

On parameter orthogonality in symmetric and skew models

M.C. Jones

Department of Mathematics & Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, U.K.

ABSTRACT

Orthogonal and partly orthogonal reparametrisations are provided for certain wide and important families of univariate continuous distributions. First, the orthogonality of parameters in location-scale symmetric families is extended to symmetric distributions involving a third parameter. This sets the scene for consideration of the four-parameter situation in which skewness is also allowed. It turns out that one specific approach to generating such four-parameter families, that of two-piece distributions with a certain parametrisation restriction, has some attractive features with regard to parameter orthogonality. This work also affords partly orthogonal parametrisations of three-parameter two-piece models.

Keywords: Extended location-scale model; Skew- t distribution; Two-piece distribution.

1. Introduction

In this paper, we consider theoretical aspects of maximum likelihood (ML) estimation of the parameters of families of univariate continuous distributions on the whole of \mathbb{R} based on an i.i.d. sample X_1, \dots, X_n taken from the model in question. Parameter orthogonality, which implies asymptotic independence between the parameter estimates concerned, occurs naturally in (two-parameter) symmetric location-scale models of the form

$$\frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right); \tag{1.1}$$

here and throughout the paper, f is symmetric about zero and $\mu \in \mathbb{R}$ and $\sigma > 0$ are location and scale parameters, respectively. However, parameter orthogonality seems to be at least partially lost in distributions with additional shape parameters. We will show how parameter orthogonality can be reclaimed — at least in principle — for all (three-parameter) symmetric distributions with an additional shape parameter. We will then investigate a particular type of (four-parameter) asymmetric extension for which most, but not all, off-diagonal elements of the expected information matrix can be made to be zero. The latter also covers certain three-parameter asymmetric distributions.

We will first consider (in Section 2) a family of symmetric distributions indexed by three parameters, location μ , scale σ and a third parameter $\delta > 0$ which, in some appropriate way, allows control over shape (in particular, this parameter will usually control tailweight):

$$\frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}; \delta\right). \tag{1.2}$$

It turns out — and this is not an especially novel observation although the general treatment here might be useful — that in model (1.2) the location parameter remains orthogonal to the scale parameter and is also orthogonal to the shape parameter. In terms of inference on the location parameter, which is often the main parameter of interest, this is itself a useful property.

Scale and shape parameters in this model, however, are not orthogonal. Indeed, when the shape parameter is a tailweight parameter we will demonstrate how a high (asymptotic) correlation between $\hat{\sigma}$ and $\hat{\delta}$ often arises where, as throughout the paper, hats over parameters denote ML estimators. This, in turn, illustrates the degree of difficulty that the data has in knowing what to assign to variations in scale and what to assign to variations in tailweight. However, since there is just one nonzero asymptotic correlation in this case, orthogonal reparametrisation is clearly possible (Cox & Reid, 1987), and this is investigated. The main purpose of this orthogonalisation is probably to make the model more amenable to well-behaved numerical likelihood maximisation.

Most experience to date suggests that the asymptotic independence of location and other parameters is lost when a further parameter, γ say, allowing and controlling asymmetry is introduced. Indeed, in almost all cases we know of, this results in the appearance of no zero entries whatsoever in the corresponding information matrix. This is certainly true of a variety of

four-parameter distributions in the author’s own work (Jones & Faddy, 2003, Jones, 2008a, Jones & Pewsey, 2009). The situation seems to be little different for four-parameter Azzalini-type skew distributions (Azzalini, 1985). For example, we can find no claim of zeroes in the information matrix of the skew t distribution of Branco & Dey (2001) and Azzalini & Capitanio (2003) (see also Azzalini & Genton, 2008). However, there is a single zero in a closely related skew- t model investigated by Gómez, Venegas & Bolfarine (2007); this zero corresponds to the term associated with skewness and tailweight parameters.

Until recently, we thought that this situation might be inevitable. However, one specific model, a “two-piece” skew- t distribution (Fernández & Steel, 1998) with a particular parametrisation of skewness (Arellano-Valle, Gómez & Quintana, 2005) appears in the literature within which this is not the case: location and skewness parameter are tied up, as are scale and tailweight parameters, but each of the former pair are asymptotically independent of each of the latter. An information matrix of this form was first produced in this special case — but without further comment — by Gómez, Torres & Bolfarine (2007). In Section 3 of this paper, we show that this is a general property of appropriately parametrised two-piece distributions. Moreover, we observe that one can go on, at least in principle, to orthogonalise parametrisations within location/skewness and scale/tail pairs, if desired. Unfortunately, combination of the two cannot be made to fully orthogonalise the set of four parameters. However, a useful partly orthogonal reparametrisation results in $\hat{\mu}$ being asymptotically independent of all three other parameters (with the scale-tail orthogonality being preserved as well).

In Section 4, two-piece distributions with skewness but without an extra shape parameter are dealt with briefly. Results here arise trivially from the work of Section 3 but such distributions include some noteworthy special cases.

This paper is essentially concerned with regular parametric likelihood asymptotics. Standard regularity conditions to achieve consistency and asymptotic normality apply in Section 2. The models of interest in Sections 3 and 4 are derived from those in Section 2 with the addition of a discontinuity in either first or, more often, second derivative with respect to μ of the likelihood at the points $\mu = X_i$, $i = 1, \dots, n$; results apply under alternative regularity conditions such as those of Huber (1967). The novelty of this paper lies not in the theoretical techniques employed but in the insights the standard theory affords.

As a running example in Sections 2 and 3, take f to be f_T , the density of the Student t distribution (which is symmetric) with unknown degrees of freedom δ (which controls tailweight):

$$f_T(x; \delta) = K_\delta \left(1 + \frac{x^2}{\delta}\right)^{-\frac{1}{2}(\delta+1)}$$

where $K_\delta = \Gamma(\frac{1}{2}(\delta + 1))/\{\sqrt{\pi\delta} \Gamma(\frac{1}{2}\delta)\}$. When δ is fixed, $\hat{\mu}$ and $\hat{\sigma}$ are indeed asymptotically independent with asymptotic variances $(\delta + 3)\sigma^2/\{(\delta + 1)n\}$ and $(\delta + 3)\sigma^2/(2\delta n)$, respectively (these formulae being special cases of those to follow).

2. The three-parameter symmetric case

2.1. Likelihood fitting in the initial parametrisation

Write $Y_i = (X_i - \mu)/\sigma$, $i = 1, \dots, n$. From (1.2), the log-likelihood is

$$\ell(\mu, \sigma, \delta) = -n \log \sigma + \sum_{i=1}^n \log f(Y_i; \delta).$$

Let primes denote differentiation of $f(x; \delta)$ with respect to x and circles denote differentiation of $f(x; \delta)$ with respect to δ . The score equations with respect to each of μ , σ and δ in turn are

$$-\frac{1}{\sigma} \sum_{i=1}^n (\log f)'(Y_i; \delta) = 0, \quad -\frac{1}{\sigma} \left\{ n + \sum_{i=1}^n Y_i (\log f)'(Y_i; \delta) \right\} = 0,$$

$$\sum_{i=1}^n (\log f)^\circ(Y_i; \delta) = 0.$$

Elements of the observed information matrix are

$$\begin{aligned} -\frac{\partial^2 \ell}{\partial \mu^2} &= -\frac{1}{\sigma^2} \sum_{i=1}^n (\log f)''(Y_i; \delta), & -\frac{\partial^2 \ell}{\partial \mu \partial \sigma} &= -\frac{1}{\sigma^2} \sum_{i=1}^n Y_i (\log f)''(Y_i; \delta), \\ -\frac{\partial^2 \ell}{\partial \mu \partial \delta} &= \frac{1}{\sigma} \sum_{i=1}^n (\log f)^\circ(Y_i; \delta), & -\frac{\partial^2 \ell}{\partial \sigma^2} &= \frac{1}{\sigma^2} \left\{ n - \sum_{i=1}^n Y_i^2 (\log f)''(Y_i; \delta) \right\}, \end{aligned}$$

$$-\frac{\partial^2 \ell}{\partial \sigma \partial \delta} = \frac{1}{\sigma} \sum_{i=1}^n Y_i (\log f)'^\circ(Y_i; \delta), \quad -\frac{\partial^2 \ell}{\partial \delta^2} = -\sum_{i=1}^n (\log f)^{\circ\circ}(Y_i; \delta).$$

Now, let $\iota_{\eta\xi} = E\{-(\partial^2 \ell / \partial \eta \partial \xi)(Y)\}$ so that the expected information matrix is n times the matrix made up of ι values. Then, the structure of the information matrix arising from the symmetry of the model is as follows:

$$\begin{aligned} \iota_{\mu\mu} &= I_m(\delta)/\sigma^2, & \iota_{\mu\sigma} &= 0, & \iota_{\mu\delta} &= 0, \\ \iota_{\sigma\sigma} &= I_s(\delta)/\sigma^2, & \iota_{\sigma\delta} &= I_c(\delta)/\sigma, & \iota_{\delta\delta} &= I_d(\delta), \end{aligned}$$

say, where the I functions are all independent of μ and σ . How so? Well, first, the symmetry of f (and hence of $\log f$) means that $(\log f)'$ is an odd function and $(\log f)''$ is an even function. Differentiation with respect to δ does not disturb the symmetry of the function (because $\log f$ depends on $|x|$ only both before and after differentiation with respect to δ). Therefore, $(\log f)^{\circ\circ}$ is an even function while $(\log f)'^\circ$ is an odd function. This accounts for the two zeroes — and no more — above. Second, the dependence on σ and δ (and vanishing of μ) arises from noting that any function h , depending on δ but not on μ and σ except through its dependence on Y_i , has expectation

$$E(h(Y; \delta)) = \int h\left(\frac{x - \mu}{\sigma}; \delta\right) \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}; \delta\right) dx = \int h(w; \delta) f(w; \delta) dw,$$

which depends only on δ and not on μ or σ . In fact, we have

$$I_m(\delta) = -E\{(\log f)''(X; \delta)\}, \quad I_s(\delta) = 1 - E\{X^2(\log f)''(X; \delta)\}, \quad (2.1)$$

$$I_c(\delta) = E\{X(\log f)'^\circ(X; \delta)\}, \quad I_d(\delta) = -E\{(\log f)^{\circ\circ}(X; \delta)\}$$

where $X \sim f(x; \delta)$. Now, hide the explicit dependence of f on δ in the following for convenience and define $\mathcal{I}_p = \int x^p \{(f')^2/f\}(x) dx$. Of course,

$$I_m(\delta) = \mathcal{I}_0 > 0,$$

the Fisher information for location. It is also the case that $\int x^2(\log f)''(x)f(x) dx = \int x^2 f''(x) dx - \mathcal{I}_2 = 2 - \mathcal{I}_2$. It follows that

$$I_s(\delta) = \mathcal{I}_2 - 1 > 0,$$

the inequality proveable using the Cauchy-Schwartz inequality. Also,

$$I_d(\delta) = \int \{(f^\circ)^2/f\}(x)dx > 0.$$

It is clear that the asymptotic independence of $\hat{\mu}$ and $\hat{\sigma}$ is maintained and is joined by asymptotic independence between $\hat{\mu}$ and $\hat{\delta}$. The asymptotic variance of $\hat{\mu}$ remains the same as for the two-parameter (known δ) symmetric case, as $\sigma^2/(\mathcal{I}_0 n)$.

However, because $\iota_{\sigma\delta} \neq 0$, the scale and shape parameters are not orthogonal. In fact, $\text{Corr}(\hat{\sigma}, \hat{\delta})$, which does not depend on (μ) or σ asymptotically, equals $-I_c(\delta)/\sqrt{I_s(\delta)I_d(\delta)}$. This correlation can be plotted as a function of δ for each specific choice of f and this will be done for the t distribution in Section 2.3. There and in other three-parameter symmetric families whose maximum likelihood estimation we have investigated (distributions with simple exponential tails in Jones, 2008a; the symmetric sinh-arcsinh distribution in Jones & Pewsey, 2009), the correlation is high for almost all practically important values of δ . This reflects the fact that one cannot tell the difference between changing scale and changing tailweight — for the shape parameter controls tailweight in the models mentioned — very easily in practice.

2.2. Orthogonal reparametrisation

With just one nonzero asymptotic correlation, orthogonal reparametrisation (Cox & Reid, 1987) of σ and δ is possible, at least in principle, and in a number of ways. Moreover, since the elements of the expected information matrix do not depend on μ , any such orthogonal reparametrisation remains orthogonal to μ . Orthogonal reparametrisations preserving one of the two original parameters work out nicely because of the structure of the information matrix in this case. The altered parameter is a product of functions of σ and δ ; the functions of σ are in each case simple and explicit; those of δ are more complicated and may not be easy to obtain in practice.

First, seek an orthogonal reparametrisation of the form $\{\mu, \sigma, \chi(\sigma, \delta)\}$. Noting that δ is thus a function of σ and χ , we need (Cox & Reid, 1987, Young & Smith, 2005, Section 9.2)

$$i_{\delta\delta} \frac{\partial\delta}{\partial\sigma} + i_{\sigma\delta} = I_d(\delta) \frac{\partial\delta}{\partial\sigma} + \frac{I_c(\delta)}{\sigma} = 0. \tag{2.2}$$

This is satisfied whenever $C(\delta) = \log \sigma + h(\chi)$ where h here and below denotes an arbitrary function and

$$C^\circ(\delta) = -I_d(\delta)/I_c(\delta). \quad (2.3)$$

One version of the resulting orthogonal reparametrisation has the form

$$\{\mu, \sigma, C(\delta) - \log \sigma\}$$

although arbitrary monotone transformations of any of these parameters are, of course, also orthogonal.

A similar development for the reparametrisation $\{\mu, \psi(\sigma, \delta), \delta\}$ is possible, leading via the requirement

$$\frac{I_s(\delta)}{\sigma^2} \frac{\partial \sigma}{\partial \delta} + \frac{I_c(\delta)}{\sigma} = 0$$

to $\sigma = h(\psi)/P(\delta)$ and thus, up to monotone transformations, the orthogonal parametrisation

$$\{\mu, \sigma P(\delta), \delta\}$$

where

$$(\log P)^\circ(\delta) = I_c(\delta)/I_s(\delta). \quad (2.4)$$

If σ is a parameter of interest, then the first of these reparametrisations should be preferred; if δ , then the second. If neither is of particular interest (as when μ alone is), then whichever is more tractable will have the edge.

2.3. Running example: the t distribution

The formulae involved in the expected information matrix for this case are

$$I_m(\delta) = \frac{\delta + 1}{\delta + 3}, \quad I_s(\delta) = \frac{2\delta}{\delta + 3}, \quad I_c(\delta) = -\frac{2}{(\delta + 1)(\delta + 3)}$$

and

$$I_d(\delta) = \frac{1}{2} \left\{ \frac{1}{2} \psi' \left(\frac{1}{2} \delta \right) - \frac{1}{2} \psi' \left(\frac{1}{2} (\delta + 1) \right) - \frac{\delta + 5}{\delta(\delta + 1)(\delta + 3)} \right\}.$$

These formulae match with asymptotic likelihood calculations for the t distribution in, for example, Lange, Little & Taylor (1989), Taylor & Verbyla (2004) and Vasconcellos & da Silva (2005).

The covariance between $\hat{\sigma}$ and $\hat{\delta}$ is therefore

$$2/((\delta + 1) [\frac{1}{2}\delta(\delta + 1)(\delta + 3) \{\psi'(\frac{1}{2}\delta) - \psi'(\frac{1}{2}(\delta + 1))\} - (\delta + 5)])^{1/2}.$$

This is plotted as the solid line in Fig. 1. The limit to which the correlation quickly rises with increasing δ is actually $2/\sqrt{7} \simeq 0.756$. The key to proving this is to show, via (6.4.12) of Abramowitz & Stegun (1965), that

$$\psi'(\frac{1}{2}\delta) - \psi'(\frac{1}{2}(\delta + 1)) \simeq \frac{2}{\delta^2} + \frac{2}{\delta^3} + O\left(\frac{1}{\delta^5}\right)$$

as $\delta \rightarrow \infty$.

* * * Fig. 1 about here * * *

The orthogonal reparameterisation of the form $\{\mu, \sigma, C(\delta) - \log \sigma\}$ is not explicitly available in this case, but the orthogonal reparameterisation of the form $\{\mu, \sigma P(\delta), \delta\}$ is. This is because $I_c(\delta)/I_s(\delta) = -1/\{\delta(\delta + 1)\}$ so that the orthogonalising factor is

$$P(\delta) \propto 1 + \frac{1}{\delta}.$$

We attempted to mention this result on p.164 of Jones & Faddy (2003) but unfortunately quoted the result wrong: that paper gives $1/P$ in place of P . We have been unable to reconcile this result with the orthogonal parametrisation put forward by Taylor & Verbyla (2004, p.98).

3. The four-parameter case

As described in the introduction, the transition from the three-parameter symmetric case to the four-parameter (asymmetric) case typically results in an absence of zero entries in the information matrix. It is, however, a different story in one particular family of four-parameter asymmetric distributions. There, it is readily possible to obtain an information matrix with four zeroes. Asymptotic independence between location, μ , and both scale, σ , and tail, δ , parameters is maintained, while one also obtains asymptotic independence between skewness, γ , and both scale and tail parameters. In other words, the four parameters $\{\mu, \gamma, \sigma, \delta\}$ fall into two groups of two, $\{\mu, \gamma\}$ and $\{\sigma, \delta\}$, with asymptotic independence between any pair of parameters with one member

in each group. A first example of such an information matrix in this type of context appears in Gómez, Torres & Bolfarine (2007).

The family in question is that of ‘two-piece’ distributions. These are simply made up by splitting a symmetric density at its centre, differentially scaling its two halves, and putting the halves back together (with a continuous and often differentiable join at the centre). Skewness is introduced and controlled by a single parameter associated with the ratio of the differential scalings. In density terms, we have

$$\frac{2}{\sigma(a(\gamma) + b(\gamma))} \left\{ f\left(\frac{x - \mu}{\sigma b(\gamma)}; \delta\right) I(x < \mu) + f\left(\frac{x - \mu}{\sigma a(\gamma)}; \delta\right) I(x \geq \mu) \right\} \quad (3.1)$$

where $I(A)$ is the indicator function of the set A . Of course, $a(\gamma), b(\gamma) > 0$ and when $a(\gamma) = b(\gamma)$, a rescaled version of (1.2) returns. This particular way of parametrising the distribution is due to Arellano-Valle, Gómez & Quintana (2005) and will be discussed further below. Two-piece distributions have a long history. In the case of normal f they go back to Fechner (1897). Some of the more prominent papers concerning such distributions include Hansen (1994), Fernández & Steel (1998), Mudholkar & Hutson (2000), Arellano-Valle et al. (2005), Bauwens & Laurent (2005) and Cassart, Hallin & Paindaveine (2008). Alternative names for such distributions include split distributions and epsilon-skew distributions.

3.1. Likelihood fitting in the Arellano-Valle et al. parametrisation

For simplicity, write $a = a(\gamma)$ and $b = b(\gamma)$. Then, the log-likelihood is

$$\begin{aligned} \ell(\mu, \gamma, \sigma, \delta) &= n \log 2 - n \log(a + b) - n \log \sigma \\ &\quad + \sum_{i=1}^n \left\{ \log f\left(\frac{Y_i}{b}\right) I(Y_i < 0) + \log f\left(\frac{Y_i}{a}\right) I(Y_i \geq 0) \right\}. \end{aligned}$$

The score equations and elements of the observed information matrix are given in the Appendix. Now take expectations of the latter to form elements of the expected information matrix. The oddness and evenness properties of derivatives of $\log f$, used in Section 2.1, also drive the following.

First, it is easy to see that

$$\iota_{\mu\sigma} = \iota_{\mu\delta} = 0$$

as in the symmetric case.

Second,

$$\iota_{\gamma\sigma} = -\frac{1}{\sigma} \left\{ (a' + b') \int_0^\infty x(xg_+''(x) + g_+'(x))f(x)dx \right\}$$

and

$$\iota_{\gamma\delta} = (a' + b') \int_0^\infty xg_+^{\circ}(x)f(x)dx.$$

These too will be zero if $a' + b' = 0$, a condition we impose from now on; that is, we take $a(\gamma) + b(\gamma) = k$ where k is a positive constant whose precise value is unimportant. For concreteness we take $k = 2$ and set $a(\gamma) = 1 - h(\gamma)$, $b(\gamma) = 1 + h(\gamma)$ where $-1 < h(\gamma) < 1$. The simplest choice $h(\gamma) = \gamma$ was made by Mudholkar & Hutson (2000) and preferred in inferential work by Arellano-Valle et al. (2005) and Cassart et al. (2008); it is an effective one. However, asymptotic independence between skewness and the scale and shape/tailweight parameters is sensitive to the skewness parametrisation used. In particular, the aesthetically pleasing choice $a(\gamma) = 1/\gamma$, $b(\gamma) = \gamma$ (Fernández & Steel, 1998, Jones, 2006) is not an appropriate parametrisation in this sense. It should also be noted that the efficacy of this parametrisation is suggested by the information matrix in the special case of the two-piece exponential power distribution with fixed δ provided by Arellano-Valle et al. (2005): it includes $\iota_{\mu\sigma} = \iota_{\gamma\sigma} = 0$.

In summary, the above yields four zeroes in the information matrix. They correspond to each of the four pairs of parameters where one member of the pair is taken from $\{\mu, \gamma\}$ and the other from $\{\sigma, \delta\}$.

The remaining, non-zero, elements of the information matrix follow. Three are precisely the same as before:

$$\iota_{\sigma\sigma} = I_s(\delta)/\sigma^2, \quad \iota_{\sigma\delta} = I_c(\delta)/\sigma, \quad \iota_{\delta\delta} = I_d(\delta).$$

(So, therefore, is the asymptotic correlation between $\hat{\sigma}$ and $\hat{\delta}$.) Also,

$$\iota_{\mu\mu} = \frac{I_m(\delta)}{a(\gamma)(2 - a(\gamma))\sigma^2}$$

which reduces to its value in the symmetric case when $a(\gamma) = 1$. Finally,

$$\iota_{\mu\gamma} = \frac{2a'(\gamma)}{a(\gamma)(2 - a(\gamma))\sigma} I_h(\delta)$$

where

$$I_h(\delta) = - \int_0^\infty \{(\log f)'(x) + x(\log f)''(x)\} f(x) dx = \int_0^\infty x \{(f')^2/f\}(x) dx > 0$$

and

$$i_{\gamma\gamma} = \frac{a'^2(\gamma)}{a(\gamma)(2 - a(\gamma))} \{I_s(\delta) + 1\}.$$

The asymptotic correlation between $\hat{\mu}$ and $\hat{\gamma}$ is

$$-2\text{sign}(a'(\gamma)) \frac{I_h(\delta)}{\sqrt{I_m(\delta)(I_s(\delta) + 1)}}.$$

This is independent of the values of μ and σ and the only role of γ is to determine its sign. Otherwise, the correlation is, again, a function of δ and, again, typically increases to high values (in absolute terms). See Section 3.3 for an example.

3.2. Orthogonal reparametrisation in the four-parameter case

For the pair of parameters $\{\sigma, \delta\}$, everything goes through precisely as in Section 2.2. Moreover, since these reparametrisations do not depend on μ or γ , the reparametrised scale/tail parameters remain orthogonal to μ and γ .

For the pair of parameters $\{\mu, \gamma\}$, things go through analogously to the work in Section 2.2. We will briefly spell out the details in the likely most practically interesting case, that of μ being the interest-preserving parameter (and $\{\mu, \chi(\mu, \gamma)\}$ being the orthogonal reparametrisation). The analogue of (2.2) is

$$i_{\gamma\gamma} \frac{\partial\gamma}{\partial\mu} + i_{\mu\gamma} = \frac{a'^2(\gamma)}{a(\gamma)(2 - a(\gamma))} \{I_s(\delta) + 1\} \frac{\partial\gamma}{\partial\mu} + \frac{2a'(\gamma)}{a(\gamma)(2 - a(\gamma))\sigma} I_h(\delta) = 0.$$

This is satisfied whenever $a(\gamma) = -M(\delta)\sigma^{-1}\mu + h(\chi)$ so that the simplest version of an orthogonal interest-preserving reparametrisation of μ, γ has the form

$$\{\mu, a(\gamma) + M(\delta)\sigma^{-1}\mu\}.$$

Here, we have the explicit formula

$$M(\delta) = 2I_h(\delta)/(I_s(\delta) + 1). \tag{3.2}$$

In fact, a very similar argument goes through if γ is preserved instead of μ . This alternative orthogonal reparametrisation is

$$\{\mu + \sigma G(\delta)a(\gamma), \gamma\}$$

where $G(\delta) = 2I_h(\delta)/I_m(\delta)$.

Observe, however, that these reparametrisations depend on both δ and σ as well as γ and μ . (This is inevitable given the dependence of $\iota_{\mu\mu}$, $\iota_{\mu\gamma}$ and $\iota_{\gamma\gamma}$ on δ and σ .) This means that orthogonalising $\{\mu, \gamma\}$ results in non-zero asymptotic correlations between the orthogonalised pair and elements of $\{\sigma, \delta\}$ whether orthogonally reparametrised or not. One cannot, therefore, provide a fully orthogonal reparametrisation of the two-piece distributions. One can, for example, eschew any $\{\mu, \gamma\}$ reparametrisation and settle for five zero elements in the expected information matrix by orthogonalising only $\{\sigma, \delta\}$. A rather better alternative, at least in inferential terms, would appear to be to implement the μ -preserving reparametrisation given at and above (3.2) together with any preferred $\{\sigma, \delta\}$ parametrisation. This yields asymptotic orthogonality between $\hat{\mu}$ and each of the other three derived parameters (as well as a fourth zero in the information matrix if $\{\sigma, \delta\}$ are orthogonally reparametrised). For example, one might employ

$$\{\mu, \chi, \psi, \delta\} = \{\mu, a(\gamma) + M(\delta)\sigma^{-1}\mu, \sigma P(\delta), \delta\}. \quad (3.3)$$

Note that the asymptotic variance of the location estimate $\hat{\mu}$ is now $1/(n\iota_{\mu\mu}) = a(\gamma)(2 - a(\gamma))\sigma^2/(nI_m(\delta)) = a(\gamma)(2 - a(\gamma))\sigma^2/(n\mathcal{I}_0)$ which reduces to $\sigma^2/(n\mathcal{I}_0)$ only under symmetry. It is clear that this asymptotic variance is smaller than that which pertained for $\hat{\mu}$ under the original non-orthogonal parametrisation, namely,

$$a(\gamma)(2 - a(\gamma))\frac{\sigma^2}{n} \frac{I_s(\delta) + 1}{\{I_m(\delta)(I_s(\delta) + 1) - 4I_h^2(\delta)\}}.$$

3.3. Example continued: the two-piece (skew) t distribution

The only formula appearing in the information matrix that was not given in Section 2.2 is

$$I_h(\delta) = 2(\delta + 1)K_\delta/(\delta + 3).$$

With this in place, the information matrix equates to that in Proposition 2.3 of Gómez et al. (2007) (except it should be noted that their formulae pertain to the scale parameter σ^2 not σ as claimed there).

The asymptotic correlation between estimated location, $\hat{\mu}$, and estimated skewness parameter, $\hat{\gamma}$, is

$$-2\text{sign}(a'(\gamma))K_\delta/\sqrt{3}.$$

Its absolute value is plotted as the dashed line in Fig. 1. As δ increases, it increases towards $4/\sqrt{6\pi} \simeq 0.921$.

The explicit reparametrisation that affords orthogonality of $\hat{\mu}$ to all other parameters together with orthogonality between its final two elements is $\{\mu, \chi, \psi, \delta\}$ is

$$\left\{ \mu, a(\gamma) + \frac{4K_\delta\mu}{3\sigma}, \sigma \left(1 + \frac{1}{\delta} \right), \delta \right\}.$$

4. The three-parameter two-piece case

This is the special case of Section 3 with no extra shape parameter δ . Again, take $a(\gamma) + b(\gamma) = 2$. Immediately, the elements of the information matrix for $\{\mu, \gamma, \sigma\}$ are as in Section 3.1 except that I_m, I_h and I_s are numbers independent of any of the parameters. Partly orthogonal reparameterisations – the novel contribution of this paper in this context, each with two zero off-diagonal elements in the expected information matrix — are given by

$$\left\{ \mu, a(\gamma) + \frac{2I_h\mu}{(I_s + 1)\sigma}, \sigma \right\} \quad \text{and} \quad \left\{ \mu + \frac{2I_h\sigma a(\gamma)}{I_m}, \gamma, \sigma \right\}.$$

The first of these is the probably more useful μ -preserving reparametrisation, which is what we shall concentrate on below. It has nonzero asymptotic correlation between $a(\hat{\gamma}) + 2\{(I_s + 1)\hat{\sigma}\}^{-1}I_h\hat{\mu}$ and $\hat{\sigma}$.

4.1. Example: the two-piece (skew) normal distribution

When $f = \phi$, the standard normal density, density (3.1) without δ reduces to the two-piece skew-normal distribution (dating to Fechner, 1897) in the parametrisation suggested by Mudholkar & Hutson (2000) (who use ϵ in place of γ and set $a(\epsilon) = 1 - \epsilon$). This is also, of course, the limiting case of the

two-piece t distribution as $\delta \rightarrow \infty$. In the two-piece normal case, $I_m = 1$, $I_h = \sqrt{2/\pi}$ and $I_s = 2$. The asymptotic variance-covariance matrix of the ML estimators of μ, γ and σ then corresponds with that given in Theorem 4.7 of Mudholkar & Hutson (2000) (except that they give it for σ^2 rather than σ). Mudholkar & Hutson go on to note that the asymptotic correlation between $\hat{\mu}$ and $\hat{\gamma}$ is $4/\sqrt{6\pi} \approx 0.921$, which is the value associated with $\delta \rightarrow \infty$ in Section 3.3.

The μ -preserving partly orthogonal reparametrisation in this case turns out to be

$$\left\{ \mu, a(\gamma) + \frac{2}{3} \sqrt{\frac{2}{\pi}} \frac{\mu}{\sigma}, \sigma \right\}.$$

(This is, again, the $\delta \rightarrow \infty$ special case of the corresponding reparametrisation for the two-piece t distribution.)

4.2. Example: the asymmetric Laplace distribution

When f is the standard Laplace (double exponential) density $e^{-|x|}/2$, (3.1) without δ yields the asymmetric Laplace density (Kotz, Kozubowski & Podgórski, 2001, Section 3). Partly orthogonal reparametrisations appear to be new. The μ -preserving one is:

$$\left\{ \mu, a(\gamma) + \frac{1}{2\sigma}, \sigma \right\}.$$

5. Closing remarks

There is often more than one way of producing a four-parameter family of distributions with some broadly similar desirable features. For example, there are a number of different “skew t ” distributions, compared briefly in Jones (2008b, including its rejoinder). In that comparison, various skew t distributions exhibit pros and cons and we regarded several of them as still “jostling for position”, with preferences being dependent on the problem at hand and, to some extent, individual investigator. We steered clear of inferential comparisons but expected them to be equally inconclusive. However, the current paper seems to suggest a possible inferential advantage of two-piece distributions over their competitors, none of which yet exhibit similar traits

of parameter orthogonalisability. (Indeed, two-piece distributions have none of the difficulties concerning non-full-rank information matrices for certain important parameter values that also beset some alternatives, e.g. Azzalini & Genton, 2008.) The question remains: is this an intrinsic (or important) advantage of this particular two-piece approach to four-parameter distributions or have researchers just not yet been able to spot similar simplifications in other contexts?

Appendix

The following formulae pertain to the four-parameter case being studied in Section 3.1. Write $g_+(y) = (\log f)(y; \delta)I(y > 0)$ and $g_-(y) = (\log f)(y; \delta)I(y < 0)$. The score equations with respect to each of μ , γ , σ and δ in turn are

$$\begin{aligned} & -\frac{1}{\sigma} \sum_{i=1}^n \left\{ \frac{1}{b} g'_- \left(\frac{Y_i}{b} \right) + \frac{1}{a} g'_+ \left(\frac{Y_i}{a} \right) \right\} = 0, \\ & -\sum_{i=1}^n Y_i \left\{ \frac{b'}{b^2} g'_- \left(\frac{Y_i}{b} \right) + \frac{a'}{a^2} g'_+ \left(\frac{Y_i}{a} \right) \right\} - n \left(\frac{a' + b'}{a + b} \right) = 0, \\ & -\frac{1}{\sigma} \left[n + \sum_{i=1}^n Y_i \left\{ \frac{1}{b} g'_- \left(\frac{Y_i}{b} \right) + \frac{1}{a} g'_+ \left(\frac{Y_i}{a} \right) \right\} \right] = 0, \\ & \sum_{i=1}^n \left\{ g_-^\circ \left(\frac{Y_i}{b} \right) + g_+^\circ \left(\frac{Y_i}{a} \right) \right\} = 0 \end{aligned}$$

and the negative second derivatives are

$$\begin{aligned} & -\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{1}{\sigma^2} \sum_{i=1}^n \left\{ \frac{1}{b^2} g_-'' \left(\frac{Y_i}{b} \right) + \frac{1}{a^2} g_+'' \left(\frac{Y_i}{a} \right) \right\}, \\ & -\frac{\partial^2 \ell}{\partial \mu \partial \gamma} = -\frac{1}{\sigma} \sum_{i=1}^n \left[\frac{b'}{b^2} g'_- \left(\frac{Y_i}{b} \right) + \frac{a'}{a^2} g'_+ \left(\frac{Y_i}{a} \right) + Y_i \left\{ \frac{b'}{b^3} g_-'' \left(\frac{Y_i}{b} \right) + \frac{a'}{a^3} g_+'' \left(\frac{Y_i}{a} \right) \right\} \right], \\ & -\frac{\partial^2 \ell}{\partial \mu \partial \sigma} = -\frac{1}{\sigma^2} \sum_{i=1}^n Y_i \left\{ \frac{1}{b^2} g_-'' \left(\frac{Y_i}{b} \right) + \frac{1}{a^2} g_+'' \left(\frac{Y_i}{a} \right) \right\}, \\ & -\frac{\partial^2 \ell}{\partial \mu \partial \delta} = \frac{1}{\sigma} \sum_{i=1}^n \left\{ \frac{1}{b} g_-^{\circ} \left(\frac{Y_i}{b} \right) + \frac{1}{a} g_+^{\circ} \left(\frac{Y_i}{a} \right) \right\}, \end{aligned}$$

$$\begin{aligned}
-\frac{\partial^2 \ell}{\partial \gamma^2} &= \sum_{i=1}^n \left[Y_i \left\{ \left(\frac{b''b - 2b'^2}{b^3} \right) g'_- \left(\frac{Y_i}{b} \right) + \left(\frac{a'' - 2a'^2}{a^3} \right) g'_+ \left(\frac{Y_i}{a} \right) \right\} \right. \\
&\quad \left. - Y_i^2 \left\{ \frac{b'^2}{b^4} g''_- \left(\frac{Y_i}{b} \right) + \frac{a'^2}{a^4} g''_+ \left(\frac{Y_i}{a} \right) \right\} \right] + n \left\{ \frac{a'' + b''}{a + b} - \left(\frac{a' + b'}{a + b} \right)^2 \right\}, \\
-\frac{\partial^2 \ell}{\partial \gamma \partial \sigma} &= -\frac{1}{\sigma} \sum_{i=1}^n \left[Y_i \left\{ \frac{b'}{b^2} g'_- \left(\frac{Y_i}{b} \right) + \frac{a'}{a^2} g'_+ \left(\frac{Y_i}{a} \right) \right\} + Y_i^2 \left\{ \frac{b'}{b^3} g''_- \left(\frac{Y_i}{b} \right) + \frac{a'}{a^3} g''_+ \left(\frac{Y_i}{a} \right) \right\} \right], \\
-\frac{\partial^2 \ell}{\partial \gamma \partial \delta} &= \sum_{i=1}^n \left[Y_i \left\{ \frac{b'}{b^2} g'^{\circ}_- \left(\frac{Y_i}{b} \right) + \frac{a'}{a^2} g'^{\circ}_+ \left(\frac{Y_i}{a} \right) \right\} \right], \\
-\frac{\partial^2 \ell}{\partial \sigma^2} &= \frac{1}{\sigma^2} \left[n - \sum_{i=1}^n Y_i^2 \left\{ \frac{1}{b^2} g''_- \left(\frac{Y_i}{b} \right) + \frac{1}{a^2} g''_+ \left(\frac{Y_i}{a} \right) \right\} \right], \\
-\frac{\partial^2 \ell}{\partial \sigma \partial \delta} &= \frac{1}{\sigma} \sum_{i=1}^n \left[Y_i \left\{ \frac{1}{b} g'^{\circ}_- \left(\frac{Y_i}{b} \right) + \frac{1}{a} g'^{\circ}_+ \left(\frac{Y_i}{a} \right) \right\} \right], \\
-\frac{\partial^2 \ell}{\partial \delta^2} &= -\sum_{i=1}^n \left\{ g^{\circ\circ}_- \left(\frac{Y_i}{b} \right) + g^{\circ\circ}_+ \left(\frac{Y_i}{a} \right) \right\}.
\end{aligned}$$

References

- Abramowitz, M., Stegun, I.A., 1965. Handbook of Mathematical Functions. Dover, New York.
- Arellano-Valle, R.H., Gómez, H.W., Quintana, F.A., 2005. Statistical inference for a general class of statistical distributions. *J. Statist. Planning Inference* 128, 427–443.
- Azzalini, A., 1985. A class of distributions which includes the normal ones. *Scand. J. Statist.* 12, 171–178.
- 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.
- Azzalini, A., Genton, M.G., 2008. Robust likelihood methods based on the skew-*t* and related distributions. *Internat. Statist. Rev.* 76, 106–129.

- Bauwens, L., Laurent, S., 1998. A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *J. Bus. Econ. Statist.* 23, 346–354.
- Branco, M.D., Dey, D.K., 2001. A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.* 79, 99–113.
- Cassart, D., Hallin, M., Paindaveine, D., 2008. Optimal detection of Fechner-symmetry. *J. Statist. Planning Inference* 138, 2499–2525.
- Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* 49, 1–39.
- Fechner, G.T., 1897. *Kollektivmasslehre*. Engleman, Leipzig.
- Fernández, C., Steel, M.F.J., 1998. On Bayesian modeling of fat tails and skewness. *J. Amer. Statist. Assoc.* 93, 359–371.
- Gómez, H.W., Torres, F.J., Bolfarine, H., 2007. Large-sample inference for the epsilon-skew- t distribution. *Commun. Statist. – Theory Meth.* 36, 73–81.
- Gómez, H.W., Venegas, O., Bolfarine, H., 2007. Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics* 18, 396–407.
- Hansen, B.E., 1994. Autoregressive conditional density estimation. *Internat. Econ. Rev.* 35, 705–730.
- Huber, P.J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In: le Cam, L.M., Neyman, J. (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, 221–233.
- Jones, M.C., 2006. A note on rescalings, reparametrizations and classes of distributions. *J. Statist. Planning Inference* 136, 3730–3733.
- Jones, M.C., 2008a. On a class of distributions with simple exponential tails. *Statist. Sinica* 18, 1101–1110.
- Jones, M.C., 2008b. The t family and their close and distant relations (with discussion). *J. Kor. Statist. Soc.* 37, 293–311.
- Jones, M.C., Faddy, M.J., 2003. A skew extension of the t distribution with applications. *J. Roy. Statist. Soc.* 65, 159–174.

- Jones, M.C., Pewsey, A., 2009. Sinh-arcsinh distributions. *Biometrika*, to appear.
- Kotz, S., Kozubowski, T.J., Podgórski, K., 2001. *The Laplace Distribution and Generalizations: a Revisit With Applications to Communications, Engineering, and Finance*. Birkhäuser, Boston.
- Lange, K.L., Little, R.J.A., Taylor, J.M.G., 1989. Robust statistical modeling using the t distribution. *J. Amer. Statist. Assoc.* 84, 881–896.
- Mudholkar, G.S., Hutson, A.D., 2000. The epsilon-skew-normal distribution for analyzing near-normal data. *J. Statist. Planning Inference*, 83, 291–309.
- Taylor, J., Verbyla, A., 2004. Joint modelling of location and scale parameters of the t distribution. *Statist. Modelling*, 4, 91–112.
- Vasconcellos, K.L.P., da Silva, S.G., 2005. Corrected estimates for Student t regression models with unknown degrees of freedom. *J. Statist. Comput. Simulation* 75, 409–423.
- Young, G.A., Smith, R.L., 2005. *Essentials of Statistical Inference*. Cambridge University Press, Cambridge.

Fig. 1. The asymptotic correlations between $\hat{\sigma}$ and $\hat{\delta}$ (solid line) in the t and two-piece t cases, and the absolute value of that between $\hat{\mu}$ and $\hat{\gamma}$ (dashed line) in the two-piece t case. Each is plotted as a function of $\log_{10} \delta$.

