A New Class of Multivariate Skew Distributions with Applications to Bayesian Regression Models

Sujit K. Sahu
Faculty of Mathematical Studies,
University of Southampton, UK
email: S.K.Sahu@maths.soton.ac.uk

Dipak K. Dey
Department of Statistics,
University of Connecticut, USA
email: dey@merlot.stat.uconn.edu

Márcia D. Branco
Department of Statistics,
University of São Paulo, Brasil
email: mbranco@ime.usp.br

August 16, 2002

A New Class of Multivariate Skew Distributions with Applications to Bayesian Regression Models

SUMMARY

This article develops a new class of distributions by introducing skewness in the multivariate elliptically symmetric distributions. The class is obtained by using transformation and conditioning. The class contains many standard families including the multivariate skew normal and t distributions. Analytical forms of the densities are obtained and distributional properties are studied. These developments are followed by practical examples in Bayesian regression models. Results on the existence of the posterior distributions and moments under improper priors for the regression coefficients are obtained. The methods are illustrated using practical examples.

KEY WORDS: Bayes factor, Elliptical distributions, Heavy tailed error distribution, Gibbs sampler, Markov chain Monte Carlo, Multivariate skewness.

1 Introduction

Advances in Bayesian computation and Markov chain Monte Carlo have extended and broadened the scope of statistical models that can be fit for practical data. Surprisingly the methodologies and techniques of data augmentation and computation can also be used for developing new sets of flexible models for data. The main motivation of this article comes from this observation. A simple but powerful method of generating a class of multivariate skew elliptical distribution is obtained with a view to finding easily implementable fitting methods.

The class of elliptical distributions, introduced by Kelker (1970), includes a vast set of known symmetric distributions, for example, normal, t and Pearson type II distributions. These ideas are quite well developed, see for example Fang *et al.* (1990). A major focus of the current paper is to propose skewed versions of these distributions which are suitable for practical implementations. A general transformation technique together with a conditioning argument is used to obtain skewed versions of the multivariate distributions. In univariate cases similar ideas have been studied by many authors, see for example, Aigner *et al.* (1977) and Chen *et al.* (1999).

The conditioning arguments on some un-observed variables used to develop the models are commonly used in regression models. The resulting models are often called the hidden truncation models, see e.g. Arnold and Beaver (2000, 2002). Consider the following motivating example. In order to gain admissions in a medical school applicants are often screened by both academic and non-academic criteria. Only the candidates meeting several academic criteria (e.g. overall grades and grades in science) are evaluated by non-academic criteria such as commitment and caring, sense of responsibility etc. A response variable, called the non-academic total obtained by summing the scores from seven such non-academic headings, is used to screen applicants for the next stage of the admission process. Thus 'meeting the academic criteria' acts as a conditioning variable for the response non-academic total. Moreover, some variables (components) in 'meeting the academic criteria' are yet un-observed since the admission process is often initiated much before the applicants take their final qualifying examinations. This example is discussed in more detail in Section 6.

The methodology developed here is also useful in modeling stock market returns. The expected rate of returns on risky financial assets like stocks, bonds, options and other securities are often assumed to be normally distributed but are subject to shocks in either positive or negative directions; positive shocks lead to positively skewed models and negative shocks lead to negatively skewed models, see e.g. Adcock (2002). In the related area of capital asset pricing models the assumption of multivariate normality is often hard to justify in real life examples (Huang and Litzenberger, 1988) and the proposed skew models can be used instead.

In many practical regression problems a suitable transformation for symmetry is often considered for skewed data. The proposed models eliminate the need for such ad-hoc transformations. Instead of transforming the data our methods transform the error distributions to accommodate skewness.

In the case of normal distributions our setup provides a new family of multivariate skew normal distributions. The distributions are different from the ones obtained by Azzalini and his colleagues, see for example, Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (1999). See also Arnold and Beaver (2000) for a generalization. They obtain the multivariate distribution by conditioning on one suitable random variable being greater than zero while we condition on as many random variables as the dimension of the multivariate distribution. Thus in the univariate case the new distributions are same as the ones obtained by Azzalini and Dalla Valle (1996). However, in the multivariate setup the two sets of distributions are quite different. Also our method extends to other distributions, for example the t and the Pearson type II distributions.

There are some other variants of skewed distributions available in the literature. For example, Jones (2000) (and references to his other work therein) provides an alternative skew t distribution which in the

limiting case is a scaled inverse χ distribution. Fernandez and Steel (1998) consider an alternative form where two t distributions (with different scale parameters) in the positive and negative domains are combined to form a skew t distribution. The distributions developed in this article, however, are much easier to work with and implement than others.

Bayesian analysis of regression problems under heavy tailed error distributions has received considerable attention in recent statistical literature. A pioneering work in this area is due to Zellner (1976), in which a study based on multivariate t distribution is considered. Extensions of those results for elliptical distributions are considered in Chib *et al.* (1988), Osiewalski and Steel (1993) and Branco *et al.* (2000). More about Bayesian regression under heavy tailed error distribution can be seen in Geweke (1993), Fernandez and Steel (1998) and references therein. However, the methodologies do not generally extend to multivariate skew distributions.

The plan of the remainder of this paper is as follows. Section 2 develops the multivariate skew elliptical distributions. Sections 3 and 4 consider the particular cases of normal and t distributions. In Section 5 we develop regression models for the skewed distributions obtained in the preceding sections. Results on the propriety of the associated posterior distributions in the univariate case are also obtained here. In Section 6.2 we illustrate our methods when the response variable is univariate. A multivariate example is discussed in Section 6.3. We give few summary remarks in Section 7. Technical proofs of our results are placed in the Appendix.

2 Multivariate Distributions

2.1 Elliptical Distribution

Let Ω be a positive definite matrix of order k and $\boldsymbol{\theta} \in \mathbb{R}^k$. Consider a k-dimensional random vector $\mathbf{X}$ having probability density function (pdf) of the form

$$f\left(\mathbf{x}|\boldsymbol{\theta}, \Omega; g^{(k)}\right) = |\Omega|^{-\frac{1}{2}} g^{(k)}\left[(\mathbf{x}-\boldsymbol{\theta})^T \Omega^{-1} (\mathbf{x}-\boldsymbol{\theta})\right], \quad \mathbf{x} \in \mathbb{R}^k \tag{1}$$

where $g^{(k)}(u)$ is a function from $\mathbb{R}^+$ to $\mathbb{R}^+$ defined by

$$g^{(k)}(u) = \frac{\Gamma(k/2)}{\pi^{k/2}} \frac{g(u;k)}{\int_0^\infty r^{k/2-1} g(r;k) dr}, \tag{2}$$

where $g(u;k)$ is a non-increasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$ such that the integral $\int_0^\infty r^{k/2-1} g(r;k) dr$ exists. In this paper we shall always assume the existence of the pdf (1). The function $g^{(k)}$ is often called the *density*

generator of the random vector $\mathbf{X}$. Note that the function $g(u;k)$ provides the kernel of $\mathbf{X}$ and other terms in $g^{(k)}$ constitute the normalizing constant for the density f. In addition the function g, hence $g^{(k)}$, may depend on other parameters which would be clear from the context. For example, in case of t distributions the additional parameter will be the degrees of freedom. The density f defined above represents a broad class of distributions called the *elliptically symmetric distribution* and we will use the notation

$$X \sim El(\boldsymbol{\theta}, \Omega; g^{(k)}),$$

henceforth in this article. Let $F\left(\mathbf{x}|\boldsymbol{\theta}, \Omega; g^{(k)}\right)$ denote the cumulative density function (cdf) of $\mathbf{X}$ where $\mathbf{X} \sim El\left(\boldsymbol{\theta}, \Omega; g^{(k)}\right)$.

We consider two examples, namely the multivariate normal and t distributions, which will be used throughout this paper.

Example 1: **Multivariate Normal**:
Let $g(u;k) = \exp(-u/2)$. Then straightforward calculation yields

$$g^{(k)}(u) = \frac{e^{-u/2}}{(2\pi)^{k/2}}.$$

Then

$$f\left(\mathbf{x}|\boldsymbol{\theta}, \Omega; g^{(k)}\right) = \frac{1}{(2\pi)^{k/2}}|\Omega|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\theta})^T\Omega^{-1}(\mathbf{x}-\boldsymbol{\theta})\right], \quad \mathbf{x} \in \mathbb{R}^k,$$

which is the pdf of the k-variate normal distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix Ω. We denote this distribution by $N_k(\boldsymbol{\theta}, \Omega)$ and the pdf by $N_k(\mathbf{x}|\boldsymbol{\theta}, \Omega)$ henceforth.

Example 2: **Multivariate** t:
Let

$$g(u;k,\nu) = \left[1 + \frac{u}{\nu}\right]^{-(\nu+k)/2}, \nu > 0. \tag{3}$$

Here g depends on the additional parameter ν, the degrees of freedom. Then straightforward calculation yields

$$g^{(k)}(u;\nu) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\ (\nu\pi)^{k/2}} g(u;k,\nu).$$

Hence

$$f\left(\mathbf{x}|\boldsymbol{\theta}, \Omega, g^{(k)}\right) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\ (\nu\pi)^{k/2}}|\Omega|^{-\frac{1}{2}}\left[1 + \frac{[\mathbf{x}-\boldsymbol{\theta}]^T\Omega^{-1}[\mathbf{x}-\boldsymbol{\theta}]}{\nu}\right]^{-(\nu+k)/2}, \quad \mathbf{x} \in \mathbb{R}^k, \tag{4}$$

which is the density of the k-variate t distribution with parameters $\boldsymbol{\theta}$, Ω and degrees of freedom ν. We denote this distribution by $t_{k,\nu}(\boldsymbol{\theta}, \Omega)$ and the density by $t_{k,\nu}(\mathbf{x}|\boldsymbol{\theta}, \Omega)$ henceforth. The subscript k will be omitted when it is equal to 1.

2.2 Skew Elliptical Distribution

Let $\boldsymbol{\epsilon}$ and $\mathbf{Z}$ denote m-dimensional random vectors. Let $\boldsymbol{\mu}$ be an m-dimensional vector and Σ be an $m \times m$ positive definite matrix. Assume that

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathbf{Z} \end{pmatrix} \sim El\left(\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix}, \Omega = \begin{pmatrix} \Sigma & 0 \\ 0 & I \end{pmatrix}; g^{(2m)}\right),$$

where 0 is the null matrix and I is the identity matrix. We consider a skew elliptical class of distributions by using the transformation

$$\mathbf{Y} = D\mathbf{Z} + \boldsymbol{\epsilon}, \tag{5}$$

where D is a diagonal matrix with elements $\delta_1, \ldots, \delta_m$, though we can work with any non-singular square matrix. Let $\boldsymbol{\delta}^T = (\delta_1, \ldots, \delta_m)$. The class is developed by considering the random variable $[\mathbf{Y}|\mathbf{Z} > \mathbf{0}]$ where $\mathbf{Z} > \mathbf{0}$ means $Z_i > 0$ for $i = 1, \ldots, m$. Note that if $\boldsymbol{\delta} = \mathbf{0}$ then we retrieve the original elliptical distribution. The construction (5) with the conditioning introduces skewness. For positive values of components of $\boldsymbol{\delta}$ we obtain positively (right) skewed distributions and for negative values we obtain negatively (left) skewed distributions. The conditional density of Y is obtained in the following theorem.

Theorem 1

Let $\mathbf{y}_* = \mathbf{y} - \boldsymbol{\mu}$. *Then the pdf of* $\mathbf{Y}|\mathbf{Z} > 0$ *is given by*

$$\begin{aligned} f\left(\mathbf{y}|\boldsymbol{\mu}, \Sigma, D; g^{(m)}\right) &= 2^m f_{\mathbf{Y}}\left(\mathbf{y}|\boldsymbol{\mu}, \Sigma + D^2; g^{(m)}\right) \times \\ &\quad F\left(\left[I - D(\Sigma + D^2)^{-1}D\right]^{-\frac{1}{2}} D(\Sigma + D^2)^{-1}\mathbf{y}_*|\mathbf{0}, I; g^{(m)}_{q(\mathbf{y}_*)}\right), \end{aligned} \tag{6}$$

where

$$g_a^{(m)}(u) = \frac{\Gamma(m/2)}{\pi^{m/2}} \frac{g(a+u; 2m)}{\int_0^\infty r^{m/2-1} g(a+r; 2m) dr}, \quad a > 0, \tag{7}$$

and

$$q(\mathbf{y}_*) = \mathbf{y}_*^T (\Sigma + D^2)^{-1} \mathbf{y}_*.$$

This density matches with the one obtained by Branco and Dey (2001) only in the univariate case. We denote the random variable $\mathbf{Y}$ by using the notation $Y \sim SE\left(\boldsymbol{\mu}, \Sigma, D; g^{(m)}\right)$. In Sections 3 and 4 we provide two examples of the density (6). In general, the cdf in (6) can be hard to evaluate. However, for practical MCMC model fitting the cdf need not be calculated, see Section 5.

In the univariate case, i.e., when $m = 1$ we take $\Sigma = \sigma^2$ and $D = \delta$. The density (6) then simplifies to

$$f\left(y|\mu,\sigma^2,\delta;g^{(1)}\right) = \frac{2}{\sqrt{\sigma^2+\delta^2}}\, g^{(1)}\left(\frac{(y-\mu)^2}{\sigma^2+\delta^2}\right) F\left(\frac{\delta}{\sigma}\frac{y-\mu}{\sqrt{\sigma^2+\delta^2}}\middle| 0,1,g_a^{(1)}\right) \tag{8}$$

where $g^{(1)}(u)$ is given in (2), $a = \frac{(y-\mu)^2}{\sigma^2+\delta^2}$ and $g_a^{(1)}(u)$ is given in (7).

Using the arguments in the proof of Theorem 1 we can obtain the marginal distribution of subsets of components of $\mathbf{Y}$. The marginal distributions are derived using the construct $[Y_i|\mathbf{Z} > 0]$ and not using the construct $[Y_i|Z_i > 0]$. Suppose that it is desired to obtain the marginal density of first m_1 components of $\mathbf{Y}$. The marginal density will be,

$$\begin{aligned} f\left(\mathbf{y}^{(1)}|\boldsymbol{\mu}^{(1)},\Sigma_{11},D_{11};g^{(m_1)}\right) &= 2^{m_1}\, f_{\mathbf{Y}^{(1)}}\left(\mathbf{y}^{(1)}|\boldsymbol{\mu}^{(1)},\Sigma_{11}+D_{11}^2;g^{(m_1)}\right) \times \\ &\quad F\left(\left[I - D_{11}(\Sigma_{11}+D_{11}^2)^{-1}D_{11}\right]^{-\frac{1}{2}} D_{11}(\Sigma_{11}+D_{11}^2)^{-1}\mathbf{y}_*^{(1)}|\mathbf{0},I;g^{(m_1)}_{q(\mathbf{y}_*^{(1)})}\right), \end{aligned}$$

where the symbols have their usual meanings. It is straightforward to observe that the above marginal density is of the same form as (6). Hence coherence with respect to marginalization is preserved under (6). The conditional density of any subset of variables can be obtained from the joint and marginal densities.

3 The Skew Normal Distribution

3.1 Density

Let $g(u;m) = \exp(-u/2)$. Then it is easy to see that $g^{(m)}(u) = (2\pi)^{-m/2}\exp(-u/2)$ and $g^{(m)}_{q(\mathbf{y}_*)}$ is free of $q(\mathbf{y}_*)$, see equation (7). Now the pdf of the skew normal distribution is given by

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\mu},\Sigma,D) &= 2^m|\Sigma+D^2|^{-\frac{1}{2}}\phi_m\left[(\Sigma+D^2)^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right] \times \\ &\quad \Phi_m\left[(I - D(\Sigma+D^2)^{-1}D)^{-\frac{1}{2}}D(\Sigma+D^2)^{-1}(\mathbf{y}-\boldsymbol{\mu})\right], \end{aligned} \tag{9}$$

where ϕ_m and Φ_m denote the density and cdf of m dimensional normal distribution with mean $\mathbf{0}$ and covariance matrix identity. (We drop the subscript m in ϕ_m and Φ_m when $m = 1$.) We denote the above distribution by $SN(\boldsymbol{\mu},\Sigma,D)$. An appealing feature of (9) is the fact that it gives independent marginals when $\Sigma = \sigma^2 I$. The density (9) then reduces to

$$f(\mathbf{y}|\boldsymbol{\mu},\sigma^2,D) = \prod_{i=1}^{m}\left[2(\sigma^2+\delta_i^2)^{-1/2}\phi\left(\frac{y_i-\mu_i}{\sqrt{\sigma^2+\delta_i^2}}\right)\Phi\left(\frac{\delta_i}{\sigma}\frac{y_i-\mu_i}{\sqrt{\sigma^2+\delta_i^2}}\right)\right]. \tag{10}$$

3.2 Moments and Skewness

For the multivariate distribution $SN(\boldsymbol{\mu}, \Sigma, D)$ we provide the first two moments. These are obtained by using the moment generating function

$$M_{\mathbf{Y}}(\mathbf{t}) = 2^m e^{\mathbf{t}^T \boldsymbol{\mu} + \mathbf{t}^T (\Sigma + D^2)\mathbf{t}/2} \Phi_m(D\mathbf{t}). \tag{11}$$

The appendix contains the derivation of this. The mean and variance of of $SN(\boldsymbol{\mu}, \Sigma, D)$ are given by,

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \left(\frac{2}{\pi}\right)^{1/2} \boldsymbol{\delta} \quad \text{and Cov}(\mathbf{Y}) = \Sigma + \left(1 - \frac{2}{\pi}\right) D^2.$$

Since the matrix D is assumed to be diagonal the introduction of skewness does not affect the correlation structure. It changes the values of correlations but the structure remains the same. Thus the mutual independence of the components, when Σ is diagonal, is preserved under (5) for the normal distribution. However, this is not true for the skew normal distribution of Azzalini and Capitanio (1999). Introduction of skewness in their setup changes the correlation structure.

As mentioned previously $SN(\boldsymbol{\mu}, \Sigma, D)$ coincides with the skew normal distribution obtained by Azzalini and Dalla Valle (1996), and Azzalini and Capitanio (1999) in the *univariate* case only. Hence, the skewness properties of the univariate distributions are not investigated here. We instead consider the bivariate distributions for comparison.

The versions of the densities to be compared are chosen so that they have identical first two moments and they differ only in skewness. We use the skewness measure $\beta_{1,2}$ introduced by Mardia (1970).

We consider the following simpler form of (9),

$$f(\mathbf{y}|\delta) = \frac{4}{1+\delta^2} \phi\left(\frac{y_1}{\sqrt{1+\delta^2}}\right) \phi\left(\frac{y_2}{\sqrt{1+\delta^2}}\right) \Phi\left(\delta \frac{y_1}{\sqrt{1+\delta^2}}\right) \Phi\left(\delta \frac{y_2}{\sqrt{1+\delta^2}}\right). \tag{12}$$

This distribution provides identical marginal distributions each with mean

$$\mu(\delta) = \delta\sqrt{\frac{2}{\pi}}, \text{ and variance } \sigma^2(\delta) = 1 + \left(1 - \frac{2}{\pi}\right)\delta^2.$$

Mardia's skewness measure is given by

$$\beta_{1,2}(\delta) = 4(4-\pi)^2 \left\{\frac{\delta^2}{\pi + \delta^2(\pi - 2)}\right\}^3.$$

A version of the bivariate skew normal distribution obtained by Azzalini and Dalla Valle (1996) is the following,

$$f(\mathbf{y}|\alpha) = 2\phi(y_1)\,\phi(y_2)\,\Phi\left(\frac{\alpha}{\sqrt{1-2\alpha^2}}(y_1 + y_2)\right), \tag{13}$$

where α is the skewness parameter. This distribution also provides identical marginal distributions each with mean

$$\mu(\alpha) = \alpha\sqrt{\frac{2}{\pi}}, \text{ and variance } \sigma^2(\alpha) = 1 - \frac{2}{\pi}\alpha^2.$$

The skewness measure $\beta_{1,2}$ is given by

$$\beta_{1,2}(\alpha) = 16(4-\pi)^2 \left\{\frac{\alpha^2}{\pi - 4\alpha^2}\right\}^3.$$

When $\delta = \alpha$ we have $\mu(\delta) = \mu(\alpha)$, but the variances $\sigma^2(\delta)$ and $\sigma^2(\alpha)$ fail to coincide. A fair graphical comparison between two densities is not possible if they have un-equal variances because the variance also affects the tail of the density. In order to have identical means and variances under the two densities we simply apply the origin and scale transformation which produces components each with mean zero and variance one. In particular, for the bivariate density (12) we transform

$$z_i = \frac{y_i - \mu(\delta)}{\sigma(\delta)},$$

while for (13) we transform

$$z_i = \frac{y_i - \mu(\alpha)}{\sigma(\alpha)}.$$

The two resulting bivariate densities provide components each with mean zero and variance one. The linear transformation, however, do not change the skewness measure $\beta_{1,2}$.

We first plot the density (12) linearly transformed to have zero mean and unit variance for each component for $\delta = 1, 3, 5$ and 10. The corresponding values of the skewness measure $\beta_{1,2}(\delta)$ are 0.04, 0.89, 1.45 and 1.82 as labelled in the plot. Clearly, the bivariate distribution gets more right skewed as δ increases.

It now remains to compare the shape of (12) with that of the Azzalini and Dalla Valle (1996) skew normal density (13). We plot the densities of the transformed random variables in Figure 2 for $\beta_{1,2}(\delta) = \beta_{1,2}(\alpha) =$ 0.71 and 0.98. The two plots in the first column correspond to the skew normal density (12) and the plots in the second column correspond to (13). The implied values of δ and α are labelled in the plot. As in Figure 1 the transformed density (12) shifts more and more probability mass to the positive quadrant from the remaining three quadrant as δ (or equivalently $\beta_{1,2}(\delta)$) becomes larger. The Azzalini and Dalla Valle (1996) version corresponding to (13) also shifts probability mass to the positive quadrant, but it keeps some substantial probability mass on the second and fourth quadrant even for the maximum allowable value of α. This phenomenon is explained by the fact that the density (13) is resulted by conditioning on *one* random variable while (12) is obtained by conditioning on *two* random variables. The two conditioning random variables in (12) allow two lines (which are almost the axes for the standardized variables when $\delta = 10$, see Figure 1) to effectively bound the left tail of the bivariate distribution while the only conditioning random variable limits the left tail of (13) by using only one line.

4 The Skew t Distribution

4.1 Density

Let

$$g(u; 2m, \nu) = \left[1 + \frac{u}{\nu}\right]^{-(\nu+2m)/2}.$$

Note that the two ingredients of (6) requires a marginal and a cumulative conditional density. For an m-dimensional marginal density we have

$$g^{(m)}(u) = \frac{\Gamma(m/2)}{\pi^{m/2}} \frac{g(u; m, \nu)}{\int_0^\infty r^{m/2-1} g(r; m, \nu) dr},$$

following Theorem 3.7 in Fang *et al.* (1990, p83). Therefore, in (6) the marginal density is $t_{m,\nu}(\mathbf{y}|\mu, \Sigma + D^2)$. For the cumulative conditional density we first obtain $g_a^{(m)}$. From (7) and ingredients in Lemma A.1 we have,

$$\begin{aligned} g_a^{(m)}(u; \nu) &= \Gamma(m/2)\pi^{-m/2} g(a+u; 2m, \nu) \left[\int_0^\infty r^{m/2-1} g(a+r; 2m, \nu) dr\right]^{-1} \\ &= \Gamma\left(\tfrac{m}{2}\right) [\pi(\nu+m)]^{-\frac{m}{2}} \left(\tfrac{\nu+m}{\nu+a}\right)^m \left(1 + \tfrac{u}{\nu+m}\tfrac{\nu+m}{\nu+a}\right)^{-(\nu+2m)/2}. \end{aligned}$$

Hence, the conditional density is of the form (A.2) and is given by

$$t_{m,\nu+m}\left(\mathbf{z}|D(\Sigma+D^2)^{-1}\mathbf{y}_*, \frac{\nu+q(\mathbf{y}_*)}{\nu+m}(I - D(\Sigma+D^2)^{-1}D)\right).$$

After standardization the cumulative conditional density is

$$T_{m,\nu+m}\left[\left(\frac{\nu+q(\mathbf{y}_*)}{\nu+m}\right)^{-\frac{1}{2}} \left(I - D(\Sigma+D^2)^{-1}D\right)^{-\frac{1}{2}} D(\Sigma+D^2)^{-1}\mathbf{y}_*\right]$$

where $T_{m,\nu+m}(\cdot)$ denote cdf of $t_{m,\nu+m}(\mathbf{0}, I)$ and $q(\mathbf{y}_*) = \mathbf{y}_*^T(\Sigma+D^2)^{-1}\mathbf{y}_*$. The generator function calculation gave two extra quantities which were not present in the multivariate skew normal distribution, namely (i) degrees of freedom of the conditional density is $\nu + m$ and (ii) the factor $\left(\frac{\nu+q(\mathbf{y}_*)}{\nu+m}\right)^{-\frac{1}{2}}$ in the argument of the cdf.

Summarizing the preceding discussion we have the density of the multivariate skew t distribution given by

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\mu}, \Sigma, D, \nu) &= 2^m\, t_{m,\nu}(\mathbf{y}|\boldsymbol{\mu}, \Sigma + D^2) \times \\ &\quad T_{m,\nu+m}\left[\left(\frac{\nu+q(\mathbf{y}_*)}{\nu+m}\right)^{-\frac{1}{2}} \left(I - D(\Sigma+D^2)^{-1}D\right)^{-\frac{1}{2}} D(\Sigma+D^2)^{-1}\mathbf{y}_*\right]. \end{aligned} \quad (14)$$

We denote this distribution by $ST_\nu(\boldsymbol{\mu}, \Sigma, D)$. For $\Sigma = \sigma^2 I$ and $D = \delta I$ the above simplifies to

$$f(\mathbf{y}|\boldsymbol{\mu}, \sigma^2, \delta, \nu) = 2^m (\sigma^2 + \delta^2)^{-m/2} \frac{\Gamma\left(\frac{\nu+m}{2}\right)}{\Gamma(\nu/2)(\nu\pi)^{m/2}} \left[1 + \frac{\mathbf{y}_*^T \mathbf{y}_*}{\nu(\sigma^2+\delta^2)}\right]^{-(\nu+m)/2} T_{m,\nu+m}\left[\left(\frac{\nu + q(\mathbf{y}_*)}{\nu+m}\right)^{-\frac{1}{2}} \frac{\delta}{\sigma} \frac{\mathbf{y}_*}{\sqrt{\sigma^2+\delta^2}}\right]. \quad (15)$$

However, unlike the skew normal case the above density cannot be written as the product of univariate skew t densities. It is to be noted that, here Y_i's are not independent but they are uncorrelated.

4.2 Moments and Skewness

The moments of the skew t distribution, $ST_\nu(\boldsymbol{\mu}, \Sigma, D)$ are not straightforward to obtain using the density (14). Here we derive the first two moments by viewing $ST_\nu(\boldsymbol{\mu}, \Sigma, D)$ as a scale mixture of $SN(\boldsymbol{\mu}, \Sigma, D)$. We obtain the following results by using the expression for the moment generating function given in the appendix.

The mean and variance of the skew t distribution $ST_\nu(\boldsymbol{\mu}, \Sigma, D)$ are given by

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \left(\frac{\nu}{\pi}\right)^{1/2} \frac{\Gamma[(\nu-1)/2]}{\Gamma(\nu/2)} \boldsymbol{\delta},$$

and

$$\mathrm{Cov}(\mathbf{Y}) = (\Sigma + D^2)\frac{\nu}{\nu-2} - \frac{\nu}{\pi}\left(\frac{\Gamma[(\nu-1)/2]}{\Gamma(\nu/2)}\right)^2 D^2,$$

when $\nu > 2$.

We have calculated the multivariate skewness measure $\beta_{1,m}$ (Mardia, 1970) in analytic form for the skew t distribution. The expression does not simplify and involves non-linear interactions between the degrees of freedom (ν) and the skewness parameter δ where $D = \delta I$. However, $\beta_{1,m}$ approaches ± 1 as $\delta \to \pm\infty$.

5 Regression Models with Skewness

5.1 Models for Univariate Response

We consider regression model where the error distribution follows the skew elliptical distribution. Let X be an $n \times p$ design matrix (with full column rank) and $\boldsymbol{\beta}$ be a vector (dimension p) of regression parameters.

Suppose that we have n independent observed one-dimensional response variables y_i. Further $y_i \sim SE(\mu_i, \sigma^2, \delta; g^{(1)})$ independently. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. For the regression model we assume that $\boldsymbol{\mu}_i = \mathbf{x}_i^T \boldsymbol{\beta}$ where $\mathbf{x}_i^T$ denote the ith row of the matrix X. Thus the assumed regression model is $\boldsymbol{\mu} = X\boldsymbol{\beta}$. The likelihood function of $\boldsymbol{\beta}, \sigma^2$ and δ and any other parameter involved in $SE(\mu_i, \sigma^2, \delta; g^{(1)})$ is given by the product of densities of the form (8). Hence we write,

$$L(\boldsymbol{\beta}, \sigma^2, \delta, g^{(1)}; y_1, \ldots, y_n) = \prod_{i=1}^{n} f\left(y_i | \mu_i, \sigma^2, \delta; g^{(1)}\right)$$

where $f\left(y|\mu, \sigma^2, \delta; g^{(1)}\right)$ is given in (8). The above likelihood may also depend on additional parameters. For example, for the t distributions the additional parameter is ν.

Often the error distribution in a regression model is taken to have mean zero. The regression model developed here can be forced to satisfy this requirement by suitably adjusting the intercept parameter. See Section 6.2 for particulars.

To completely specify the Bayesian model we need to specify prior distributions for all the parameters. As a default prior for $\boldsymbol{\beta}$ we take the constant prior $\pi_{\boldsymbol{\beta}} \propto 1$ in $\mathbb{R}^p$. For $\tau = 1/\sigma^2$ we assume a gamma prior distribution $\Gamma(\kappa, \kappa)$ where the parameterization has mean 1 and κ is assumed to be a known parameter. In other words σ^2 is given an inverse gamma distribution. The parameter δ is given a normal prior distribution. Specific parameter values of this prior distribution will be discussed in particular examples. When the t models are considered we need prior distribution for the degrees of freedom parameter ν. For this, we use the exponential distribution with parameter 0.1 truncated in the region $\nu > 2$, so that the underlying t distribution (skew or not) has finite mean and variance.

Now the joint posterior density is given by,

$$\pi(\boldsymbol{\beta}, \sigma^2, \delta, \nu | y_1, \ldots, y_n) \propto L(\boldsymbol{\beta}, \sigma^2, \delta, g^{(1)}; y_1, \ldots, y_n)\, \pi_{\boldsymbol{\beta}}\, \pi_{\sigma^2}\, \pi_\delta \pi_\nu, \tag{16}$$

where $\pi_\cdot$ on the right hand side denote the prior density of its argument. Note that the parameter ν is omitted for the normal distributions.

In many practical examples it is possible to decide the type of skewed distributions appropriate for data a-priori. For example either the positively skewed or the negatively skewed distributions may be considered to be appropriate for data. Thus it is reasonable to assume proper prior distributions for the skewness parameter.

In many examples, however, we may not have precise information about $\boldsymbol{\beta}$ and σ^2 and we will be required to use the default prior distributions, as is often done in practice. A natural question in such a case is whether

the full posterior distribution is proper. In the following theorem we answer this in affirmative for the skew normal or skew-t error distributions.

Theorem 2

Suppose that π_δ and π_ν are proper distributions and $\pi_{\boldsymbol{\beta}} \propto 1$. Then the posterior (16) is proper under the skew normal or skew t model if $n > p$.

In the appendix we provide a proof of this theorem. In fact a more general theorem is proved and the above result is obtained under the special cases of normal and t distributions. As a consequence of the proof of Theorem 2 we have the following result on the existence of the posterior moments of σ^2.

Theorem 3

Suppose that π_δ and π_ν are proper distributions and and $\pi_{\boldsymbol{\beta}} \propto 1$. Then $E[(\sigma^2)^k|\mathbf{y}]$ exists under the skew normal or skew t model if $n - p > 2k$.

A similar result is obtained by Geweke (1993) for the t model with unequal variance assumption. Theorem 3 extends his result to the skewed models which include several other distributions.

5.2 MCMC Specification

In order to specify the model (5) for MCMC computation we use the hierarchical setup of $f(\mathbf{y}|\mathbf{z})$ and $f(\mathbf{z})I(\mathbf{z} > \mathbf{0})$ where f is used as a generic notation denoting the density of the random variable in the argument. We obtain these two distributions from (A.3) which is,

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \sim El\left(\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix}, \Omega = \begin{pmatrix} \Sigma + D^2 & D \\ D & I \end{pmatrix}; g^{(2m)}\right). \tag{17}$$

Here we have

$$\mathbf{Y}|\mathbf{Z} = \mathbf{z} \sim El\left(\boldsymbol{\mu} + D\mathbf{z}, \Sigma; g^{(m)}_{q(z)}\right)$$

where $q(z) = \mathbf{z}^T\mathbf{z}$. For the skew normal model this is simply a multivariate normal distribution with mean $\boldsymbol{\mu} + D\mathbf{z}$ and covariance matrix Σ, since $g^{(m)}_{q(z)}$ is independent of $q(z)$. However, for the skew t model this is not so and

$$\mathbf{Y}|\mathbf{Z} = \mathbf{z} \sim t_{m,\nu+m}\left(\boldsymbol{\mu} + D\mathbf{z}, \frac{\nu + \mathbf{z}^T\mathbf{z}}{\nu + m}\Sigma\right).$$

The marginal specification for $\mathbf{Z}$ for the skew normal case is simply the $N_m(\mathbf{0}, I)$ distribution. The same for the skew t case is $t_{m,\nu}(\mathbf{0}, I)$. Lastly, the distribution of $\mathbf{Z}$ is truncated in the space $\mathbf{z} > \mathbf{0}$.

5.3 Multivariate Response

Regression models for multivariate response variables are constructed as follows. Let $\mathbf{Y}_i \sim SE\left(\boldsymbol{\mu}_i, \Sigma, D; g^{(m)}\right)$ for $i = 1, \ldots, n$. For each data point with covariate information assumed in a $p \times m$ matrix X_i, we can specify the linear model

$$\boldsymbol{\mu}_i = X_i^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a p-vector of regression coefficients. The coefficients are given a multivariate normal $N_p(\boldsymbol{\beta}_0, \Lambda)$ prior distribution, where Λ is a known positive definite matrix and $\boldsymbol{\beta}_0$ is a vector of constants to be chosen later. The matrix Σ is assigned independent conjugate Wishart prior distribution as follows:

$$\Sigma^{-1} = Q \sim W_m(2r, 2\kappa)$$

where $2r$ is the assumed prior degrees of freedom ($\geq m$) and κ is a positive definite matrix. We say that $\mathbf{X}$ has the Wishart distribution $W_m(k, A)$ if its density is proportional to

$$|A|^{k/2}|y|^{\frac{1}{2}(k-p-1)}e^{-\frac{1}{2}\mathrm{tr}(A\mathbf{x})} \tag{18}$$

if $\mathbf{x}$ is an $m \times m$ positive definite matrix. (Here $\mathrm{tr}(A)$ is the trace of a matrix A.) This is the parameterization used by for example, the `BUGS` (Spiegelhalter *et al.*, 1996) software. The skewness parameters in D, vectorized as $\boldsymbol{\delta}$, are given a normal prior distribution $N_m(\mathbf{0}, \Gamma)$ where Γ is a positive definite matrix.

In the remainder of this section we develop computational procedure for the multivariate skew t distribution and obtain the methods for multivariate skew normal as a special case. The full likelihood specification is given as follows. We introduce n i.i.d. random variables w_i for each data point to obtain the t models. For the normal distributions each of these will be set at 1.

$$\begin{aligned}
\mathbf{Y}_i|\mathbf{z}_i, \boldsymbol{\beta}, X_i, \Sigma, D, w_i &\sim N_m\left(X_i^T\boldsymbol{\beta} + D\mathbf{z}_i, \tfrac{\Sigma}{w_i}\right) \\
\mathbf{Z}_i &\sim N_m(\mathbf{0}, I)I(\mathbf{z} > \mathbf{0}) \\
\boldsymbol{\beta} &\sim N_p(\boldsymbol{\beta}_0, \Lambda) \\
Q = \Sigma^{-1} &\sim W_m(2r, 2\kappa) \\
\boldsymbol{\delta} &\sim N_m(\mathbf{0}, \Gamma) \\
w_i &\sim \Gamma(\nu/2, \nu/2) \\
\nu &\sim \Gamma(1, 0.1)I(\nu > 2),
\end{aligned}$$

the last two distributional specifications are omitted in the normal distribution case. All of the full conditional distributions for Gibbs sampling are straightforward to derive and sample from except for $\mathbf{z}_i$ and $\boldsymbol{\delta}$. Their full conditional distributions are given by

$$\begin{aligned}
\mathbf{Z}_i|\cdots &\sim N_m(A_i^{-1}\mathbf{a}_i, A_i^{-1})I(\mathbf{z}_i > \mathbf{0}) \\
\boldsymbol{\delta}|\cdots &\sim N_m(B^{-1}\mathbf{b}, B^{-1})
\end{aligned}$$

where

$$A_i = I + w_i DQD \text{ and } \mathbf{a}_i = w_i DQ(\mathbf{y}_i - X_i^T\boldsymbol{\beta}),$$

$$B = \Gamma^{-1} + \sum_{i=1}^{n} \text{diag}(\mathbf{z}_i)\, Q\, \text{diag}(\mathbf{z}_i) \text{ and } \mathbf{b} = \sum_{i=1} \text{diag}(\mathbf{z}_i)\, Q\, (\mathbf{y}_i - X_i^T\boldsymbol{\beta}),$$

where diag($\mathbf{a}$) is a diagonal matrix with diagonal elements being the components of $\mathbf{a}$.

6 Examples

6.1 Interview Data

In order to gain admissions in a certain medical school the applicants are screened using both academic qualifications and non-academic characteristics. Each applicant meeting some observed and some predicted academic criteria receives a non-academic total which is the sum of seven scores. The seven scores are assigned on the basis of work experience, sense of responsibility, commitment and caring, motivation, study skills, interest and referees comments. Applicants are subsequently selected for interviews based on their non-academic totals. The interviewed applicants are given scores which are sums of two individual scores given by each member of a two-member interview committee.

In our univariate skewed regression model set-up the non-academic totals are considered to be realizations of the response variable. Here the academic scores of applicants work as the un-observed conditioning variables leading to our regression model. The true academic scores of applicants are yet un-observed because the admission process takes place much before the applicants sit for the final qualifying examination called the A-level examination in Great Britain.

The data set to be analyzed here is obtained as part of a large cohort data giving the details of candidates who applied for a medical degree in a certain medical school in Great Britain. For the univariate analysis we have the non-academic totals for 584 applicants, while for our bivariate analysis we have the non-academic totals and interview scores for only 328 candidates who were selected for and subsequently attended the interviews.

6.2 Univariate Regression

The response variable non-academic total is influenced by several academic, socio-economic and demographic factors as expected. In our current study we only consider the influence of race and gender of the applicants;

these characteristics of the applicants were known to their evaluators. Although the applicants are classified to come from 6 combined ethnic types viz. white, black, Indian, Pakistani and Bangladeshi, other Asian and others, for our purposes we classify candidates whether they were white or non-white. We are then interested to compare four groups of applicants: white female, white male, non-white female and non-white male. The boxplot of the data in Figure 3 indicates that the first group has higher average non-academic totals than the other groups. Simple t tests on the data also show significant differences between the groups.

The data are not expected to be heavily skewed since the individual data points are sum of components as mentioned previously. However, the observations are sum of only seven components so the central limit theorem may not ensue for such a small sample size. This is indeed true, see Figure 4. The left tail of the underlying distribution descends more slowly than the right tail. Our explicit skew-regression models will estimate and test for the skewness in the data more formally.

Let y_i denote the non-academic total of the ith applicant for $i = 1, \ldots, n = 584$. In order to compare between the four groups, white female, white male, non-white female and non-white male we code three binary regressors taking the values 0 and 1 as described below. The first regressor takes the value 1 for white male, the second takes the value 1 for non-white female, while the third takes the value 1 for non-white female. The resulting regression co-efficients allow comparison of the last three groups keeping the white female as the base group. Thus we have the regression model

$$y_i = \alpha' + \sum_{j=1}^{3} \beta_j x_{ij} + \delta\, z_i + \epsilon_i, i = 1, \ldots, n. \tag{19}$$

We calculate the true intercept $\alpha = \alpha' - \delta E(z_i)$ which corresponds to the regression model where the error distribution has mean zero. For the normal and t models expressions for $E(z_i)$ are given in the preceding sections.

Throughout we assume independent diffuse prior distribution $N(0, 10^4)$ for the regression parameters α' and β_j. For $\tau = 1/\sigma^2$, we assume a limiting non-informative gamma prior distribution $\Gamma(0.01, 0.01)$ where the parameterization has mean 1. When δ is not assumed to be zero, it is given a normal prior distribution with mean zero and variance 100. Thus δ is assigned a proper prior distribution which is a requirement of Theorems 2 and 3.

1. Normal linear model: We take $\delta = 0$ and $\epsilon \sim N(0, \sigma^2)$.

2. Skew Normal model: We assume that z_i given all the other random quantities in the model follows the standard half-normal distribution.

3. t-model: We assume that $\epsilon \sim t_\nu$ where t_ν is the standard t-distribution with ν degrees of freedom. We further assume $\delta = 0$. Although the parameter ν is traditionally taken as an integer, it can be treated as a continuous parameter taking positive values since the associated densities are well defined in this case. We assume a truncated ($\nu > 2$) exponential distribution with parameter 0.1. The truncation assures the finiteness of the mean and variance of the associated t error distribution.

4. Skew t-model: We assume that $\epsilon \sim t_\nu$ where t_ν is the $t-$distribution with ν degrees of freedom. We also assume that z_i follows i.i.d. $|t_\nu|$ conditional on other random quantities in the model. The prior distributions for the remaining parameters are assumed the same as in the previous cases.

The Gibbs sampler has been implemented using the `BUGS` software, the codes are available from the authors upon request. We use the 10,000 iterates after discarding first 5000 iterates to make inference. The regression model has an intercept α' and three regression parameters: β_1 for white male applicants, β_2 for non-white female applicants and β_3 for the non-white male applicants. Resulting parameter estimates (posterior means) are given in Table 1.

The estimates of the regression parameters across the models agree broadly. All three regression parameters β_j are significant in all the models since the associated 95% probability intervals do not include the value zero. The negative estimates show that the base group of white female receives significantly higher average non-academic totals than the remaining three groups. The difference between the white female and non-white male is the most significant. Thus the latter group seems to perform most poorly non-academically even though they all met the academic criteria.

The estimates of the parameter σ^2 are smaller for the corresponding skewed model. This is expected because high variability, heaviness of the tails and skewness are interchangeable to a certain extent. The non-skewed symmetric error models endeavor to capture skewness by having larger tails. The important question is that whether high variability can completely replace skewness. In the next paragraph we answer this in negative.

The skewness parameter δ is estimated to be negative in both the skew normal and skew-t model; this confirms the exploratory left skewed histogram of the response variable in Figure 4. Moreover, δ is significant under the skew normal model since the 95% probability interval is (–3.28, –1.77). Thus we can conclude that significant skewness is required to model the data.

The parameter δ does not turn out to be significant under the skew-t model. This is explained as follows. Observe that the fitted symmetric t-error distribution is lighter tailed (estimated df =13.85) with larger

dispersion parameter σ^2 than the fitted skew-t model (estimated df =9.99). With such heavy tailed error distribution, it was not possible to see significant skewness in the data. This, however, does not necessarily reduce the predictive power of the skew-t model.

To compare the four models informally we compute the effective number of parameters ρ_D and the deviance information criterion *(DIC)* as presented by Spiegelhalter *et al.* (2002). They claim that the *(DIC)* as implemented in the BUGS software can be used to compare complex models and large differences in the criterion can be attributed to real predictive differences in the models, although there are many critics. Using the *DIC* values shown in Table 2, we see that the skewed models improve the corresponding symmetric models; the symmetric normal and t models are very similar; the skew-t model is the best model for the data. For the symmetric normal and t models the effective number of parameters ρ_D roughly indicates the number of parameters in the regression model. Spiegelhalter *et al.* (2002) mention that ρ_D can be negative for non-log-concave densities, the present example with skewed distributions provides a case in point. The same conclusions, e.g. the skewed models are better and the skew-t model is the best are also arrived at using more formal Bayesian predictive model choice criteria, e.g. the Bayes factors (DiCiccio *et al.*, 1997). We, however, omit the details. Instead, we compare the residuals from the symmetric and the corresponding skew models to examine if indeed the skew models were able to improve upon the symmetric models. In Figure 5 we plot kernel density estimates of the standardized residuals with the same smoothing parameter. Clearly the density plots for the skewed models have thinner tails than the corresponding symmetric models.

6.3 A Multivariate Illustration

The non-academic totals and interview scores of 328 candidates are plotted in Figure 6. From the plot it is clear that symmetric distributions should not be fitted to this data set. We proceed with the multivariate models of Section 5.3. We adopt the following values of the hyper-parameters. Let $\boldsymbol{\xi}$ be the two component vector where each element is the mid-point of the corresponding component of the bivariate data. Further, let R denote the diagonal matrix where each diagonal entry is the squared range of the corresponding component in the data. Since we do not consider any covariate for this example the regression parameter $\boldsymbol{\beta}$ is the mean parameter $\boldsymbol{\mu}$. For this we assume a normal prior distribution with mean $\boldsymbol{\beta}_0 = \boldsymbol{\xi}$ and covariance matrix $\Lambda = 100 \times R$. The degrees of freedom parameter $2r$ in the Wishart distribution is set at 3 which corresponds to the non-informative prior distribution, see the Wishart density in (18). The matrix κ in the Wishart distribution is taken as $100/(2r)R^{-1}$. Finally, each component of $\boldsymbol{\delta}$ is given an independent normal prior distribution with mean zero and variance 100.

	Mean	sd	2.5%	97.5%
		Normal model		
α	26.18	0.14	25.90	26.47
β_1	-0.65	0.24	-1.13	-0.18
β_2	-0.87	0.36	-1.58	-0.17
β_3	-0.98	0.39	-1.74	-0.21
σ^2	6.46	0.38	5.75	7.25
		Skew normal model		
α	30.40	0.62	28.96	31.44
β_1	-0.63	0.24	-1.10	-0.15
β_2	-0.89	0.35	-1.58	-0.19
β_3	-1.11	0.39	-1.86	-0.33
σ^2	3.88	0.65	2.77	5.34
δ	-2.64	0.38	-3.28	-1.77
		t model		
α	26.21	0.14	25.93	26.49
β_1	-0.57	0.24	-1.04	-0.10
β_2	-0.89	0.35	-1.58	-0.19
β_3	-0.97	0.37	-1.70	-0.25
σ^2	5.39	0.44	4.54	6.29
ν	13.85	5.61	6.71	28.74
		Skew-t model		
α	26.47	2.75	23.11	30.80
β_1	-0.56	0.24	-1.04	-0.07
β_2	-0.89	0.35	-1.57	-0.21
β_3	-1.00	0.38	-1.72	-0.23
σ^2	4.10	0.98	2.28	5.72
δ	-0.16	1.58	-2.70	1.76
ν	9.99	3.95	5.17	20.69

Table 1: Posterior mean, sd and 95% probability intervals for the parameters.

	p_D	DIC
Normal	4.96	2750.65
Skew-N	217.82	2658.06
t	5.62	2742.06
Skew-t	–187.5	2387.4

Table 2: The effective number of parameters, p_D and DIC for the four fitted models.

In Table 3 we provide the estimates of the marginal likelihood using the approach of Gelfand and Dey (1994). The skew bivariate normal model is a large improvement over the bivariate normal model and a heavy tailed bivariate skew t distribution seems to be the most appropriate model for the data.

Normal	Skew Normal	t	Skew-t
–1776.0	–1671.7	–1722.5	–1631.8

Table 3: Marginal likelihood for the bivariate example.

7 Conclusion

The new class of skewed distributions obtained in this article is very general, quite flexible and widely applicable. The skewed distributions are shown to provide an alternative to symmetric distributions often assumed in regression. Although the associated density functions are quite difficult to handle, we show that the models can be easily fit using MCMC methods. Moreover, the univariate models are fitted using publicly available software `BUGS`. This makes our approach quite powerful and accessible to the practicing statisticians. Other variants of skewed distributions currently available are not so easy to implement.

In this article we obtain the skewed distributions by transformation and then conditioning on the same number random variables, m in Theorem 1. As mentioned in the introduction, Azzalini and Capitanio (1999) condition on one random variable being positive. It is certainly possible to impose the non-negativity condition on any other number of random variables, although we have not pursued this.

Observe that the exact form of the densities of skewed distributions obtained in Theorem 1 need not be calculated if the sole purpose is to perform model fitting. However, model comparison using the Bayes factors can be performed easily if it was possible to calculate the density. The augmented variables used in model fitting can be ignored when calculating the marginal likelihood since the marginal density of the data

is available analytically.

Although we have not discussed, the Bayes factors can be used to solve the associated problems of variable selection. Moreover, other existing Bayesian techniques of variable selection and model averaging can be implemented with the models developed here.

Acknowledgement

The authors would like to thank specially Adelchi Azzalini and Chris Jones for many insightful discussions.

Appendix: Theorem Proofs

Before proving Theorem 1 we consider the following Lemma which is Theorem 2.18 in Fang *et al.* (1990, p45). Partition $\mathbf{X}, \boldsymbol{\theta}, \Omega$ into

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{pmatrix}, \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$$

where $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are respectively k_1 and k_2 dimensional random vectors, and $k_1 + k_2 = k$. The parameters $\boldsymbol{\theta}$ and Ω are partitioned accordingly.

Lemma A.1

Let $\mathbf{X} \sim El(\boldsymbol{\theta}, \Omega; g^{(k)})$. *Then*

$$\mathbf{X}^{(1)} | \mathbf{X}^{(2)} = \mathbf{x}^{(2)} \sim El\left(\boldsymbol{\theta}_{1.2}, \Omega_{11.2}; g^{(k_1)}_{q(\mathbf{x}^{(2)})}\right).$$

where

$$\begin{aligned} \boldsymbol{\theta}_{1.2} &= \boldsymbol{\theta}^{(1)} + \Omega_{12}\Omega_{22}^{-1}\left(\mathbf{x}^{(2)} - \boldsymbol{\theta}^{(2)}\right) \\ \Omega_{11.2} &= \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} \\ q(\mathbf{x}^{(2)}) &= \left(\mathbf{x}^{(2)} - \boldsymbol{\theta}^{(2)}\right)^T \Omega_{22}^{-1}\left(\mathbf{x}^{(2)} - \boldsymbol{\theta}^{(2)}\right) \\ g_a^{(k_1)}(u) &= \frac{\Gamma(k_1/2)}{\pi^{k_1/2}} \frac{g(a+u;k)}{\int_0^\infty r^{k_1/2-1} g(a+r;k) dr}. \end{aligned}$$

We give an example which illustrates this Lemma.

Example 3: **Conditional distribution of multivariate** t:

Consider the multivariate t distribution $t_{k,\nu}(\boldsymbol{\theta}, \Omega)$ as defined in (4). It is well known, see for example Bernardo and Smith (1994, p140), that

$$\mathbf{X}^{(1)}|\mathbf{X}^{(2)} = \mathbf{x}^{(2)} \sim t_{k_1,\nu+k_2}\left(\boldsymbol{\theta}_{1.2}, \frac{\nu + q(\mathbf{x}^{(2)})}{\nu + k_2}\,\Omega_{11.2}\right). \tag{A.1}$$

We get the same result using Lemma A.1 as follows. Note that here $g(u;k,\nu) = \left[1+\frac{u}{\nu}\right]^{-(\nu+k)/2}$ as given in (3). We first obtain $g_a^{(k_1)}(u)$ as

$$\begin{aligned} g_a^{(k_1)}(u) &= \Gamma\left(\tfrac{k_1}{2}\right)\pi^{-\frac{k_1}{2}} g(a+u;k,\nu)\left[\int_0^\infty r^{\frac{k_1}{2}-1} g(a+r;k,\nu)dr\right]^{-1} \\ &= \Gamma\left(\tfrac{k_1}{2}\right)\pi^{-\frac{k_1}{2}}\left(1+\tfrac{a+u}{\nu}\right)^{-(\nu+k)/2}\left[\int_0^\infty r^{\frac{k_1}{2}-1}\left(1+\tfrac{a+r}{\nu}\right)^{-(\nu+k)/2} dr\right]^{-1} \\ &= \Gamma\left(\tfrac{\nu+k}{2}\right)\Gamma^{-1}\left(\tfrac{\nu+k_2}{2}\right)[\pi(\nu+k_2)]^{-\frac{k_1}{2}}\left(\tfrac{\nu+k_2}{\nu+a}\right)^{k_1/2}\left(1+\tfrac{u}{\nu+k_2}\tfrac{\nu+k_2}{\nu+a}\right)^{-(\nu+k)/2}, \end{aligned}$$

after some obvious manipulations. Hence the density of $\mathbf{X}^{(1)}|\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ is given by,

$$\begin{aligned} f(\mathbf{x}^{(1)}|\mathbf{x}^{(2)}) &= \Gamma\left(\frac{\nu+k}{2}\right)[\pi(\nu+k_1)]^{-\frac{k_1}{2}}\Gamma^{-1}\left(\frac{\nu+k_2}{2}\right)\left|\frac{\nu+a}{\nu+k_2}\Omega_{11.2}\right|^{-\frac{1}{2}} \times \\ &\quad \left(1+\frac{(\mathbf{x}^{(1)}-\boldsymbol{\theta}_{1.2})^T\left(\frac{\nu+a}{\nu+k_2}\Omega_{11.2}\right)^{-1}(\mathbf{x}^{(1)}-\boldsymbol{\theta}_{1.2})}{\nu+k_2}\right)^{-(\nu+k_2+k_1)/2} \end{aligned} \tag{A.2}$$

where $a = q(\mathbf{x}^{(2)}) = \left(\mathbf{x}^{(2)}-\boldsymbol{\theta}^{(2)}\right)^T\Omega_{22}^{-1}\left(\mathbf{x}^{(2)}-\boldsymbol{\theta}^{(2)}\right)$. Clearly (A.2) is the density of (A.1).

Proof of Theorem 1: Consider the transformation

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{V}\end{pmatrix} = \begin{pmatrix} I & D\\ 0 & I\end{pmatrix}\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathbf{Z}\end{pmatrix}.$$

where 0 is the null matrix. Using Theorem 2.16 of Fang *et al.* (1990) we see that

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{V}\end{pmatrix} \sim El\left(\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix}, \Omega = \begin{pmatrix}\Sigma+D^2 & D\\ D & I\end{pmatrix}; g^{(2m)}\right). \tag{A.3}$$

From this joint distribution we aim to obtain the conditional density of $\mathbf{Y}|\mathbf{V}>\mathbf{0}$. Since $\mathbf{V} \sim El(\mathbf{0}, I; g^{(m)})$ marginally, we have that $Pr(\mathbf{V}>\mathbf{0}) = 2^{-m}$. By using standard arguments,

$$f(\mathbf{y}|\mathbf{V}>\mathbf{0}) = 2^m f_{\mathbf{Y}}\left(\mathbf{y}|\mu, \Sigma+D^2; g^{(m)}\right) Pr(\mathbf{V}>\mathbf{0}|\mathbf{y}).$$

In order to calculate $Pr(\mathbf{V}>\mathbf{0}|\mathbf{y})$, we first obtain the conditional density of $\mathbf{V}|\mathbf{Y}=\mathbf{y}$ from the joint distribution (A.3). Using Lemma A.1 we have that

$$\mathbf{V}|\mathbf{Y}=\mathbf{y} \sim El\left(\boldsymbol{\theta} = D(\Sigma+D^2)^{-1}\mathbf{y}_*, \Omega = I - D(\Sigma+D^2)^{-1}D; g^{(m)}_{q(\mathbf{y}_*)}\right)$$

where $q(\mathbf{y}_*) = \mathbf{y}_*^T(\Sigma + D^2)^{-1}\mathbf{y}_*$. Consider the standardization

$$\mathbf{W} = (I - D(\Sigma + D^2)^{-1}D)^{-\frac{1}{2}}(\mathbf{V} - D(\Sigma + D^2)^{-1}\mathbf{y}_*).$$

Now $\mathbf{W}|\mathbf{Y} = \mathbf{y} \sim El\left(\mathbf{0}, I; g^{(m)}_{q(\mathbf{y}_*)}\right)$. Therefore,

$$\begin{aligned} Pr(\mathbf{V} > \mathbf{0}|\mathbf{y}) &= Pr\left(\mathbf{W} > -(I - D(\Sigma + D^2)^{-1}D)^{-\frac{1}{2}}D(\Sigma + D^2)^{-1}\mathbf{y}_*\right) \\ &= Pr\left(\mathbf{W} < (I - D(\Sigma + D^2)^{-1}D)^{-\frac{1}{2}}D(\Sigma + D^2)^{-1}\mathbf{y}_*\right) \\ &= F\left((I - D(\Sigma + D^2)^{-1}D)^{-\frac{1}{2}}D(\Sigma + D^2)^{-1}\mathbf{y}_*|\mathbf{0}, I; g^{(m)}_{q(\mathbf{y}_*)}\right). \end{aligned}$$

Hence the proof is complete. □

Lemma A.2

If $\mathbf{Y} \sim SN(\boldsymbol{\mu}, \Sigma, D)$, *then it has the moment generating function (11).*

Proof of Lemma A.2: Note that

$$M_{\mathbf{Y}}(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu}}M_{\mathbf{X}}(\mathbf{t}), \quad \text{where } \mathbf{X} \sim SN(\mathbf{0}, \Sigma, D).$$

Let $Q = (\Sigma + D^2)^{-1}$ and $B = I - DQD$ for notational convenience. Now,

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= 2^m \textstyle\int_{\mathbb{R}^m} |Q|^{1/2}(2\pi)^{-m/2}e^{-\frac{1}{2}\mathbf{x}^TQ\mathbf{x}+\mathbf{t}^T\mathbf{x}}\Phi_m\left(B^{-1/2}DQ\mathbf{x}\right)d\mathbf{x} \\ &= e^{\mathbf{t}^TQ^{-1}\mathbf{t}/2}2^m \textstyle\int_{\mathbb{R}^m} |Q|^{1/2}(2\pi)^{-m/2}e^{-\frac{1}{2}(\mathbf{x}-Q^{-1}\mathbf{t})^TQ(\mathbf{x}-Q^{-1}t)}\Phi_m\left(B^{-1/2}DQ\mathbf{x}\right)d\mathbf{x} \\ &= e^{\mathbf{t}^TQ^{-1}\mathbf{t}/2}2^m \textstyle\int_{\mathbb{R}^m} |Q|^{1/2}(2\pi)^{-m/2}e^{-\frac{1}{2}\mathbf{z}^TQ\mathbf{z}}\Phi_m\left\{B^{-1/2}DQ(\mathbf{z} + Q^{-1}\mathbf{t})\right\}d\mathbf{z} \\ &= e^{\mathbf{t}^TQ^{-1}\mathbf{t}/2}2^m\Phi_m(D\mathbf{t}) \end{aligned}$$

using the following result. □

Proposition 1

If $\mathbf{Z} \sim N_m(\mathbf{0}, \Sigma)$, *then*

$$E_{\mathbf{Z}}\left[\Phi_m(\mathbf{a} + G\mathbf{Z})\right] = \Phi_m\left\{(I + G\Sigma G^T)^{-1/2}\mathbf{a}\right\}.$$

The following Lemma obtains the moment generating function of scale mixture of skew normal distribution.

Lemma A.3

Let $\mathbf{X} \sim SN(\mathbf{0}, \Sigma, D)$ *and* $\mathbf{Y} = w^{-1/2}\mathbf{X}$ *given* $W = w$ *where* $W \sim \Gamma(\nu/2, \nu/2)$ *where the parameterization has mean 1. Then the moment generating function of the marginal distribution of* $\mathbf{Y}$ *is given by*

$$M_{\mathbf{Y}}(\mathbf{t}) = 2^m \int_0^\infty e^{\mathbf{t}^TQ^{-1}\mathbf{t}/(2w)}\Phi_m(Dw^{-1/2}\mathbf{t})dG(w),$$

where $G(w)$ *denote the cumulative distribution function of* $\Gamma(\nu/2, \nu/2)$.

Theorem A.1

Assume that

$$h(w,\mathbf{y},X)=\int_{\mathbb{R}^p}\prod_{i=1}^{n}g^{(1)}\left(\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{w}\right)d\pi_{\boldsymbol{\beta}}\leq M(\mathbf{y},X)w^{p/2},$$

where $M(\mathbf{y},X)$ is a constant free of w, δ and ν. Also assume that π_δ and π_ν are proper distributions. Then the posterior (16) is proper under the skew normal or skew t model if $n>p$.

Lemma A.4

Under the skew normal model

$$h(w,\mathbf{y},X)\leq M(\mathbf{y},X)w^{p/2}.$$

Proof of Lemma A.4: Here $g^{(1)}(u)=e^{-u/2}/(2\pi)^{1/2}$. Hence $\prod_{i=1}^{n}g^{(1)}(u_i)=(2\pi)^{-n/2}e^{-\frac{1}{2}\sum_{i=1}^{n}u_i}$. However, note that

$$\begin{aligned}\textstyle\sum_{i=1}^{n}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2 &= (\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})\\ &= (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(X^TX)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+\mathbf{y}^T(I-X(X^TX)^{-1}X^T)\mathbf{y},\end{aligned}$$

where $\hat{\boldsymbol{\beta}}=(X^TX)^{-1}X^T\mathbf{y}$. Let $H=I-X(X^TX)^{-1}X^T$ and $Q=X^TX$. By assumption on X we have that Q is a positive definite matrix. Also H is idempotent (and hence a non-negative definite) matrix. Let $\pi_{\boldsymbol{\beta}}=1$. Now we have,

$$\begin{aligned}h(w,\mathbf{y},X) &= \textstyle\int_{\mathbb{R}^p}\prod_{i=1}^{n}g^{(1)}\left(\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{w}\right)d\pi_{\boldsymbol{\beta}}\\ &= \textstyle\int_{\mathbb{R}^p}(2\pi)^{-n/2}e^{-\frac{1}{2w}\left\{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^TQ(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+\mathbf{y}^TH\mathbf{y}\right\}}d\boldsymbol{\beta}\\ &= \textstyle\frac{(2\pi)^{p/2}}{(2\pi)^{n/2}}w^{p/2}|Q|^{-1/2}e^{-\frac{1}{2w}\mathbf{y}^TH\mathbf{y}}\\ &= w^{p/2}\times M(\mathbf{y},X),\end{aligned}$$

since $\mathbf{y}^TH\mathbf{y}$ is a non-negative definite quadratic form. Hence the proof is complete. □

Lemma A.5

Under the skew t model

$$h(w,\mathbf{y},X)\leq M(\mathbf{y},X)w^{p/2}.$$

Proof of Lemma A.5: Let

$$J=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\,(\nu\pi)^{1/2}}.$$

Here

$$g^{(1)}(u;\nu)=J\left[1+\frac{u}{\nu}\right]^{-(\nu+1)/2}.$$

Now,

$$\begin{aligned}
\textstyle\prod_{i=1}^n g^{(1)}(u_i) &= J^n \left[\textstyle\prod_{i=1}^n \left\{1+\frac{u_i}{\nu}\right\}\right]^{-(\nu+1)/2} \\
&\leq J^n \left[1+\frac{\sum_{i=1}^n u_i}{\nu}\right]^{-(\nu+1)/2} \\
&= J^n \left[1+\frac{\sum_{i=1}^n (y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{\nu w}\right]^{-(\nu+1)/2} \\
&= J^n \left[1+\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T Q(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+\mathbf{y}^T H\mathbf{y}}{\nu w}\right]^{-(\nu+1)/2} \\
&\leq J^n \left[1+\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T Q(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\nu w}\right]^{-(\nu+1)/2},
\end{aligned}$$

where the first inequality follows by considering $\prod_{i=1}^n (1+u_i) \geq 1+\sum_{i=1}^n u_i$ for $u_i > 0, i = 1, \ldots, n$. Now we have,

$$\begin{aligned}
h(w, \mathbf{y}, X) &= \textstyle\int_{\mathbb{R}^p} \prod_{i=1}^n g^{(1)}\left(\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{w}\right) d\pi_{\boldsymbol{\beta}} \\
&\leq \textstyle\int_{\mathbb{R}^p} J^n \left[1+\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T Q(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\nu w}\right]^{-(\nu+1)/2} d\boldsymbol{\beta} \\
&= J^n \frac{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{p/2}}{\Gamma\left(\frac{\nu+p}{2}\right)} w^{p/2} |Q|^{-1/2} \\
&= w^{p/2}\, a(\nu)\, |Q|^{-1/2},
\end{aligned}$$

where

$$\begin{aligned}
a(\nu) &= \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{1/2}}\right]^n \frac{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{p/2}}{\Gamma\left(\frac{\nu+p}{2}\right)} \\
&\leq C
\end{aligned}$$

where C is a constant free of ν. The last in-equality follows by using the following bound for the gamma function, see for example Whittaker and Watson (1927, chapter 12),

$$\Gamma(z) = (2\pi)^{1/2} z^{z-1/2} e^{-z+b(z)}, \text{ for } z > 0$$

with $0 < b(z) < K/z$ for some positive constant K. Hence the proof is complete. □

Proof of Theorem A.1: We first consider the integral of the likelihood times the prior. Let

$$A = \int \cdots \int L(\boldsymbol{\beta}, \sigma^2, \delta, \nu; \mathbf{y}) d\pi_{\boldsymbol{\beta}} d\pi_{\sigma^2} d\pi_\delta d\pi_\nu.$$

In the following derivation the value of the constant C may not be the same in every line.

$$\begin{aligned}
A &= C \textstyle\int \cdots \int \frac{2^n}{(\sigma^2+\delta^2)^{-n/2}} \prod_{i=1}^n \left\{ g^{(1)}\left(\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{\sigma^2+\delta^2}\right) F\left(\frac{\delta}{\sigma}\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{\sqrt{\sigma^2+\delta^2}} | 0, 1, g_a^{(1)}\right)\right\} d\pi_{\boldsymbol{\beta}} d\pi_{\sigma^2} d\pi_\delta d\pi_\nu \\
&\leq C \textstyle\int \cdots \int (\sigma^2+\delta^2)^{-n/2} \prod_{i=1}^n \left\{ g^{(1)}\left(\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{\sigma^2+\delta^2}\right)\right\} d\pi_{\boldsymbol{\beta}} d\pi_{\sigma^2} d\pi_\delta d\pi_\nu \\
&\leq C \textstyle\int\int\int (\sigma^2+\delta^2)^{-n/2} \left[\int_{\mathbb{R}^p} \prod_{i=1}^n \left\{ g^{(1)}\left(\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{\sigma^2+\delta^2}\right)\right\} d\pi_{\boldsymbol{\beta}}\right] d\pi_{\sigma^2} d\pi_\delta d\pi_\nu \\
&\leq C \textstyle\int\int\int (\sigma^2+\delta^2)^{-n/2} (\sigma^2+\delta^2)^{p/2} d\pi_{\sigma^2} d\pi_\delta d\pi_\nu \\
&\leq C \textstyle\int\int\int (\sigma^2+\delta^2)^{-\frac{n-p}{2}} d\pi_{\sigma^2} d\pi_\delta d\pi_\nu \\
&\leq C \textstyle\int\int\int (\sigma^2)^{-\frac{n-p}{2}} d\pi_{\sigma^2} d\pi_\delta d\pi_\nu \\
&\leq C \textstyle\int\int \left[\int_0^\infty (\sigma^2)^{-\left(\frac{n-p}{2}+\kappa+1\right)} e^{-\kappa/\sigma^2} d\sigma^2\right] d\pi_\delta d\pi_\nu.
\end{aligned}$$

Now as $\kappa \to 0$ the innermost integral is finite if $n > p$. Also it is assumed that π_δ and π_ν are proper. Hence A is finite and the result follows. □

Proof of Theorem 2: The proof follows by using Lemma A.4, Lemma A.5 and Theorem A.1. □

Proof of Theorem 3: Follows from the last displayed inequality in the proof of Theorem A.1. □

REFERENCES

Adcock, C. J. (2002). Asset pricing and portfolio selection based on the multivariate skew-Student distribution. *Technical Report*, University of Sheffield.

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

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.

Arnold, B. C. and Beaver, R. J. (2000). Hidden Truncation Models. *Sankhyā*, A, 23–35.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society,* B, **61**, 579-602.

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

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: John Wiley & Sons.

Branco, M., Bolfarine, H. Iglesias, P. and Arellano-Valle, R. B. (2000). Bayesian analysis of calibration problem under elliptical distributions. *Journal of Statistical Planning and Inference*, **90**, 69–85.

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

Chen, M.-H., Dey, D. and Shao, Q-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, **94**, 1172–1186.

Chib, S., Tiwari, R.C. and Jammalamadaka, S.R. (1988). Bayes prediction in regressions with elliptical errors. *Journal of Econometrics,* **38**, 349–360.

DiCiccio, T. J., Kass, R. E., Raftery, A. Wasserman, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, **92**, 903–915.

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

Fernandez,C. and Steel,M.F.J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association,* **93**, 359–371 .

Gelfand, A.E. and Dey, D.K. (1994). Bayesian Model Choice - Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society,* B **56**, 501–514.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics,* **8**, 519–540.

Huang, O. and Litzenberger, R. H. (1988). *Foundations for Financial Economics.* New York: North Holland.

Jones, M. C. (2000). A skew t distribution. In *Recent Advances in Probability and Statistics; in Honor of T. Cacoullos* (eds C. A. Charalambides, M. V. Koutras and N. Balakrishnan), 269–78. London: Chapman and Hall.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā*, **32**, 419–430.

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

Osiewalski J. and Steel, M.F.J. (1993). Robust Bayesian inference in elliptical regression models. *Journal of Econometrics*, **57**, 345–363.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, B, **64**.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1996). Computation on Bayesian graphical models. In *Bayesian Statistics 5*, (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press, pp. 407–426.

Whittaker, E. T. and Watson, G. N. (1927). *A Course of Modern Analysis: An Introduction to the General Theory of Infinite Processes and of Analytic Functions.* Cambridge, U.K.: Cambridge University Press.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error term. *Journal of the American Statistical Association*, **71**, 400–405.

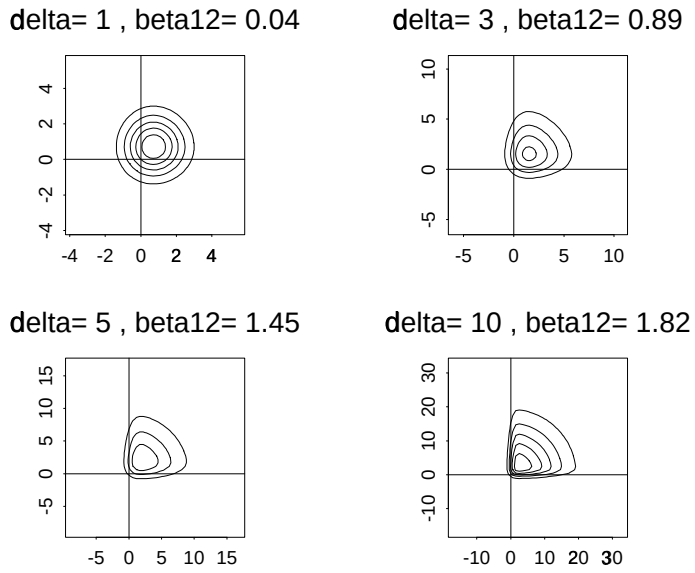

Figure 1: Contour plots of bivariate skew normal distributions.

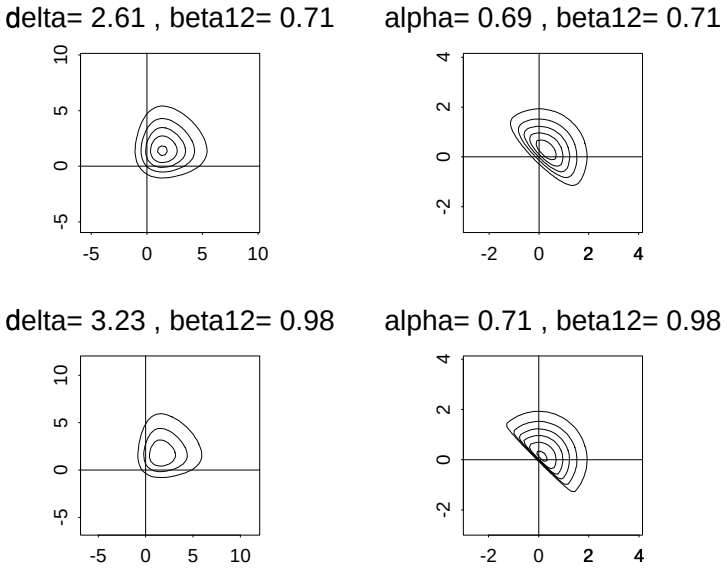

Figure 2: Contour plots of bivariate skew normal distributions.

Figure 3: Box plot of interview scores.

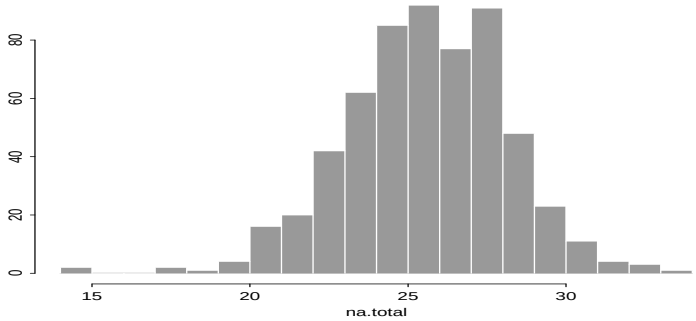

Figure 4: Box plot of interview scores.

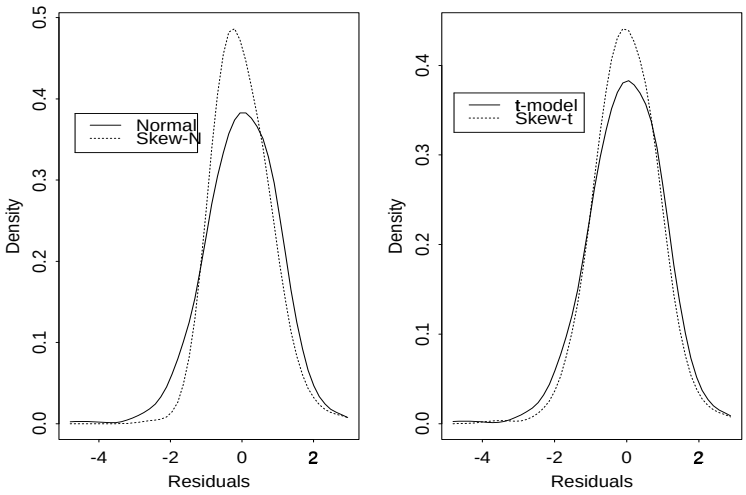

Figure 5: Kernel density estimates of the residuals under four regression models.

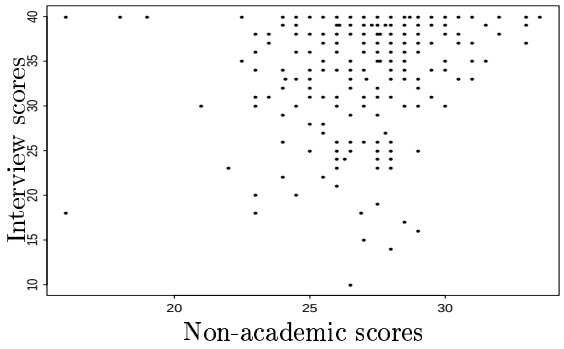

Figure 6: Scatter plot of the bivariate data used in model fitting.