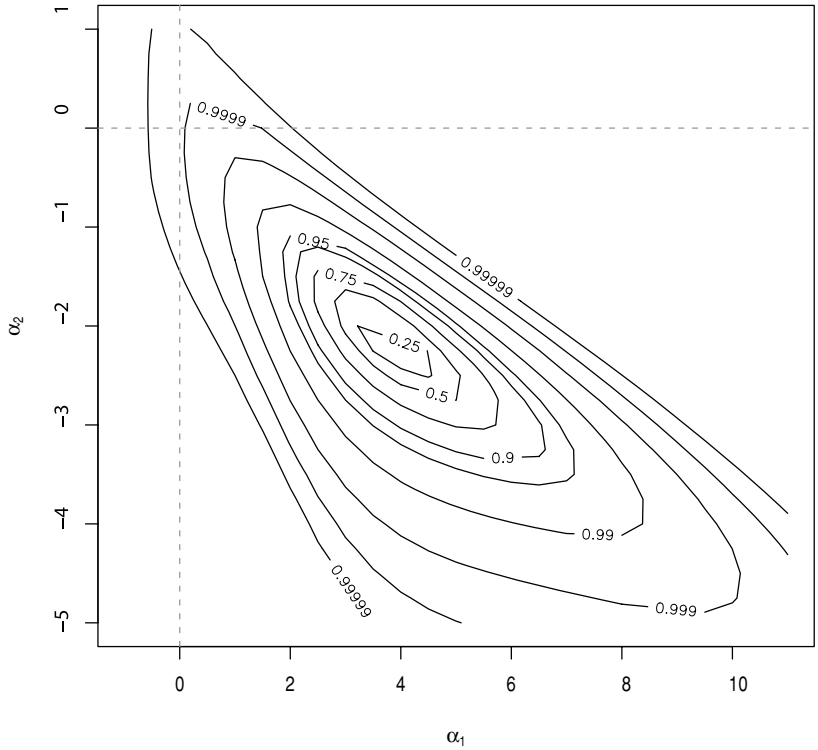

Figure 2. *Profile log-likelihood for the shape parameters α_1 and α_2 of the bivariate ST distribution fitted to the variables (Wt, Ht) of the AIS data. Note that there is no indication of a stationary point at (0,0).*

in Figure 2 which refers to the weight (*Wt*) and height (*Ht*) of the AIS data already used for Figure 1. Consider the profile log-likelihood function

$$\ell^*(\alpha) = \max_{\xi,\Omega,\nu} \ell(\xi, \Omega, \alpha, \nu; Ht, Wt)$$

where $\alpha \in \mathbb{R}^2$ and the corresponding deviance function

$$D(\alpha) = 2\,\{\ell^*(\hat{\alpha}) - \ell^*(\alpha)\}$$

where $\hat{\alpha}$ denotes the MLE of α. This function is displayed in Figure 2 in the form of a set of contour levels. Since these contour levels correspond to percentage points of the χ^2_2 distribution, then each contour level selects a confidence region at the quoted confidence level. The set of confidence levels considered here has been stretched far beyond common values so that the outer curve includes the origin, to illustrate the lack of a stationary point of the "deviance" function at the origin. Also, notice that for more standard values of the confidence level (typically up to 0.95) the shape of the regions is reasonably close to ellipsoids.

3 Monte Carlo Simulation Study

We investigate the finite-sample performance of the MLEs of the univariate and bivariate ST distribution by means of a Monte Carlo simulation study. As indicated in the Introduction, the likelihood function of the SN distribution may sometimes reach its maximum at the boundary of the shape parameter space. For small-to-moderate sample sizes, this problem occurs with

positive probability given in the univariate case with a single shape parameter by

$$\Pr(Z > 0)^n + \Pr(Z < 0)^n, \tag{15}$$

where $Z \sim SN_1(0, 1, \alpha)$ and n is the sample size. For the univariate ST distribution, this problem occurs with the same probability since

$$\Pr(X > 0) = \Pr(Z/D > 0) = \Pr(Z > 0), \tag{16}$$

where $X \sim ST_1(0, 1, \alpha, \nu)$ and D is a suitable positive random variable.

Currently, there are essentially three approaches to deal with this issue. The first one, which will be implemented in this study for the ST distribution, is similar to the procedure proposed by Azzalini & Capitanio (1999). It is based on the fact that when either of the parameters α or ν is large, then the shape of the ST density function remains almost unchanged if either parameter is substantially increased. This suggests replacing the MLE of (α, ν), in samples where it occurred on the boundary, by the smallest value (α_0, ν_0) such that $H_0 : (\alpha, \nu) = (\alpha_0, \nu_0)$ is not rejected by a likelihood ratio test statistic based on the χ^2_{d+1} distribution at a fixed level, q say. This approach, based on the deviance function, is fairly easy to implement, even in the multivariate case.

As a variant form of this "deviance approach," one could decide to apply the same scheme in all cases, not only when the MLE hits the boundary of the parameter space. This option has the advantage of being a more homogeneous approach, at the cost of additional complication even for the larger portion of cases when it is not needed. Indeed, both ways are conceivable; however, for simplicity reasons, we opted for using the MLE when the deviance approach is not needed.

A second approach to the boundary problem of the MLE has recently been proposed by Sartori (2006), and it consists in constructing a modified score function as an estimating equation. The resulting modified maximum likelihood estimator for the shape parameter of the univariate SN distribution has been proved to be always finite. The method has been applied to the univariate ST distribution although only for fixed degrees of freedom. No proof of the finiteness of the resulting shape estimator has been provided in this case. One drawback of this approach is that the multivariate case has currently not been addressed.

Finally, a third approach arises from the Bayesian paradigm and consists in computing Jeffreys' prior distribution for the shape parameter. Liseo & Loperfido (2006) have shown that, in the case of the univariate SN distribution with a single shape parameter, the Jeffreys' prior is proper. In that setting, Bayes & Branco (2007) have shown that the Jeffreys' prior is well approximated by a Student's t distribution. This guarantees the finiteness of the shape estimator resulting from the mode of the posterior distribution. Unfortunately, neither the ST distribution case nor the multivariate setting seems to have been investigated so far.

3.1 Univariate ST Distribution

In this simulation experiment, the parameters of the univariate ST distribution were set to $\xi = 0$, $\omega = 1$, $\alpha = 2, 5$, $\nu = 5, 10$, the sample sizes to $n = 100, 200, 400$, and the levels to $q = 0$, 0.5, 0.9. The "deviance approach" is implemented for situations where either of $|\hat{\alpha}|$ or $\hat{\nu}$ is larger than 100. We used 1,000 simulation replicates. Because the estimators of the shape and degrees of freedom (df) parameters are sometimes infinite, we measure the accuracy of the estimator $\hat{\theta}$ of the parameter θ by

$$m(\hat{\theta}) = \text{median}_{i=1,\ldots,B}\{\hat{\theta}_i - \theta\}, \tag{17}$$

and

$$iqr(\hat{\theta}) = \text{interquartile range}_{i=1,\ldots,B}\{\hat{\theta}_i - \theta\}, \tag{18}$$

Table 1
Univariate ST distribution with $\nu = 5$: simulations with 1,000 replicates of maximum likelihood estimation adjusted with the "deviance approach".

α	n	q	$m(\hat{\xi})$ $iqr(\hat{\xi})$	$m(\hat{\omega})$ $iqr(\hat{\omega})$	$m(\hat{\alpha})$ $iqr(\hat{\alpha})$	$m(\hat{\nu})$ $iqr(\hat{\nu})$	% refit
2	100	0	−0.024	0.039	0.204	0.776	—
			0.234	0.459	1.388	6.631	
2	100	0.5	−0.008	0.015	0.138	0.776	10.6
			0.229	0.449	1.326	6.534	
2	100	0.9	0.008	−0.008	0.043	0.776	10.6
			0.246	0.438	1.386	6.163	
2	200	0	−0.009	0.015	0.090	0.292	—
			0.160	0.308	0.917	3.244	
2	200	0.5	−0.007	0.015	0.087	0.292	3.0
			0.160	0.307	0.902	3.244	
2	200	0.9	−0.006	0.012	0.077	0.292	3.0
			0.162	0.301	0.904	3.244	
2	400	0	−0.006	0.007	0.034	0.130	—
			0.109	0.198	0.633	2.055	
2	400	0.5	−0.006	0.007	0.034	0.130	0.6
			0.109	0.198	0.633	2.055	
2	400	0.9	−0.006	0.007	0.033	0.130	0.6
			0.109	0.198	0.631	2.055	
5	100	0	−0.005	0.034	0.466	0.692	—
			0.109	0.198	0.631	2.055	
5	100	0.5	−0.003	0.031	0.338	0.692	11.8
			0.117	0.337	3.606	6.665	
5	100	0.9	0.002	0.017	0.201	0.668	11.8
			0.117	0.321	3.444	6.149	
5	200	0	0.001	0.011	0.213	0.348	—
			0.079	0.219	2.171	3.099	
5	200	0.5	0.001	0.011	0.194	0.348	3.5
			0.078	0.219	2.157	3.099	
5	200	0.9	0.001	0.010	0.175	0.348	3.5
			0.078	0.218	2.173	3.099	
5	400	0	−0.004	0.011	0.067	0.248	—
			0.051	0.151	1.477	1.745	
5	400	0.5	−0.004	0.011	0.067	0.248	1.4
			0.051	0.151	1.477	1.745	
5	400	0.9	−0.004	0.011	0.067	0.248	1.4
			0.051	0.151	1.477	1.745	

where the number of replicates is $B = 1{,}000$ in our setting. Tables 1 and 2 report the results of the simulation study along with the percentages of refit in the last column, which therefore indicates the fraction of cases where the MLE was modified according to the "deviance approach". This percentage decreases with increasing sample size, but it increases with larger shape and df parameters. There is an appreciable bias only for df, when $n = 100$ or 200. The "deviance approach" sometimes alleviates the bias problem for α and ν for small sample size. The bias on location and scale parameters is usually small. The difference between the "deviance approach" using $q = 0.5$ and $q = 0.9$ is small when the df is small, but becomes more important for large df.

3.2 Bivariate ST Distribution

In this simulation experiment, the parameters of the bivariate ST distribution were set to $\xi_1 = \xi_2 = 0$, $\Omega_{11} = \Omega_{22} = 1$, $\Omega_{12} = 0$, $(\alpha_1, \alpha_2) = (2, 2), (2, 5), (5, 5)$, $\nu = 5$, the sample sizes to $n = 100, 200, 400$, and the levels to $q = 0, 0.5, 0.9$. The "deviance approach" is implemented for

Table 2
Univariate ST distribution with $\nu = 10$: simulations with 1,000 replicates of maximum likelihood estimation adjusted with the "deviance approach".

α	n	q	$m(\hat{\xi})$ $iqr(\hat{\xi})$	$m(\hat{\omega})$ $iqr(\hat{\omega})$	$m(\hat{\alpha})$ $iqr(\hat{\alpha})$	$m(\hat{\nu})$ $iqr(\hat{\nu})$	% refit
2	100	0	−0.016	0.052	0.185	5.920	—
			0.225	0.352	1.349	11989.310	
2	100	0.5	0.018	−0.031	−0.024	2.373	32.9
			0.243	0.396	1.286	15.963	
2	100	0.9	0.069	−0.135	−0.318	0.329	32.9
			0.360	0.439	1.511	6.904	
2	200	0	−0.012	0.030	0.086	1.722	—
			0.164	0.279	0.892	24.271	
2	200	0.5	−0.002	0.014	0.023	1.654	18.5
			0.158	0.285	0.848	13.617	
2	200	0.9	0.015	−0.036	−0.080	0.714	18.5
			0.190	0.292	0.926	7.804	
2	400	0	−0.012	0.019	0.080	1.268	—
			0.122	0.212	0.692	10.102	
2	400	0.5	−0.011	0.017	0.061	1.268	9.0
			0.116	0.209	0.653	9.881	
2	400	0.9	−0.007	0.014	0.040	1.197	9.0
			0.118	0.203	0.625	8.033	
5	100	0	−0.004	0.032	0.535	4.754	—
			0.110	0.275	3.394	12238.742	
5	100	0.5	0.011	0.005	−0.051	4.101	33.8
			0.111	0.311	3.137	30.522	
5	100	0.9	0.026	−0.029	−0.580	1.851	33.8
			0.121	0.296	3.162	15.835	
5	200	0	−0.006	0.021	0.285	2.032	—
			0.071	0.204	2.231	20.259	
5	200	0.5	0.001	0.016	0.095	2.032	16.4
			0.070	0.218	2.151	17.461	
5	200	0.9	0.004	0.007	−0.024	1.696	16.4
			0.072	0.205	2.163	13.502	
5	400	0	−0.002	0.013	0.180	1.394	—
			0.056	0.147	1.490	9.374	
5	400	0.5	−0.002	0.012	0.122	1.394	7.0
			0.056	0.148	1.476	9.374	
5	400	0.9	−0.001	0.010	0.089	1.394	7.0
			0.055	0.147	1.508	9.121	

situations where either of the quantities $\|\hat{\alpha}\|$ or $\hat{\nu}$ is larger than 100. We used 1,000 simulation replicates. Table 3 reports the results of the simulation study along with the percentages of refit in the last column. The percentage of refit is smaller than in the univariate case and there is no refit for $n = 400$, suggesting again that the problem of boundary values disappears for large sample sizes. The overall behaviour of the "deviance approach" in Table 3 is similar to the univariate case, although there seems to be some bias for α in small sample sizes.

4 Examples

4.1 Stack-Loss Data

The stack-loss data presented by Brownlee (1960, pp. 491–500) represent a classical benchmark for the performance of multiple regression procedures, largely in connection with robustness, presence of outliers and related issues. In this context, the data have been used for numerical illustration by a very large number of authors, as documented in detail by Dodge

Table 3
Bivariate ST distribution with $\nu = 5$: simulations with 1,000 replicates of maximum likelihood estimation adjusted with the "deviance approach".

α_1	α_2	n	q	$m(\hat{\xi}_1)$	$m(\hat{\xi}_2)$	$m(\hat{\Omega}_{11})$	$m(\hat{\Omega}_{12})$	$m(\hat{\Omega}_{22})$	$m(\hat{\alpha}_1)$	$m(\hat{\alpha}_2)$	$m(\hat{\nu})$	% refit
2	2	100	0	−0.017	−0.019	0.047	−0.015	0.046	0.349	0.359	0.592	—
2	2	100	0.5	−0.017	−0.019	0.047	−0.015	0.046	0.232	0.267	0.622	4.2
2	2	100	0.9	−0.017	−0.019	0.047	−0.015	0.046	0.232	0.267	0.622	4.2
2	2	200	0	−0.010	−0.002	0.008	−0.007	0.018	0.122	0.143	0.328	—
2	2	200	0.5	−0.010	−0.002	0.008	−0.007	0.018	0.121	0.140	0.328	0.1
2	2	200	0.9	−0.010	−0.002	0.008	−0.007	0.018	0.121	0.140	0.328	0.1
2	2	400	0	−0.002	0.004	0.012	−0.008	0.012	0.044	0.046	0.118	—
2	2	400	0.5	−0.002	0.004	0.012	−0.008	0.012	0.044	0.046	0.118	0
2	2	400	0.9	−0.002	0.004	0.012	−0.008	0.012	0.044	0.046	0.118	0
2	5	100	0	−0.002	−0.001	0.050	−0.001	0.023	0.469	1.099	0.645	—
2	5	100	0.5	−0.002	−0.001	0.050	−0.001	0.023	0.115	0.208	0.880	12.3
2	5	100	0.9	−0.002	−0.001	0.050	−0.001	0.023	0.115	0.208	0.880	12.3
2	5	200	0	0.004	−0.007	0.019	−0.004	0.019	0.135	0.356	0.212	—
2	5	200	0.5	0.004	−0.007	0.019	−0.004	0.019	0.122	0.331	0.215	0.8
2	5	200	0.9	0.004	−0.007	0.019	−0.004	0.019	0.122	0.331	0.215	0.8
2	5	400	0	−0.001	0.002	0.009	0.004	−0.003	0.058	0.106	0.007	—
2	5	400	0.5	−0.001	0.002	0.009	0.004	−0.003	0.058	0.106	0.007	0
2	5	400	0.9	−0.001	0.002	0.009	0.004	−0.003	0.058	0.106	0.007	0
5	5	100	0	0.004	−0.019	0.019	−0.005	0.048	1.562	1.625	0.597	—
5	5	100	0.5	0.004	−0.019	0.019	−0.005	0.048	−0.250	−0.094	1.510	21.0
5	5	100	0.9	0.004	−0.019	0.019	−0.005	0.048	−0.250	−0.094	1.510	21.0
5	5	200	0	0.007	−0.003	0.003	−0.008	0.018	0.400	0.513	0.257	—
5	5	200	0.5	0.007	−0.003	0.003	−0.008	0.018	0.307	0.408	0.271	1.7
5	5	200	0.9	0.007	−0.003	0.003	−0.008	0.018	0.307	0.408	0.271	1.7
5	5	400	0	0.002	−0.002	−0.003	−0.002	0.003	0.208	0.194	0.107	—
5	5	400	0.5	0.002	−0.002	−0.003	−0.002	0.003	0.208	0.194	0.107	0
5	5	400	0.9	0.002	−0.002	−0.003	−0.002	0.003	0.208	0.194	0.107	0

(1996). The data refer to $n = 21$ days of observations on a chemical process for which a variable of interest, $y =$ `Stack-loss`, is related to three other chemical variables, namely $x_1 =$ `Air flow`, $x_2 =$ `Water temperature` and $x_3 =$ `Acid concentration`.

While it is possible to argue that a suitable formulation leads to a generalized linear model with no evidence of outliers (Nelder, 2000), here we are concerned with the comparison of the behaviour of the proposed methodology and similar ones when the fitted models are assumed to be of linear regression type.

Consider a regression model with explanatory variables $X = (1, x_1, x_2, x_3)$ and a response variable y of type

$$y = X\beta + \varepsilon, \tag{19}$$

with error component ε having an ST distribution. All numerical computations, including those described in subsequent sections, have been performed in the `R` computing environment (R Development Core Team, 2006). Model fitting of the ST and SN regression models has been accomplished using the `R` package `sn` (Azzalini 2006).

The parameter estimates and standard errors for model (19) are reported in Table 4. These estimates, except clearly the shape parameter, are very close to those obtained by Lange *et al.* (1989) fitting a similar model with an error term having a symmetric Student's t distribution. Besides the results of Lange *et al.* (1989), a number of other papers mentioned by Dodge (1996) indicate a strongly heavy-tailed distribution of the error term. Since the above value of the shape parameter is close to 0, the other parameters are not changed appreciably with respect to the symmetric t case. The discrepancy of $\hat{\alpha}$ from 0 is so small compared to its standard error that a

Table 4
Stack-loss data: Parameter estimates and standard errors for model (19).

	Constant	Air flow	Water temp.	Acid conc.	Scale	Shape	Df
Estimate	−38.05	0.86	0.48	−0.08	0.98	0.28	1.14
Standard error	3.75	0.06	0.15	0.06	0.44	0.65	0.53

Table 5
Stack-loss data: Summary index of discrepancy Q(p) between observed and fitted values for a linear regression model with three covariates. The minimum for each p is in bold font.

p	0.5	1	2
LS	30.1	49.7	**178.8**
Huber	28.3	46.1	190.5
LTS	25.4	49.4	322.6
ST	**25.0**	**43.4**	240.0

formal test does not seem necessary. While one could remark that for this data set the addition of a skewness parameter has not led to any appreciable improvement of fit with respect to the usual t distribution, it is required that the wider model with ST errors is first fit to the data to reach this conclusion.

In computing the fitted values $\hat{y}$ corresponding to a model of the above type, the skewness of the error component must be taken into account via an expression of type

$$\hat{y} = X\hat{\beta} + M_\varepsilon$$

where $\hat{\beta}$ denotes the estimates of β and M_ε is a suitable quantity which reflects the lack of centring of the error term. In a sense, the obvious choice for M_ε is to set it equal to $\mathbb{E}\{\varepsilon\}$ computed at the estimated values of the fitted ST distribution, since this option leads to an unbiased estimate of the expected value of y for given X, up to the approximation due to replacing the true parameters ω, α, ν, by their estimates. This route is the one taken by Azzalini & Capitanio (2003, Section 5.2) and by Sahu *et al.* (2003, Section 6.2).

There are, however, reasons to consider other options: (1) when the estimate $\hat{\nu}$ of the degrees of freedom does not exceed 1, the above choice of M_ε does not produce a finite value; (2) even if $\hat{\nu}$ is above 1 but it is not much larger, the corresponding estimate of $\mathbb{E}\{\varepsilon\}$ fluctuates very widely. The case under consideration is of the latter type; if $\mathbb{E}\{\varepsilon\}$ is computed for a range of values of ν within one standard error from $\hat{\nu}$, the result ranges from 0.457 to ∞. On the basis of these remarks, a more stable and generally admissible choice for M_ε is advisable. In the subsequent numerical work, M_ε has been taken to be the median of the fitted distribution of ε. For the stack-loss data, the corresponding range of M_ε is from 0.245 to 0.301, when ν varies within one standard error from $\hat{\nu}$.

To compare the performance of the present methodology with some of the main alternatives currently in use, the summary index

$$Q(p) = \sum_{i=1}^{n} |y_i - \hat{y}_i|^p, \qquad (p = 0.5,\ 1,\ 2), \tag{20}$$

has been adopted. The resulting numerical values are shown in Table 5 which, besides the ST fit, reports the corresponding values for other methods, namely classical least squares (denoted LS in the subsequent tables), an M-estimate using the Huber's psi-function (denoted Huber), and the least trimmed sum of squares (denoted LTS) studied by Rousseeuw & Leroy (1987). The

Table 6
Austrian bank interest rates data: Parameter estimates and standard errors for model (21).

	β_0	β_1	Scale	Shape	Df
Estimate	0.18	0.98	0.08	0.15	1.13
Standard error	0.19	0.02	0.02	0.28	0.24

numerical fitting of the Huber and LTS estimates has been accomplished with the aid of the R package MASS associated to Venables & Ripley (2002), keeping the tuning parameters of the procedures at their default values. Clearly, when $p = 2$, the least-squares fit is known beforehand to be superior, but it is useful to have a perception of the relative loss of other methods. Inspection of Table 5 indicates an overall satisfactory behaviour of ST with respect to other methods, when the previous remark is taken into account.

4.2 Austrian Bank Interest Rates Data

These data consist of $n = 91$ monthly interest rates of an Austrian bank. They have been analyzed already by Künsch (1984) and Ma & Genton (2000) in the context of robust time series analysis. Indeed, the series contains three large outliers for the months 18, 28 and 29, and this severely affects the classical model-fitting techniques.

Consider an autoregressive model of order one, AR(1), for $Y(t)$, the interest rate at month t, of type

$$Y(t) = \beta_0 + \beta_1 Y(t-1) + \varepsilon(t)\,, \tag{21}$$

with $\beta_0 \in \mathbb{R}$, $|\beta_1| < 1$, and i.i.d. error components $\varepsilon(t)$ having an ST distribution. This error distribution models possible innovation outliers (Denby & Martin, 1979), but can also downweight the effect of other types of outliers as we show in this section. The corresponding parameter estimates and standard errors are reported in Table 6.

The shape parameter is close to 0, indicating negligible asymmetry in the distribution of the errors. The small df parameter reflects heavy tails due to the outliers. Figure 3 depicts PP-plots for a normal and ST fit to the error components $\varepsilon(t)$. The PP-plot for the ST fit follows the diagonal line very closely, indicating a much better fit than the normal one. Moreover, the inappropriateness of the normal distribution is apparent not only for a few points related to outliers but for the whole set of data points, whereas the ST fit accommodates all points. In computing the fitted values $\hat{Y}(t)$ corresponding to the AR(1) model above, the skewness of the error component must be taken into account via an expression of type

$$\hat{Y}(t) = \hat{\beta}_0 + \hat{\beta}_1 Y(t-1) + M_\varepsilon\,,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimates of β_0 and β_1, and M_ε is the median of the fitted distribution of $\varepsilon(t)$, similarly to the stack-loss data analysis.

The classical least squares (LS) method for fitting an AR(1) model of type (21) to the Austrian bank interest rates data yields $\hat{\beta}_0 = 1.93$ and $\hat{\beta}_1 = 0.79$, which are quite different from the ST fit. Note that compared to the ST fit, the LS estimate of β_1 is pushed towards 0 because of the presence of outliers, a well-known weakness of the LS estimator in the AR(1) time series model; see Genton & Lucas (2003) for further discussions. Künsch (1984) uses optimal robust estimators in the sense that, among all M-estimators with a bound on their influence function, they minimize the trace of the corresponding asymptotic covariance matrix. His method produces the estimates $\hat{\beta}_0 = 0.37$ and $\hat{\beta}_1 = 0.96$, which are quite close to the ST fit. Figure 4 depicts the

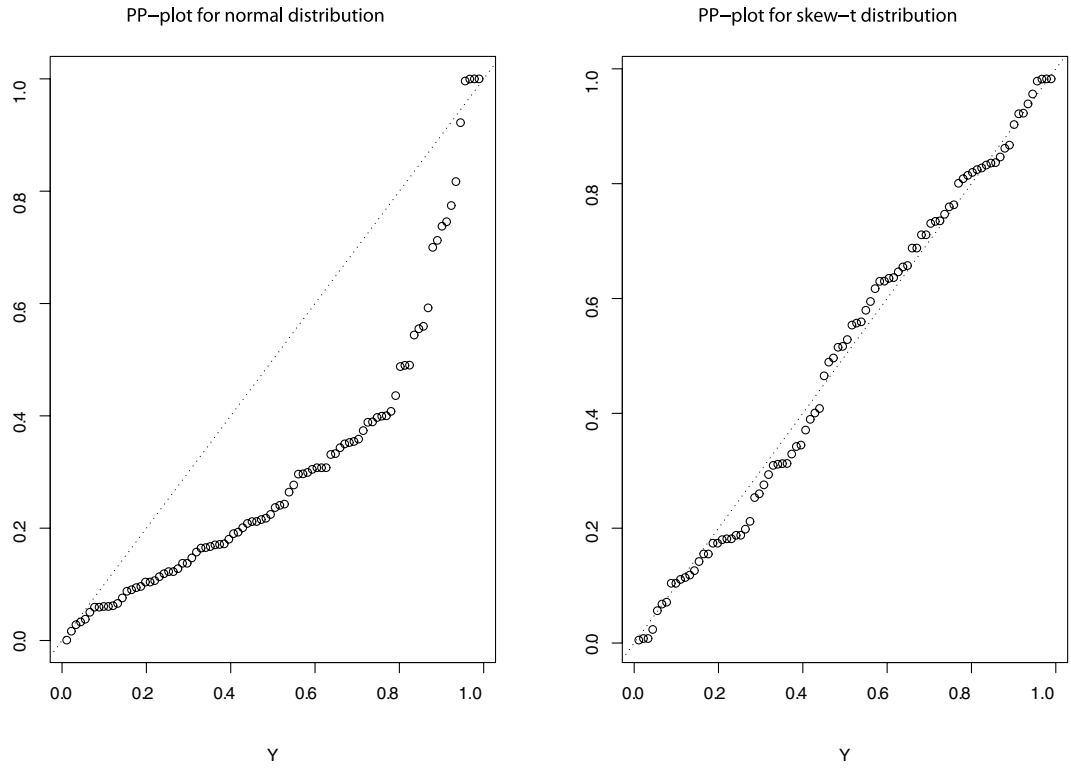

Figure 3. *Austrian bank interest rates data: PP-plots for normal and ST fit.*

autoregression lines fitted by LS, Künsch's optimal robust estimator, and ST. In this case, the ST and the Künsch estimates behave very similarly, while the LS autoregression line is severely influenced by two leverage points in the middle-right area of the plot. From this figure, it is clear that the LS fit does not represent the "main body" of the data well.

To compare the performance of the ST methodology with some of the main alternatives currently in use, the summary index $Q(p)$ defined in (20) has been adopted. The resulting numerical values are reported in Table 7 which, besides the ST fit, reports the corresponding values for Künsch's optimal robust estimator and for least squares. Inspection of Table 7 indicates an overall satisfactory behaviour of ST with respect to other methods.

4.3 Wind Speed Data

Although its contribution to global power production is still slight, the use of wind to generate power is increasingly important from both the financial as well as ecological perspectives. In particular, an ability to forecast wind power production is extremely important to those managing the supply of power over national grids, see Genton & Hering (2007) for a recent discussion. Gneiting *et al.* (2006) proposed a regime-switching space–time model for short-term forecasts of wind speed based on the alternation of westerly and easterly winds near the Stateline wind energy centre in the US Pacific Northwest. We propose to study the spatial distribution of wind speed by means of the ST distribution.

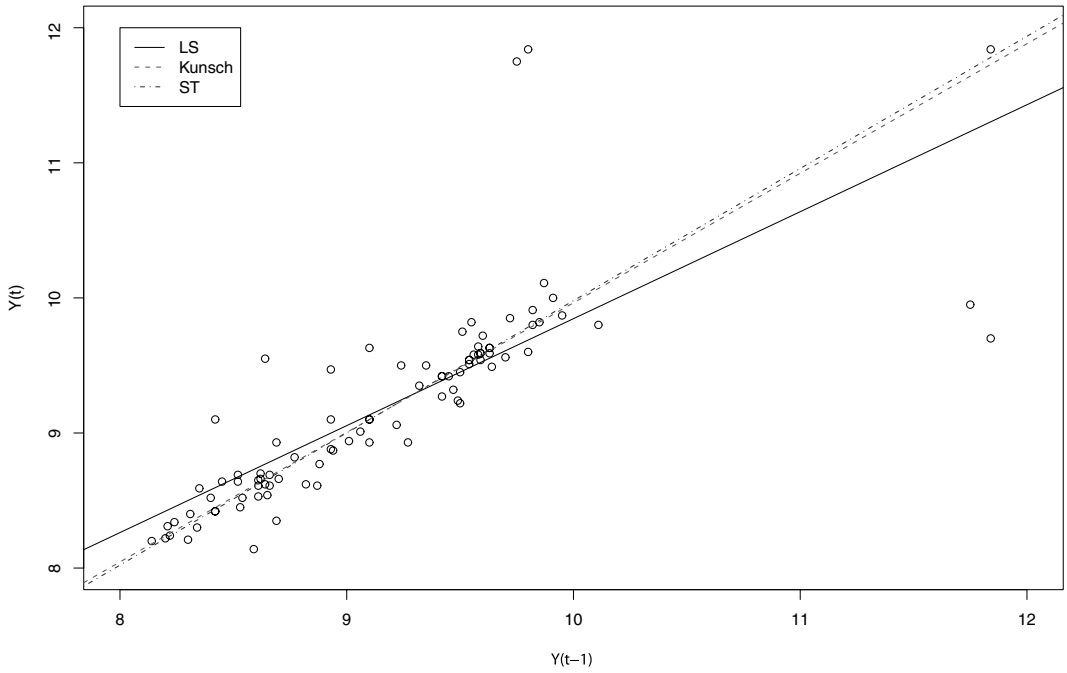

Figure 4. *Austrian bank interest rates data: autoregression lines produced by different methods.*

Table 7
Austrian bank interest rates data: Summary index of discrepancy Q(p) between observed and fitted values for an AR(1) model. The minimum for each p is in bold font.

p	0.5	1	2
LS	37.5	21.5	**17.5**
Künsch	32.1	**18.9**	18.8
ST	**31.8**	**18.9**	19.1

The wind speed data consists of hourly average wind speed collected at three meteorological towers: Goodnoe Hills (gh), Kennewick (kw), and Vansycle (vs). Those towers are approximately located on a line and ordered from west to east. We consider the wind speed from 25 February to 30 November 2003 recorded at midnight, a time when wind speeds tend to peak. In this region, wind patterns are mostly westerly and sometimes easterly. This information is coded in the sign of the wind speed data: a positive sign represents a westerly wind direction, with an angle in the interval $(-\pi/2, \pi/2)$ radians, and a negative sign represents an easterly wind direction, with an angle in the interval $(\pi/2, 3\pi/2)$ radians.

Denote by $Y(t)$ the three-dimensional vector of wind speed at the towers recorded at time $t = 1, \ldots, 278$. A Ljung–Box test indicates some serial correlation at the Goodnoe Hills tower, but not at the other two towers. Consequently, we decide to treat the observations as being independent and identically distributed. We use an ST model for the distribution of $Y(t)$. Figure 5 depicts bivariate scatter plots of the wind speed data at the three towers along with the contours of the fitted ST distribution. The plots reveal both skewness and heavy tails. The indices of skewness of the three univariate distributions obtained by marginalization of the fitted trivariate distribution are $\hat{\gamma}_1 = -3.45$, $\hat{\gamma}_2 = -1.45$, and $\hat{\gamma}_3 = -2.20$ indicating negative skewness at all

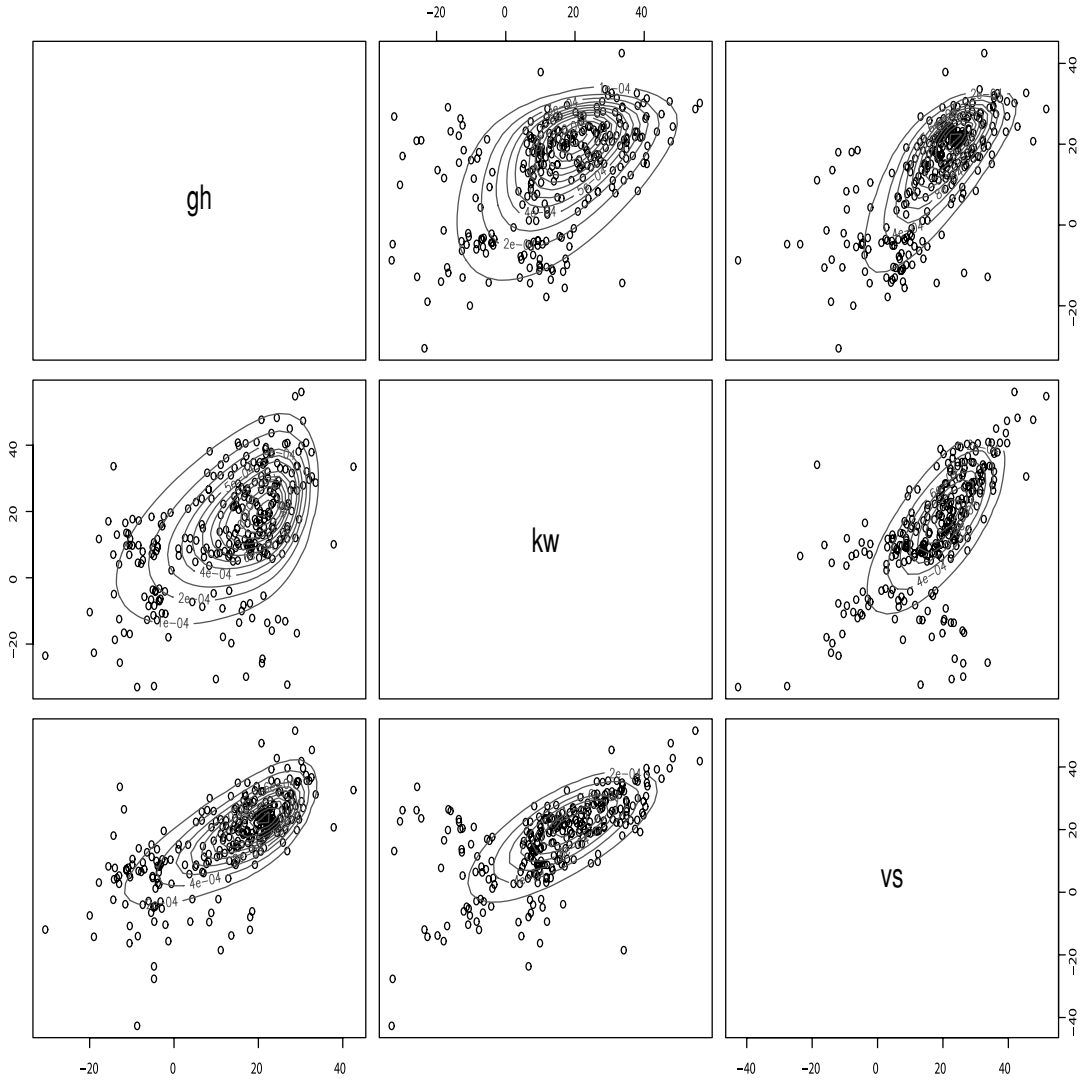

Figure 5. *Wind speed data: bivariate scatter plots and fitted ST contours.*

three towers, the strongest being at Goodnoe Hills. This asymmetry, due to prevailing westerly winds, has also been noted by Gneiting *et al.* (2006). The heavy tails behaviour is confirmed by the estimated degrees of freedom $\hat{\nu} = 4.0$, indicating the presence of extreme wind speeds. Figure 6 presents QQ-plots for a normal and an ST fit. The plots indicate that the ST model brings significant improvements over the normal distribution.

4.4 Vowel Recognition Data

Another field of potential application of the SN and ST distributions is the context of classification methods. Some initial work in this direction has been conducted by Azzalini & Capitanio (1999) based on SN distribution. For similarity with the traditional linear discriminant analysis, they have kept the scale matrix Ω and the skewness vector parameter α constant among

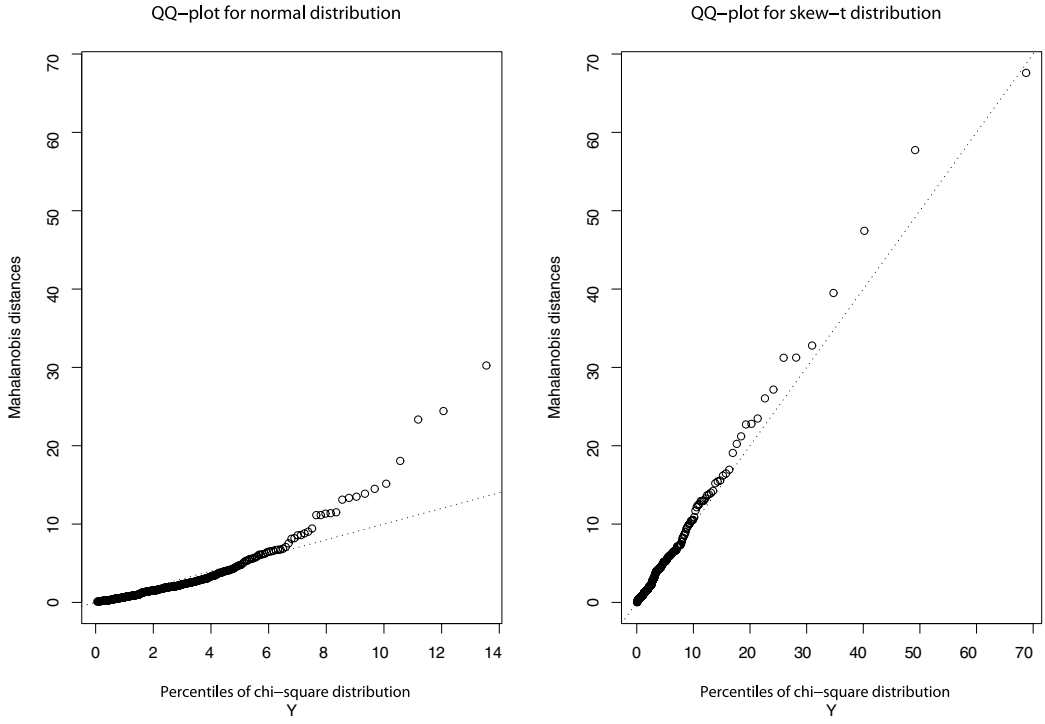

Figure 6. *Wind speed data: QQ-plots for normal and ST fit.*

sub-populations, varying only the location parameter ξ_k, say, with the subpopulation index $k = 1, \ldots, K$. Hence the discriminant function in use has been of the form

$$\log f_{\mathrm{SN}}(x; \xi_k, \Omega, \alpha) + \log \pi_k, \qquad x \in \mathbb{R}^d, \tag{22}$$

where f_{SN} denotes the SN density and π_k the prior probability of the k-th subpopulation.

Obviously, one is not compelled to keep Ω and α constant for all subpopulations, but relaxing this assumption inflates substantially the number of parameters to be estimated. A more attractive option seems to be replacing the SN density in (22) by the analogous ST density, with only one extra parameter ν to be estimated.

This approach has been experimented on the vowel recognition data which consist of a set of records on 10 continuous variables generated by the pronunciation of the 11 vowel sounds of the English language, plus the label of the vowel category. The 11 vowel sounds have been recorded from the pronunciation of various speakers, producing two tables of 528 and 426 rows, respectively. These data represent a popular and quite severe benchmark for classification methods, using the first table for training a given method to recognize the vowel on the basis of the 10 continuous variables and the second table for testing its performance. Among others, these data have been considered by Hastie *et al.* (1994) who have used them to compare the performance of a range of classification methods; also, they have provided a more detailed description of the genesis of the data.

There are therefore $K = 11$ subpopulations in $d = 10$ dimensions involved, requiring to estimate $11 \times 10 + 10 \times 11/2 + 10 = 175$ parameters for the SN formulation (22) and 176 parameters for the similar ST formulation, in addition to the prior probabilities. For the latter component, we

Table 8
Vowel recognition data: Percentages of misclassified cases.

Percentage of error	On training data	On test data
LDA	32	56
SN	29	51
ST	27	47
LINR	48	67
LOGR	22	51
QDA	1	53

have adopted the standard option of using the frequencies in the training set to estimate the π_k's. Estimation of the f_{SN} and f_{ST} parameters via maximum likelihood has required some numerical care due to high dimensionality of the parameter space. This fact has affected especially ST estimation, since for the SN case the search can effectively be reduced to consider a parameter space in 120 dimensions, using the technique of Section 2.1 based on (9)–(11).

The parameters estimated from the training data have been inserted in the discriminant function of form (22), in its variants with the SN and ST densities, and applied to the test data. The observed percentages of misclassified cases are given in Table 8, which also includes the corresponding value using the normal distribution, i.e. using linear discriminant analysis (LDA).

It was expected that the use of SN and ST distribution leads to some improvement with respect to the normal associated to LDA, because of the nesting of the parametric classes. It is, however, good to see that in practical terms the effect is non-negligible. In addition, it is worth noticing that the addition of a single extra parameter leading from the SN class to the ST class produces a noticeable improvement.

Table 8 also reports misclassification rates from Hastie *et al.* (2001, p. 85) for linear regression (LINR) of a class indicator matrix, logistic regression (LOGR), and quadratic discriminant analysis (QDA). The ST discriminant function on the test data still produces smaller error rates than for those three methods. It is interesting to note that QDA requires to estimate $11 \times 10 + 11 \times 10 \times 11/2 = 715$ parameters compared to only 176 for the ST formulation.

To get a more comprehensive view of the performance of the method, consider the top part of table 5 of Hastie *et al.* (1994) which includes several competing methods. The error rate of 47% of the ST discriminant function on the test data is not as good as the similar figure produced by the best classification methods, which score down to a 39% error rate. One must, however, take into account that these methods are substantially more complex and targeted specifically to classification. Indeed, the ST classification compares favourably even with many of those, and loses only with the most sophisticated ones.

5 Discussion

We have advocated the use of flexible parametric classes as an attractive alternative to the classical robustness approach. In particular, we have focused on the multivariate skew-*t* distribution along with the maximum likelihood method, and we have shown that the multivariate ST distribution has better inferential properties than the SN and SEP2 families. For the SN family, a more regular behaviour of the log-likelihood function can be achieved by a suitable re-parameterization, which has also the advantage of simpler interpretation of the parameters; for a discussion of this point, see, for instance Section 2.4 of Azzalini (2005). In fact, as for interpretability of parameters, a similar form of reparameterization is convenient even for ST distribution.

We have illustrated the advantages of our approach on various data sets and settings. In particular, the multivariate ST distribution is a flexible and parsimonious parametric alternative

to multivariate nonparametric density estimation. The latter is well known to suffer from the curse of dimensionality when several variables are considered simultaneously, for instance, such as in the example of the vowel recognition data. The flexibility of the ST distribution compared to the SN is due to a single additional parameter, namely the degrees of freedom, which results also in the parsimony of the model.

One potential disadvantage, though, is that there is only one parameter that regulates the tail behaviour of all variables. If, for example, one variable has Gaussian tails whereas another has Cauchy tails, then the single degrees of freedom parameter has to provide a compromise between those two tail behaviours. One general approach would be to consider some other multivariate t distributions with multiple degrees-of-freedom parameters, such as those proposed by Miller (1968), as the "base" function f_0 in (1). Unfortunately, those proposals do not have appealing parametric forms and rely on complicated hypergeometric functions that would prevent their use in routine applied work. A relatively simpler formulation, at least for the bivariate case, is the one of Jones (2002); see also additional proposals reviewed in chapters 4 and 5 of Kotz & Nadarajah (2004). These various formulations tend, apparently inevitably, to be appreciably more complicated from the formal viewpoint, often already in the expression of the density, let alone higher-order moments. While there certainly exist cases for which this extra complication is definitely required, it seems to us that our previous remark still holds, namely that the multivariate ST distribution provides a reasonable compromise between flexibility and mathematical tractability, making it a particularly attractive general-purpose tool.

Acknowledgements

We thank the editor and two referees for comments that improved this paper. This work was started while the first author was visiting the Department of Statistics, Texas A&M University, USA, whose kind hospitality is gratefully acknowledged. The work was partially supported by NSF grant DMS-0504896, by a Swiss National Science Foundation grant, and by MIUR, Italy, with grant PRIN 2006132978.

References

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*, **46**, 199–208.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussion by Marc G. Genton and a rejoinder by the author). *Scand. J. Statist.*, **32**, 159–200.

Azzalini, A. (2006). R package `sn`: The skew-normal and skew-t distributions (version 0.4-2). URL http://azzalini.stat.unipd.it/SN

Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. Roy. Statist. Soc. Ser. B*, **61**, 579–602. Full-length version of this paper is available from http://azzalini.stat.unipd.it/SN

Azzalini, A. & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *J. Roy. Statist. Soc. Ser. B*, **65**, 367–389. Full-length version of this paper is available from http://azzalini.stat.unipd.it/SN

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

Azzalini, A. & Genton, M.G. (2007). On Gauss's characterization of the normal distribution. *Bernoulli*, **13**, 169–174.

Bayes, C.L. & Branco, M.D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Braz. J. Probab. Statist.*, in press.

Box, G.E.P. & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA., London, Don Mills, ON.: Addison-Wesley.

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

Branco, M.D. & Dey, D.K. (2002). Regression model under skew elliptical error distribution. *J. Math. Sci.*, **1**, 151–168.

Brownlee, K.A. (1960, 2nd ed. 1965). *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.

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

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

Denby, J. & Martin, R.D. (1979). Robust estimation of the first-order autoregressive parameter. *J. Amer. Statist. Assoc.*, **74**, 140–146.

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

Dodge, Y. (1996). The guinea pig of multiple regression. In *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*, Lecture Notes in Statistics 109. New York: Springer-Verlag.

Fernández, C. & Steel, M.F.J. (1999). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika*, **86**, 153–168.

Genton, M.G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, edited volume. Boca Raton, FL: Chapman & Hall/CRC.

Genton, M.G. & Hering, A.S. (2007). Blowing in the wind. *Significance*, **4**, 11–14.

Genton, M.G. & Lucas, A. (2003). Comprehensive definitions of breakdown-points for independent and dependent rvations. *J. Roy. Statist. Soc. Ser. B*, **65**, 81–94.

Gneiting, T., Larson, K., Westrick, K., Genton, M.G. & Aldrich, E. (2006). Calibrated probabilistic forecasting at the Stateline wind energy center: the regime-switching space-time method. *J. Amer. Statist. Assoc.*, **101**, 968–979.

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

Hastie, T., Tibshirani, R. & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.*, **89**, 1255–1270.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*, New York: Springer.

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

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

Jones, M.C. (2002). A bivariate *t* distribution with marginals on different degrees of freedom. *Statist. Probab. Lett.*, **56**, 163–170.

Jones, M.C. & Faddy, M.J. (2003). A skew extension of the *t*-distribution, with applications. *J. Roy. Statist. Soc. Ser. B*, **65**, 159–174.

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

Kotz, S. & Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge: Cambridge University Press.

Künsch, H. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.*, **12**, 843–863.

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

Liseo, B. & Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *J. Statist. Planning Infer.*, **136**, 373–389.

Ma, Y. & Genton, M.G. (2000). Highly robust estimation of the autocovariance function. *J. Time Series Anal.*, **21**, 663–684.

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

Ma, Y., Genton, M.G. & Tsiatis, A.A. (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *J. Amer. Statist. Assoc.*, **100**, 980–989.

Ma, Y. & Hart, J.D. (2007). Constrained local likelihood estimators for semiparametric skew-normal distributions. *Biometrika*, **94**, 119–134.

Miller, K.S. (1968). Some multivariate *t*-distributions. *Ann. Statist.*, **39**, 1605–1609.

Nelder, J.A. (2000). There are no outliers in the stack-loss data. *Student*, **3**, 211–216.

Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution. *J. Appl. Statist.*, **27**, 859–870.

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

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

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

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew *t* distributions. *J. Statist. Plann. Inference*, **136**, 4259–4275.
Stigler, S.M. (1977). Do robust estimators work with real data? *Ann. Statist.*, **5**, 1055–1098.
Subbotin, M.T. (1923). On the law of frequency of error. *Mathematicheskii Sbornik*, **31**, 296–301.
Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*, 4th Ed. New York: Springer.

Résumé

Le problème de la robustesse est attaqué en adoptant une classe paramétrique de distributions qui sont suffisamment flexibles pour représenter le comportement des observations. Dans une variété de cas pratiques, une option raisonnable est de considérer des distributions qui incluent des paramètres pour régler leur asymétrie et leur aplatissement. Comme représentant spécifique de cette approche, la distribution *t* asymétrique est explorée plus en détail et des raisons sont apportées pour adopter cette option comme un compromis judicieux et à tous usages entre la robustesse et la simplicité du traitement et de l'interprétation des résultats. Quelques arguments théoriques, les résultats de simulations et divers exemples sur des données réelles sont fournis afin de soutenir cette affirmation.

Appendix

Proof of Proposition 1: (1) Consider the log-likelihood function (6) associated to the sample $y_1, \ldots, y_n$. The partial derivatives of order one of the log-likelihood are

$$\frac{\partial \ell}{\partial \xi} = -\frac{1}{\omega}\left[\sum_{i=1}^{n} \frac{f_0'(z_i)}{f_0(z_i)} + \sum_{i=1}^{n} P_K'(z_i)\frac{g\{P_K(z_i)\}}{G\{P_K(z_i)\}}\right],$$

$$\frac{\partial \ell}{\partial \omega} = -\frac{1}{\omega}\left[n + \sum_{i=1}^{n} z_i \frac{f_0'(z_i)}{f_0(z_i)} + \sum_{i=1}^{n} z_i P_K'(z_i)\frac{g\{P_K(z_i)\}}{G\{P_K(z_i)\}}\right],$$

$$\frac{\partial \ell}{\partial \alpha_j} = \sum_{i=1}^{n} z_i^j \frac{g\{P_K(z_i)\}}{G\{P_K(z_i)\}}, \quad j = 1, 3, \ldots, K,$$

where $z_i = \omega^{-1}(y_i - \xi)$ and $g = G'$. Setting $v_i = f_0'(z_i)/f_0(z_i)$ and $w_i = g\{P_K(z_i)\}/G\{P_K(z_i)\}$, the solutions to the score equations satisfy

$$-\bar{v} = \alpha_1 \bar{w} + 3\alpha_3 \overline{z^2 w} + \cdots + K\alpha_K \overline{z^{K-1} w}, \tag{23}$$

$$0 = 1 + \overline{zv} + \alpha_1 \overline{zw} + 3\alpha_3 \overline{z^3 w} + \cdots + K\alpha_K \overline{z^K w}, \tag{24}$$

$$0 = \overline{z^j w}, \quad j = 1, 3, \ldots, K \tag{25}$$

where a notation of the form $\overline{z^j w}$ denotes the average of the component-wise evaluation of the sample values of $z^j w$.

From (24) and (25), $\overline{zv} = -1$ for any solution. For $\alpha_1 = \alpha_3 = \cdots = \alpha_K = 0$ to be a solution to the score equations requires $\bar{v} = 0$ from (23), that is,

$$\sum_{i=1}^{n} \frac{f_0'\{(y_i - \bar{y})/\omega\}}{f_0\{(y_i - \bar{y})/\omega\}} = 0. \tag{26}$$

Consequently, $\bar{w} = 2g(0)$, $\xi = \overline{yw}/\bar{w} = \bar{y}$ and $\omega = \xi\bar{v} - \overline{vy} = -\overline{vy}$, which for $f_0 = \phi$, the standard normal density, simplifies to $\omega = s$. This is true whatever the choice of G.

(2) Setting $u_i = f_0''(z_i)/f_0(z_i)$ and $t_i = g'\{P_K(z_i)\}/G\{P_K(z_i)\}$, straightforward but tedious computations yield expressions for the second-order partial derivatives of the log-likelihood function (6). At $\alpha_1 = \alpha_3 = \cdots = \alpha_K = 0$, they simplify to

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \xi^2} &= -\frac{n}{\omega^2}[\overline{v^2} - \bar{u}],\\
\frac{\partial^2 \ell}{\partial \omega^2} &= -\frac{n}{\omega^2}[\overline{z^2 v^2} - \overline{z^2 u} + 1],\\
\frac{\partial^2 \ell}{\partial \alpha_j^2} &= 2n\overline{z^{2j}}[g'(0) - 2g^2(0)], \quad j = 1, 3, \ldots, K,\\
\frac{\partial^2 \ell}{\partial \xi\, \partial \omega} &= -\frac{n}{\omega^2}[\overline{z v^2} - \overline{z u}],\\
\frac{\partial^2 \ell}{\partial \xi\, \partial \alpha_j} &= -\frac{n}{\omega}[2g(0) j \overline{z^{j-1}}], \quad j = 1, 3, \ldots, K,\\
\frac{\partial^2 \ell}{\partial \omega \partial \alpha_j} &= -\frac{n}{\omega}[2g(0) j \overline{z^{j}}], \quad j = 1, 3, \ldots, K,\\
\frac{\partial^2 \ell}{\partial \alpha_j\, \partial \alpha_k} &= 2n\overline{z^{j+k}}[g'(0) - 2g^2(0)], \quad j, k = 1, 3, \ldots, K.
\end{aligned}$$

When $f_0 = \phi$, we have $\bar{u} = 0$, $\overline{v^2} = \overline{z^2} = 1$, $\overline{z v^2} = \overline{z u} = \overline{z^3}$ and $\overline{z^2 u} + 1 = \overline{z^2 v^2} = \overline{z^4}$, and thus the following additional simplifications occur:

$$\frac{\partial^2 \ell}{\partial \xi^2} = -\frac{n}{\omega^2}, \qquad \frac{\partial^2 \ell}{\partial \omega^2} = -\frac{2n}{\omega^2}, \qquad \frac{\partial^2 \ell}{\partial \xi\, \partial \omega} = 0.$$

Thus, the first and third rows of the observed Fisher information matrix, corresponding to ξ and α_1, are

$$n\Big(\quad \tfrac{1}{\omega^2} \quad 0 \quad \tfrac{1}{\omega} 2g(0) \quad \tfrac{1}{\omega} 2g(0) 3\overline{z^2} \quad \tfrac{1}{\omega} 2g(0) 5\overline{z^4} \quad \cdots \quad \tfrac{1}{\omega} 2g(0) K \overline{z^{K-1}} \quad \Big),$$

$$n\Big(\quad \tfrac{1}{\omega} 2g(0) \quad 0 \quad 4g^2(0) \quad 4g^2(0)\overline{z^4} \quad 4g^2(0)\overline{z^6} \quad \cdots \quad 4g^2(0)\overline{z^{K+1}} \quad \Big).$$

Taking expectations and using the well-known fact that $\mathbb{E}\{j\overline{z^{j-1}}\} = \mathbb{E}\{\overline{z^{j+1}}\}$ for the Gaussian distribution, we see that those two rows are proportional by the factor $\omega 2g(0)$, and thus the expected Fisher information matrix is singular. Note that for $K = 1$, the observed Fisher information matrix is already singular, as noted by Pewsey (2006). This is true whatever the choice of G, and thus completes the proof of the statement.

[*Received February 2007, accepted April 2007*]