J. R. Statist. Soc. B (2003)
65, *Part* 1, *pp.* 159–174

A skew extension of the *t*-distribution, with applications

M. C. Jones

The Open University, Milton Keynes, UK

and M. J. Faddy

University of Birmingham, UK

[Received March 2000. Final revision July 2002]

Summary. A tractable skew t-distribution on the real line is proposed. This includes as a special case the symmetric t-distribution, and otherwise provides skew extensions thereof. The distribution is potentially useful both for modelling data and in robustness studies. Properties of the new distribution are presented. Likelihood inference for the parameters of this skew t-distribution is developed. Application is made to two data modelling examples.

Keywords: Beta distribution; Likelihood inference; Robustness; Skewness; Student's t-distribution

1. Introduction

Student's t-distribution occurs frequently in statistics. Its usual derivation and use is as the sampling distribution of certain test statistics under normality, but increasingly the t-distribution is being used in both frequentist and Bayesian statistics as a heavy-tailed alternative to the normal distribution when robustness to possible outliers is a concern. See Lange *et al.* (1989) and Gelman *et al.* (1995) and references therein.

It will often be useful to consider a further alternative to the normal or t-distribution which is both heavy tailed and skew. To this end, we propose a family of distributions which includes the symmetric t-distributions as special cases, and also includes extensions of the t-distribution, still taking values on the whole real line, with non-zero skewness. Let $a > 0$ and $b > 0$ be parameters. Then, the density function of this new distribution is

$$f(t) = f(t; a, b) = C_{a,b}^{-1} \left\{1 + \frac{t}{(a+b+t^2)^{1/2}}\right\}^{a+1/2} \left\{1 - \frac{t}{(a+b+t^2)^{1/2}}\right\}^{b+1/2} \qquad (1)$$

where

$$C_{a,b} = 2^{a+b-1}\, B(a, b)(a+b)^{1/2}$$

and $B(\cdot, \cdot)$ denotes the beta function. When $a = b$, f reduces to the t-distribution on $2a$ degrees of freedom. When $a < b$ or $a > b$, f is negatively or positively skewed respectively. In fact, $f(t; b, a) = f(-t; a, b)$. Note that a and b are positive real numbers and need not be integer or half-integer.

Address for correspondence: M. C. Jones, Department of Statistics, Faculty of Mathematics and Computing, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.
E-mail: m.c.jones@open.ac.uk

 1369–7412/03/65159

A preliminary account of this distribution is in Jones (2001a), where two derivations of it are provided. The first is a mathematical manipulation in which the symmetric t-density function is factorized into two parts and those parts are taken to different powers. The skew t-densities thus emulate the symmetric t-densities in forming a mathematical sequence tending to the normal density, albeit with two parameters $a, b \to \infty$ rather than one.

The second derivation is to note that, if B has the beta distribution on (0, 1) with parameters a and b, then

$$T = \frac{\surd(a+b)(2B-1)}{2\surd\{B(1-B)\}} \tag{2}$$

has density (1). The $a=b$ special case of this relationship was known to Fisher (1915).

An equivalent formulation comes from the relationship between a beta random variable and a pair of independent χ^2 random variables. This yields

$$T = \frac{\surd(a+b)(U-V)}{2\surd(UV)} \tag{3}$$

where U and V are independent with χ^2_{2a}- and χ^2_{2b}-distributions respectively. The $a=b$ special case of this relationship is also known (Cacoullos, 1965). From equation (3), yet another representation of T is as $\frac{1}{2}\surd(a+b)\ (W^{1/2}-W^{-1/2})$ where $W=aF/b$ and F has the F-distribution with $2a$ and $2b$ degrees of freedom. Simple methods for obtaining random variates having density (1) are thus available by transforming variates from standard distributions.

Fig. 1 shows a variety of skew t-densities. In Fig. 1(a), $b=4$ and, in Fig. 1(b), $b=1$. In each frame, seven densities are shown: for $a=b$, the symmetric t_{2b}-density, and, in increasing order of skewness, for $a=2^i b$, $i=1,\ldots,6$. The raw densities have been recentred to have mean 0 and rescaled to the same scaling as the standard symmetric t-density in each case. In Fig. 1(a), this is done by using the formulae for the mean and variance given in Section 2.1; in Fig. 1(b), the variance does not exist, and so, purely as a matter of expediency, the densities are matched in terms of mean and inter-point-of-inflection distance. Fig. 1(a) displays a range of fairly gentle skewings of the t_8-distribution. Fig. 1(b) suggests that a little more severe skewness can be attained based on the t_2 starting-point.

Some properties of the distribution are presented in Section 2; these are its distribution function, moments and mode (Section 2.1), its large skewness limit (Section 2.2), some measures of skewness (Section 2.3) and the distribution's tail behaviour (Section 2.4). Proofs of two results from Section 2 are given in Appendix A.

Although one potential application of the skew t-distribution is in robustness studies, it is the more important robust data modelling aspect of the skew t-distribution, as a model for data exhibiting skewness and/or heavy tails, that we concentrate on in this paper. This is accomplished in practice by incorporating location and scale parameters, quite possibly depending on covariates. Likelihood inference for the resulting four parameters of the skew t-distribution is discussed in Section 3, first derivatives of the log-likelihood and elements of the observed information matrix being given in Appendix B. These are written in terms of a reparameterization of the skew t-distribution which is a central concern of Section 3, with an outline proof of its properties in Appendix C.

We go on, in Section 4, to present two illustrative examples in which the usefulness of the skew t-distribution is apparent. The current skew t proposal is briefly compared with other recent skew t-distributions in Section 5. Finally, in Section 6, some closing comments are made.

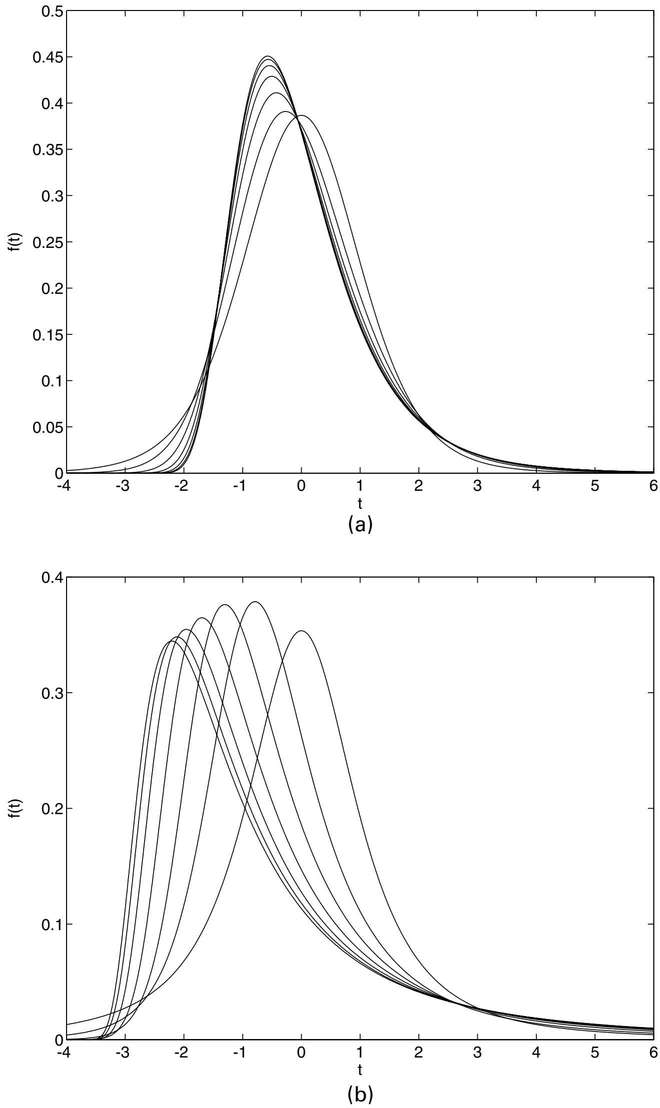


Fig. 1. Standardized densities (1) for $a = 2^i b$, $i = 0, \ldots, 6$, having increasing amounts of skewness, in the cases (a) $b = 4$ and (b) $b = 1$

2. Basic properties

2.1. Distribution function, moments and mode

Formula (2) is invertible and thus provides the distribution function of T as

$$F(t; a, b) = I_{\{1+t/\sqrt{(a+b+t^2)}\}/2}(a, b)$$

where $I_x(\cdot, \cdot)$ denotes the incomplete beta function ratio.

The following equivalent expressions for $E(T^r)$ can be obtained, as in Appendix A. Note the conditions on a and b for the moments to exist.

Result 1. Provided that $a > r/2$ and $b > r/2$,

$$E(T^r) = \frac{(a+b)^{r/2}}{B(a,b)} \sum_{i=0}^{r} \binom{r}{i} 2^{-i}(-1)^i \, B\left(a + \frac{r}{2} - i, b - \frac{r}{2}\right) \tag{4a}$$

$$= \frac{(a+b)^{r/2}}{2^r \, B(a,b)} \sum_{i=0}^{r} \binom{r}{i} (-1)^i \, B\left(a + \frac{r}{2} - i, b - \frac{r}{2} + i\right). \tag{4b}$$

It is immediate that

$$E(T) = \frac{(a-b)\sqrt{(a+b)}}{2} \frac{\Gamma(a-\frac{1}{2})\,\Gamma(b-\frac{1}{2})}{\Gamma(a)\,\Gamma(b)},$$

where $\Gamma(\cdot)$ is the gamma function, and

$$E(T^2) = \frac{(a+b)}{4} \frac{(a-b)^2 + a - 1 + b - 1}{(a-1)(b-1)}.$$

These expressions reduce to the values for the symmetric t-distribution on $2a$ degrees of freedom when $a = b$, namely 0 and $a/(a-1)$ respectively.

By differentiating equation (1), it is easy to see that f is unimodal, with mode at

$$\frac{(a-b)\sqrt{(a+b)}}{\sqrt{(2a+1)}\sqrt{(2b+1)}}. \tag{5}$$

2.2. *Limiting distributions*

It was mentioned in Section 1 that, if a and b become large together, the skew t-distribution tends to the normal distribution.

A different limiting result is obtained if we let b remain fixed and consider the case where $a \to \infty$. Provided that we normalize density (1) by accounting for its increasing location and scale, a limiting distribution arises, and this reflects the situation in Fig. 1. This limiting distribution is the distribution of $\sqrt{2}$ over the square root of a χ^2 random variable on $2b$ degrees of freedom.

Result 2. As $a \to \infty$ and $b > 1$ remains fixed, $\sigma_{a,b}\, f(\sigma_{a,b}t + \mu_{a,b}; a, b) \to \sigma_b\, f_b(\sigma_b t + \mu_b)$ where

$$f_b(t) = 2\{\Gamma(b) t^{2b+1}\}^{-1} \exp(-1/t^2). \tag{6}$$

Here $\mu_{a,b}$ and $\sigma_{a,b}$ are the mean and standard deviation of T given in Section 2.1, and $\mu_b = \Gamma(b-\frac{1}{2})/\Gamma(b)$ and

$$\sigma_b = \left[\frac{1}{b-1} - \left\{\frac{\Gamma^2(b-\frac{1}{2})}{\Gamma^2(b)}\right\}\right]^{1/2}$$

are the mean and standard deviation of the limiting distribution.

See Appendix A for the proof. Result 2 normalizes using mean and variance, and it is this that introduces the $b > 1$ restriction; other location and scale measures could have been used which would not require this restriction.

2.3. *Skewness*

The classical skewness measure based on the third moment is available for $a, b > \frac{3}{2}$ from the moment formulae of Section 2.1. An attractive alternative skewness measure is $1 - 2\,F(\text{mode})$

(Arnold and Groeneveld, 1995). This exists for any $a, b > 0$ and has a simple expression:

$$1 - 2\, I_{(a+1/2)/(a+b+1)}(a, b).$$

For $a, b > \frac{1}{2}$, the third L-moment ratio or L-skewness (Hosking, 1990),

$$\int F(x)\{2\,F(x) - 1\}\{1 - F(x)\}\,\mathrm{d}x \bigg/ \int F(x)\{1 - F(x)\}\,\mathrm{d}x,$$

can be readily calculated numerically.

Numerical experiments suggest that all three skewness measures are monotone increasing functions of a for fixed b and monotone decreasing functions of b for fixed a. We have not been able to prove this. Broadly speaking, this means that high absolute values of skewness within the class are associated with small values of the parameters, as suggested by Fig. 1.

2.4. *Tail behaviour*

Student's t_{2b}-density has tails that behave as $|t|^{-(2b+1)}$ as $t \to \pm\infty$. For $a > b$, the skew t-distribution has a right-hand tail which remains as $t^{-(2b+1)}$, which is large for small b. The left-hand tail of the skew t-distribution is reduced in weight to order $|t|^{-(2a+1)}$ as $t \to -\infty$. This well reflects the tail behaviour that is observable in Fig. 1. For $a < b$, the tails are reversed.

3. Likelihood inference

For fitting to independent and identically distributed data, $X_1, \ldots, X_n$, we consider the general, four-parameter, version of the skew t-density (1) given by $\sigma^{-1}\ f\{\sigma^{-1}(x - \mu); a, b\}$ where μ and σ are additional location and scale parameters. The skew t-distribution affords full tractability of quantities involved in asymptotic likelihood theory. See, for example, the elements of the observed information matrix (after reparameterization, for which see below) given in Appendix B. A Fisher scoring algorithm would therefore be available for likelihood maximization.

We have used likelihood inference in the examples to follow mostly in a straightforward way using numerical optimization of the log-likelihood function and standard χ^2-approximations to the distributions of log-likelihood ratios. In general, such likelihood inference seems very successful. In our many investigations, we have never come across a likelihood surface that was not unimodal; see Section 4.2 for a brief report of some simulation results to this effect. The only situation in which likelihood inference breaks down corresponds to a problem that is well documented in likelihood fitting of symmetric t-distributions with unknown degrees of freedom, namely the difficulties that are encountered when $a + b$ is rather less than 1 (Fraser (1979), chapter 2, Lange *et al.* (1989) and Fernández and Steel (1999)). We are pragmatic about this: such distributions have *extremely* heavy tails and the whole business of directly modelling data containing many extreme outliers is not to be recommended.

There remains, however, the issue of reparameterization. As a starting-point, we might reparameterize in terms of $\nu = a + b$ and $\lambda = a - b$. The former has the same degrees of freedom role as ν in the symmetric t_ν-distribution, and λ is a parameter controlling skewness, the symmetric t-distribution corresponding to $\lambda = 0$. That said, a little thought shows that λ is not very satisfactorily tied to skewness: a normalization of λ with respect to ν would appear to be more meaningful and will be provided below.

In general, there are no zeros in the expected information matrix nor therefore in the asymptotic covariance matrix of the maximum likelihood estimates $(\hat{\nu}, \hat{\lambda}, \hat{\mu}, \hat{\sigma})$. This contrasts with the symmetric t case for which the only non-zero asymptotic correlation is between $\hat{\nu}$ and

$\hat{\sigma}$. Indeed, in the symmetric t case, it is possible to derive an orthogonal reparameterization (Cox and Reid, 1987) such as $\{\nu, \mu, \nu\sigma/(1+\nu)\}$. However, there does not seem to be any such orthogonal reparameterization in the skew t case. We are not, however, too dismayed by this; an orthogonal parameterization, although attractive if available, is a luxury that is not available to many useful statistical models.

It is, however, possible to produce a reparameterization which induces regular estimation for a and/or b infinite. The reparameterization that we suggest can also be found in Prentice (1975):

$$\begin{aligned} p &= \frac{2}{a+b} = \frac{2}{\nu}, \\ q &= \frac{a-b}{\sqrt{\{ab(a+b)\}}} = \frac{2\lambda}{\sqrt{\{(\nu^2-\lambda^2)\nu\}}}. \end{aligned} \tag{7}$$

As values of (a, a), (∞, a), (a, ∞) and $(-\infty, \infty)$ for (a, b) correspond respectively to t_{2a}, $\sqrt{(2/\chi^2_{2a})}$, $-\sqrt{(2/\chi^2_{2a})}$ and normal distributions, so the same four distributions correspond to (p, q) taking values of $(1/a, 0)$, $(0, 1/\sqrt{a})$, $(0, -1/\sqrt{a})$ and $(0, 0)$ respectively. The advantage of this is that the limits as p and q tend to 0 both singly and together of $\partial l/\partial p$ and $\partial l/\partial q$ are finite and non-zero, where l denotes the contribution to the log-likelihood based on the normalized density $\sigma_{a,b}\, f(\sigma_{a,b}t + \mu_{a,b})$ from an observation x, where $t = (x-\mu)/\sigma$. This result requires a substantial amount of Taylor series manipulation which is outlined in Appendix C. In particular,

$$\begin{aligned} \lim_{p,q\to 0}\left(\frac{\partial l}{\partial q}\right) &= -\frac{5}{12}(t^3 - 3t), \\ \lim_{p,q\to 0}\left(\frac{\partial l}{\partial p}\right) &= \frac{1}{8}(t^4 - 6t^2 + 3). \end{aligned} \tag{8}$$

It follows that the score test for normality within this family is based on $(nS^3)^{-1}\,\Sigma\,(X_i - \bar{X})^3$ and $(nS^4)^{-1}\,\Sigma\,(X_i - \bar{X})^4 - 3$, the usual sample skewness and kurtosis.

Reparameterization (7) first arose in Prentice's (1975) work on likelihood fitting of the log-F-distribution. It turns out that a yet more general family of distributions to be investigated elsewhere suggests strong analogies between the skew t- and log-F-distributions. However, the skew t-family has polynomial tails and the log-F-family has exponential tails, so the fact that the same reparameterization works for both is by no means obvious.

4. Examples

4.1. Example 1: strengths of glass fibres

Our first example concerns a random sample of data which would appear to be well fitted by the skew t-distribution. The data set is 'sample 1' of Table 1 of Smith and Naylor (1987) concerning

Table 1. Estimates and approximate standard errors

	$\hat{p}$	$\hat{q}$	$\hat{\mu}$	$\hat{\sigma}$
Estimate	0.627	−0.360	1.698	0.179
Standard error (observed)	0.346	0.143	0.077	0.035
Standard error (expected)	0.285	0.148	0.079	0.029

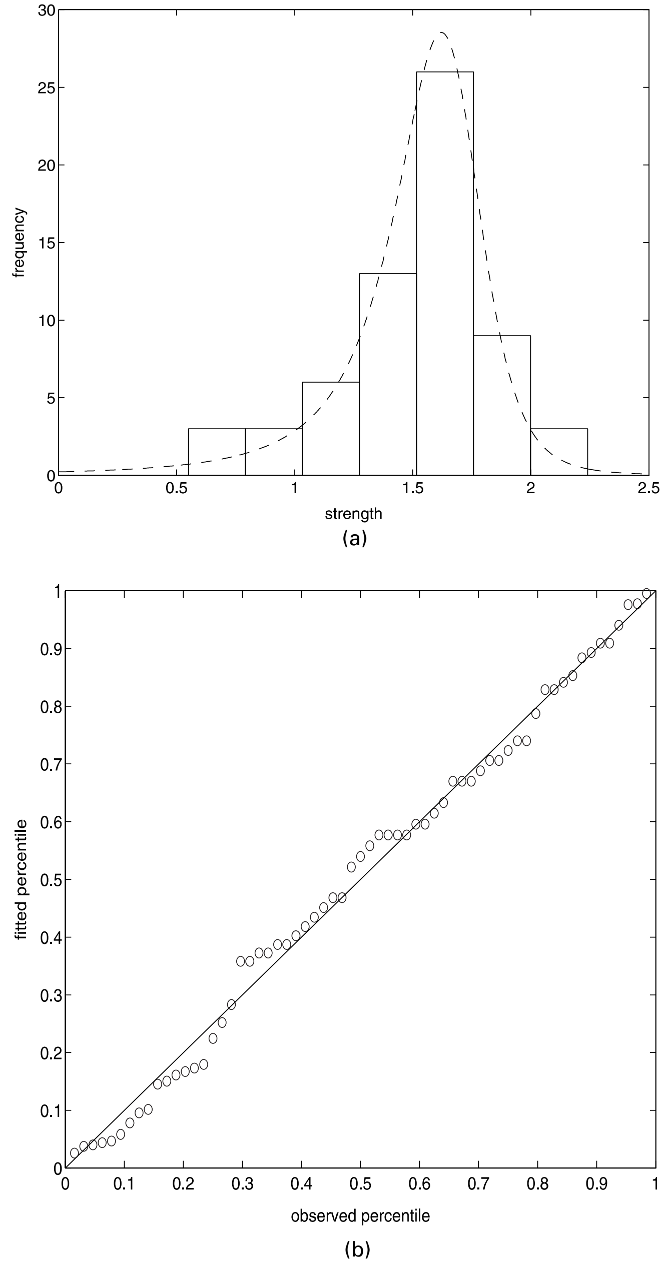
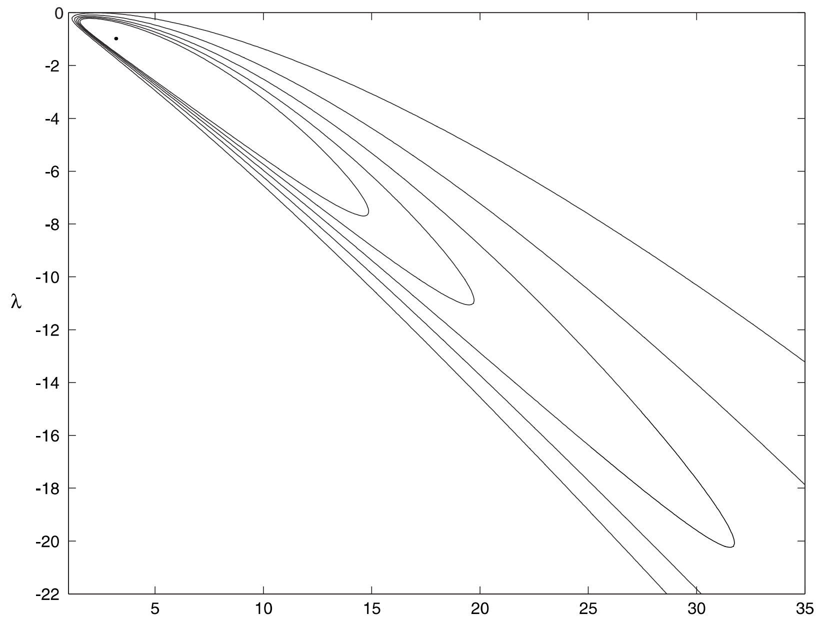


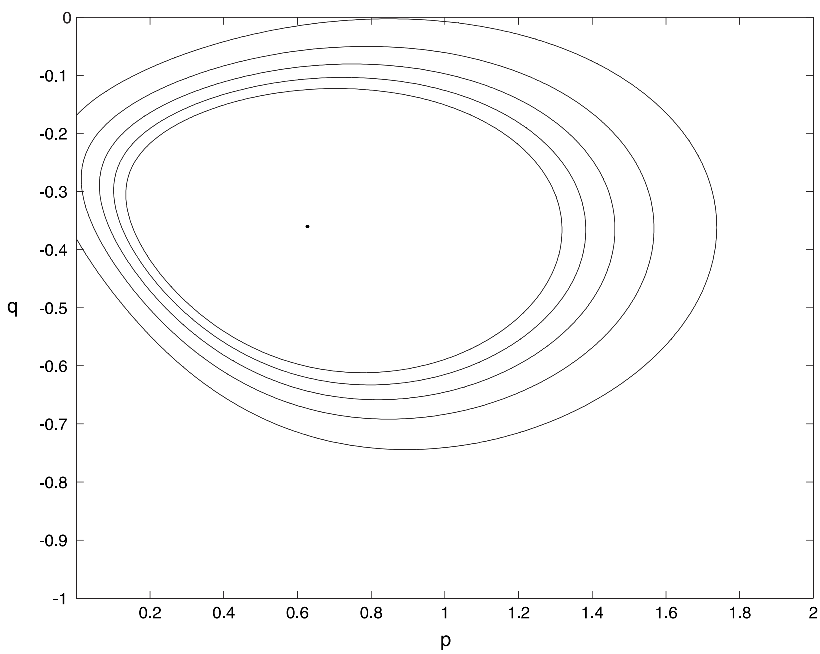
Fig. 2. (a) Histogram of the Smith and Naylor (1987) data (———), together with the fitted skew *t*-distribution (– – –) and (b) *PP*-plot for the same data and fit

the breaking strengths of $n = 63$ glass fibres of length 1.5 cm, originally obtained by workers at the UK National Physical Laboratory.

A histogram is presented in Fig. 2(a) along with the fitted skew t-distribution obtained by maximum likelihood. The latter has $\hat{p} = 0.627$, $\hat{q} = -0.360$, $\hat{\mu} = 1.698$ and $\hat{\sigma} = 0.179$. The left skewness of this data set is apparent from $\hat{q} < 0$, and there is evidence in favour of this



(a)



(b)

Fig. 3. 75%, 80%, 85%, 90% and 95% profile log-likelihood confidence regions for the parameters (a) (ν, λ) and (b) (p, q), under different parameterizations of the skew t-distribution for the Smith and Naylor (1987) data: •, positions of the maximum likelihood estimates

asymmetry according to twice the log-likelihood ratio of 6.08 on 1 degree of freedom (p-value approximately 0.01). In addition, the fitted skew t-distribution has fairly heavy tails, particularly to the left, $\hat{a} = 1.107$, but also to the right, $\hat{b} = 2.084$. A PP-plot of the fitted distribution is given in Fig. 2(b); the good fit of the model to the data is apparent.

Table 1 shows approximate standard errors of the estimates, obtained from both the observed and the expected information matrices. The latter were obtained from formulae which we do not give in the paper. In Fig. 3, we contrast the profile log-likelihood confidence regions in the (ν, λ) and (p, q) parameterizations. Fig. 3(a) exhibits considerable asymmetry in the profile log-likelihoods for ν and λ, with a strong negative correlation between the estimates of these two parameters. However, reparameterization (7) resulted in more elliptically shaped confidence regions, as shown in Fig. 3(b). Approximate orthogonality of p and q seems apparent here. The parameter q might then be a preferred measure of skewness over λ itself. Estimation of μ and σ appears to be more precise than that of p and q with the approximate standard errors, shown in Table 1, being relatively smaller (considerably so for μ). This will be of particular value in the regression examples to follow.

Smith and Naylor (1987) fitted a three-parameter Weibull distribution to these data. One parameter was a lower cut-off point, estimated to be -1.6 by maximum likelihood, but more reasonably estimated to be 0.027 or 0.172 by two Bayesian methods. Although the last two estimates point to a (defensible) lower limit of 0, it is unclear what a non-zero result would mean, and the difficulties associated with estimation involving such a parameter are considerable (witness the whole paper devoted to this fitting!). Confirmation that the combination of skewness and heavy tails that is provided by the skew t-distribution is necessary is provided by the inadequacy of a skew *normal* distribution fitted to these data in unpublished work of A. R. Pewsey.

4.2. Simulations

Simulations of data from distributions with a range of different parameter values suggest that the log-likelihood surface is of the unimodal shape typified by Fig. 3. Confidence regions for (λ, ν), or contours of the log-likelihood surface, become less concentrated around the maximum likelihood estimate as the 'degrees of freedom' parameter ν increases, or equivalently the confidence regions for (p, q) become more concentrated around their maximum likelihood estimate as p decreases. Confidence regions for (p, q) are typified by Fig. 3(b), and thus these simulations lend support to the reparameterization (7) achieving approximate orthogonality of parameters.

4.3. Example 2: blood flow data

Lange *et al.* (1989), example 3, described calibration of blood flow data by using a non-linear regression model. Here, the response variable Y is the blood flow in the canine myocardium measured non-invasively by using positron emission tomography (PET) and the covariate x is this blood flow measured invasively by using radioactively labelled microspheres. Two PET measurements were considered: one from scans taken up to 60 s and the other from scans taken up to 510 s, and the regression model was based on $E(Y) = x\{1 - \theta_1 \exp(-\theta_2/x)\}$.

Fig. 4(a) shows the first of these PET measurements plotted against x, together with the fitted regressions from using both a symmetric t-distribution for the residuals (as in Lange *et al.* (1989)) and the skew t-distribution. The skew t-distribution gives a significant improvement over the symmetric t (twice the log-likelihood ratio statistic is 19.50), although the fitted regression lines are quite similar. The parameter estimates (with asymptotic standard errors in parentheses) are $\hat{\theta}_1 = 0.62$ (0.013), $\hat{\theta}_2 = 104.6$ (8.1), $\hat{p} = 0$ and $\hat{q} = 0.321$ (0.066), so the fitted

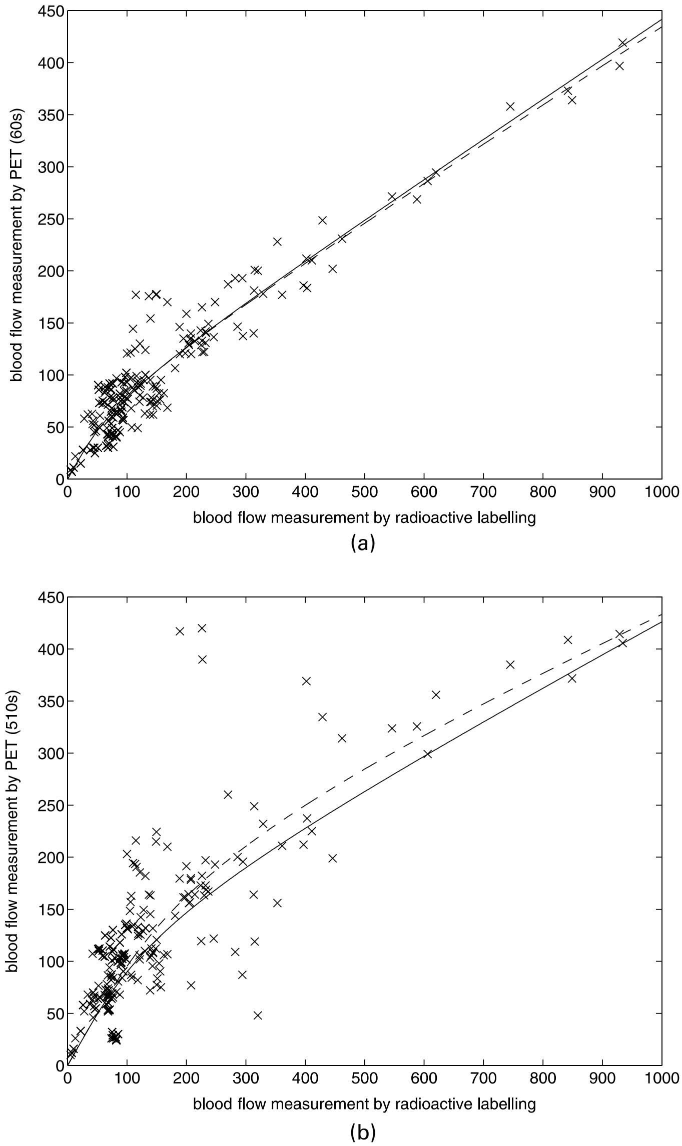


Fig. 4. Blood flow data (×), skew t (———) and symmetric t (– – –) regression fits: (a) first data set; (b) second data set

residual distribution is of the limiting form (6) with $\hat{\sigma} = 393.0$ (169.3). Residual plots show that this distribution adequately describes the variation.

In Fig. 4(b) is a plot of the second PET measurement data set with its symmetric t and skew t regression lines. Again, there is a significant improvement in the fit using the skew t-distribution for the residual variation, twice the log-likelihood ratio statistic being 21.43. But, here, the fitted regression lines are more than a little different, and the fitted skew t-distribution has $\hat{p} = 1.21$ (0.17) so the tails are heavy and result in infinite variance. The other estimated parameters are $\hat{\theta}_1 = 0.69$ (0.029), $\hat{\theta}_2 = 191.3$ (26.8), $\hat{q} = -0.197$ (0.053) and $\hat{\sigma} = 23.12$ (2.19). Residual plots from this skew t fit point to some inadequacies in the model. This is quite possibly due to the clear shift of the fitted regression line, compared with the symmetric t fit, in the direction of the skewness, Fig. 4(b), and is a consequence of modelling the *mean* in conjunction with a highly skewed residual distribution. It might be preferable to model an alternative location measure, such as the *mode*, when there is such a high level of residual variation. If this is done, then a regression line between those shown in Fig. 4(b) is obtained, with some improvement in the residual plots as well as a significant improvement over the symmetric t fit.

5. Comparison with other skew *t*-distributions

There are several proposals in the literature that can be regarded as competing skew t-distributions. In this section, we make brief comparisons with perhaps the two simplest and most popular alternatives. Other skew t-type distributions which are more complicated and more difficult to work with include the non-central t-distribution (e.g. Johnson *et al.* (1995), chapter 31), the Pearson type IV distribution (e.g. Skates (1993)), a special case of the generalized hyperbolic distribution (Barndorff-Nielsen and Shephard, 2001) and other somewhat arbitrary mathematical constructs (e.g. Butler *et al.* (1990)).

5.1. A first alternative skew t-distribution

Let l and L be the density and distribution functions of any distribution on the real line, symmetric about zero. A first general method of skewing l is to define

$$g(t;\psi) = 2\,L(\psi t)\,l(t). \tag{9}$$

Here, $\psi = 0$ is the symmetric case; positively skew distributions arise for $\psi > 0$, and $g(t;-\psi) = g(-t;\psi)$. For general and normal cases, see O'Hagan and Leonard (1976) and Azzalini (1985). For Bayesian deployments of the skew t case, see Mukhopadhyay and Vidakovic (1995) and DiCiccio *et al.* (1997). For a closely related variation on the skew t version of equation (9), see Azzalini and Capitanio (2003) and Sahu *et al.* (2002).

Although equation (9) is easy to write down, there is, in the t case, an incomplete beta function in the density, leading to relative intractability of the distribution function. The distribution is always unimodal (Azzalini, 1985) but it is not possible to find an analytic expression for the mode. The even moments are those of the t_ν-distribution but the odd moments are intractable. For $\psi > 0$, the left-hand tail of the t version of equation (9) goes as $|t|^{-(2\nu+1)}$ and the right-hand tail as $t^{-(\nu+1)}$ for $|t| \to \infty$. As $\psi \to \infty$ for fixed ν, $g(t;\psi)$ tends to the half-t_ν-distribution. The intractability of certain moments extends to intractability of the expected information matrix and thus the unavailability of a Fisher scoring algorithm for likelihood maximization.

Azzalini and Capitanio (2003) also fitted his skew t-distribution to the glass fibre data of Section 4.1. We would strengthen Azzalini's statement that conclusions from fitting the two skew t models are 'broadly similar' to being very similar.

The skew t-distributions of this section do, however, have extensions to the multivariate case (e.g. Azzalini and Capitanio (2003)) that seem to be more useful than current multivariate extensions of our skew t-distribution (e.g. Jones (2001b)).

5.2. A second alternative skew t-distribution

An alternative approach to skewing symmetric distributions is to piece together two differently scaled halves of the symmetric base distribution in a continuous manner:

$$h(t;\phi) = \frac{2\phi}{1+\phi}\{l(t)\, I(t \geqslant 0) + l(\phi t)\, I(t < 0)\}. \tag{10}$$

Here, $I(\cdot)$ is the indicator function and $\phi > 0$. In equation (10), $\phi = 1$ is the symmetric case, positively skew distributions arise from $\phi > 1$ and their negatively skew companions from $0 < \phi < 1$. The two-piece normal distribution seems to originate from Gibbons and Mylroie (1973) and has recently been discussed by Garvin and McClean (1997) and Mudholkar and Hutson (2000). Application to the t-distribution has been made by Fernández and Steel (1998).

Expression (10) is tractable, if a little clumsy, using the t_ν base density. The distribution function and moments can be written down and the distribution is unimodal with mode at zero. Both tails match those of the t_ν-distribution. As $\phi \to \infty$ for fixed ν, $l(t;\phi)$ also tends to the half-t_ν-distribution.

Note, however, that whereas odd derivatives of $l(t;\phi)$ are 0 at the origin even derivatives, including importantly the second, are discontinuous there. This has the disadvantage of making standard asymptotic likelihood theory inapplicable. Fernández and Steel (1998) resorted to Bayesian fitting of this skew t-distribution.

6. Closing remarks

To illustrate the potential of the skew t-distribution for data analysis, we have presented single-sample and non-linear regression examples that were sufficiently simple for the usefulness of the skew t-distribution to be apparent. Of course, the skew t-distribution is equally applicable to linear regression and to all the more complicated modelling situations such as multiple regression and time series modelling, and will be useful as an additional distribution in robust statistical modelling. Our skew t-distribution is sufficiently tractable that likelihood theory for it can be developed fully, an advantage that this distribution has over competing distributions. (This also affords a Fisher scoring algorithm for maximization of the likelihood function in more complex cases.) It is also appealing that modelling using the skew t-distribution is performed on the original scale of the data.

Acknowledgements

Thanks are due to Ken Lange for making available the data that were used as our example 2. We are also very grateful to the referees and Joint Editor for prompting much improvement in the paper in revision.

Appendix A

A.1. *Proof of result 1*

From equation (2), we have

$$\begin{aligned} E(T^r) &= 2^{-r}(a+b)^{r/2}\, E\left[\frac{(2B-1)^r}{\{B(1-B)\}^{r/2}}\right] \\ &= \frac{(a+b)^{r/2}}{2^r\, B(a,b)} \int_0^1 (2y-1)^r y^{a-r/2-1} (1-y)^{b-r/2-1}\, \mathrm{d}y \\ &= \frac{(a+b)^{r/2}\, B(a-r/2, b-r/2)}{2^r\, B(a,b)} E\{(2Y_r-1)^r\} \end{aligned}$$

where Y_r has the beta distribution on (0, 1) with parameters $a - r/2$ and $b - r/2$. This is the source of the requirement that a and b both be greater than $r/2$. Formulae (4a) and (4b) follow via binomial expansion of $(2Y_r - 1)^r$, directly in the case of equation (4a) and after writing $2Y_r - 1 = Y_r - (1 - Y_r)$ for equation (4b).

A.2. Proof of result 2

First, note that, as $a \to \infty$, $\Gamma(a - \frac{1}{2})/\Gamma(a) \sim a^{-1/2}$, so that $\mu_{a,b} \sim a\mu_b/2$ and $\sigma_{a,b} \sim a\sigma_b/2$. To $O(1)$:

$$\log(\sigma_{a,b}) \simeq \log(a) + \log(\sigma_b/2),$$

$$-\tfrac{1}{2}\log(a+b) \simeq -\tfrac{1}{2}\log(a),$$

$$-\log\{B(a,b)\} \simeq -\log\{\surd(2\pi)\} - (b - \tfrac{1}{2})\log(b-1) + b\log(a) + b - 1$$

(Abramowitz and Stegun, 1965),

$$(a + \tfrac{1}{2})\log(u_+) \simeq (a + \tfrac{1}{2})\log(2) - \chi^{-2}(t; b)$$

and

$$(b + \tfrac{1}{2})\log(u_-) \simeq -(b + \tfrac{1}{2})\log(a) + (b + \tfrac{1}{2})\log(2) - (2b+1)\log\{\chi(t; b)\},$$

where $u_\pm = [1 \pm (\sigma_{a,b}t + \mu_{a,b})/\{a + b + (\sigma_{a,b}t + \mu_{a,b})^2\}^{1/2}]$ and $\chi(t; b) = \sigma_b t + \mu_b$. Summing all these gives the approximation to $\log\{\sigma_{a,b}\, f(\sigma_{a,b}t + \mu_{a,b})\}$. The coefficients of a and $\log(a)$ are 0, and the leading term is

$$\log(\sigma_b) + \log(2) - \log\{\surd(2\pi)\} + b - 1 - (b - \tfrac{1}{2})\log(b-1) - \chi^{-2}(t; b) - (2b+1)\log\{\chi(t; b)\}$$

which, when exponentiated and with Stirling's formula applied, yields equation (6).

Appendix B

Write $u_k = \{2\sigma^2 + p(X_k - \mu)^2\}^{-1/2}\surd p(X_k - \mu)$, $R = \surd(q^2 + 2p)$ and $\psi(x) = \mathrm{d}[\log\{\Gamma(x)\}]/\mathrm{d}x$. The first derivatives of the log-likelihood using the reparameterized form, in the order associated with p then q, μ and σ, can be shown to be

$$\begin{aligned}\frac{-1}{2p^2}\Bigg(&4n\,\psi\Big(\frac{2}{p}\Big) - np - 4n\log(2) \\ &- \Big(1 + \frac{q}{R} + \frac{p}{2}\Big)\sum_{k=1}^{n} u_k(1 - u_k) + \Big(1 - \frac{q}{R} + \frac{p}{2}\Big)\sum_{k=1}^{n} u_k(1 + u_k) \\ &+ 2\Big\{1 + \frac{q}{R^3}(R^2 + p)\Big\}\Big[\sum_{k=1}^{n}\log(1 + u_k) - n\,\psi\Big\{\frac{1}{p}\Big(1 + \frac{q}{R}\Big)\Big\}\Big] \\ &+ 2\Big\{1 - \frac{q}{R^3}(R^2 + p)\Big\}\Big[\sum_{k=1}^{n}\log(1 - u_k) - n\,\psi\Big\{\frac{1}{p}\Big(1 - \frac{q}{R}\Big)\Big\}\Big]\Bigg),\end{aligned}$$

$$\frac{2}{R^3}\Big(-n\Big[\psi\Big\{\frac{1}{p}\Big(1 + \frac{q}{R}\Big)\Big\} - \psi\Big\{\frac{1}{p}\Big(1 - \frac{q}{R}\Big)\Big\}\Big] + \sum_{k=1}^{n}\{\log(1 + u_k) - \log(1 - u_k)\}\Big),$$

$$\frac{1}{\sigma\surd(2p)}\Big\{\Big(1 - \frac{q}{R} + \frac{p}{2}\Big)\sum_{k=1}^{n}\surd(1 - u_k^2)(1 + u_k) - \Big(1 + \frac{q}{R} + \frac{p}{2}\Big)\sum_{k=1}^{n}\surd(1 - u_k^2)(1 - u_k)\Big\}$$

and

$$-\frac{n}{\sigma} + \frac{1}{p\sigma}\Big\{\Big(1 - \frac{q}{R} + \frac{p}{2}\Big)\sum_{k=1}^{n} u_k(1 + u_k) - \Big(1 + \frac{q}{R} + \frac{p}{2}\Big)\sum_{k=1}^{n} u_k(1 - u_k)\Big\}.$$

Next, we present the elements j^k of the observed information matrix $j = \Sigma_{k=1}^n j^k$. Additional notation is $\psi'(x) = \mathrm{d}^2[\log\{\Gamma(x)\}]/\mathrm{d}^2 x$. Ignoring the dependence on k for clarity, these elements are

$$j_{pp} = \frac{1}{p^4}\Big[\Big\{1+\frac{q(R^2+p)}{R^3}\Big\}^2\psi'\Big\{\frac{1}{p}\Big(1+\frac{q}{R}\Big)\Big\}+\Big\{1-\frac{q(R^2+p)}{R^3}\Big\}^2\psi'\Big\{\frac{1}{p}\Big(1-\frac{q}{R}\Big)\Big\}$$
$$+\frac{2q}{R^3}(R^2+p)up - 4\,\psi'\Big(\frac{2}{p}\Big)-\frac{p^2}{2}-u^2p+\frac{p}{2}\Big\{\frac{q}{R}(u^3-3u)+2(p+1)u^2-(p+2)u^4\Big\}\Big],$$

$$j_{pq} = \frac{-2}{p^2R^3}\Big[\Big\{1+\frac{q(R^2+p)}{R^3}\Big\}\psi'\Big\{\frac{1}{p}\Big(1+\frac{q}{R}\Big)\Big\}-\Big\{1-\frac{q(R^2+p)}{R^3}\Big\}\psi'\Big\{\frac{1}{p}\Big(1-\frac{q}{R}\Big)\Big\}+up\Big],$$

$$j_{p\mu} = \frac{\surd\{2(1-u^2)\}}{p^{3/2}\sigma}\Big[\frac{-q(R^2+p)}{R^3}+u-\frac{(1-u^2)}{2}\Big\{(p+2)u-\frac{q}{R}\Big\}\Big],$$

$$j_{p\sigma} = \frac{2u}{p^2\sigma}\Big[\frac{-q(R^2+p)}{R^3}+u-\frac{(1-u^2)}{2}\Big\{(p+2)u-\frac{q}{R}\Big\}\Big],$$

$$j_{qq} = \frac{4}{R^6}\Big[\psi'\Big\{\frac{1}{p}\Big(1+\frac{q}{R}\Big)\Big\}+\psi'\Big\{\frac{1}{p}\Big(1-\frac{q}{R}\Big)\Big\}\Big],$$

$$j_{q\mu} = \frac{2\surd(2p)\surd(1-u^2)}{R^3\sigma},$$

$$j_{q\sigma} = \frac{4u}{R^3\sigma},$$

$$j_{\mu\mu} = \frac{1-u^2}{2\sigma^2}\left\{\Big(1-\frac{q}{R}+\frac{p}{2}\Big)(1+u)(1-2u)+\Big(1+\frac{q}{R}+\frac{p}{2}\Big)(1-u)(1+2u)\right\},$$

$$j_{\mu\sigma} = \frac{(1-u^2)^{3/2}}{\sigma^2\surd(2p)}\left\{\Big(1-\frac{q}{R}+\frac{p}{2}\Big)(1+2u)-\Big(1+\frac{q}{R}+\frac{p}{2}\Big)(1-2u)\right\},$$

$$j_{\sigma\sigma} = -\frac{1}{\sigma^2}+\frac{u}{p\sigma^2}\left\{\Big(1-\frac{q}{R}+\frac{p}{2}\Big)(1+u)(2+u-2u^2)-\Big(1+\frac{q}{R}+\frac{p}{2}\Big)(1-u)(2-u-2u^2)\right\}.$$

Appendix C

Preliminary Taylor series expansions include

$$\log\{\Gamma(k)\} \simeq k\log(k)-k-\frac{1}{2}\log(k)+\frac{1}{2}\log(2\pi)+\frac{1}{12k}-\frac{1}{360k^3}$$

(Abramowitz and Stegun (1965), page 257) so that

$$G_k \equiv \frac{\Gamma(k-\frac{1}{2})}{2\,\Gamma(k)} \simeq \frac{1}{2\surd k}\left(1+\frac{3}{8k}+\frac{25}{128k^2}+\frac{105}{1024k^3}\right)$$

and

$$\psi(k) \simeq \log(k)-\frac{1}{2k}-\frac{1}{12k^2}$$

so that

$$\psi(k+L)-\psi(k) \simeq \frac{L}{k}-\frac{L(L-1)}{2k^2}$$

as $k \to \infty$ with L fixed.

With l as described before equations (8), currently written as a function of a, b, $\mu_{a,b}$ and $\sigma_{a,b}$, we have

$$\frac{\partial l}{\partial a} = \frac{\sigma'_{a,b}}{\sigma_{a,b}} + \psi(a+b) - \psi(a) - \frac{1}{2(a+b)} - \log(2) + \log(u_+) + \frac{(a+\frac{1}{2})u'}{u_+} - \frac{(b+\frac{1}{2})u'}{u_-}$$

where primes signify differentiation with respect to a, $u_\pm$ is given in Appendix A and $u' \equiv u'_+$. For large a, $\mu_{a,b} \simeq aG_b - (b/2 - 3/8)G_b \equiv aG_b + A$, say, and $\sigma_{a,b} \simeq a(\sigma_b/2) + \{\sigma_b^2 - 1 + (4b+1)G_b^2\}/4\sigma_b \equiv \sigma_b/2 + B$, say, where σ_b is defined in result 2. Write $y = \sigma_b/2 + G_b$ and $K = At + B$. Then,

$$u_\pm \simeq 1 \pm \left\{1 - \frac{1}{2ay^2} + \frac{1}{a^2}\left(\frac{3}{8y^4} - \frac{b}{2y^2} + \frac{K}{y^3}\right)\right\}$$

and u' can be written accordingly. It follows that

$$\frac{\partial l}{\partial a} \simeq \frac{1}{a^2}\left\{\frac{b(1-3b)}{2} - \frac{2A}{\sigma_b} - \frac{5}{32y^4} - \frac{K}{2y^3} + \frac{2b+1}{2y^2} + \frac{(2b+1)K}{y}\right\}. \tag{11}$$

Similarly,

$$\begin{aligned}\frac{\partial l}{\partial b} &= \frac{\sigma^\sharp_{a,b}}{\sigma_{a,b}} + \psi(a+b) - \psi(b) - \frac{1}{2(a+b)} - \log(2) + \log(u_-) + \frac{(a+\frac{1}{2})u^\sharp}{u_+} - \frac{(b+\frac{1}{2})u^\sharp}{u_-} \\ &\simeq \frac{\sigma_b^\sharp}{\sigma_b} - \psi(b) - 2\log(2) - 2\log(y) + \frac{y^\sharp}{2y^3} - (2b+1)\frac{y^\sharp}{y},\end{aligned} \tag{12}$$

where the hash symbols signify differentiation with respect to b. This is $O(1)$ in a.

Now, much as in Prentice (1975), page 612,

$$\begin{aligned}\lim_{p\to 0}\left(\frac{\partial l}{\partial p}\right) &= \lim_{a\to\infty}\left(-\frac{a^2}{2}\frac{\partial l}{\partial a} - \frac{3b^2}{2}\frac{\partial l}{\partial b}\right), \\ \lim_{p\to 0}\left(\frac{\partial l}{\partial q}\right) &= \lim_{a\to\infty}\left(-2b^{3/2}\frac{\partial l}{\partial b}\right).\end{aligned} \tag{13}$$

These are both, clearly, $O(1)$ in a, explicit expressions being available by the insertion of expressions (11) and (12) above.

It remains to obtain the limits of $\partial l/\partial p$ and $\partial l/\partial q$ as q also tends to 0, by letting $b \to \infty$ above. The elements of expression (11) are functions of σ_b and G_b. The former is also, in fact, a function of G_b, but it remains useful to give its asymptotic approximation as

$$\sigma_b \simeq \frac{1}{2b}\left(1 + \frac{15}{16b} + \frac{439}{512b^2}\right).$$

Much further manipulation shows terms of order b^2, $b^{3/2}$ and b in the bracketed right-hand side of expression (11) to be 0. Leading terms turn out to be

$$a^2\frac{\partial l}{\partial a} \simeq \frac{5}{8}\sqrt{b}(3t - t^3) - \frac{5}{32}(7t^4 - 12t^2 + 1). \tag{14}$$

Similar manipulations apply to the elements of equation (12) which are functions of σ_b and G_b and their derivatives with respect to b. Terms of order 1, $1/\sqrt{b}$ and $1/b$ are 0, and we find that

$$\frac{\partial l}{\partial b} \simeq -\frac{5}{24b^{3/2}}(3t - t^3) + \frac{1}{96b^2}(27t^4 - 12t^2 - 19). \tag{15}$$

Entering expressions (14) and (15) into equations (13) yields equations (8).

References

Abramowitz, M. and Stegun, I. A. (eds) (1965) *Handbook of Mathematical Functions*. New York: Dover Publications.

Arnold, B. C. and Groeneveld, R. A. (1995) Measuring skewness with respect to the mode. *Am. Statistn*, **49**, 34–38.

Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scand. J. Statist.*, **12**, 171–178.
Azzalini, A. and Capitanio, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Statist. Soc.* B, **65**, in the press.
Barndorff-Nielsen, O. E. and Shephard, N. (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion). *J. R. Statist. Soc.* B, **63**, 167–241.
Butler, R. J., McDonald, J. B., Nelson, R. D. and White, S. B. (1990) Robust and partially adaptive estimation of regression models. *Rev. Econ. Statist.*, **72**, 321–327.
Cacoullos, T. (1965) A relation between t and F distributions. *J. Am. Statist. Ass.*, **60**, 528–531; correction, 1249.
Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.
DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *J. Am. Statist. Ass.*, **92**, 903–915.
Fernández, C. and Steel, M. J. F. (1998) On Bayesian modelling of fat tails and skewness. *J. Am. Statist. Ass.*, **93**, 359–371.
Fernández, C. and Steel, M. J. F. (1999) Multivariate Student-t regression models: pitfalls and inference. *Biometrika*, **86**, 153–167.
Fisher, R. A. (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**, 507–521.
Fraser, D. A. S. (1979) *Inference and Linear Models*. New York: McGraw-Hill.
Garvin, J. S. and McClean, S. I. (1997) Convolution and sampling theory of the binormal distribution as a prerequisite to its application in statistical process control. *Statistician*, **46**, 33–47.
Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*. London: Chapman and Hall.
Gibbons, J. F. and Mylroie, S. (1973) Estimation of impurity profiles in ion-implanted amorphous targets using joined half-Gaussian distributions. *Appl. Phys. Lett.*, **22**, 568–569.
Hosking, J. R. M. (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Statist. Soc.* B, **52**, 105–124.
Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, vol. 2, 2nd edn. New York: Wiley.
Jones, M. C. (2001a) A skew t distribution. In *Probability and Statistical Models with Applications* (eds C. A. Charalambides, M. V. Koutras and N. Balakrishnan), pp. 269–277. London: Chapman and Hall.
Jones, M. C. (2001b) Multivariate t and beta distributions associated with the multivariate F distribution. *Metrika*, **54**, 215–231.
Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modelling using the t distribution. *J. Am. Statist. Ass.*, **84**, 881–896.
Mudholkar, G. S. and Hutson, A. (2000) The epsilon-skew-normal distribution for analyzing near-normal data. *J. Statist. Planng Inf.*, **83**, 291–309.
Mukhopadhyay, S. and Vidakovic, B. (1995) Efficiency of linear Bayes rules for a normal mean: skewed priors class. *Statistician*, **44**, 389–397.
O'Hagan, A. and Leonard, T. (1976) Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, **63**, 201–202.
Prentice, R. L. (1975) Discrimination among some parametric models. *Biometrika*, **62**, 607–614.
Sahu, S. K., Dey, D. K. and Branco, M. D. (2002) A new class of multivariate skew distributions with applications to Bayesian regression models. To be published.
Skates, S. J. (1993) On secant approximations to cumulative distribution functions. *Biometrika*, **80**, 223–235.
Smith, R. L. and Naylor, J. C. (1987) A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Appl. Statist.*, **36**, 358–369.