A New Skew-Elliptical Distribution and Its Properties

Sujit K. Sahu, High S. Chai

Abstract

This article generalizes a multivariate skew-elliptical distribution and describes its many interesting properties. The univariate version of the new distribution is compared with two other currently used distributions. The use of the new distribution is illustrated with a real data example suitable for regression modelling. The new model provides a better model fit than its two rivals as evaluated by some suitable Bayesian model selection criteria.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/19

A NEW SKEW-ELLIPTICAL DISTRIBUTION AND ITS PROPERTIES

SUJIT K. SAHU[1] AND HIGH S. CHAI[2]

[1]School of Mathematics, S^{3}RI, University of Southampton, UK, S.K.Sahu@maths.soton.ac.uk.

[2]Mayo Clinic College of Medicine, Division of Cardiovascular Diseases, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA.

ABSTRACT

This article generalizes a multivariate skew-elliptical distribution and describes its many interesting properties. The univariate version of the new distribution is compared with two other currently used distributions. The use of the new distribution is illustrated with a real data example suitable for regression modeling. The new model provides a better model fit than its two rivals as evaluated by some suitable Bayesian model selection criteria.

Key Words: Bayesian inference; Gibbs sampler; kurtosis; skewness; skew elliptical distributions.

1 Introduction

Recently there has been renewed interest in the statistical literature towards robust statistical methods in order to represent features of the data as adequately as possible and reduce unrealistic assumptions. This remark is reflected in the substantial growth in the number of distributional families developed, studied and used for data modeling as alternatives to the normal theory statistics. See the edited volume by Genton (2004) for a snapshot of recent activities. Some families of distributions which allow for skewness and contain the normal distribution as a proper member or as a limiting case have played an important role in these developments. Among them are the skew-normal distribution (Azzalini, 1985, 1986), the multivariate skew-normal distribution (Azzalini and Dalla Valle, 1996), the two-piece normal distribution (John, 1982), the epsilon-skew-normal distribution (Mudholkar and Hutson, 2000), the skew-t distribution (Jones and Faddy, 2003), the generalized skew-t

distribution (Theodossiou, 1998), the two-piece t distribution (Fernandez and Steel, 1998), the skew-elliptical distribution (Sahu, Dey and Branco, 2003), and the generalized skew-elliptical distribution (Genton and Loperfido, 2002).

The term skew normal distribution was first introduced by Azzalini in 1985 as a natural extension of the normal density to accommodate asymmetry. A random variable Z is said to have a skew normal distribution with parameter $\lambda \in \Re$ if it has probability density function

$$f(z|\lambda) = 2\phi(z)\Phi(\lambda z), \qquad z \in \Re \tag{1}$$

where here and henceforth we denote the standard normal density and distribution function by $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. The above density is positively skewed when $\lambda > 0$, negatively skewed when $\lambda < 0$, and symmetric when $\lambda = 0$ (in which case it coincides with the standard normal distribution). Therefore it is reasonable to regard λ as the skewness parameter.

The above skew normal distribution can be physically justified by considering its genesis. For example, Arnold *et al.* (1993) derive it as the marginalization of a hidden truncated bivariate normal density; and Loperfido (2002) has shown it as emerging from selective reporting. The distinct genesis representations may also be useful for simplifying some computations such as the moments and random variate generation. Another attractive implication of the various genesis methods is that they can be fruitfully employed for extending the basic skew normal distribution to more general settings. Although this opens the way to the study of particular cases, this paper will only use a conditioning and truncation method described in Sahu *et al.* (2003).

This article extends the previous version of skew-elliptical (SE) distribution, introduced by Sahu *et al.* (2003). The extended class is distinct from the one obtained by Branco and Dey (2001) but contains the Sahu *et al.* (2003) family as a special case. Branco and Dey (2001) develop their multivariate SE distributions by conditioning on one suitable random variable being positive while Sahu *et al.* (2003) impose the non-negativity condition on the same number of random variables. Heuristically, we generalize their ideas by releasing the dimensionality restriction on the conditioned variables.

In Section 2 we derive the density function of this new SE distribution. The family is then used in Section 3 to define a class of univariate skew normal distributions, which is the main focus for the remainder of the paper. Central moments of the skew-normal distribution are obtained, along with a discussion of some related properties. Section 4 compares three distinct versions of univariate skew normal distributions. Linear regression models are developed using the new distribution in Section 5. A real data example is given in Section 6. The paper concludes with a few summary remarks in Section 7. An appendix contains the proofs of the theoretical results.

2 Derivation of the skew-elliptical distributions

The present section utilizes a general method for introducing skewness into any symmetric distributions and applies it on the elliptical distributions. To this end, consider two independent random vectors $\mathbf{U}$ and $\mathbf{V}$, both with unimodal and symmetric densities. Now a class of skew distributions can be generated via the following formulation

$$\mathbf{Z} = \mathbf{D}\mathbf{U} + \mathbf{V}, \qquad \mathbf{U} > \mathbf{0} \tag{2}$$

where $\mathbf{D}$ is a fixed matrix. For the univariate setting in which U and V are chosen to be independent and identically standard normal random variables, a simple convolution computation shows that $Z/\sqrt{D^2+1}$ indeed has a basic skew normal distribution (1) with $\lambda = D$. The paradigm (2) provides a very simple way of generalizing the basic skew normal density.

The replacement of normal variate in the development of model (1) by other statistical distributions has become quite popular. For example, Arnold and Beaver (2000, 2002) have substituted the normal component by a suitable heavy tail alternative to obtain the skew Cauchy density. A broader class of multidimensional models, hinted by Azzalini and Capitanio (1999), can be elicited if the normal distribution is replaced by an elliptical distribution. Adcock and Shutes (2005) use the exponential distribution as a skewing function to obtain a multivariate skew-normal distribution. Some other results along these lines can be found in Branco and Dey (2001) and Sahu *et al.* (2003). The probability distribution proposed in this section extends the previous version induced by Sahu *et al.* (2003). Before presenting the new skew elliptical distribution, it is useful to recall the definition of the elliptical distributions.

2.1 Elliptical distributions

The *elliptical distribution*, originally defined by Kelker (1970), represents a natural generalization of the concept of symmetry to the multivariate setting. A comprehensive review of the distribution can be found in Fang *et al.* (1990). A random vector $\mathbf{X}$ with values in $\Re^k$ has an elliptical distribution with location vector $\boldsymbol{\mu} \in \Re^k$ and scale matrix Σ if its density function is of the form

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma; g^{(k)}) = |\Sigma|^{-1/2} g^{(k)}[(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})] \tag{3}$$

for some density generator function defined by

$$g^{(k)}(u) = \frac{\Gamma(k/2)}{\pi^{k/2}} \frac{g(u;k)}{\int_0^\infty r^{k/2-1} g(r;k) dr}, \qquad u \geq 0$$

where $g(u;k)$ is a non-increasing function ensuring that the integral $\int_0^\infty r^{k/2-1} \; g(r;k)dr$ exists. For simplicity, Σ is assumed to be positive definite. In what follows the notation $El(\boldsymbol{\mu}, \Sigma; g^{(k)})$ will be used to describe the above probability distribution.

The choice of generator function $g^{(k)}(\cdot)$ will determine the distribution of X. Its flexibility enables the elliptical class to acknowledge many well-known symmetrical distributions as proper members, e.g. the multivariate normal, uniform, Student's t, exponential power, and Pearson type II distributions. These densities have a wide range of tail shapes, but the general specification of X being elliptically distributed does not imply either light or heavy tailed distribution. Hence, to some extent, it is admissible to consider (3) as a universal model for summarizing kurtosis of a symmetric data. The particular case of normal distribution $N_k(\boldsymbol{\mu}, \Sigma)$ is obtained by defining $g(u;k) = e^{-u/2}$. Note that the function $g(u;k)$ may depend on other parameters. As an example, the Student's t distribution is obtained by taking $g(\cdot) = [v+u]^{(-v+k)/2}$.

Elliptical distribution, however, imposes the restriction on symmetry, which does not facilitate the analysis of the effects of skewness. It is accepted that, in real applications, kurtosis and skewness are often observed characteristics of empirical data. Accordingly, statistics employed by assuming ellipticity are not always valid and can be of little value for summarizing the structure in a body of data. The ability to incorporate these pervasive features simultaneously is therefore an important practical consideration. Hence it seems reasonable and appropriate to acquire a skewed version of elliptical distribution so as to enable a trustworthy analysis of non-normal data.

2.2 Skew-elliptical distributions

The general procedure of skewing a symmetric unimodal distribution presented at the beginning of this section provides a simple yet powerful way for generating new distributions. The following theorem applies the previous results to develop a general class of skewed multivariate distributions. The proof of the theorem, provided in the Appendix, rests mainly on the properties of the elliptical distributions, see Chapter 2 of Fang *et al.* (1990).

Theorem 1 Let $\mathbf{U}$ and $\mathbf{V}$ be two independent random vectors distributed as

$$\mathbf{U} \sim El(\mathbf{0}, \mathbf{I}; g^{(p)}) \qquad \text{and} \qquad \mathbf{V} \sim El(\boldsymbol{\mu}, \Sigma; g^{(m)}).$$

Here $\mathbf{0}$ is the zero vector and $\mathbf{I}$ is the identity matrix. Let $\mathbf{Z}_{m\times 1} = \mathbf{D}_{m\times p}\mathbf{U}_{p\times 1} + \mathbf{V}_{m\times 1}$. The conditional density of $[\mathbf{Z}|\mathbf{U} > \mathbf{0}]$ will be of the form

$$h(\mathbf{z}|\boldsymbol{\mu}, \Sigma, \mathbf{D}; g^{(m)}) = 2^p f(\mathbf{z}|\boldsymbol{\mu}, \Sigma + \mathbf{D}\mathbf{D}^T; g^{(m)}) Pr(\mathbf{W} > \mathbf{0}|\mathbf{z}), \tag{4}$$

where $f(\cdot)$ is the elliptical density function as that in (3), and

$$\mathbf{W}|\mathbf{z} \sim El(\mathbf{D}^T[\Sigma + \mathbf{D}\mathbf{D}^T]^{-1}\mathbf{z}_*, \mathbf{I} - \mathbf{D}^T[\Sigma + \mathbf{D}\mathbf{D}^T]^{-1}\mathbf{D}; g^{(p)}_{q(\mathbf{z}_*)}),$$

where $q(\mathbf{z}_*) = \mathbf{z}_*^T[\Sigma + \mathbf{D}\mathbf{D}^T]^{-1}\mathbf{z}_*$, $\mathbf{z}_* = \mathbf{z} - \boldsymbol{\mu}$ and

$$g_a^{(p)}(u) = \frac{\Gamma(p/2)}{\pi^{p/2}} \frac{g(a+u; m+p)}{\int_0^\infty r^{p/2-1} g(a+r; m+p)dr}.$$

Non-singularity of the distributional parameter $\Sigma + DD^T$ is a prerequisite for ensuring the existence of the resulting density (4). The matrix $\mathbf{D}$, in a broad sense, controls the degree of asymmetry of the density via the probability function $Pr(\mathbf{W} > \mathbf{0}|\mathbf{z})$. Henceforth $\mathbf{D}$ will be interpreted as the skewness parameter and $Pr(\mathbf{W} > \mathbf{0}|\mathbf{z})$ as the skewing function. It is clear that the particular case $\mathbf{D} = \mathbf{0}$ corresponds to the one of the elliptical distribution $El(\boldsymbol{\mu}, \Sigma; g^{(m)})$. Consequently, the random vector $\mathbf{Y} = [\mathbf{Z} > \mathbf{0}]$ can reasonably be regarded as having an m-dimensional *skew-elliptical distribution*. For brevity, the symbols $SE(\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p}; g^{(m)})$ are employed to denote the sampling density in (4). Note that, in general, the quantities $\boldsymbol{\mu}$ and Σ are not the mean and the scale matrix of $\mathbf{Y}$ as the density may not be symmetric with respect to $\boldsymbol{\mu}$.

The use of an elliptical model in the development of (4) is motivated by its desirable property of including thin and thick tailed distributions as special cases. As a result, in addition to the obvious increased flexibility in skewness, the family $SE(\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p}; g^{(m)})$ should allow for a variety of tail thickness. In the case where $\mathbf{D} = \delta\mathbf{I}$, the proposed class closely parallels to the one given in Branco and Dey (2001). Moreover, it agrees with the skew elliptical densities mentioned in Sahu *et al.* (2003) when $\mathbf{D}$ is diagonal of order m. Therefore the present class includes the earlier version obtained by Sahu *et al.* (2003) as a special case. Another appealing feature of the skew elliptical in (4) is its coherence under marginalization operation, i.e. it has marginal distributions that still belong to the same family. This is essentially an implicit result in the genesis of the distribution. Although the skewing function $Pr(\mathbf{W} > \mathbf{0}|\mathbf{z})$ may prove to be hard to evaluate, it need not be computed for practical model fitting using the popular Markov chain Monte Carlo methods, see Section 5. Summing up, this new skew distribution should be valuable in modeling multivariate random phenomena which display both skewness and kurtosis.

3 Skew-normal distributions

As pointed out in the last section, construction (2) is a vigorous technical tool for transforming a symmetric distribution into a skewed one. Clearly, joint consideration of asymmetry and tail behavior can now be achieved by applying this method to a suitable fat or thin tailed distribution. From the inferential viewpoint it means that the resulting skew distribution is made up of two components, $\mathbf{DU}$ and $\mathbf{V}$ in the preceding notations. Skewness is driven only by a single vector $\mathbf{U}$ and its sensitivity is dependent on $\mathbf{D}$. Although it is not obvious in the context, operation (2) does have an effect on other distributional characteristics. The principal purpose of the current section is to examine how procedure (2) influences the shape of the skewed density. Since normal distribution has been the standard point of reference for many characteristic measurements, attention will be held on its skewed counterpart from this time onwards. After presenting the density function of the m-dimensional version, this section will focus on the general univariate case. Specifically, mathematical moments and some properties of the latter will be presented in streamlined form.

3.1 Multivariate skew-normal distributions

As an immediate use of Theorem 1, consider the particular case $g(u;m) = e^{-u/2}$. Now, since the generator function simplifies to $g^{(m)}(u) = (2\pi)^{-m/2}e^{-u/2}$ and $g^{(p)}_{q(\mathbf{z}_*)}(u)$ is free of $q(\mathbf{z}_*)$, it is straightforward to verify that the joint density of $\mathbf{Y} = [\mathbf{Z}|\mathbf{U} > \mathbf{0}]$ is of the form

$$h(\mathbf{y}|\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p}) = 2^p|\Sigma + \mathbf{DD}^T|^{-1/2}\phi_m([\Sigma + \mathbf{DD}^T]^{-1/2}(\mathbf{y} - \boldsymbol{\mu}))Pr(\mathbf{W} > \mathbf{0}|\mathbf{y}), \quad (5)$$

where ϕ_m is the multivariate normal density of $N_m(\mathbf{0}, \mathbf{I})$, and

$$\mathbf{W}|\mathbf{Y} = \mathbf{y} \sim N_p(\mathbf{D}^T[\Sigma + \mathbf{DD}^T]^{-1}(\mathbf{y} - \boldsymbol{\mu}), \mathbf{I}_{p\times p} - \mathbf{D}^T[\Sigma + \mathbf{DD}^T]^{-1}\mathbf{D}).$$

It follows that $\mathbf{Y}$ has a multivariate skew normal distribution, indicated henceforth by the notation $\mathbf{Y} \sim SN_m(\boldsymbol{\mu}, \Sigma, \mathbf{D}_{m\times p})$. As expected the original normal density is retrieved when $\mathbf{D} = \mathbf{0}$. Conversely, deviation of the parameter $\mathbf{D}$ from $\mathbf{0}$ measures the departure of the distribution from normality. Therefore the above family nests the normal distribution as a proper member and permits a continuous departure from normality to non-normality.

3.2 Univariate skew-normal distributions

3.2.1 Density function

Specifying $m = 1$ in (5), the matrix D becomes a column vector $\boldsymbol{\delta}^T = (\delta_1, ..., \delta_p) \in \Re^p$ and Σ reduces to a scalar σ^2. In this case, Y is a univariate skew normal variate with density function given by

$$h(y|\mu, \sigma^2, \boldsymbol{\delta}) = \frac{2^p}{\sqrt{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2}} \phi \left[\frac{y - \mu}{\sqrt{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2}} \right] Pr(\mathbf{W} > \mathbf{0}|y), \tag{6}$$

where

$$\mathbf{W}|Y = y \sim N_p \left(\frac{y - \mu}{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2} \boldsymbol{\delta}, \mathbf{I}_{p \times p} - \frac{1}{\sigma^2 + \delta_1^2 + \cdots + \delta_p^2} \boldsymbol{\delta}\boldsymbol{\delta}^T \right).$$

In what follows, (6) will be referred to the general form of the univariate skew normal distribution.

3.2.2 Moments

As mentioned previously computation of the skewing function $Pr(\mathbf{W} > \mathbf{0}|y)$ can be obstructive. As a consequence, direct evaluation of the moments of the general univariate skew normal distribution will not be straightforward. A convenient way of proceeding is the following one. According to the representation (2), $Y \sim SN(\mu, \sigma^2, \boldsymbol{\delta})$ is the upshot of a linear combination of independent normal and standard half normal random variables. That is

$$Y = \boldsymbol{\delta}^T \mathbf{U}_{p \times 1} + V \qquad\qquad \mathbf{U} > \mathbf{0}, \tag{7}$$

where V and $U_1, \cdots, U_p$ independent and $V \sim N(\mu, \sigma^2)$ and each U_i follows the standard half normal distribution. Using this fact as well as the properties of moment generating function, expressions for the mean and the central moments of orders two through four are explicitly evaluated as follows.

Result 1 Let $m_i(Y) == E[\{Y - E(Y)\}^i]$. The random variable Y has

$$E(Y) = \mu + (\delta_1 + \cdots + \delta_p)\sqrt{\frac{2}{\pi}}, \quad m_2(Y) = \sigma^2 + (\delta_1^2 + \cdots + \delta_p^2)\left(1 - \frac{2}{\pi}\right),$$

$$m_3(Y) = (\delta_1^3 + \cdots + \delta_p^3)\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi} - 1\right),$$

$$\begin{aligned} m_4(Y) &= 3\sigma^4 + (\delta_1^4 + \cdots + \delta_p^4)\left[3 - \tfrac{4}{\pi}\left(\tfrac{3}{\pi} + 1\right)\right] \\ &\quad + 6\left[(\delta_1^2 + \cdots + \delta_p^2)\left(1 - \tfrac{2}{\pi}\right)\sigma^2 + \delta_1^2 \cdots \delta_p^2\left(1 - \tfrac{2}{\pi}\right)^2\right]. \end{aligned}$$

The proof of this result is placed in the Appendix.

In order to illustrate the influence of the parameter $\boldsymbol{\delta}$, it is necessary to adopt some suitable measures of skewness and kurtosis. To this end, a natural choice is the two classical measurements defined as follows. The skewness and kurtosis measures of a random variable X are respectively defined as the third and fourth standardized central moments of X, i.e.

$$Sk(X) = \frac{E[\{X - E(X)\}^3]}{[Var(X)]^{3/2}}, \qquad \text{and} \qquad Ku(X) = \frac{E[\{X - E(X)\}^4]}{[Var(X)]^2} - 3.$$

Thus skewness and kurtosis of Y can be readily obtained from the moments reported in Result 1. Intuitively, symmetrical distributions have skewness measures equal to zero, positive values correspond to distributions skewed to the right and negative values to those skewed to the left. Kurtosis, on the other hand, measures the degree of flatness of a density. Intrinsically positive kurtosis indicates peaked center and negative one signifies flat center relative to the normal curve.

An elementary calculation demonstrates that the skewness approaches its supremum (infimum) as $\delta_i \to \infty(-\infty)$, $i = 1, \cdots, p$, with

$$\sup[Sk(Y)] = -\inf[Sk(Y)] = \sqrt{2}(4 - \pi)(\pi - 2)^{-3/2} \simeq 0.9953.$$

Similarly, the bounds of the kurtosis can be shown to be

$$0 \le Ku(Y) \le (3\pi^2 - 4\pi - 12)(\pi - 2)^{-2} - 3 \quad (\simeq 0.8692).$$

Therefore we conclude that, with other parameters fixed, $\boldsymbol{\delta}$ in (6) can only produce more central peakedness than those in the original distribution.

In addition to creating some savings in moment calculations, relation (7) leads to an efficient algorithm for computer generation of skew normal random samples. The method can be described as follows. First, sample a p-dimensional vector $\mathbf{U}$ from $N_p(\mathbf{0}, \mathbf{I})$ and a scalar V from $N(\mu, \sigma^2)$. Then a random number Y from density (6) is obtained by setting

$$Y = \boldsymbol{\delta}^T |\mathbf{U}_{p\times 1}| + V.$$

This construction avoids rejection of sampling. The role played by $\boldsymbol{\delta}$ will be further highlighted in the coming sections.

3.2.3 Some simple properties

Note that there is no closed form expression for the distribution function of Y. Here we list other basic properties of the general univariate skew normal distribution.

Property 1 The density (6) reduces properly to the $N(\mu, \sigma^2)$ density when $\boldsymbol{\delta} = \mathbf{0}$.

Property 2 Reversing the sign of $\boldsymbol{\delta}$ and μ in (6) yields the density of $-Y$, i.e. the distribution $SN(-\mu, \sigma^2, -\boldsymbol{\delta})$ is the reflection of the distribution of $SN(\mu, \sigma^2, \boldsymbol{\delta})$ about $y = 0$.

Property 3 The way in which $\boldsymbol{\delta}$ intervenes in the central moments implies that

$$Sk(Y|\sigma^2, -\boldsymbol{\delta}) = -Sk(Y|\sigma^2, \boldsymbol{\delta}) \text{ and } Ku(Y|\sigma^2, -\boldsymbol{\delta}) = Ku(Y|\sigma^2, \boldsymbol{\delta}).$$

Property 4 The parameter $\boldsymbol{\delta}$ regulates skewness, which is positive if $\wedge > 0$ and negative if $\wedge < 0$ where $\wedge = \sum_{i=1}^{p} \delta_i^3$. Clearly, symmetric distribution can be obtained by taking $\wedge = 0$.

Property 5 The skewness $Sk(Y)$ is an increasing function of δ_i while the kurtosis $Ku(Y)$ is an increasing function of $|\delta_i|$, $i = 1, \cdots, p$.

Property 6 Large $\boldsymbol{\delta}$ will have momentous impact on the spread on (6) as $Var(Y)$ grows without bound with the absolute value of δ_i, $i = 1, \cdots, p$.

Property 7 Since $\frac{d \log h(y)}{dy}$ is a decreasing function of y it follows that the density (6) is unimodal.

Property 8 The mode of $SN(\mu, \sigma^2, \boldsymbol{\delta})$ is at the right of μ when $\sum_{i=1}^{p} \delta_i^3 > 0$ and vice versa. Except for the symmetric cases, it is in general not possible to find the mode analytically.

3.3 Two specific cases of univariate skew-normal distributions

Generally speaking, a one-parameter distribution can model only one empirical characteristic while greater flexibility is necessarily accompanied by increasing complexity in probability distribution. Therefore, the choice of p in (6) should depend on the level of difficulty in modeling the distributional characteristics in a body of data. From a pragmatic perspective, normal distribution ($p = 0$) is often sufficient for reflecting the structure underlying a population distribution. Other selections of p can be useful for analyzing data with the presence of possible skewness or kurtosis. For ease of exposition, only two particular cases of (6) are examined extensively in the rest of this paper.

The case $p = 1$ is of special interest, since it coincides with the univariate skew normal distribution obtained by Sahu *et al.* (2003). After some straightforward computations, it follows that the density of Y is of the form

$$h(y|\mu, \sigma^2, \delta) = \frac{2}{\sqrt{\sigma^2 + \delta^2}} \phi \left[\frac{y - \mu}{\sqrt{\sigma^2 + \delta^2}} \right] \Phi \left[\frac{\delta}{\sigma} \frac{y - \mu}{\sqrt{\sigma^2 + \delta^2}} \right]. \tag{8}$$

We write $Y \sim \mathrm{SN}_{\mathrm{sdb}}(\mu, \sigma^2, \delta)$ for future reference. For this density the effect of increasing δ is to magnify both the dispersion and asymmetry of the distribution. See Sahu *et al.* (2003) for various possible shapes of this distribution.

Considering $p = 2$, it is straightforward to verify that

$$h\left(y|\mu, \sigma^2, \begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}\right) = \frac{4}{\sqrt{\sigma^2 + \delta_1^2 + \delta_2^2}} \phi \left[\frac{y - \mu}{\sqrt{\sigma^2 + \delta_1^2 + \delta_2^2}} \right] F(\mathbf{0}), \tag{9}$$

where F stands for the cumulative density function of the bivariate normal distribution

$$N_2 \left(-\frac{y - \mu}{\sigma^2 + \delta_1^2 + \delta_2^2} \begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}, \frac{1}{\sigma^2 + \delta_1^2 + \delta_2^2} \begin{pmatrix} \sigma^2 + \delta_2^2 & -\delta_1\delta_2 \\ -\delta_1\delta_2 & \sigma^2 + \delta_1^2 \end{pmatrix} \right).$$

Henceforth, we shall denote this distribution by $\mathrm{SN}_{\mathrm{new}}(\mu, \sigma^2, \binom{\delta_1}{\delta_2})$. Examples of the densities for different combinations of values for δ_1 and δ_2 are presented in Figure 1. Observe that the graphs plotted below the diagonal are duplications of those above. This is a consequence of the exchangeability property of δ_1 and δ_2.

Besides the discrepancy in the level of algebraic complications, densities (8) and (9) differ in the sense that skewness of the former is driven by differences in $(\delta_1, -\delta_2)$, while the latter by a single parameter δ. Trivially, density (9) reduces to the one in (8) when $\delta_1 = 0$ or when $\delta_2 = 0$. As it might be expected, the major improvement involving the presence of an additional parameter is the flexibility in kurtosis variation. The three-parameter density (8) imposes some restraints on the kurtosis as soon as the skewness is fixed. In contrast, a broad range of the kurtosis of (9) can be covered by appropriate choices of δ_1 and δ_2 for any degree of skewness. To gain more insight in the impact of (2) on the kurtosis, consider Figure 2 which provides plots of (9) when δ_1 is set to equal $-\delta_2$ with variance of Y appointed at unity. All graphs in the diagram illustrate greater peakedness around the center as compared to the normal density.

4 Graphical comparisons

The specific aim of the current section is to compare three different variants of skew normal distributions by means of some graphical plots. We shall compare these distributions with a popular skew normal distribution called the two-piece skew normal distribution. This distribution has been studied by many authors, including Gibbons and Mylroie (1973), John (1982) and Kimber (1985). See also Fernandez and Steel (1998) for a generalization. The density function of the two piece skew normal distribution is given by:

$$h(y|\mu,\sigma^2,\delta) = \frac{2}{\sigma(\delta+1/\delta)}\left(\phi\left[\frac{y^*}{\sigma\delta}\right] I(y^* \geq 0) + \phi\left[\frac{\delta(y^*)}{\sigma}\right] I(y^* < 0)\right), \qquad (10)$$

where $y^* = y - \mu$ and I is an indicator function with $I(Q) = 1$ if Q is true and equals 0 otherwise. For convenience in notation, we say that the random variable Y is of the class $\mathrm{SN}_{\mathrm{tpn}}(\mu, \sigma^2, \delta)$ henceforth. The parameter $\delta \in (0, \infty)$ controls the allocation of probability mass to each side of the mode. It can be shown that $Pr(Y \geq \mu)/Pr(Y < \mu) = \delta^2$. The normal distribution is a special case ($\delta = 1$) and the half normal distribution is a limiting case. The mode of the distribution (10) is retained at μ for any value of δ. See Fernandez and Steel (1998) for the first four moments of this distribution.

The three distributions we compare are:

1. $\mathrm{SN}_{\mathrm{sdb}}(\mu, \sigma^2, \delta)$: the Sahu *et al.* (2003) distribution displayed in (8).

2. $\mathrm{SN}_{\mathrm{new}}(\mu, \sigma^2, \binom{\delta_1}{\delta_2})$: the new distribution specified by (9).

3. $\mathrm{SN}_{\mathrm{tpn}}(\mu, \sigma^2, \delta)$: the two piece skew normal distribution with density (10).

The disparities of these skewed models in distributional structure may be illustrated effectively by drawing their densities in the same diagram, in which they admit the same amount of skewness. Yet, in order to have a fair comparison, it is necessary to require common mean and variance across the distributions. Figure 3 pictures the densities for a selection of values of mean, variance and skewness. Note that, unlike $\mathrm{SN}_{\mathrm{sdb}}$ and $\mathrm{SN}_{\mathrm{tpn}}$ distributions, there is a series of $\mathrm{SN}_{\mathrm{new}}$ densities complying with the imposition. In spite of that, only two of these densities are plotted in the figure so as to enhance visualization. The supplementary flexibilities of $\mathrm{SN}_{\mathrm{new}}$ over $\mathrm{SN}_{\mathrm{sdb}}$ in terms of height and tails controls should be apparent from the graphic display. As one can anticipate, the behavior of $\mathrm{SN}_{\mathrm{tpn}}$ exhibits a manifest difference from the other densities. This is because $\mathrm{SN}_{\mathrm{tpn}}$ has a much lighter tail and its distinct style of descending from the mode. Similar figures could be constructed for other combinations of characteristic measures.

It is of interest to acquire a visual summary of the relationship between the skewing parameter and the degree of asymmetry of the distributions. On this basis, Figures 4 and 5 delineate the level of skewness measure $Sk(Y)$ as the parameter $\boldsymbol{\delta}$ changes for representative values of σ^2. Although Figure 5 only copes with positive values of δ_1, analogous plots for the negative domain can be obtained by a simple 180° rotation. It appears from Figure 4 that $\mathrm{SN}_{\mathrm{tpn}}$ is very sensitive to variations in δ. In fact, a wide range of its skewness can be covered for δ varying in $(0.2, 2)$. The rate with which $\mathrm{SN}_{\mathrm{sdb}}$ diverges from symmetry as $|\delta|$ increases is substantially influenced by the parameter σ^2. Smaller values of σ^2 will be associated with greater steeply sloped skewness curves and vice versa. Evidently the parameter has a similar impact on $\mathrm{SN}_{\mathrm{new}}$, as shown in Figure 5. It is important to note that a rise in $|\delta_1^3 + \delta_2^3|$ does not necessarily mean a greater asymmetry, especially when $\mathrm{SN}_{\mathrm{new}}$ is already highly skewed.

Figure 6 gives an additional insight into the achievable kurtosis $Ku(Y)$ as a function of skewness $Sk(Y)$. For $\mathrm{SN}_{\mathrm{tpn}}$ and $\mathrm{SN}_{\mathrm{sdb}}$ distributions, greater asymmetry will inevitably result in larger values for the kurtosis. Similarly, smaller magnitude of skewness will correspond to less central peakedness. Nonetheless the two distributions depart from normality in a quite different manner. The gap between the dotted and solid lines shows the advantage of $\mathrm{SN}_{\mathrm{new}}$ over $\mathrm{SN}_{\mathrm{sdb}}$ in the form of kurtosis variation. It is seen that the advantage is most perceptible for near normal cases and gradually melted away as asymmetry increases.

5 Application in linear regression models

In this section we illustrate the skew normal distribution $\mathrm{SN}_{\mathrm{new}}$ defined in (9) to model the conditional distribution of a response variable given the covariates. Suppose the observed data y_i, $i = 1, \cdots, n$, are independent samples generated from the regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{11}$$

where $\mathbf{x}_i \in \Re^k$ are the values of k explanatory variables for the i-th observation and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)^T$ is a vector of regression parameter associated with these variables and residuals, ε_i, are independently, identically distributed random variables having the distribution:

$$\varepsilon_i \sim \mathrm{SN}_{\mathrm{new}} \left(0, \sigma^2, \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \right).$$

An important implication of this assumption is that the conditional mean of $Y_i|\mathbf{x}_i$ will be equal to $\mathbf{x}_i^T \boldsymbol{\beta}$ plus the average value of the error distribution. To parallel the conventional regression analysis, the error distribution can be forced to take mean zero by suitably adjusting β_1 which is the intercept parameter of the regression model, see Section 6.1 for more details.

5.1 The posterior distribution

Since the observations are assumed to be independent given $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, the likelihood function of the model parameters is obtained as the product of the individual density of the observable y_i yielding

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}|\mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} h\left(y_i|\boldsymbol{\beta}, \sigma^2, \begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}, \mathbf{x}\right),$$

where $h(\cdot)$ is stated in (9). Henceforth in this section, we condition on $\mathbf{x}_i$ without explicit mention. Presently, prior distribution for the model parameters is needed to complete the specification of a Bayesian model. In our investigation, components of both $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are assigned independent normal prior distributions, and a gamma distribution $G(\nu, \nu)$ (having mean 1) with a small positive choice of ν is used as prior for the precision $\tau = 1/\sigma^2$. More formally, the adopted joint prior distribution is given by

$$p(\boldsymbol{\beta}, \tau, \boldsymbol{\delta}) = p(\boldsymbol{\beta}) \times p(\tau) \times p(\boldsymbol{\delta}) = N_k(\tilde{\boldsymbol{\beta}}, \Omega) \times G(\nu, \nu) \times N_2(\mathbf{0}, \Psi)$$

where Ω and Ψ are diagonal matrices, $\tilde{\boldsymbol{\beta}} = (\bar{y}, 0, \cdots, 0)$ ($\bar{y}$ is the sample mean), and $\nu = 0.001$. We choose the diagonal elements of Ω to be 10^4 and the diagonal elements of Ψ to be 100. See Sahu *et al.* (2003) for a detailed discussion regarding the choice and sensitivity of the prior distributions.

The joint posterior distribution of $\boldsymbol{\beta}$, τ and $\boldsymbol{\delta}$ is simply proportional to the likelihood function times the joint prior distribution

$$p(\boldsymbol{\beta}, \tau, \delta_1, \delta_2|\mathbf{y}) \propto \prod_{i=1}^{n} h\left(y_i|\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2, \begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}\right) \times p(\boldsymbol{\beta}, \tau, \boldsymbol{\delta}). \tag{12}$$

Due to the complexity of the likelihood function, it is not possible to evaluate the marginal posterior distributions of the model parameters by analytical means. Hence, we resort to Gibbs sampling. The necessary conditional distributions for use of the Gibbs sampler are placed in the appendix.

Before proceeding with the examples, there is still a technical issue needing to be addressed. The roles played by the two skewness parameters in $\mathrm{SN}_{\mathrm{new}}$ are exchangeable, thus allowing their Gibbs realizations to travel from one target distribution to the other. This sort of behavior will cause identifiability problems in determining the true marginal distributions of δ_1 and δ_2. Bimodality will, not surprisingly, be a typical characteristic in the estimated posterior densities. As a consequence, point and interval estimations based upon such marginal distributions can be rather misleading about the actual distributional structures. To resolve this we adopt the following simple strategy. When the estimated marginal posterior distributions are bimodal with relatively negligible probability mass in between the modes we recommend using the following steps:

1. At the jth cycle of the Gibbs sampler, re-arrange the MCMC simulated values by setting

$$\delta_1^{(j)} = \max(\delta_1^{(j)}, \delta_2^{(j)}) \text{ and } \delta_2^{(j)} = \min(\delta_1^{(j)}, \delta_2^{(j)}),$$

2. Carry out all marginal calculations using the resulting samples.

This re-labeling proposal should be exercised prudently as it can have an adverse effect in other situations. For example, it is quite plausible that we might have skewness parameters with identical true value. Two unimodal marginal distributions (estimated using the original Gibbs output) will then be encountered. These should themselves provide a good approximation to the underlying distributions, thus there is no benefit in using the proposed scheme. In the case where there are two distinct but close modes in the original marginal distributions, the above approach will inevitably lead to densities with obvious truncations. So, neither the original nor the re-estimated marginal distributions represent the actual distributions. Use of the joint distribution to make probability statements will be more sensible in this situation. However, visual inspection of the original marginal distributions is recommended before choosing an appropriate method of output analysis.

6 A REAL DATA EXAMPLE

The particular data set that we consider here concerns admission to a Welsh medical institution in 1996 first reported in Sahu *et al.* (2003). Non-academic scores for home applicants meeting the school academic criteria were recorded after reading the corresponding Universities & Colleges Admissions Service (UCAS) application forms. (UCAS is the central organization that processes applications for full-time undergraduate courses at British universities and colleges.) Candidates who met the non-academic standards would be screened in the next level of selection process. The objective of the investigation is to determine the group of students that is least likely to be successful in this stage of their application. Accordingly, our response variable Y is the non-academic scores of $n = 777$ individuals, and the covariates of interest include: number of GCSE A grades x_2, race (white or non-white) x_3, age in years x_4, and predicted/achieved A Level examination points x_5. Interaction between number of GCSE A grades and age $x_6 = x_2 x_4$ is also embraced in the analysis, since it is a significant predictor in classical normal regression. Note that GCSE, stands for General Certificate of Secondary Education, is a national school leaving examination in Britain. As we shall see below there is considerable amount of skewness present in the response variable.

A total of four sampling models have been fitted: SN_{new} in (9), SN_{sdb} in (8), SN_{tpn} in (10) and the usual normal model. Legitimate comparison is elicited by allocating $\boldsymbol{\beta}$ and τ in the latter models the same prior distributions as those specified in Section 5. The remaining parameter δ is given the $N(0, 100)$ prior under the SN_{sdb} model whilst a-priori $\delta \sim N(0, 100)I(\delta > 0)$ is specified for the SN_{tpn} model. The Gibbs sampler outlined earlier has been executed by using the `WinBUGS` software, (Spiegelhalter *et al.*, 1996). Inferences are based on $200,000$ sequential version of Gibbs realizations, following a burn-in period of $10,000$ iterations to mitigate the impact of starting points.

6.1 Results

Table 1 reports the parameter estimates for all four models under consideration. Inspection of the table indicates little alteration in the estimates of β_3 and β_5, but inferences on β_2, β_4 and β_6 are noticeably affected by allowing for skewness. A closer examination reveals that the posterior mean of the latter parameters are very close to zero under the skew normal models. In other words, the covariate effects are attenuated by assuming skewed models. Further

	β_1	β_2	β_3	β_4	β_5	β_6	σ^2	δ or δ_1	δ_2
Normal	25.1	1.51	-0.91	0.40	0.05	-0.05	9.36	–	–
	(0.123)	(0.377)	(0.278)	(0.106)	(0.017)	(0.020)	(0.478)		
SN_{sdb}	28.2	1.07	-0.87	0.29	0.05	-0.03	3.71	-3.90	–
	(0.217)	(0.387)	(0.260)	(0.111)	(0.016)	(0.020)	(0.550)	(0.253)	
SN_{tpn}	26.4	1.04	-0.94	0.29	0.05	-0.03	8.21	0.75	–
	(0.260)	(0.388)	(0.261)	(0.111)	(0.016)	(0.021)	(0.487)	(0.039)	
SN_{new}	26.92	0.93	-0.89	0.27	0.05	-0.03	1.16	-4.20	1.88
	(0.793)	(0.403)	(0.253)	(0.110)	(0.017)	(0.021)	(1.184)	(0.327)	(1.140)

Table 1: Parameter estimates and the associated standard deviations (given in parentheses) for the non-academic scores example.

insight into the behavior of these parameters is obtained through graphical representations of their marginal densities in Figure 7. As shown in the diagrams, the marginal posterior distributions based on SN_{sdb} and SN_{tpn} modeling are remarkably cohesive. Interestingly, there seems to be evidence of association between sampling model flexibility and marginal locality (simpler sampling model possesses marginal distributions that are farther from zero). One concern in the plots is the effect of the interaction variable on the analysis. The skewed sampling assumptions induce β_6 to have substantial posterior mass around zero, thus lessening the interaction variable's momentousness in predicting the non-academic scores. Therefore, according to the skewed normal models, there is an improvement in regression additivity in the sense that the main covariates are emphasized relative to the interaction.

On the basis of the 95% credible intervals, it appears that modeling using different versions of error distribution can have considerable influences on the posterior of β_1. This is not surprising because β_1 is not the true regression intercept in the skew normal cases. We can express the actual intercept parameter α as: $\alpha = \beta_1 + \delta\sqrt{2/\pi}$ for SN_{sdb}; $\alpha = \beta_1 + \sigma(\delta - 1/\delta)\sqrt{2/\pi}$ for SN_{tpn}; and $\alpha = \beta_1 + (\delta_1 + \delta_2)\sqrt{2/\pi}$ for SN_{new}. Thus a meaningful location comparison should be obtained via α instead of β_1. Markov chain simulations of the intercept are readily computed from the existing Gibbs output. The resulting α estimates for $\mathrm{SN}_{\mathrm{sdb}}$, $\mathrm{SN}_{\mathrm{tpn}}$ and $\mathrm{SN}_{\mathrm{new}}$, together with their estimated standard deviations in parentheses, are given by 25.1 (0.122), 25.0 (0.126), and 25.1 (0.122) respectively. As expected, these values are in good agreement and are consistent with the findings in the normal model.

Consider now the inferences on the shape parameters: σ^2 and $\boldsymbol{\delta}$. Table 1 shows notably different estimates of σ^2 under the normal model as compared to the skewed models. This is justifiable since the parameter has dissimilar interpretations for all these models. Variability of the data is represented solely by σ^2 in the normal case, but non-zero skewness parameter(s) also share part of the variability in the skew normal cases. This explains the differences in the estimates. Posterior means of δ for both $\mathrm{SN}_{\mathrm{tpn}}$ and $\mathrm{SN}_{\mathrm{sdb}}$ show that moderate right skewness is present in the data. Statistical significance of the parameter under the two models reinforces the fact that normal family would be unsuitable for modeling the original non-academic scores. The reported estimates of δ_1 and δ_2 are also significant and lead to similar conclusion.

6.2 Model comparisons

To assess model adequacy, Figure 8 displays the data histogram with superimposed posterior predictive densities under each of the four models. All skewed models seem to provide an adequate fit to the non-academic scores, with the predictive distribution from $\mathrm{SN}_{\mathrm{new}}$ most closely resembles the histogram. Observe that the predictive distribution under the normal model needs to be shifted to the left in order to account for the skewness in the data. This has an adverse effect on the model ability in capturing the peak of the histogram. A formal model comparison can be conducted through the use of Bayes factors. We compute the criterion by exercising the methods advocated by Meng and Wong (1996). Table 2 lists the resulting Bayes factors. The upshots indicate a dramatic improvement in the skew normal fits over the normal fit. In addition, $\mathrm{SN}_{\mathrm{new}}$ is substantially better than $\mathrm{SN}_{\mathrm{sdb}}$, which is in turn definitely preferable to $\mathrm{SN}_{\mathrm{tpn}}$. Hence, the Bayes factor approach selects $\mathrm{SN}_{\mathrm{new}}$ as the best model for the empirical data.

	SN_{new}	SN_{sdb}	SN_{tpn}	Normal
SN_{new}	1	4.76	5.92E3	4.87E8
SN_{sdb}	–	1	1.24E3	1.02E8
SN_{tpn}	–	–	1	8.23E4
Normal	–	–	–	1

Table 2: Bayes factors based on the Laplace-bridge method for non-academic scores data. Entry (i, j) indicates the evidence in favor of model i versus model j. (Note: $\Re E \aleph \equiv \Re \times 10^{\aleph}$.)

6.3 Conclusions

The analysis based on our best model SN_{new} suggests that non-academic scores are strongly related to number of GCSE A grades, race, age and number of predicted/achieved A Level points. Individual scores improve with GCSE results at approximately 0.93 credit for each A grade. Good non-academic outcomes are more prevalent among white students, who have 0.89 higher scores than non-white candidates generally. Age of the applicants also have a positive impact on the non-academic totals. The relative increment is about 0.27 unit per age year. As for the GCSE results, number of predicted/achieved A level points is positively related to the non-academic outcome. However, the magnitude of influence is much smaller, approaching 0.05 score for each A level point gained. The analysis indicates no evidence of association between the response and the interaction effect, contradicting the upshot under the normal regression model. Putting these results together, we conclude that young non-white students with unfavorable GCSE and A level outcomes are those most probable to achieve inferior non-academic scores.

7 Discussion

There are many possible ways of generalizing the skew elliptical distribution (4). Some suggestions are: (i) From representation (2), it may be claimed that skewness is instigated by some unobserved additive random effects $\mathbf{U}$ which were truncated at a specific threshold. This suggests that further flexibility should be annexed to the model (4) by adopting a more general threshold or permitting broader style of truncation on $\mathbf{U}$. (ii) The random variables $\mathbf{U}$ and $\mathbf{V}$ used in the development of elliptical distribution were assumed to have come from the same standard family. Allowing a combination of assorted distributions will result in new classes of skewed distributions. (iii) A natural way of extending (4) is to utilize a comprehensive transformation mechanism admitting the representation: $\mathbf{DU}+\mathbf{BV}$. Obviously, a joint density for $\mathbf{U}$ and $\mathbf{V}$ can be used instead, for the sake of releasing the independence assumption, which in turn will lead to extra level of generalization.

In this paper a new class of univariate skew normal distributions is obtained by using simple transformation and conditioning. The family represents a mathematically tractable extension of the normal density, with the addition of a vector of parameters $\boldsymbol{\delta}$ to regulate distributional shape. Our focus in this paper has been concentrated on the scalar and the 2 dimensional $\boldsymbol{\delta}$ cases. We find the latter case quite appealing for some of its attractive features. It contains the normal distribution by strict inclusion, thus allowing a smooth transition from normality to non-normality. It admits the Sahu *et al.* (2003) skew normal density (equivalent to the scalar $\boldsymbol{\delta}$ case) as a proper member, but possesses an extra parameter to account for kurtosis. It is a flexible unimodal density that is able to reflect practical values of skewness and some levels of non-normal peakedness. Therefore, the proposed four-parameter distribution is potentially useful for data modeling, statistical analysis and robustness studies of normal theory methods. We have illustrated this with an example in linear regression.

APPENDIX

Proof of Theorem 1: To derive (4), we need the following well-known results, see Fang *et al.* (1990). Suppose that $\mathbf{X} \sim El(\boldsymbol{\mu}, \Sigma; g^{(n)})$. Now partition $\mathbf{X}$ into $\mathbf{X}^T = (\mathbf{X}_{(1)}^T, \mathbf{X}_{(2)}^T)$ of dimensions m and $n-m$ respectively, with the corresponding partitions of $\boldsymbol{\mu}$ and Σ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Lemma 1 (Theorem 2.16 of Fang *et al.*, 1990.) If $\mathbf{B}$ is a non-singular $n \times m$ matrix and $\mathbf{v}$ is an $m \times 1$ vector, then

$$\mathbf{v} + \mathbf{B}^T\mathbf{X} \sim El(\mathbf{v} + \mathbf{B}^T\boldsymbol{\mu}, \mathbf{B}^T\Sigma\mathbf{B}; g^{(m)}).$$

Lemma 2 (Corollary of Fang *et al.*, 1990: page 43.) The marginal distributions of $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ are given by:

$$\mathbf{X}_{(1)} \sim El(\boldsymbol{\mu}_{(1)}, \Sigma_{11}; g^{(m)}), \quad \mathbf{X}_{(2)} \sim El(\boldsymbol{\mu}_{(2)}, \Sigma_{22}; g^{(n-m)}).$$

Lemma 3 (Theorem 2.18 of Fang *et al.*, 1990.) The conditional distribution $\mathbf{X}_{(1)}|\mathbf{X}_{(2)}$ is given by

$$\mathbf{X}_{(1)}|\mathbf{X}_{(2)} = \mathbf{x}_{(2)} \sim El(\boldsymbol{\mu}_{1.2}, \Sigma_{11.2}; g^{(m)}_{q(\mathbf{x}_{(2)})})$$

where

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}), \quad \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$
$$q(\mathbf{x}_{(2)}) = (\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)})^T\Sigma_{22}^{-1}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)}), \quad g_a^{(m)}(u) = \frac{\Gamma(m/2)}{\pi^{m/2}} \frac{g(a+u; m+n)}{\int_0^\infty r^{m/2-1} g(a+r; m+n)dr}.$$

An alternative and convenient expression for (2) is the following

$$\begin{pmatrix} \mathbf{Z}_{m\times1} \\ \mathbf{W}_{p\times1} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{D}_{m\times p} \\ \mathbf{0}_{p\times m} & \mathbf{I}_{p\times p} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{m\times1} \\ \mathbf{U}_{p\times1} \end{pmatrix},$$

from which the probability density function (4) can be obtained by computing the conditional density $\mathbf{Z}|\mathbf{W} > \mathbf{0}$.

It can be easily verified from Lemma 1 that

$$\begin{pmatrix} \mathbf{Z}_{m\times1} \\ \mathbf{W}_{p\times1} \end{pmatrix} \sim El\left(\begin{pmatrix} \boldsymbol{\mu}_{m\times1} \\ \mathbf{0}_{p\times1} \end{pmatrix}, \begin{pmatrix} \Sigma_{m\times m} + \mathbf{DD}^T_{m\times m} & \mathbf{D}_{m\times p} \\ \mathbf{D}^T_{p\times m} & \mathbf{I}_{p\times p} \end{pmatrix}; g^{(m+p)} \right).$$

It follows easily using Lemma 2 that

$$\mathbf{Z} \sim El(\boldsymbol{\mu}, \Sigma + \mathbf{DD}^T; g^{(m)}),$$
$$\mathbf{W} \sim El(\mathbf{0}, \mathbf{I}; g^{(p)}).$$

Symmetry of the elliptical distribution and the Bayes theorem implies that

$$h(\mathbf{z}|\mathbf{W} > \mathbf{0}) = 2^p f(\mathbf{z}|\boldsymbol{\mu}, \Sigma + (\mathbf{DD}^T); g^{(m)}) Pr(\mathbf{W} > \mathbf{0}|\mathbf{z}).$$

The proof is completed by specifying the conditional density of $\mathbf{W}|\mathbf{Z}$ using Lemma 1:

$$\mathbf{W}|\mathbf{Z} = \mathbf{z} \sim El(\mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{z}_*, \mathbf{I}_{p\times p} - \mathbf{D}^T(\Sigma + \mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}; g^{(p)}_{q(\mathbf{z}_*)}),$$

where $\mathbf{z}_* = \mathbf{z} - \boldsymbol{\mu}$, $q(\mathbf{z}_*) = \mathbf{z}_*^T[\Sigma + \mathbf{D}\mathbf{D}^T]^{-1}\mathbf{z}_*$, and

$$g^{(p)}_a(u) = \frac{\Gamma(p/2)}{\pi^{p/2}} \frac{g(a+u; m+p)}{\int_0^\infty r^{p/2-1} g(a+r; m+p)dr}.$$

□

Proof of Result 1: The derivation of the above equations is straightforward but lengthy, the basic steps are presented as follows. For convenience, we provide the details for the case $p = 2$, the proof for the general case is similar. Now, as a result of (7), the moment generating function of Y can be written as

$$M_Y(t) = M_{Z_1}(t) M_{Z_2}(t) M_V(t).$$

Here $Z_i = \delta_i U_i$, $i = 1, 2$. In general, $M_Y(t)$ has no closed form expression since $M_{Z_1}(t)$ and $M_{Z_2}(t)$ do not lend themselves to explicit computation. Nevertheless, the above formulation can still be used as an indirect tool to obtain the moments.

Together with the fact $\frac{d^r}{dt^r} M_X(t)\Big|_{t=0} = E(X^r)$, it is easy to check that the central moments of Y satisfy the relationships

$$\begin{aligned}
E(Y) &= E(Z_1) + E(Z_2) + E(V), \\
Var(Y) &= Var(Z_1) + Var(Z_2) + Var(V), \\
m_3(Y) &= m_3(Z_1) + m_3(Z_2) + m_3(V), \\
m_4(Y) &= m_4(Z_1) + m_4(Z_2) + m_4(V) + \\
&\qquad 6[Var(Z_1)Var(Z_2) + Var(Z_1)Var(V) + Var(Z_2)Var(V)],
\end{aligned}$$

where $m_i(X) = E[\{X - E(X)\}^i]$. These results essentially reduce the problem to the evaluation of moments of normal and standard half normal distributions.

The r-th non-central moments of Z_i, $i = 1, 2$, is

$$E(Z_i^r) = \begin{cases} \delta^r \frac{1}{\pi} 2^{r/2} \left(\frac{r-1}{2}\right)! & \text{when } r \text{ is odd,} \\ \delta^r 2^{-(r-2)/2} 1 \cdot 3 \cdot 5 \cdots (r-1) & \text{when } r \text{ is even.} \end{cases}$$

The r-th central moments of V is

$$E[(V-\mu)^r] = \begin{cases} 0 & \text{when } r \text{ is odd,} \\ \frac{r!}{(r/2)!}\frac{\sigma^r}{2^{r/2}} & \text{when } r \text{ is even.} \end{cases}$$

The proof follows immediately by direct substitution.

□

FULL CONDITIONAL DISTRIBUTIONS FOR GIBBS SAMPLING

Because the skewing function of the sampling distribution (9) does not possess a closed form expression, it is not possible to obtain the required conditional distributions directly from (12). A better way to proceed is the following. Notice that the derivation of the sampling distribution allows us to alternatively express the skew normal linear regression model in (11) as

$$Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\delta}^T\mathbf{Z}_i + \epsilon_i, \qquad Z_i > 0$$

with

$$\mathbf{Z}_i = \begin{pmatrix} Z_{1i} \\ Z_{2i} \end{pmatrix} \sim N_2(\mathbf{0}, \mathbf{I}) \qquad \text{and} \qquad \epsilon_i \sim N(\mu, \sigma^2),$$

where $\mathbf{Z}_i$ and ϵ_i are independent, and $\boldsymbol{\delta} = (\delta_1, \delta_2)^T$. Evidently, by treating the auxiliary variables $\mathbf{Z}_i$ as covariates, model (11) is seen to have an underlying normal linear regression model on the observations $\mathbf{y} = (y_1, \cdots, y_n)^T$. Thus casting the model in this form eliminates the need for the skewing function evaluations which in turn greatly facilitates the computation of the conditional distributions.

Now the complete Bayesian model of all the unknowns ($\mathbf{Z}$, $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\delta}$) can be written hierarchically as

$$Y_i|\mathbf{z}_i, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \sim N(\mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\delta}^T\mathbf{z}_i, \sigma^2), \quad \mathbf{Z}_i \sim N_2(\mathbf{0}, \mathbf{I})I(\mathbf{z}_i > 0),$$

$$\boldsymbol{\beta} \sim N_k(\tilde{\boldsymbol{\beta}}, \Omega), \quad \tau = \frac{1}{\sigma^2} \sim G(\nu, \nu), \boldsymbol{\delta} \sim N_2(\mathbf{0}, \Psi).$$

The expression for the joint posterior density is then

$$p(\mathbf{z}, \boldsymbol{\beta}, \tau, \boldsymbol{\delta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}, \tau, \boldsymbol{\delta})p(\mathbf{z})p(\boldsymbol{\beta})p(\tau)p(\boldsymbol{\delta}).$$

Now straightforward calculations yield the conditional distributions of the regression parameters given by:

$$\beta_j|\{\beta_s\},\sigma^2,\boldsymbol{\delta},\mathbf{z},\mathbf{y} \sim N\left(\frac{\Omega_{jj}\sum_{i=1}^n[(y_i-\mathbf{x}_i^T\boldsymbol{\beta}+x_{ij}\beta_j-\boldsymbol{\delta}^T\mathbf{z}_i)x_{ij}]+\sigma^2\tilde{\beta}_j}{\Omega_{jj}\sum_{i=1}^n x_{ij}^2+\sigma^2},\quad \frac{\sigma^2\Omega_{jj}}{\Omega_{jj}\sum_{i=1}^n x_{ij}^2+\sigma^2}\right)$$

for $s \neq j, j = 1, \cdots, k$. For the scale and skewness parameters we obtain

$$\tau = \frac{1}{\sigma^2}|\boldsymbol{\beta},\boldsymbol{\delta},\mathbf{z},\mathbf{y} \sim G\left(\frac{n}{2}+\nu, \frac{1}{2}\sum_{i=1}^n(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\boldsymbol{\delta}^T\mathbf{z}_i)^2+\nu\right)$$

and

$$\delta_j|\boldsymbol{\beta},\sigma^2,\{\delta_s : s\neq j\},\mathbf{z},\mathbf{y} \sim N\left(\frac{\Psi_{jj}\sum_{i=1}^n(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\boldsymbol{\delta}^T\mathbf{z}_i+\delta_j(\mathbf{z}_j)_i)(\mathbf{z}_j)_i}{\Psi_{jj}\sum_{i=1}^n(\mathbf{z}_j)_i^2+\sigma^2},\quad \frac{\sigma^2\Psi_{jj}}{\Psi_{jj}\sum_{i=1}^n(\mathbf{z}_j)_i^2+\sigma^2}\right),$$

for $j = 1, 2$. Finally, $\mathbf{Z}_i$'s have full conditional distributions defined by

$$\mathbf{Z}_i|\boldsymbol{\beta},\sigma^2,\boldsymbol{\delta},y_i \sim N_2\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{\delta_1^2+\delta_2^2+\sigma^2}\begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix},\quad \frac{1}{\delta_1^2+\delta_2^2+\sigma^2}\begin{pmatrix}\delta_2^2+\sigma^2 & -\delta_1\delta_2\\ -\delta_1\delta_2 & \delta_1^2+\sigma^2\end{pmatrix}\right)I(\mathbf{Z}_i>0).$$

These distributions are all of standard functional forms in which sample generation is relatively straightforward. So, Gibbs sampling can be easily implemented.

BIBLIOGRAPHY

Adcock, C. J. and Shutes, K. (2005). A Multivariate Skew-Normal Exponential Distribution. *Technical Report*, Management School, University of Sheffield, UK.

Arnold, B. C. and Beaver, R. J. (2000). The skew-Cauchy distribution. *Statistics & Probability Letters*, **49**, 285–290.

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

Arnold, B. C., Beaver, R. J., Groeneveld R. A. and Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, **58**, 471–488.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, **46**, 199–208.

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.

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

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.

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

Genton, M. G. and Loperfido, N. (2002). Generalized skew-elliptical distributions and their quadratic forms. *Institute of Statistics Mimeo Series* **2539**, *Annals of the Institute of Statistical Mathematics.*

Gibbons, J. F. and Mylroie, S. (1973). Estimation of impurity profiles in ion-implanted amorphous targets using joined half-Gaussian distributions. *Applied Physics Letters*, **22**, 568–572.

Henze, N. (1986). A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics*, **13**, 271–275.

John, S. (1982). The three parameter two-piece normal family of distributions and its fitting. *Communications in Statistics – Theory and Methods*, **11**, 879–885.

Jones, M. C. and Faddy, M. J. (2003). A skew extension of the t distribution, with applications. *Journal of the Royal Statistical Society*, B, **65**, 159–174.

Kelker, D. (1970). Distribution thoery of spherical distributions and a location-scale parameter generalization. *Sankhyā*, **32**, 831–860.

Kimber, A. C. (1985). Methods for the two-piece normal distribution. *Communications in Statistics – Theory and Methods*, **14**, 235–245.

Loperfido, N. (2002). Statistical implications of selectively reported inferential results. *Statistics & Probability Letters*, **56**, 13–22.

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

Mudholkar, G. S. and Hutson, A. D. (2000). The epsilon-skew-normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference*, **83**, 291–309.

Sahu, S. K., Dey, D. K. and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regressian models. *The Canadian Journal of Statistics*, **31**, 129–150.

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.

Theodossiou, P. (1998). Financial data and the skewed generalized T distribution. *Management Science*, **44**, 1650–1661.

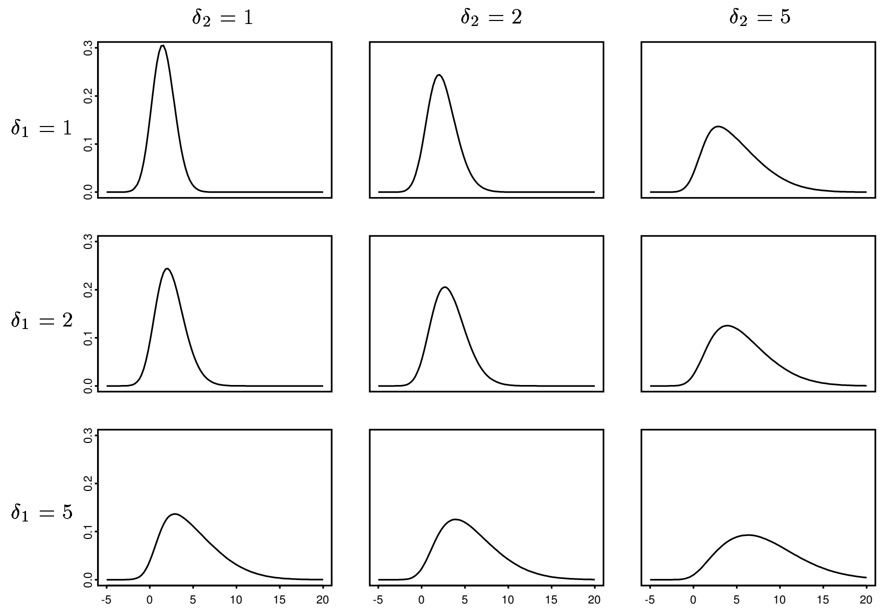

Figure 1: Plot of the density functions of $\mathrm{SN}_{\mathrm{new}}(0, 1, \binom{\delta_1}{\delta_2})$ for $\delta_1, \delta_2 = 1, 2, 5$.

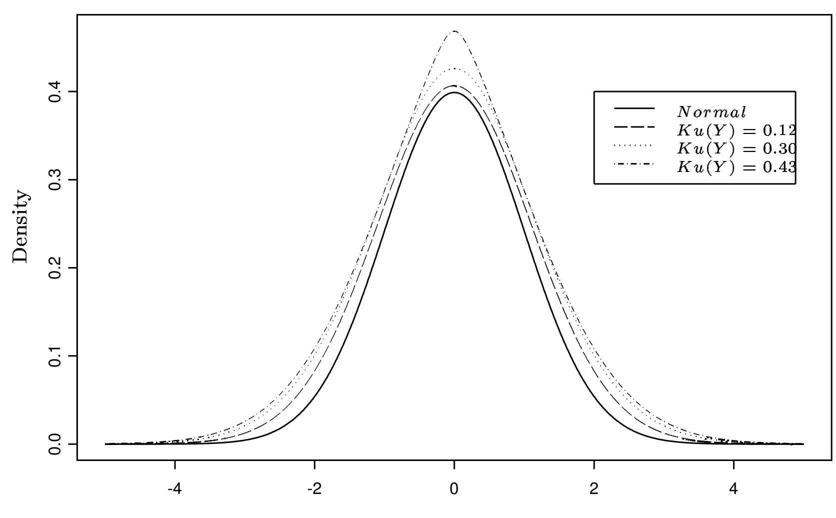

Figure 2: Plot of the density functions of $\mathrm{SN}_{\mathrm{new}}(0, \sigma^2, \binom{\delta_1}{\delta_2})$ where $\delta_1 = -\delta_2$ and $Var(Y) = 1$.

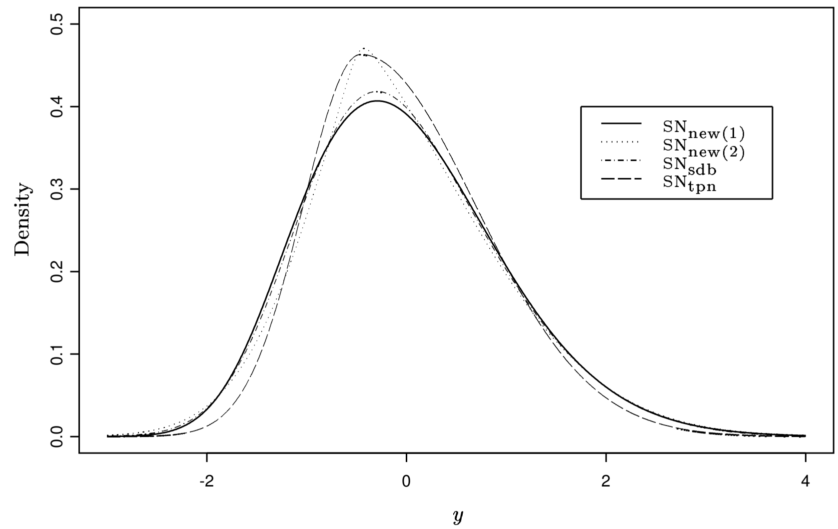

Figure 3: Plots of the density functions of various skew-normal distributions. All distributions are scaled to have zero mean, unit variance and $Sk(Y) = 0.5$.

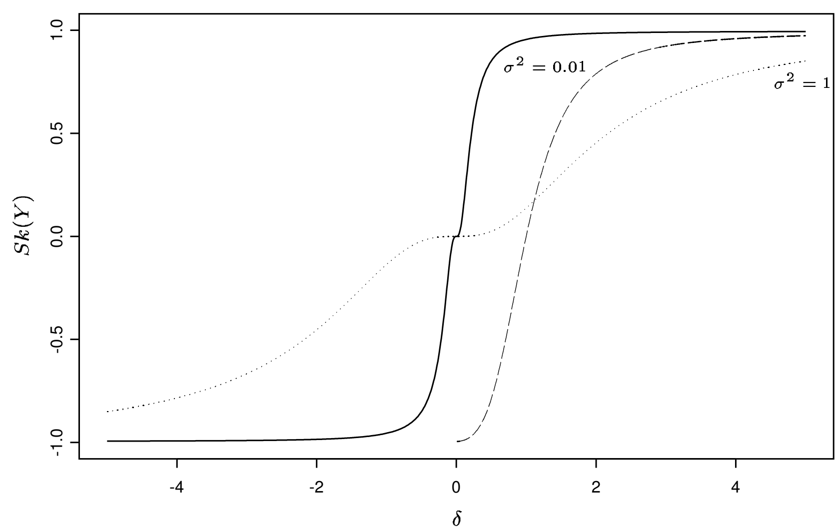

Figure 4: Plots of the skewness measure $Sk(Y)$ against δ. Dashed line is for $\mathrm{SN}_{\mathrm{tpn}}$; solid line is for $\mathrm{SN}_{\mathrm{sdb}}$ with $\sigma^2 = 0.01$ and the dotted line is for $\sigma^2 = 1$. Note that skewness of $\mathrm{SN}_{\mathrm{tpn}}$ does not depend on σ^2.

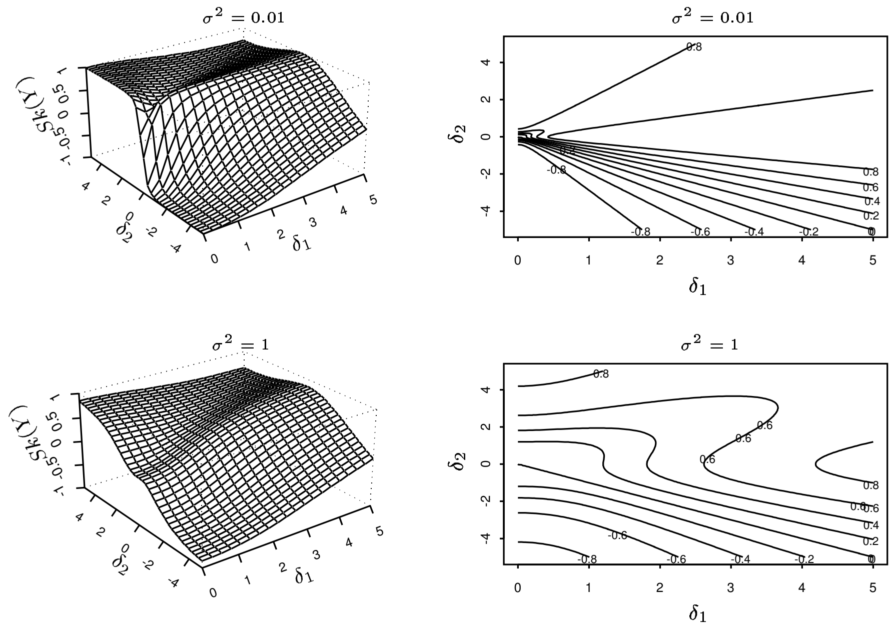

Figure 5: Surface and contour plots of the skewness measure $Sk(Y)$ of $\mathrm{SN}_{\mathrm{new}}$ against δ_1 and δ_2 for two different values of σ^2.

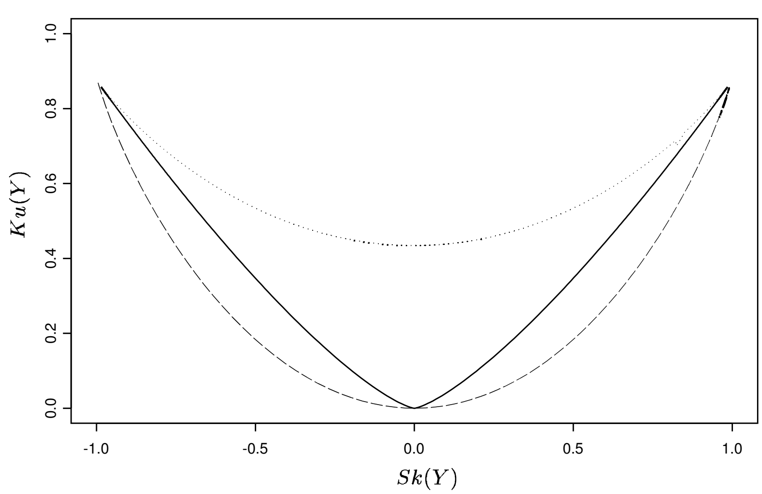

Figure 6: Plots of kurtosis versus skewness $Sk(Y)$; dashed line is for $\mathrm{SN}_{\mathrm{tpn}}$, solid line is for $\mathrm{SN}_{\mathrm{sdb}}$, and dotted line is for the maximum achievable kurtosis of $\mathrm{SN}_{\mathrm{new}}$.

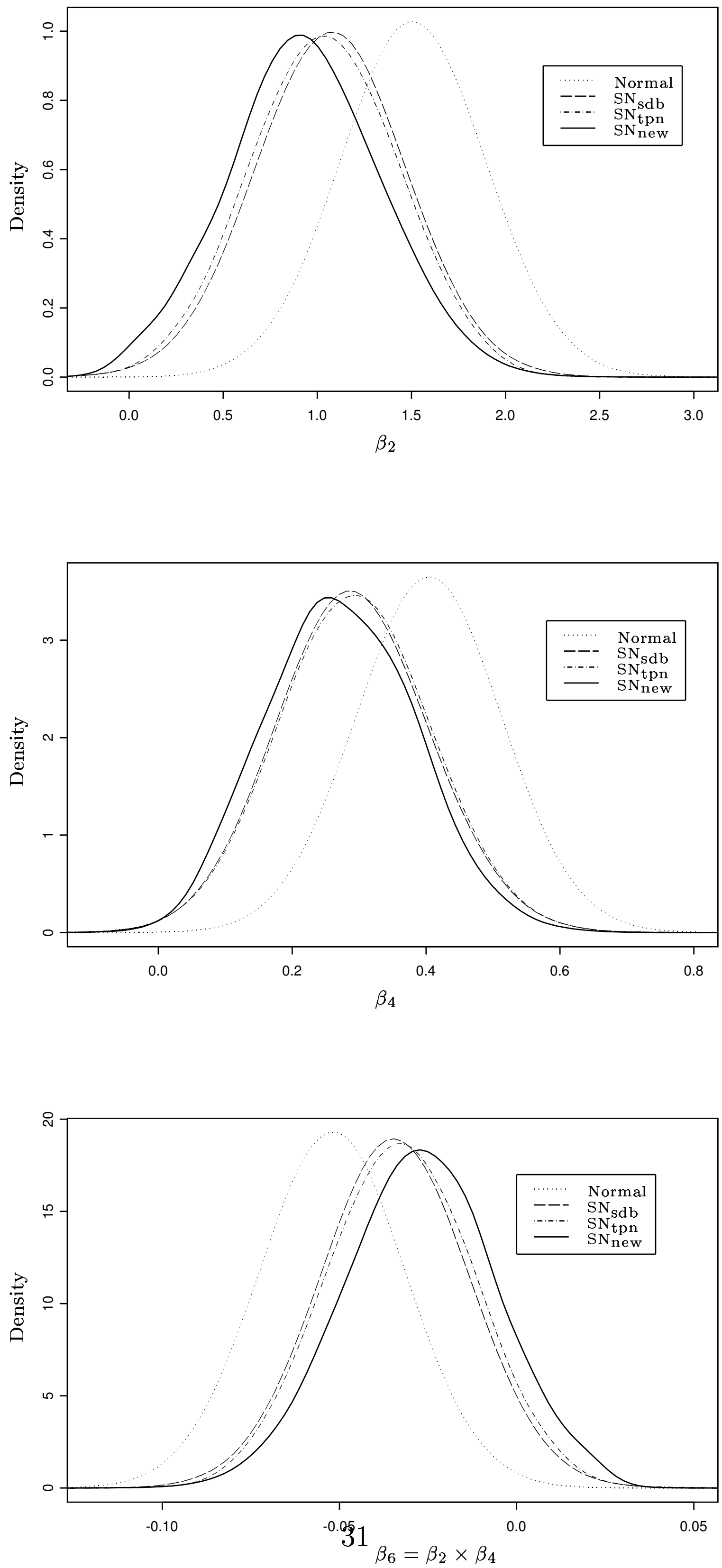

Figure 7: Marginal posterior densities of β_2, β_4 and β_6 for the non-academic scores example.

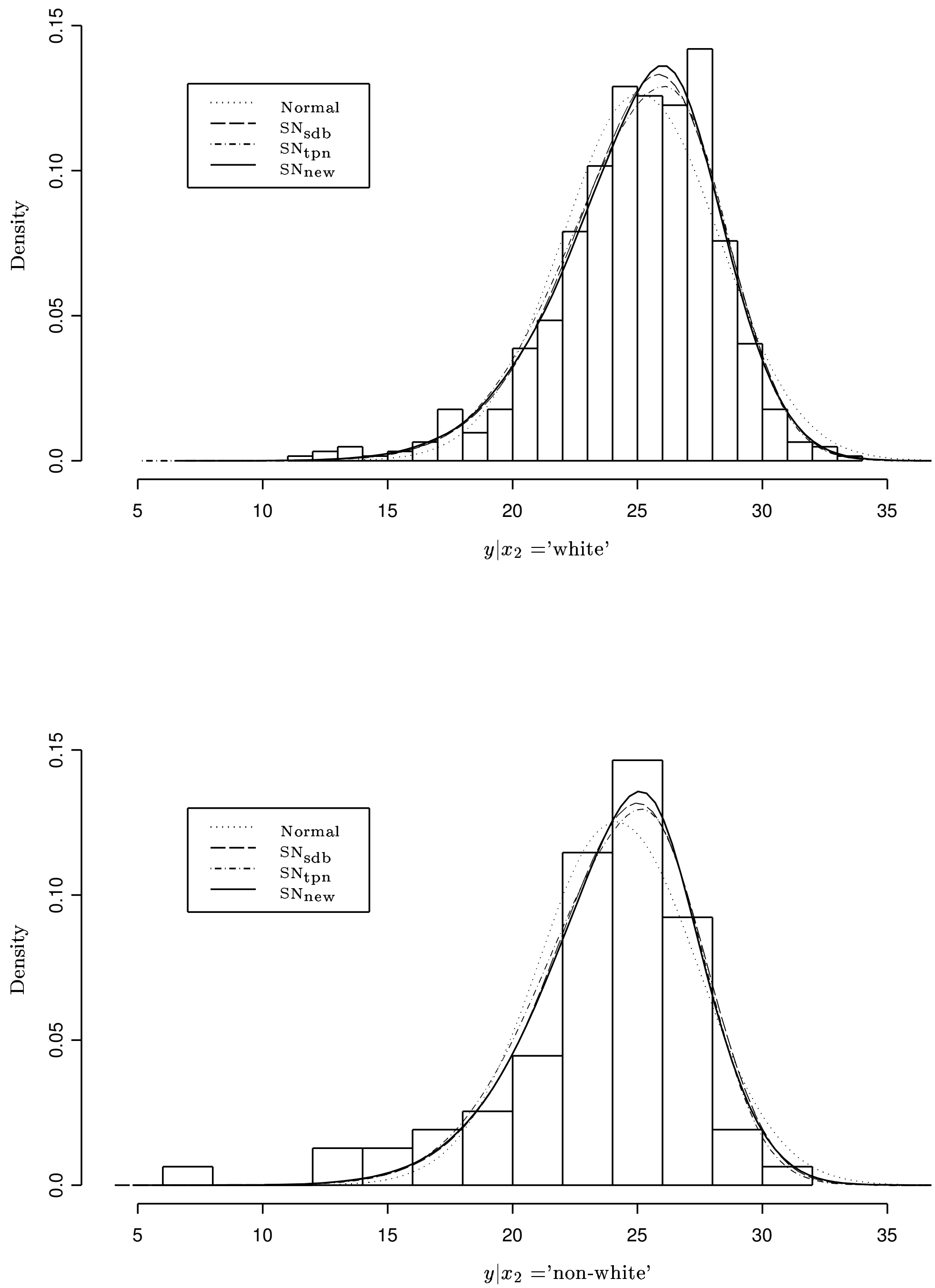

Figure 8: Histograms of the non-academic scores for both the races with superimposed posterior predictive densities under the $\mathrm{SN}_{\mathrm{new}}$, $\mathrm{SN}_{\mathrm{sdb}}$, $\mathrm{SN}_{\mathrm{tpn}}$ and normal models.