Bayesian Multivariate Regression Analysis with a New Class of Skewed Distributions

José T.A.S. Ferreira[1] and Mark F.J. Steel[2*]
Endeavour Capital Management, London[1] and Dept. of Statistics, University of Warwick[2]

tome.ferreira@endeavour-capital.com M.F.Steel@stats.warwick.ac.uk

Abstract

In this paper, we discuss a novel class of skewed multivariate distributions and, more generally, a method of building such a class on the basis of univariate skewed distributions. The method is based on a general linear transformation of a multidimensional random variable with independent components, each with a skewed distribution. The proposed class of multivariate skewed distributions has a simple form for the pdf and moment existence only depends on that of the underlying symmetric univariate distributions. In addition, we can freely allow for any mean and covariance structure in combination with any magnitude and direction of skewness. In order to deal with both skewness and fat tails, we introduce multivariate skewed regression models with fat tails, based on Student distributions. We present two main classes of such distributions, one of which is novel even under symmetry. Under standard non-informative priors on both regression and scale parameters, we derive conditions for propriety of the posterior and for existence of posterior moments. We describe MCMC samplers for Bayesian inference and analyse an application to biomedical data.

Key words and phrases: Asymmetric distributions; Heavy tails; Linear regression model; Mardia's measure of skewness; Orthogonal matrices; Posterior propriety.

1 Introduction

The contribution of this article is twofold. We start by introducing a general method for the definition of multivariate skewed distributions. Then we use this new class of distributions in a multivariate regression context and propose Bayesian inference procedures.

So far, the literature on skewed distributions has mainly dealt with univariate cases. Azzalini and Dalla Valle (1996) is one of the first multivariate proposals. Based on the univariate skew-Normal distribution analysed in detail by Azzalini (1985), this method can be interpreted as defining a multivariate skew-Normal density by conditioning on an unobserved argument. Such conditioning models, also known as hidden truncation models (Arnold and Beaver 2000), have been generalised further. Still conditioning on one unobserved variable, Branco and Dey (2001) introduced a class of multivariate skew-elliptical distributions, and Arnold and Beaver (2002) made these models more general by allowing for non-elliptical skew distributions. Within the class of hidden truncation models, but conditioning on as many arguments as observed variables, Sahu, Dey and Branco (2003) (SDB,

[*] *Address for correspondence*: Mark Steel, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. Tel.: +44-24-7652 3369; Fax: +44-24-7652 4532; Email M.F.Steel@stats.warwick.ac.uk.

hereafter) generated a very general class of multivariate skew-elliptical distributions. A different approach to multivariate skewed distributions was proposed by Jones (2002): starting with spherically symmetry, this replaces the marginal distribution of some of the variables by a skewed distribution.

The class that we discuss in this article is based on a general linear transformation of a multi-dimensional random variable with independent components, each having a skewed distribution, with probability density function (pdf) constructed using the method introduced in Fernández and Steel (1998), hereafter denoted by FS. Our proposal for multivariate skewed distributions has a simple form for the pdf and moment existence is only dependent on the existence of the moments of the underlying symmetric univariate distributions. In addition, we can freely allow for any mean and covariance structure in combination with any magnitude and direction of skewness.

Despite focusing on this class, we highlight that it is possible to use any other general method for generating univariate skewed distributions for the independent components, such as the univariate distributions introduced in Azzalini (1985), Azzalini and Capitanio (2003) or Jones and Faddy (2003).

A proposal for multivariate skewed distributions using a linear combination of independent univariate skewed distributions has appeared before in Bauwens and Laurent (2005). However, the one we present here is fundamentally different, as will be explained in the sequel.

Subsequently, we introduce multivariate skewed regression models with fat tails, by considering a linear regression structure with skewed and heavy-tailed error terms. In order to allow for heavy tails we use skewed versions of Student-t distributions. We consider standard non-informative priors on both regression and scale parameters. Skewness and tail behaviour are not fixed but inferred from the data. We derive conditions that make Bayesian analysis feasible (*i.e.* lead to a proper posterior), under the improper prior structure. In addition, we provide results on the existence of posterior moments of the regression coefficients and the determinant of the scale matrix.

We introduce two different Student-based multivariate regression models. One can be represented as a scale mixture of multivariate Normals, and is, thus, characterized by one single mixing variable. Therefore, this leads to the skewed analogue of the multivariate Student-t regression model in Fernández and Steel (1999). In the latter symmetric model, there is no need to use the orthogonal transformations that we introduce in this paper, since the model is based on a multivariate Normal, which is spherical. The moment we introduce skewness such an orthogonal transformation becomes crucial as a means of specifying the directions of the skewness. The other class of heavy-tailed models that we introduce here is based on a transformation of independent Student-t distributed random variables. As this class of distributions is no longer based on a spherical class, we need to use the orthogonal transformations introduced in the sequel, even under symmetry. Thus, the present paper also introduces an, as yet unexplored, class of symmetric heavy-tailed distributions and sheds light on its properties regarding Bayesian inference.

Inference in our regression setup is performed using hybrid Markov chain Monte Carlo (MCMC) samplers using data augmentation. We illustrate the flexibility of the proposed framework in a regression problem using biomedical data from the Australian Institute of Sport.

Section 2 briefly recalls the univariate skewed distributions of FS, while Section 3 introduces the multivariate skewed distributions, together with some properties. Section 4 provides a useful parameterisation of these distributions and Section 5 presents key examples. Section 6 develops the Bayesian multivariate skewed regression models, studies the effect of skewness on the existence of posterior moments, and assesses the feasibility of inference under asymmetric, heavy-tailed sampling.

Section 7 is devoted to the numerical implementation employed to conduct inference. In Section 8 we present the application. Finally, Section 9 provides some concluding remarks. Proofs can be obtained upon request from the authors.

2 Skewed Distributions: The univariate case

FS propose a method for introducing skewness into a unimodal distribution symmetric around zero. The basic idea is to introduce inverse scale factors in the positive and negative half real lines. Let $f(\cdot)$ be a univariate pdf that is symmetric around zero, such that $f(s)$ is decreasing in $|s|$. Let γ be a scalar in $\Re_+$. Then, the skewed distribution on the real line is given by the pdf

$$p(\epsilon|\gamma, f) = \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ f\left(\frac{\epsilon}{\gamma}\right) I_{[0,\infty)}(\epsilon) + f(\gamma\epsilon) I_{(-\infty,0)}(\epsilon) \right\} = \frac{2}{\gamma + \frac{1}{\gamma}} f\left(\epsilon\gamma^{-\text{sign}(\epsilon)}\right), \tag{1}$$

where $I_S(\cdot)$ is the indicator function on S, and $\text{sign}(\cdot)$ is the usual sign function in $\Re$.

If the skewness parameter γ is unity, then we retrieve the original symmetric density. The mode of the density is unchanged, remaining at zero irrespective of the particular value of γ. Also, the probability mass assigned to each side of the mode is independent of $f(\cdot)$ and given by $P(\epsilon > 0 \,|\, \gamma, f) = \gamma^2/(1+\gamma^2)$, allowing γ to parameterise the complete range of mass on each side of the origin. The existence of moments of $p(\epsilon|\gamma, f)$ does not depend on γ, but only on the existence of moments of the initial, symmetric density $f(\cdot)$. Furthermore, the rth moment ($r \in \Re$) is obtained as

$$E(\epsilon^r|\gamma, f) = M_r \frac{\gamma^{r+1} + \frac{(-1)^r}{\gamma^{r+1}}}{\gamma + \frac{1}{\gamma}} \quad \text{where} \quad M_r(f) = \int_0^\infty s^r 2 f(s) ds. \tag{2}$$

Finally, simple manipulation reveals that $p(\epsilon|\gamma, f) = p(-\epsilon|1/\gamma, f)$.

3 Skewed Distributions: The multivariate case

3.1 Definition

The construction of multivariate skewed distributions presented here is based on linear transformations of univariate skewed distributions. Let m be the dimension of the random variable $\epsilon = (\epsilon_1, \ldots, \epsilon_m)' \in \Re^m$ and $\gamma = (\gamma_1, \ldots, \gamma_m)' \in \Re_+^m$. Further, let $f = (f_1(\cdot), \ldots, f_m(\cdot))'$ denote a vector of m unimodal and symmetric univariate pdfs. The pdf of the multivariate skewed distribution with independent components is given by

$$p(\epsilon|\gamma, f) = \prod_{j=1}^{m} p(\epsilon_j|\gamma_j, f_j), \tag{3}$$

where each $p(\epsilon_j|\gamma_j, f_j)$ is as in (1).

Following an affine linear transformation, given a vector $\mu = (\mu_1, \ldots, \mu_m)'$ and a non-singular matrix $A \in R^{m \times m}$, the variable $\eta = (\eta_1, \ldots, \eta_m)' \in R^m$, defined as

$$\eta = A'\epsilon + \mu \tag{4}$$

has a general multivariate skewed distribution, with parameters μ, A, γ and f, denoted by $Sk_m(\mu, A, \gamma, f)$. The pdf for η is given by

$$p(\eta|\mu, A, \gamma, f) = \|A\|^{-1} \prod_{j=1}^{m} p[(\eta - \mu)' A_{\cdot j}^{-1} | \gamma_j, f_j], \tag{5}$$

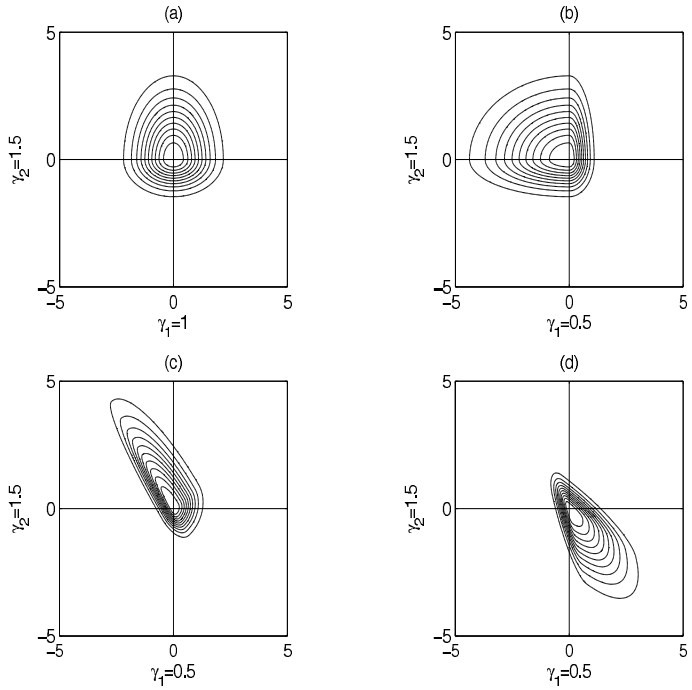

Figure 1: Contour plots of four different bivariate distributions. Plots (a) and (b) correspond to $A = I$, whereas A for plots (c) and (d) was chosen to lead to the same $A'A$ matrix.

where $A^{-1}_{\cdot j}$ denotes the j-th column of A^{-1}, $\|A\|$ denotes the absolute value of the determinant of A, and $p(\cdot|\gamma_j, f_j)$ is as in (1). The distribution of η is unimodal with mode μ, A introduces the dependence between the components of η, while γ determines the skewness of the independent components of ϵ. The distribution in (5) is intended as a flexible (yet parametric) tool for modelling multivariate data, but is not based on any underlying physical interpretation, such as can be assigned to hidden truncation models (see Arnold and Beaver 2002 and SDB).

Figure 1 presents contour plots for four different bivariate skewed distributions, with both $f_1(\cdot)$ and $f_2(\cdot)$ equal to $\phi(\cdot)$, the univariate standard Normal pdf, and μ set to the zero vector. Figure 1 (a) represents the density of a distribution with independent components, where only one of these is (positively) skewed. The remaining three plots were all obtained using the same values for the skewness components, namely $\gamma = (0.5, 1.5)'$. By varying the transformation matrix A it is possible to obtain a diverse set of shapes for the density. If A equals the identity matrix, then the effect of γ is evident (see Figure 1 (b)). In the context of skewed distributions with independent components, γ values larger than one always correspond to a positively skewed marginal, and the reverse happens for values of $\gamma \in (0, 1)$. Figures 1 (c) and (d) represent skewed distributions with dependent components. It can be seen that the shape of the contours varies extensively, even with the same γ, highlighting the flexibility of the method we introduce. To further illustrate the role of the matrix A, we have generated plots (c) and (d) with the same matrix $A'A = [\frac{1}{2} \quad -\frac{1}{2}; -\frac{1}{2} \quad 1]$. As discussed in Subsection 3.3, $A'A$ is all that would matter without skewness.

3.2 Moments

Calculation of the moments of η can be based on the moments of the univariate pdfs $f_j(\cdot)$, $j = 1, \ldots, m$. Further, like in the univariate case, the existence of the moments of η depends only on $f_j(\cdot)$ and not

on the skewness parameters. Due to the linear transformation used in (4), the existence of the rth positive moment of η depends exclusively on the existence of the first r moments of the distributions with density $f_j(\cdot)$. As an illustration, assuming a common $f_j(\cdot) = f(\cdot),\ j = 1, \ldots, m$, the mean vector and the covariance matrix of η are given by

$$E(\eta) = \mu + M_1 A' \begin{pmatrix} \gamma_1 - \frac{1}{\gamma_1} \\ \ldots \\ \gamma_m - \frac{1}{\gamma_m} \end{pmatrix}, \quad \text{and}$$

$$Var(\eta) = A' \left\{ Diag \left[\left(M_2 - M_1^2\right) \left(\gamma_j^2 + \frac{1}{\gamma_j^2} \right) + 2M_1^2 - M_2 \right]_{j=1,\ldots,m} \right\} A,$$

as long as M_1 and M_2, given by (2), both exist.

Thus, even though $E(\eta)$ and $Var(\eta)$ depend on γ directly, their values are not restricted by it. We can obtain any desired mean and covariance values for the distribution even after setting γ, simply by choosing μ and A appropriately. We feel this is an advantage of this class of skewed distributions, when compared to proposals such as the ones introduced by Azzalini and Dalla Valle (1996) or SDB. In these, the set of covariances obtainable after setting the skewness parameters is restricted. In contrast, our framework allows for independent modelling of mean, covariance and skewness.

Figure 2 illustrates how we can fix the covariance and generate quite different distributions by changing both A and γ. All the contour plots in Figure 2 represent distributions with identity covariance matrix, but with quite different shapes.

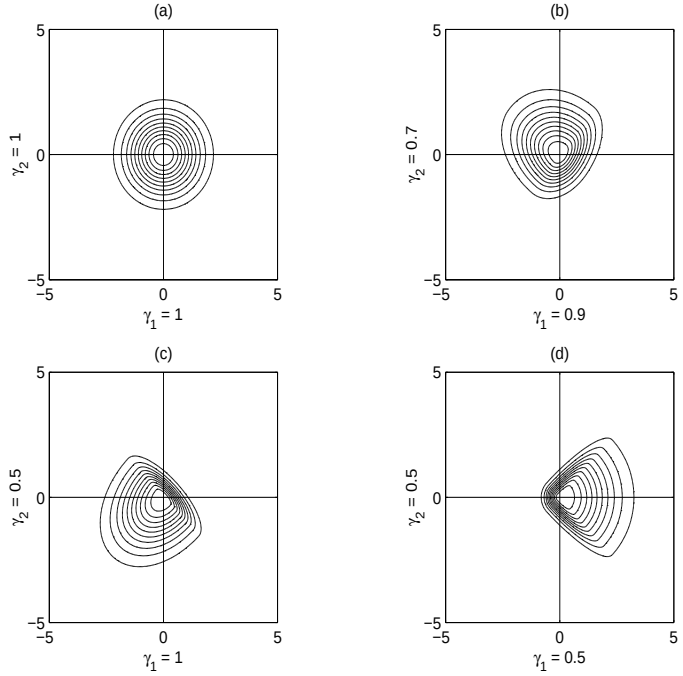

Figure 2: Contour plots of four pdfs with identical covariance matrix (equal to the identity matrix).

This ability of our class of skewed distributions to cover all possible mean and covariance structures in linked with one potential drawback, and that is the fact that the class of distributions is not closed under marginalisation. This results from the fact that a linear combination of random quantities with pdf as in (1) does not necessarily have a density of the same form. In a bivariate context, Moran

(1967) remarks that a necessary condition for classes of distributions with fixed marginals to cover the entire range of values for the correlation coefficient is that the marginals are symmetric.

Not many measures of multivariate skewness have been proposed in the literature. One measure of multivariate skewness is $\beta_{1,m}$, introduced by Mardia (1970) and given by

$$\beta_{1,m}(\eta) = \sum_{r,s,t}^{m} \sum_{r',s',t'}^{m} \sigma^{rr'}\sigma^{ss'}\sigma^{tt'} E\left[(\eta_r - \alpha_r)(\eta_s - \alpha_s)(\eta_t - \alpha_t)\right] E\left[(\eta_{r'} - \alpha_{r'})(\eta_{s'} - \alpha_{s'})(\eta_{t'} - \alpha_{t'})\right],$$

where α_j and $\sigma^{jj'},\ j,\ j' = 1, \ldots, m$ denote the elements of the mean vector and precision matrix of η, respectively. Two main characteristics of $\beta_{1,m}$ make it interesting for use: it equals zero for any symmetric distribution, with unimodal asymmetric distributions being characterised by values of the measure larger than zero, and it is invariant under non-singular affine transformations.

As $\beta_{1,m}$ is invariant under non-singular affine transformations, the calculation of its value for a multivariate skewed distribution generated using the construction we propose is trivial. Let $\eta \sim Sk_m(\mu, A, \gamma, f)$, then by making use of an alternative affine transformation of the original variables ϵ it is possible to obtain a set of variables $\psi \sim Sk_m(\mu^*, A^*, \gamma, f)$, with A^* diagonal, such that $Var(\psi)$ equals the identity matrix and $E(\psi)$ is zero. Now, as A^* is diagonal, by (5), the components of ψ are independent and Mardia's skewness measure is given by

$$\beta_{1,m}(\eta) = \beta_{1,m}(\psi) = \sum_{j=1}^{m} \left[E\left(\psi_j^3\right)\right]^2, \tag{6}$$

which, from (2), is straightforward to calculate and does not depend on μ or A. This ease of calculating Mardia's measure of skewness, for any pdfs $f_j(\cdot)$, is not shared by any of the methods based on conditioning on unobserved arguments or marginal replacement. The expression in (6) also shows that each particular γ_j has a contribution to the measure that is independent of the remaining elements of γ. As a consequence, if γ_j is set to one, its contribution to (6) vanishes. The existence of $\beta_{1,m}$ depends exclusively on the existence of $M_3(f_j),\ j = 1, \ldots, m$ defined in (2).

Figure 3 plots $\beta_{1,2}$ as a function of $\gamma \in (0, 1] \times (0, 1]$ with $f_j(\cdot) = \phi(\cdot),\ j = 1, 2$. As expected, $\beta_{1,2}$ is a continuous, strictly decreasing function of γ_j in $(0, 1]$, $j = 1, 2$. Other values of γ are covered by the fact that the value of $\beta_{1,2}$ is unaffected by inverting either γ_1, γ_2 or both. The value of $\beta_{1,2}$ is bounded below by zero (symmetric case) and

$$\lim_{\gamma \to 0} \beta_{1,2} = 4\frac{(4-\pi)^2}{(\pi-2)^3} \approx 1.96. \tag{7}$$

These same bounds are obtained in SDB for their definition of the skew-Normal distribution. For the skew-Normal distribution of Azzalini and Dalla Valle (1996), the upper bound is half the value in (7).

3.3 The importance of orthogonal transformations

In the sequel, we make use of the following result on the decomposition of nonsingular matrices.

Lemma 1. If A is any $m \times m$ real non-singular matrix, there exists an orthogonal matrix O_U such that $A = O_U U$, where U is a real upper triangular matrix with positive diagonal elements. Likewise, there exists another orthogonal matrix O_L such that $A = L O_L$, where L is a real lower triangular matrix with positive diagonal elements. Both representations are unique.

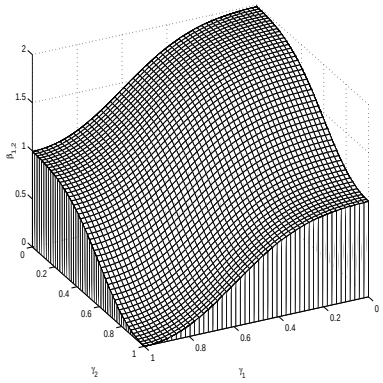

Figure 3: Plot of $\beta_{1,2}$ for a bivariate skew-Normal distribution as a function of γ.

In order to gain further insight, suppose that $A = LO$ (as in Lemma 1) and, for simplicity, assume that $\mu = 0$. From (4) we then have $\eta = O'L'\epsilon$, indicating that ϵ is first subjected to a linear transformation, and then to a rotation if $|O| = 1$ or a rotoinversion if $|O| = -1$. If O is the identity matrix, the j-th component of η is a linear combination of the last $m - j + 1$ components of ϵ. The effect of O is that it rotates and/or reflects the axes along which the joint distribution is a linear combination of the last $m-j+1$ components of ϵ. Figure 4 exemplifies the effect of O, using bivariate skewed distributions. In Figure 4(a) $A = LO_a$, while in Figure 4(b) $A = LO_b$, with $\mu = 0$ and

$$L = \begin{pmatrix} 1 & 0 \\ \frac{1}{4} & 1 \end{pmatrix}, O_a = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, O_b = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } \gamma = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \end{pmatrix}.$$

Along the axes e_1 the distribution is given as a linear combination of two independent univariate skewed distributions with skewness parameters $3/4$ and $3/2$. Similarly, along the axes e_2 the distribution is a univariate skewed distribution with skewness parameter equal to $3/2$. The contours in Figure 4 (b) can be obtained from the ones in Figure 4 (a) by reflecting them about any of the axes and rotating them. Inspired by the representation $A = LO$, we define the *basic axis* $e_j,\ j = 1, ...m$ as the axis along

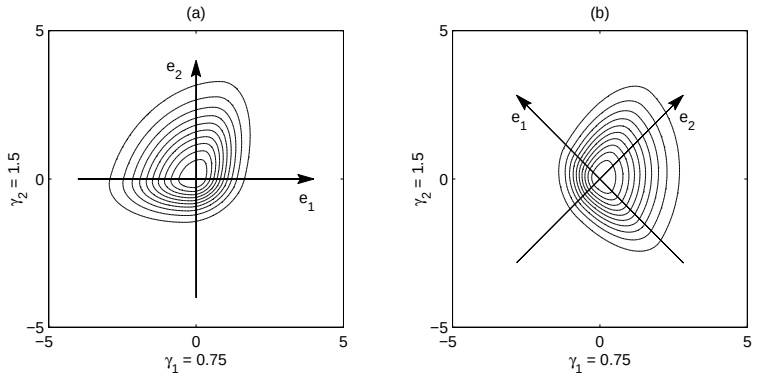

Figure 4: Contour plots of two bivariate skewed pdfs, together with their basic axes.

which the distribution is a linear combination of $m-j+1$ independent univariate skewed distributions with skewness parameters $\gamma_k,\ k = j, \ldots, m$. Changing the orthogonal matrix O is then equivalent to performing a rotation of the basic axes, possibly after performing a reflection about some of them. As is evident from Figure 4 these axes define the direction of the skewness of the distribution.

A well-known fact from the theory of multivariate distributions is that if η is given by (4) and the distribution of ϵ belongs to the spherical class (*e.g.* the Normal distribution), then A only needs to be

known up to $\Sigma = A'A$ or, equivalently, A can be either upper or lower triangular. Sphericity can be defined as distributional invariance with respect to orthogonal transformations (see *e.g.* Fang, Kotz and Ng, 1990, p. 27), so that $\epsilon \stackrel{d}{=} O\epsilon$ for any orthogonal matrix O under spherical ϵ. It is then obvious that for $A = OU$ the transformed variable η in (4) has the same distribution as $U'\epsilon + \mu$, so that the choice of O is irrelevant. Equivalently, continuous spherical distributions on ϵ are characterised by a pdf that only depends on $\epsilon'\epsilon$, so it is clear that the induced pdf on η will only depend on $(\eta-\mu)'\Sigma^{-1}(\eta-\mu)$ and only $\Sigma = A'A = U'U$ matters. However, if ϵ is outside the spherical class then knowledge of Σ alone is no longer sufficient. For example, this is the case when the components of ϵ have independent Student distributions or if their distributions are skewed. Thus, the orthogonal matrix O plays an important role. Not taking O into account (*e.g.*, by implicitly taking $O = I$) in skewed cases would imply favouring specific directions for the asymmetry of the distribution. When defining a general class of skewed distributions, we feel it is important to also introduce parameters that specify the direction of the skewness, in our case by specifying the basic axes through A.

The skewed distribution of SDB introduces skewness into symmetric distributions along the coordinate axes. Bauwens and Laurent (2005) use a regression framework with a linear transformation as in (4), where ϵ has a similar distribution as in (3), but fix $A = \Sigma^{1/2}$, the spectral decomposition of Σ. The latter formulation does not allow for a separate choice of the directions of the asymmetry of the distribution, and fixes it to be a function of Σ.

4 Unique parameterisation of skewed distributions

Subsection 3.3 shows that defining A via Σ is no longer sufficient for the class of distributions that we introduce in this article. Here we provide a unique parameterisation of our skewed distributions. However, even if the components of γ are set to unity, the parameterisation is still suitable, *i.e.* is unique, if the distribution of ϵ in (4) is not spherical.

Let $\eta \sim Sk_m(\mu, A, \gamma, f)$. Recall that A is a non-singular matrix, and that $\gamma \in \Re_+^m$. However, this parameter space is not fully adequate in the sense that the same distribution of η can be defined using different sets of parameter values, which is an undesirable feature, especially for inference.

Let $r = (r_1, \ldots, r_m)'$ be a permutation of the first m positive integers, A_r be the $m \times m$ matrix where the jth row is the r_jth row of A, and γ_r be the m-dimensional vector where the jth element is the r_jth element of γ and define f_r similarly. Then, it follows directly from (4) that $Sk_m(\mu, A, \gamma, f) = Sk_m(\mu, A_r, \gamma_r, f_r)$. There are $m!$ different permutations r.

Also, let $s = (s_1, \ldots, s_m)'$ be a vector whose components are in $\{-1, 1\}$, A_s be the $m \times m$ matrix where the jth row equals the jth row of A times s_j, and let γ_s be the m-dimensional vector where component j is given by the jth element of γ to the power s_j. Then, it follows directly from the property of the univariate skew distributions stated at the end of Subsection 2.1 that $Sk_m(\mu, A, \gamma, f) = Sk_m(\mu, A_s, \gamma_s, f)$. The number of different vectors s is 2^m.

Combining both transformations gives all the $m!2^m$ parameter values that define the same distribution. These values are distinct if the components of γ are all distinct and different from unity.

There are several ways of reducing the parameter space in order to achieve a one-to-one parameterisation of the class of skewed distributions. Here we present one that is valid except for a set that, under most commonly used probability measures, will have zero mass. Using Lemma 1, through $A = OU$, we first reparameterise from (μ, A, γ, f) to (μ, O, U, γ, f) where O is an orthogonal matrix

and U an upper triangular matrix with strictly positive diagonal elements. We can now create a one-to-one parameterisation by restricting the matrix $O = (O_{ij}),\ i, j = 1, \ldots, m$ to have

$$O_{11} > -O_{m1} > -O_{(m-1)1} > \ldots > |O_{21}| > 0 \text{ and } |O| = (-1)^{m+1}, \tag{8}$$

and by adjusting γ and f accordingly. The set of all such matrices O will be denoted by $\mathcal{O}^m$. This set of restrictions provides a one-to-one parameterisation, for all distributions with distinct components of γ and matrices A without zeros in the first column. Indeed, if A_1 and A_2 differ by a signed permutation of rows (*i.e.* are equivalent), then $A_1 = O_1 U$ and $A_2 = O_2 U$, where O_1 and O_2 differ by the same signed permutation of rows. The two conditions above ensure that one and only one of such signed permutations is allowed in the parameter space.

In the special case where we fix all skewness parameters to unity and, in addition, we choose a spherical distribution for ϵ (so we simply have an elliptical distribution for η), our parameterization will be too rich and we will be able to update all parameters but O, for which we will simply retrieve the (proper) prior distribution. Thus, inference is still possible in this case, which will, moreover, not occur under continuous priors on γ.

5 Examples of multivariate skewed distributions

Even though it is possible to use symmetric, unimodal pdfs $f_j(\cdot),\ j = 1, \ldots, m$ from different parametric families, we will mainly focus on cases where for any $j = 1, \ldots, m,\ f_j(\cdot)$ can be written as $f_{\nu_j}(\cdot)$, with ν_j in some set $\mathcal{N}$. We then identify the multivariate skewed distribution generated by (5) with the name of the multivariate distribution that would result if $\gamma_j = 1,\ j = 1, \ldots, m$.

5.1 Skew-Normal

The multivariate skew-Normal distribution is obtained when $f_{\nu_j}(\cdot) = \phi(\cdot),\ j = 1, \ldots, m$, *i.e.* the pdf of the univariate standard Normal distribution. In this case, $\nu_j,\ j = 1, \ldots, m$, is vacuous.

5.2 Skew-Independent Student

The multivariate skew-Independent Student (skew-IStudent) with degrees of freedom (df) vector $\nu = (\nu_1, \ldots, \nu_m)'$ is generated when $f_{\nu_j}(\cdot)$ is the univariate Student pdf with $\nu_j \in \Re_+$ df, given by

$$f_{\nu_j}(x) = \frac{\Gamma\left(\frac{\nu_j+1}{2}\right)}{\Gamma\left(\frac{\nu_j}{2}\right)(\pi\nu_j)^{1/2}}\left[1 + \frac{x^2}{\nu_j}\right]^{-\frac{\nu_j+1}{2}} \tag{9}$$

5.3 Skewed mixture of Normals

Mixtures of Normals are an important class of distributions. Using a slight extension of the framework in (4), scale mixtures of Normals can be created by

$$\eta = \lambda^{-\frac{1}{2}} A'\epsilon + \mu \tag{10}$$

where ϵ follows a multivariate standard Normal distribution and a mixing distribution is assigned to λ. Skewed mixtures of Normals are defined in a similar way, by taking ϵ as in (3), with $f_j(\cdot) = \phi(\cdot),\ j = 1, \ldots, m$. A particular case that will be used in the sequel is the skew-Student distribution with ν^* df, obtained if λ has a Gamma distribution with both shape and precision parameter set to $\nu^*/2$.

6 Regression Modelling

We assume that we have n observations from an underlying process, given by pairs $(x_i, y_i),\ i = 1, \ldots, n$, where $x_i \in \Re^k$ is a vector of explanatory variables and $y_i \in \Re^m$ is the variable of interest. Throughout, we condition on x_i without explicit mention. The n observations are grouped in $X \in \Re^{n\times k}$ and $Y \in \Re^{n\times m}$, with each row corresponding to one observation.

Let us assume the observables $y_i \in \Re^m,\ i = 1, \ldots, n$, are generated from

$$y_i = g_i(B) + \lambda_i^{-\frac{1}{2}} A'\epsilon_i \tag{11}$$

where $g_i(\cdot)$ is a known measurable function in $\Re^m$, B parameterises the location, $A = OU$ is the transformation matrix for y_i, with $U \in \mathcal{U}^m$, the set of upper triangular $m \times m$ matrices with positive diagonal elements and $O \in \mathcal{O}^m$, the set of $m \times m$ orthogonal matrices that satisfy (8). $\lambda_i,\ i = 1, \ldots, n$ are independently drawn from some distribution P_λ on $\mathcal{L}$. We assume $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})',\ i = 1, \ldots, n$ to be i.i.d. conditionally on parameters $\nu \in \mathcal{N}^m$, and $\gamma = (\gamma_1, ...\gamma_m)'$ in $\Re_+^m$, with pdf as in (3).

Whenever we mention "tail behaviour" in models generated through (11), we will refer to the tails of the latent quantities $\lambda_i^{-1/2}\epsilon_i$, rather than the tails of the observables.

6.1 Existence of moments under improper priors

We now consider the impact of introducing skewness into the multivariate sampling distribution on the existence of the posterior distribution and moments in the context of this general regression model.

Let $\lambda = (\lambda_1, \ldots, \lambda_n)'$. We adopt the following prior product structure:

$$P_{B,O,U,\lambda,\gamma,\nu} = P_{B,O,U} \times P_\lambda \times P_\gamma \times P_\nu. \tag{12}$$

The usual non-informative prior for regression modelling with elliptically distributed errors is an improper prior on B and $\Sigma = A'A$ given by

$$p(B, \Sigma) = p(B)p(\Sigma) \propto |\Sigma|^{-\frac{m+1}{2}}. \tag{13}$$

We define a non-informative prior on B, O and U that is compatible with (13). From (13) and transforming from $\Sigma = A'A = U'U$ to U we have that $p(\Sigma) \propto |\Sigma|^{-\frac{m+1}{2}} \Leftrightarrow p(U) \propto \prod_{j=1}^m u_{jj}^{m-j}|U|^{-m}$. In addition, we take $p(O)$ such that its distribution on $\mathcal{O}^m$ is invariant to linear orthogonal transformations (see (20)-(21) and Appendix A.4). The prior on B is as in (13). Finally, we assume that P_λ and P_γ and P_ν are all proper distributions on $\mathcal{L}^n$, $\Re_+^m$ and $\mathcal{N}^m$, respectively. The full prior distribution is given by (12) with $P_{B,O,U}$ corresponding to

$$p(B, O, U) \propto p(O) \prod_{j=1}^m u_{jj}^{m-j}|U|^{-m}. \tag{14}$$

Theorem 1. Consider n independent replications from the sampling distribution given in (11) and the prior in (12) and (14). Denoting the lth element of B by B_l, $l = 1, \ldots, p$, and given $r_1, \ldots, r_p$ and $r \geq 0$, we obtain that for any P_γ

$$E\left(|\Sigma|^{r/2} \prod_{l=1}^p |B_l|^{r_l} \;\middle|\; Y \right) < \infty,$$

if and only if the same holds for inference with symmetrically distributed disturbances.

The result in Theorem 1 states that the existence of posterior moments of B and of non-negative posteriors moment of $|\Sigma|$ is unaffected by the extra vector of unknowns γ under any proper prior P_γ. Propriety of the posterior distribution is therefore not influenced by incorporating skewness in the sampling, as can be assessed by setting $r = r_1 = \cdots = r_p = 0$. This result extends Theorem 1 in FS to the case of multivariate skewed distributions.

We now completely specify two Bayesian models that account for both skewness and fat tails. Further, we provide results on posterior inference with these models. We define $g_i(B) = B'x_i$, where $B \in \Re^{k\times m}$ (so, $p = mk$, and we shall now denote the elements of B by $B_{lj}, l = 1,\ldots,k, j = 1,\ldots,m$). This corresponds to the commonly used linear regression model. The complete design matrix $X = (x_1,\ldots,x_n)'$ will always be assumed to be of full column rank, implying that $n \geq k$.

6.2 Inference under skew-Student sampling

The first of the models that we introduce here is the linear regression model, assuming that the errors have a skew-Student distribution, defined in Subsection 5.3. In particular we consider the following special case of the model in (11), (12) and (14):

- $f_{\nu_j}(\cdot) = \phi(\cdot),\ j = 1,\ldots,m.$
- For $i = 1,\ldots,n$, λ_i, given a positive parameter $\nu^* \in \mathcal{N}^*$, has a Gamma distribution with both parameters equal to $\nu^*/2$. The prior distribution on ν^*, P_{ν^*}, is proper.

Thus, we assume n independent replications of the sampling density

$$p(y_i|B,O,U,\nu^*,\gamma) = |U|^{-1}\prod_{j=1}^{m}\frac{2}{\gamma_j+\frac{1}{\gamma_j}}\int_{\Re_+}\lambda_i^{\frac{m}{2}}\ \phi\left(\lambda_i^{\frac{1}{2}}d_{ij}\ \gamma_j^{-\mathrm{sign}(d_{ij})}\right)p_G\left(\lambda_i\middle|\frac{\nu^*}{2},\frac{\nu^*}{2}\right)d\lambda_i, \qquad (15)$$

with $d_{ij} = [O(U')^{-1}]_{j\cdot}(y_i - B'x_i)$.

The sampling distribution given in (15) will be denoted as m-dimensional skew-Student with location $B'x_i$, transformation OU, skewness parameter γ and ν^* df. Matrix B is usually of primary interest as it represents the regression coefficients. Also of common practical importance will be $\Sigma = U'U$ as it contains information about the dispersion of y. The remaining parameters have a well-defined purpose. Skewness is controlled jointly by γ and OU, while $\nu^* \in \Re_+$ determines the thickness of the tails of the multivariate distribution.

This subsection again extends results from FS to the multivariate case.

Theorem 2. Consider n independent replications from the sampling model in (15) under the prior in (12) and (14). Then the posterior distribution is proper if and only if $n \geq m + k$, for any P_{ν^*} and P_γ.

The extra model flexibility introduced by modelling tail behaviour and skewness is thus seen not to affect the propriety of the posterior distribution. As a consequence, the well-known result under Normal sampling holds in our much more general framework. Throughout, we assume $n \geq m + k$.

The following definition from Fernández and Steel (2000), concerning the design matrix X, is required to adequately characterise the existence of the marginal posterior moments of B.

Definition 1. Given an $n \times k$ full column-rank matrix X, the singularity index for column $l = 1, 2, \ldots, k$ is defined as the largest number p_l $(0 \leq p_l \leq n-k)$ such that there exists a $(k-1+p_l) \times k$ submatrix of X of rank $k-1$ that remains of rank $k-1$ after removing its lth column.

If X contains rows of zeros, then p_l is at least equal to the number of such rows for all $l = 1, 2, \ldots, k$. Furthermore, $\max\{p_l,\ j = 1, 2, \ldots, k\} = 0$ if and only if every $k \times k$ submatrix of X is non-singular.

Theorem 3. Consider the Bayesian model given in (12), (14) and (15) and $r > 0$. Let $\mathcal{N}^* = (\nu_0^*, \infty)$, $\nu_0^* \geq 0$. Then $E\left(|B_{lj}|^r | Y\right) < \infty$ if $r < \min\{n-m-k+1, m(n-k-p_l)+\nu_0^*\}$, with p_l the singularity index for column l of the design matrix X.

We can also show that if $r \geq n-m-k+1$ there is no possibility for the moment to exist, regardless of the properties of the design matrix or the prior P_{ν^*}. Such a result is due to the uncertainty about B and Σ. However, both X and P_{ν^*} intervene in the sufficient condition stated in Theorem 3.

We now turn our attention to the posterior moments of $|\Sigma|$ of order $r/2 \geq 0$. For this quantity, the order up to which the posterior moments are finite does not depend on the design matrix or the distributions of P_γ and P_{ν^*} as is stated in the following theorem.

Theorem 4. Consider the Bayesian model given in (15), under the prior in (12) and (14), and $r \geq 0$. Then, $E\left(|\Sigma|^{\frac{r}{2}} | Y\right) < \infty$ if and only if $r < n-m-k+1$.

Note that if we impose a Dirac distribution on $\lambda_i = 1$, *i.e.* $p(\lambda_i = 1) = 1,\ i = 1, \ldots, n$, we obtain a regression model with skew-Normal disturbances. The results above apply to this model in the limit as $\nu_0^* \to \infty$. Also, if we set the components of γ equal to one, we obtain the symmetric versions of the distributions. We know from Theorem 1 that the results derived here also apply to the case of symmetric Student sampling. In that case, as explained in Subsection 3, the matrix O is no longer necessary for inference, and therefore we can set it to $O = I_m$, the $m \times m$ identity matrix.

6.3 Inference under skew-IStudent sampling

The regression framework introduced in the previous subsection implies a common tail behaviour for ϵ along all directions. Here we relax that assumption by allowing $f_{\nu_j}(\cdot),\ j = 1, \ldots, m$ to have different tail behaviour. In particular, we adopt the Bayesian regression model in (11)-(12) and (14) with

- For $j = 1, \ldots, m$, $f_{\nu_j}(\cdot)$ is the univariate Student distribution with ν_j df.
- P_{λ_i} is a Dirac distribution on $\lambda_i = 1,\ i = 1, \ldots, n$.
- $P_\nu = \prod_{j=1}^m P_{\nu_j}$.

The sampling distribution is then given by

$$p(y_i | B, O, U, \nu, \gamma) = |U|^{-1} \prod_{j=1}^{m} \frac{2}{\gamma_j + \frac{1}{\gamma_j}} f_{\nu_j}\left(d_{ij}\ \gamma_j^{-\text{sign}(d_{ij})}\right), \tag{16}$$

where $f_{\nu_j}(\cdot)$ is given in (9) and d_{ij} is as in (15). This defines the m-dimensional skew-IStudent with location $B'x_i$, transformation OU, skewness parameter γ and df vector $\nu = (\nu_1, \ldots, \nu_m)'$.

Theorem 5. Consider n independent replications from the sampling model in (16) under the prior in (12) and (14). If for any $j = 1, \ldots, m$, $P(\nu_j \leq m-1) = 0$ and $n \geq m + k$, then the posterior distribution is proper for any P_γ.

The requirement that $P(\nu_j \leq m-1) = 0$ can be restrictive, especially if the dimension of the problem is large. However, for reasonably small m, the restriction is unlikely to cause much harm, as only distributions with extremely heavy tails are excluded. In addition, if we also want the existence of third moments, required for the calculation of the Mardia measure of skewness, then we need $\nu_j > 3$, implying no extra restriction for problems up to and including dimension 4. In what follows, we always assume that P_ν complies with the sufficient condition in Theorem 5.

Theorem 6. Consider the Bayesian model given in (12), (14) and (16) and $r > 0$. Let $\mathcal{N} = (\nu_0, \infty)$ be the common support of P_{ν_j}, $j = 1, \ldots, m$. Then we obtain that $E\left(|B_{lj}|^r | Y\right) < \infty$ if $r < \min\{n - m - k + 1, m(n - k - p_l - 1) + \nu_0 + 1\}$, with p_l the singularity index for column l of X.

Theorem 7. Consider the Bayesian model given in (16), under the prior in (12) and (14), and $r \geq 0$. Then, $E\left(|\Sigma|^{\frac{r}{2}} | Y\right) < \infty$ if $r < n - m - k + 1$.

If we consider the special case where the components of γ are set to one, we obtain sampling under the Independent Student (IStudent) distribution. However, as the product of univariate Student distributions is not in the spherical class, it is still necessary to consider the orthogonal matrix O. To our knowledge, this sampling model has not been analysed in the literature, even under symmetry. Thus, this subsection also introduces a novel class of symmetric heavy-tailed distributions and analyses its properties in a Bayesian regression context.

6.4 Completing the prior specification

Having already specified the prior structure and the prior distributions for B and O and U, the Bayesian models become fully specified by assigning proper prior distributions for γ, ν^* and ν.

We assume that the components of $\gamma \in \Re_+^m$ are independently distributed according to a common logNormal distribution, *i.e.* $\log(\gamma_j) \sim N(0, s^2)$, $j = 1, \ldots, m$. This centers the prior over symmetry and implies that for any two constants such that $1 < \gamma_a < \gamma_b$, we have $P[\gamma_j \in (\gamma_a, \gamma_b)] = P\left[\gamma_j \in \left(\frac{1}{\gamma_b}, \frac{1}{\gamma_a}\right)\right]$, thus treating positive and negative skewness symmetrically in the prior.

For ν^* and the components of ν we use an Exponential prior with parameter $d > 0$, restricted to $\mathcal{N}^* = \mathcal{N} = (\max\{3, m-1\}, \infty)$, allowing at the same time the use of improper priors and calculation of the third moments, necessary to calculate the Mardia measure of skewness. In addition, it will, in most practical situations, avoid the problems of posterior nonexistence with point observations pointed out in Fernández and Steel (1999) for symmetric multivariate Student regression models.

Finally, we choose $s = 1$, corresponding to a rather vague prior on γ, and $d = 0.1$ as in FS.

7 Numerical Implementation

Inference with the Bayesian models introduced in Section 6 requires numerical methods. Here we conduct inference using Markov chain Monte Carlo (MCMC) methods, in particular hybrid samplers with both Metropolis-Hastings and Gibbs components.

As the univariate Student-t distribution can be expressed as a mixture of Normals, (16) can be written as

$$p(y_i|B,O,U,\nu,\gamma) = |U|^{-1} \prod_{j=1}^{m} \frac{2}{\gamma_j + \frac{1}{\gamma_j}} \int_{\Re_+} \lambda_{ij}^{\frac{1}{2}} \, \phi\left(\lambda_{ij}^{\frac{1}{2}} \, d_{ij} \, \gamma_j^{-\text{sign}(d_{ij})}\right) p_G\left(\lambda_{ij} \middle| \frac{\nu_j}{2}, \frac{\nu_j}{2}\right) d\lambda_{ij}. \qquad (17)$$

This illustrates a fundamental difference between the skew-Student and the skew-IStudent sampling models. In the skew-Student in (15), each observation y_i has its corresponding mixing parameter $\lambda_i,\ i = 1,\ldots,n$, pairwise i.i.d. given ν^*. For the skew-IStudent model, each observation y_i has its vector of independent mixing parameters $\lambda_i = (\lambda_{i1},\ldots,\lambda_{im})'$, with different distributions for each element. Thus, even if $\nu_j = \nu^*, j = 1\ldots,m$, the two models are still quite different. For the skew-IStudent model, we conduct inference on $(B,O,U,\gamma,\lambda,\nu \mid Y)$, where $\lambda = (\lambda_1,\ldots,\lambda_n)'$ while for skew-Student sampling, we merely replace ν by ν^*.

Most steps in the sampler are fairly standard, with one exception: the step to draw O, which will be explained below. For both models, the components of λ are independent given the other parameters, and can directly be sampled from Gamma distributions. Drawings from the conditional posterior distributions for ν and ν^* are generated using a rejection sampler. We use individual random-walk Metropolis-Hastings samplers for B, O, U and γ, common to both models. For the components of B, the off-diagonal elements of U and the logarithm of the components of γ we use a Normal proposal, while we use a half-Normal proposal distribution for the diagonal elements of U. Throughout, we update one component at a time.

7.1 Sampling O

Sampling orthogonal matrices $O \in \mathcal{O}^m$ directly is complicated. Thus, we use a reparameterisation of O, more suitable for sampling. We first define $(\theta^2,\ldots,\theta^m)' \in \Theta^2 \times \ldots \Theta^m$, where $\theta^j = (\theta_1^j,\ldots,\theta_{j-1}^j)$ and Θ^j is as defined in Appendix A.4.

The appendix shows that if the $v^j = (v_1^j,\ldots v_j^j)'$ are such that $v_1^j = \sin(\theta_1^j)$, $v_i^j = \prod_{l=1}^{i-1}\cos(\theta_l^j) \times \sin(\theta_i^j), i < j$ and $v_j^j = \prod_{l=1}^{j-1}\cos(\theta_l^j)$, the matrix $H_{\theta^j} = I_j - 2v^j\left(v^j\right)'$ and $O_{\theta^j}^m$ is defined as

$$O_{\theta^j}^m = \begin{pmatrix} I_{m-j} & 0 \\ 0 & H_{\theta^j} \end{pmatrix},$$

then we can express any $m \times m$ orthogonal matrix $O \in \mathcal{O}^m$ as $O = O_{\theta^m}^m \times \ldots \times O_{\theta^2}^m$.

We can then sample O easily by sampling in turn each component of each $\theta^j,\ j = 2,\ldots,m$. For each one of those, we sample from a Normal random walk proposal distribution, restricted so that $\theta^j \in \Theta^j$ and, thus, the proposed orthogonal matrix is in $\mathcal{O}^m$.

8 Example on Biomedical Data

Along with the most general models, incorporating both skewness and fat tails - skew-Student and skew-IStudent, we also consider simpler alternatives: Student, IStudent, skew-Normal and Normal. The Student, IStudent and the Normal models assume symmetry ($\gamma = 1$), while the latter model and the skew-Normal do not allow for heavy tails. The prior distributions for the parameters of these models are compatible with those in the more general ones.

Here, we do not present any comparison with regression models based on other classes of skewed distributions. To our knowledge, no other method has been shown to allow for inference under an

improper prior structure compatible with (12) and (14). We refer the interested reader to Ferreira and Steel (2004), where a comparison between our methodology and the one in SDB is presented under a proper prior using firm size data.

Inference was conducted using every tenth of 100,000 realisations from the Markov chain described in Section 4, after discarding the first 20,000 samples (a burn-in sufficient for convergence in all cases). Model comparison is provided through Bayes factors. Estimates of marginal likelihoods are obtained using the p_4 measure in Newton and Raftery (1994), with their δ set to 0.1. As pointed out by a referee, this method can be less optimal than *e.g.* bridge sampling, as in Meng and Wong (1996) and DiCiccio, Kass, Raftery and Wasserman (1997). However, the latter is hard to implement in our example with many, often highly dependent parameters. We have verified the obtained Bayes factors (which are overwhelming in most cases) through Savage-Dickey density ratios (for skewed versus symmetric models) and plots of the sampled likelihood values (not reported due to space limitations).

We use a dataset from the Australian Institute of Sport. In particular, we study the distribution of four biomedical variables: body mass index (BMI), sum of skin folds (SSF), percentage of body fat (PBF), and lean body mass (LBM). The data were collected for 202 athletes at the Australian Institute of Sport and are described in Cook and Weisberg (1994). Besides a constant term we use information on three covariates: red cell count (RCC), white cell count (WCC) and plasma ferritin concentration (PFC). In order to compare the influence of the covariates, the data was normalised to have mean zero and variance one. These data, without the covariates, have been used previously in the context of skewed distributions in, *i.a.*, Azzalini and Capitanio (2003).

Figure 5 (a) presents the marginal posterior pdfs of the elements of γ for the model that proved to be most adequate - the skew-IStudent model, together with the prior pdf. For all but one component of γ, the pdfs have low mass near unity, implying that the data requires that at least three components in the linear transformation are skewed. Figure 5 (b) exhibits the posterior pdf of Mardia's measure of skewness for the skew-IStudent and also the prior distribution for that quantity. The posterior pdf of $\beta_{1,4}$ has most of its mass concentrated away from zero, implying that the distribution of the quantities of interest is asymmetric. We note that by assuming our fairly uninformative prior on γ, the prior on $\beta_{1,4}$ puts substantial mass on asymmetric distributions, but also retains mass on low values corresponding to symmetry. The posteriors of γ and $\beta_{1,4}$ are fairly robust to changes of the prior.

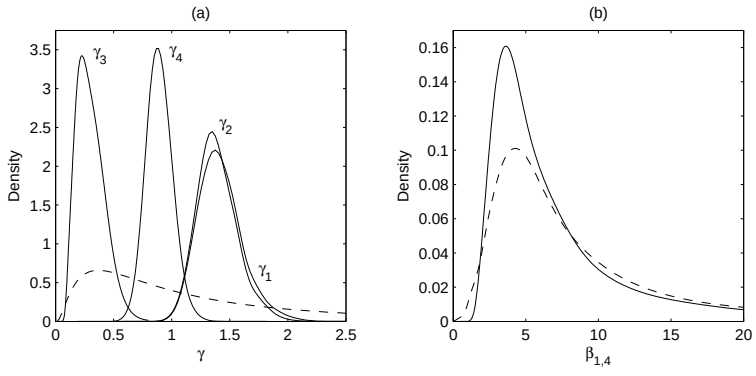

Figure 5: (a) marginal posterior pdfs for the components of γ for the skew-IStudent model (solid) together with the prior pdf (dashed); (b) posterior pdf of Mardia's measure of skewness for the skew-Student (solid) and the prior pdf for the same quantity (dashed)

Figures 6 (a)-(d) present the posterior distributions of the coefficients of B for the intercept,

RCC, WCC and PFC, respectively. In most cases, the covariates are shown to have an effect on the distribution of the variables, particularly for RCC. BMI does not seem to need any of the covariates, but all regressors intervene crucially in modelling SSF. The posterior distribution of the regression coefficients is quite different from the prior distribution, which is improper uniform.

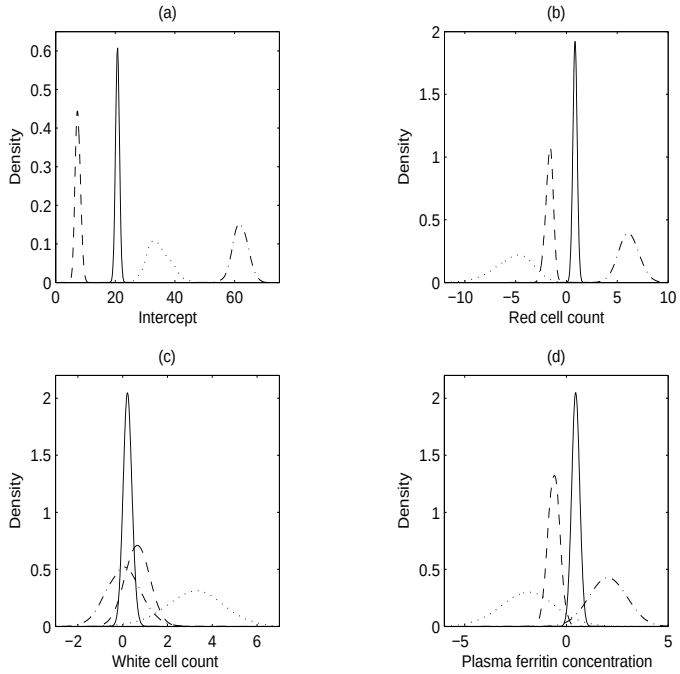

Figure 6: (a)-(d) posterior distributions of the coefficients of B for the intercept, RCC, WCC and PFC, respectively, corresponding to BMI (solid), SSF (dotted), BFAT (dashed) and LBM (dot-dashed), evaluated for the skew-Student model

The distribution of the biomedical measurements has heavier than Normal tails. Figure 7 presents the posterior density for the df for the skew-IStudent model, together with the prior distribution. Some components (of ϵ) require much heavier tails than others with the medians of ν_j given by 10.7, 16.2, 5.7 and 13.4, $j = 1, \ldots, 4$. The skew-Student model leads to a median value of ν^* equal to 15.8. Both models lead to heavier tails when we impose symmetry. Thus, if we (wrongly) impose symmetry, the skewness in the data is partly misinterpreted as fat tails.

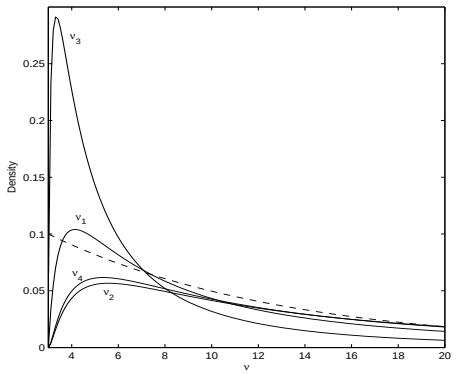

Figure 7: Posterior density for $\nu_1, \ldots, \nu_4$ for the skew-IStudent model (solid line), together with the prior pdf (dashed). Note the truncation at three.

Table 1: Bayes factors for Australian Institute of Sport data. Entries indicate support in favour of the model in the row versus that in the column

	skew-Student	Student	skew-Normal	Normal	skew-IStudent	IStudent
skew-Student	1	> 1000	2.6	> 1000	< 0.001	> 1000
Student		1	< 0.001	26	< 0.001	< 0.001
skew-Normal			1	> 1000	< 0.001	> 1000
Normal				1	< 0.001	< 0.001
skew-IStudent					1	> 1000

Table 1 compares the models using Bayes factors. We conclude that the skew-IStudent model is the most favoured model for these data, with the skew-Student model a distant second. A large difference exists between the adequacy of the skewed models and the others. There is also strong evidence in favour of heavy tails, but interestingly, the IStudent tails receive a lot more support from the data than the Student tails. This is partly due to the differences between the ν_js in Figure 7, but, as explained in Section 7, other differences exist between these models. In summary, both skewness and heavy tails are strongly supported, which is in line with the findings of Azzalini and Capitanio (2003) in the context of a location-scale model.

9 Conclusion

In this article we present a novel general method for the construction of skewed multivariate distributions. Based on linear transformations of univariate skewed distributions, the method we introduce is quite flexible, *i.e.* it imposes few restrictions on the form of the distribution.

In particular, we use linear transformations of independent and univariate random quantities with skewed distributions as in FS. The generated class of distributions has a number of appealing characteristics. Moment calculation is always straightforward if the moments of the underlying univariate distribution are available, and mean, variance and skewness can be modelled independently. Also, unlike other classes of skewed distributions proposed in the literature, our method makes no use of conditioning arguments, which require the calculation of cumulative distributions functions. This aspect can be quite relevant, especially for certain distributions and for high dimensions. A drawback of our proposal is that the class of skewed distributions is in general not closed under linear transformations or marginalisation.

We provide results on inference with these skewed distributions in a Bayesian regression model, under commonly used improper priors, and show that the extra flexibility induced by skewness does not have any impact on the existence of the posterior distribution, or even on the existence of posterior moments of the parameters. Further, we introduce two classes of skewed and heavy-tailed multivariate regression models, skew-IStudent and skew-Student, and establish results on posterior propriety and existence of posterior moments. One of these classes (the skew-IStudent) is novel even under symmetry.

We introduce MCMC samplers to conduct inference with these skewed and heavy-tailed linear regression models and examine an application to biomedical measurements, where we find strong evidence in favour of skewness and heavy tails.

Acknowledgements: We thank the Editors, an Associate Editor and a Referee for constructive comments. Ferreira was supported by grant SFRH BD 1399 2000 from Fundação para a Ciência e Tecnologia, Portugal and was affiliated to the University of Warwick during this research.

A Orthogonal Matrices

In the sequel, O denotes an $m \times m$ orthogonal matrix with determinant equal to $(-1)^{m+1}$. The set of all such matrices will be denoted as O^m.

A.1 Householder matrices

Definition 2. Let v be a vector in $\Re^m$. Then $H(v) = I_m - 2\frac{vv'}{v'v}$ is a m-dimensional Householder matrix (e.g., Golub and van Loan, 1989, p. 38)). For any $v \in \Re^m$, $H(v)$ is an orthogonal, symmetric matrix of determinant -1 for which $H(v) = H(-v) = H(av)$, where a is a scalar. The latter property implies that if $v \in S_m^{1/2}$, a unit half-sphere in $\Re^m$, then v provides a one-to-one parameterisation of the set of m-dimensional Householder matrices. Writing $v_\theta = (v_1, \ldots, v_m)'$ in polar coordinates, *i.e.*,

$$v_1 = \sin(\theta_1), \quad v_j = \prod_{l=1}^{j-1} \cos(\theta_l) \times \sin(\theta_j), j < m \quad \text{and} \quad v_m = \prod_{l=1}^{m-1} \cos(\theta_l)$$

and selecting $\theta = (\theta_1, \ldots, \theta_{m-1})'$ to be in Θ^m, defined as $(-\frac{\pi}{2}, \frac{\pi}{2})$ if $m = 2$ and $(0, \pi/2) \times (-\pi/2, \pi/2)^{m-3} \times (-\pi, \pi)$ if $m > 2$, implies that $v_\theta \in S_m^{1/2}$. Thus, by defining $H_\theta = H(v_\theta)$ we can uniquely parameterise the set of m-dimensional Householder matrices using $\theta \in \Theta^m$.

A.2 Decomposing O using Householder matrices

We now use Householder matrices to decompose any orthogonal matrix O, with $|O| = (-1)^{m+1}$. Let $\theta^j = (\theta_1^j, \ldots, \theta_{m-1}^j)' \in \Theta^j$ and for $j = 1, \ldots, m$ define $O_{\theta^j}^m$ as

$$O_{\theta^j}^m = \begin{pmatrix} I_{m-j} & 0 \\ 0 & H_{\theta^j} \end{pmatrix}. \tag{18}$$

Lemma 2. (Golub and Van Loan, 1989) Any matrix $O \in O^m$ can be written uniquely as

$$O = O_{\theta^m}^m \times \ldots \times O_{\theta^2}^m, \tag{19}$$

where $\theta^j \in \Theta^j$, $j = 2, \ldots, m$.

Thus, $O \in O^m$ can be parameterised uniquely by a set of $m-1$ vectors $\theta^j \in \Theta^j$, $j = 2, \ldots, m$.

A.3 Distribution on O^m invariant to linear orthogonal transformations

Stewart (1980) uses the decomposition in (19) to describe an algorithm for generating random orthogonal matrices from the invariant (with respect to linear orthogonal transformations) distribution on O^m (*i.e.* the Haar measure with respect to the orthogonal group). Using similar arguments, we can write the invariant distribution of O on O^m as

$$p(O) = p(\theta^2, ..., \theta^m) = \prod_{j=2}^{m} p(\theta^j), \tag{20}$$

where $p(\theta^j)$ is the pdf on $\theta^j \in \Theta^j$ that generates H_{θ^j} with first column uniformly distributed on S_j, $j = 2, \ldots, m$. Calculation reveals that the i-th element of the first column of H_{θ^j} is given by $\cos(2\theta_1^j)$ if $i = 1$, $-\sin(2\theta_1^j)\prod_{l=2}^{i-1}\cos(\theta_l^j)\sin(\theta_i^j)$ if $2 < i \leq j-1$ and $-\sin(2\theta_1^j)\prod_{l=2}^{m-1}\cos(\theta_l^j)$ if $i = j$.

Variable transformation shows that, if

$$p(\theta^j) \propto |\sin(2\theta_1^j)^{j-2}\prod_{l=1}^{j-3}\cos(\theta_{l+1}^j)^{j-2-l}|,\ j = 2, \ldots, m, \tag{21}$$

then the first column of H_{θ^j} has an uniform distribution on the j-dimensional unit sphere S_j.

Equations (20)-(21) provide the necessary distribution of θ^j, $j = 2, \ldots, m$, such that the distribution on O defined as in (19) is invariant on O^m.

A.4 Invariant distribution on $\mathcal{O}^m$

The invariant distribution on $\mathcal{O}^m$ is easily obtained as a restriction of the invariant distribution on O^m. Let O be an orthogonal matrix belonging to $\mathcal{O}^m$ as defined in (8). Using $O = O_{\theta^m}^m \times \ldots \times O_{\theta^2}^m$, all the restrictions in (8) are translated into restrictions on θ^m. Manipulation of the first column of H_{θ^m} shows that θ^m has to meet the following requirements:

$\mathbf{m} = \mathbf{2} : \theta^2 \in \left(-\frac{\pi}{8}, \frac{\pi}{8}\right)$

$$\mathbf{m} > \mathbf{2} : \theta^m \in \left\{ \left(\theta_1^m, \ldots, \theta_{m-1}^m\right)' : \theta_{m-1}^m \in (a, \pi/4) \ \wedge \ \theta_j^m \in \left(0, \operatorname{atan}\left[\sin\left(\theta_{j+1}^m\right)\right]\right),\ j = 2, \ldots, m-2 \wedge \right.$$

$$\left. \wedge\ \theta_1^m \in \left(0, \frac{\operatorname{acot}\left[\prod_{j=2}^{m-1}\cos\left(\theta_j^m\right)\right]}{2}\right) \right\},\ a = -\frac{\pi}{4} \text{ if } m = 3,\ a = 0 \text{ otherwise.}$$

For $j = 2, \ldots, m-1$, $\theta^j \in \Theta^j$ as defined previously. As a consequence, the parameter space of $\theta^2, \ldots, \theta^m$ is always connected, which facilitates inference.

References

Arnold, B. C. and Beaver, R. J. (2000). Hidden truncation models. *Sankhyā A* **62**, 23–35.

Arnold, B. C. and Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting (with discussion). *Test* **11**, 7–54.

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**, 367–389.

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

Bauwens, L. and Laurent, S. (2005). A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *J. Bus. Economic Statist.* **23**, 346–354.

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

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

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulations and asymptotic approximations. *J. Am. Statist. Assoc.* **92**, 903–915.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London.

Fernández, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *J. Am. Statist. Assoc.* **93**, 359–371.

Fernández, C. and Steel, M. F. J. (1999). Multivariate student-t regression models: Pitfalls and inference. *Biometrika* **86**, 153–167.

Fernández, C. and Steel, M. F. J. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory* **16**, 80–101.

Ferreira, J. T. A. S. and Steel, M. F. J. (2004). Bayesian multivariate skewed regression modelling with an application to firm size. *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Ed. M. G. Genton, 175–190. CRC Chapman & Hall, Boca Raton.

Gantmacher, F. R. (1959). *The Theory of Matrices, vol. 1*. Chelsea, New York.

Golub, G. H. and van Loan, C. F. (1989). *Matrix Computations*, 2nd edn. John Hopkins University Press, Baltimore.

Jones, M. C. (2002). Marginal replacement in multivariate densities, with application to skewing spherically symmetric distributions. *J. Multivariate Anal.* **81**, 85–99.

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

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530.

Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6**, 831–860.

Moran, P. A. P. (1967). Testing for correlation between non-negative variates. *Biometrika* **54**, 385–394.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. R. Statist. Soc.* B **56**, 3–48.

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

Stewart, G. W. (1980). The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM J. Numerical Anal.* **17**, 403–409.