Model Comparison of Coordinate-Free Multivariate Skewed Distributions with an Application to Stochastic Frontiers*

José T.A.S. Ferreira and Mark F.J. Steel†

Department of Statistics
University of Warwick, UK

Abstract

We consider classes of multivariate distributions which can model skewness and are closed under orthogonal transformations. We review two classes of such distributions proposed in the literature and focus our attention on a particular, yet quite flexible, subclass of one of these classes. Members of this subclass are defined by affine transformations of univariate (skewed) distributions that ensure the existence of a set of coordinate axes along which there is independence and the marginals are known analytically. The choice of an appropriate m-dimensional skewed distribution is then restricted to the simpler problem of choosing m univariate skewed distributions. We introduce a Bayesian model comparison setup for selection of these univariate skewed distributions. The analysis does not rely on the existence of moments (allowing for any tail behaviour) and uses equivalent priors on the common characteristics of the different models. Finally, we apply this framework to multi-output stochastic frontiers using data from Dutch dairy farms.

Keywords: Coordinate-free distributions, dairy farm, multivariate skewness, orthogonal transformation, stochastic frontier.

JEL classification: C11; C16; C52

1 Introduction

Probability distributions that can model the presence of skewness in the distribution of a phenomenon have been the focus of interest in recent years (see Genton, 2004 for a review). Some of the classes of multivariate skewed distributions present in the literature introduce skewness along a pre-determined set of directions. Here, we are interested in classes that do not make such assumptions. We consider two classes of such distributions proposed in the literature and focus our attention on a particular subclass of one of them.

A class of multivariate distributions is defined to be coordinate-free if it is closed under orthogonal transformations. A simple example illustrates the importance of dealing with a coordinate-free class of distributions, say $\mathcal{S}$. Suppose that a process is measured using an orthogonal set of coordinates $X = (x_1, \ldots, x_m)'$ (*i.e.* x_i is perpendicular to $x_j,\ i \neq j,\ i, j = 1, \ldots, m$) and that the process can be

*The work of the first author was supported by grant SFRH BD 1399 2000 from Fundação para a Ciência e Tecnologia, Ministério para a Ciência e Tecnologia, Portugal.

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

described by a distribution $S_X \in \mathcal{S}$. Now consider a change to a different orthogonal set of coordinates $Y = (y_1, \ldots, y_m)'$, spanning the same space. The class $\mathcal{S}$ is coordinate-free if the process can also be described by a distribution $S_Y \in \mathcal{S}$, for any set of coordinates Y.

One of the many interesting features of elliptical distributions (Kelker, 1970) is that they are closed under orthogonal transformations (see Fang *et al.*, 1990 for details). When going from elliptical distributions to skewed distributions, coordinate-free classes become even more valuable. For elliptical classes, the only characteristic that changes with direction is spread. For skewed distributions both asymmetry and spread can vary with the direction. As a consequence, classes of skewed distributions that are not coordinate-free necessarily impose that skewness is manifested along particular directions. For example, the class of distributions introduced in Sahu *et al.* (2003) is not coordinate-free and introduces skewness into a symmetric elliptical distribution along the original coordinates.

We consider in some detail two main classes of multivariate skewed distributions that are closed under orthogonal transformations. The first is the class of skew-elliptical distributions, initially introduced through its special case of the multivariate skew-Normal distribution of Azzalini and Dalla-Valle (1996), generalised by Branco and Dey (2001), and extended by a number of authors (see Genton, 2004 for further details). The members of this class can be interpreted as generated by conditioning on an unobserved variable, so they are multivariate "hidden truncation" distributions. More recently, a different class of coordinate-free distributions has been suggested in Ferreira and Steel (2004a), henceforth FS, based on linear affine transformations of multivariate random variables with independent components, each having an univariate skewed distribution.

For reasons that will become clear in the sequel, the latter class of distributions is the main focus of our attention here. FS allows for any non-singular affine linear transformation. In this article, we restrict the set of transformations by imposing that for any distribution there is one set of orthogonal coordinates along which the components are independent and have known univariate distributions. In the (rather different) context of bivariate symmetric distributions with different kurtosis, Hoggart *et al.* (2003) introduced a class of distributions with a similar characteristic.

In FS, the authors point out that the skewed distributions of the univariate components in the transformation can be freely chosen, but focus on distributions that are generated by transforming originally symmetric distributions through inverse scale factors in the positive and the negative orthant (Fernández and Steel, 1998). This method can be viewed as a particular example of a general mechanism for transforming univariate symmetric distributions (see Ferreira and Steel, 2004b). Here, in addition to distributions generated by inverse scale factors, we analyse others generated by three distinct methods: hidden truncation (see *e.g.* Azzalini, 1985 and Arnold and Beaver, 2002), order statistics (Jones, 2004) and a construct (Ferreira and Steel, 2004b).

Given the flexibility of this class of multivariate distributions, and in particular the possibility of using different distributions for the univariate components, one important question is how to select appropriate forms for a specific problem. We analyse this issue for a general Bayesian regression setup. Prior specification is of special importance and we tackle the problem by using the same priors on common parameters. This requires, however, that these parameters share the same interpretation across models, which we ensure through normalisation of the skewed univariate distributions. The latter normalisation is based on robust measures of location and spread and does not rely on moment existence. For parameters that are specific to particular models, the skewness parameters, we propose prior matching ideas, where the priors on the parameters are not elicited directly but through a prior

on a quantity common to all models. We achieve this by specifying a prior on a measure of skewness and deriving equivalent priors on the skewness parameters for each model.

In addition to modelling skewness, it is often important to model tail behaviour of the distribution. We accommodate varying tail behaviour in our analysis by using two different types of heavy tailed distributions that differ in whether they assume a common tail behaviour for all dimensions. We specify Bayesian regression models using a proper prior structure. This implies we do not need conditions on moment characteristics to ensure posterior existence and, as such, we do not place any restriction on tail weight. Consequently, our analysis allows for distributions with (extremely) heavy tails.

We take two different approaches to model comparison. We use Bayes factors and we also compare predictive quality using log predictive scores.

The regression framework is then applied to a multivariate stochastic frontier problem. Such problems are traditionally dealt with through a composed error framework (as introduced in Aigner *et al.*, 1977 and Meeusen and van den Broeck, 1977) with separate measurement and inefficiency error terms, but here we use a skewed distribution to model the composed error directly. One important advantage of this approach is that it immediately generalises to the analysis of multi-output production, in contrast to the composed error framework. We apply this to a dataset of Dutch dairy farms with two outputs, milk production and non-milk outputs.

Section 2 introduces the class of multivariate skewed distributions and presents a number of results for the class. In Section 3, we review four different alternatives for introducing skewness in the distribution of the univariate components and we describe the normalisation. Section 4 introduces the Bayesian regression models considered here. Equivalent priors on the skewness parameters for the different models are determined in Section 5. In Section 6 details about the model comparison procedures are presented. Section 7 describes the application to the stochastic frontier problem. Finally, Section 8 gives some concluding remarks.

2 Coordinate-Free Distributions

In this section we briefly review the complete class of distributions introduced in FS and define the subclass that will be the focus of our attention in this paper. We also briefly review the skew-elliptical class.

2.1 Complete Class of Distributions

The class introduced in FS is constructed using linear transformations of univariate skewed distributions. Let m be the dimension of the random variable $\epsilon = (\epsilon_1, \ldots, \epsilon_m)' \in \Re^m$. In addition, let $f = (f_1, \ldots, f_m)'$ denote a vector of m univariate symmetric densities on $\Re$ and $\psi = (\psi_1, \ldots, \psi_m)'$ be a vector of parameters $\psi_j,\ j = 1, \ldots, m$. We then say that ϵ has a multivariate distributions with independent components with parameters f and ψ if its density is given by

$$p(\epsilon|f, \psi) = \prod_{j=1}^{m} s_j(\epsilon_j),$$

where, for $j = 1, \ldots, m$, s_j denotes a density obtained from f_j via some skewness-inducing transformation, indexed by ψ_j. Different transformations will be introduced in Section 3.

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

$$\eta = A'\epsilon + \mu \tag{1}$$

has a general multivariate skewed distribution, with density

$$p(\eta|\mu, A, f, \psi) = \|A\|^{-1} \prod_{j=1}^{m} s_j[(\eta - \mu)' A_{\cdot j}^{-1}], \tag{2}$$

with $A_{\cdot j}^{-1}$ denoting the j-th column of A^{-1} and $\|A\|$ the absolute value of the determinant of A. The random variable η is then said to follow distribution $Sk_m(\mu, A, f, \psi)$. The dependence between the components of η is modelled by the matrix A, with μ determining the location of the distribution. Evidently, $\epsilon \sim Sk_m(0_m, I_m, f, \psi)$, where 0_m and I_m denote the m-dimensional zero vector and identity matrix of size m, respectively.

FS illustrates that if s_j is not symmetric for all $j = 1, \ldots, m$, then the multivariate distributions of ϵ and η are skewed. By varying the parameters we then generate a flexible class of skewed distributions that is closed under orthogonal transformations.

For this class of skewed distributions, a number of results can be obtained. Here we mention two results on modality of the distribution and moment characteristics. Proofs are deferred to the Appendix.

For general f and ψ, it is not possible to derive any results for the modality of the distribution of η, particularly as the specific forms of $s_j,\ j = 1, \ldots, m$ are yet to be specified. Nevertheless, a useful result can be derived by imposing a rather plausible restriction on $s_j,\ j = 1, \ldots, m$.

Property 1. Let f and ψ be such that the densities $s_j,\ j = 1, \ldots, m$, are unimodal and have mode at zero. Then, for any μ and A the distribution $Sk_m(\mu, A, f, \psi)$ is unimodal with mode at μ.

When modelling real phenomena, is it not uncommon to assume that the distributions are unimodal. Property 1 shows that the unimodality of the multivariate distribution depends solely on the unimodality of univariate distributions, which is often much simpler to ensure.

The second result is on the existence of moments of η.

Property 2. Let $\eta \sim Sk_m(\mu, A, f, \psi)$. Further, let $r_1, \ldots, r_m$ be non-negative integers and $r = \sum_{j=1}^{m} r_j$. If $E\left[\epsilon_j^r\right]$ exists for $j = 1, \ldots, m$, then $E\left[\prod_{j=1}^{m} \eta_j^{r_j}\right]$ also exists.

Property 2 states that the mechanism that generates the multivariate distribution does not restrict the existence of positive moments. Non-negative integer moments of the distribution $Sk_m(\mu, A, f, \psi)$ depend solely on non-negative integer moments of the univariate distributions with densities $s_j,\ j = 1, \ldots, m$.

2.2 Restricted Class of Distributions

In order to gain further insight into the full effect of the transformation matrix A, it is useful to recall that any nonsingular matrix A can be written as the product of a lower triangular matrix L with positive diagonal elements and an orthogonal matrix O. Without loss of generality, assume μ to be the zero vector. The effect of the transformation matrix $A = LO$ is then clear. From (1), $\eta = O'L'\epsilon$, indicating that ϵ is first subjected to a linear transformation L' introducing dependence between the

variables and modifying scales, and then to linear orthogonal transformation O', a rotation if $|O| = 1$ or a rotoinversion if $|O| = -1$. The set of coordinate axes is modified by O', in effect changing the correlation in the original variables η. The orthogonal matrix O rotates and/or reflects the axes along which the joint distribution is a linear combination of the last $m - j + 1$ components of ϵ. FS defined these axes as the *basic axes* $e_j,\ j = 1, \ldots, m$ of the multivariate distribution.

The subclass of distributions that is the focus of interest here is obtained by imposing that along each one of the basic axes e_j, the distribution is a scaled version of the one with density $s_j,\ j = 1, \ldots, m$. This requirement can be straightforwardly transferred to A. All that is necessary is to replace L by D, a diagonal matrix with strictly positive diagonal elements. The effect of D and O is immediate. Matrices D and O parameterise scale and the orientation of the basic axes, respectively. To denote a distribution that is a member of this subclass we will use $Sk_m(\mu, D, O, f, \psi)$.

By restricting the complete class of distributions we necessarily obtain a less flexible set of distributions. The main restriction, when comparing to the complete class of distributions, is that now scale and skewness are introduced along the same directions given by the basic axes, with the correlation of η parameterised by O alone. However, this subclass is more easily interpretable, which is often of primary importance in the context of applications. In addition, it has the advantage that the marginals along the basic axes are of known form.

One other advantage of confining the attention to distributions of the form $Sk_m(\mu, D, O, f, \psi)$ is that these are closer to elliptical distributions. The latter class can be generated as in (1), with ϵ having a symmetric distribution (though not necessarily with independent components), and its parameterisation is in terms of the scale matrix $\Sigma = A'A = O'D'DO = O'D^2O$. By the singular value decomposition, $\Sigma = O'D^2O$ can cover all possible covariance structures. Even in the case when $Sk_m(\mu, D, O, f, \psi)$ represents an elliptical distribution (*e.g.* s_j is the standard normal density, $j = 1, \ldots, m$), there is no parameter redundancy, as there would be for the complete class (see FS, Section 2.2.3).

Like for the complete class defined in FS, $E(\eta)$ and $Var(\eta)$, if they exist, can take any value in $\Re^m$ and in the set of covariance matrices, respectively. Thus, under the conditions of Property 2, it is always possible to model the expected value and the covariance of the multivariate distribution of η, irrespective of f and ψ. Another characteristic common to both the complete class and the subclass is the range of achievable skewness values, as quantified using the multivariate skewness measure $\beta_{1,m}$ of Mardia (1970).

In order to achieve a unique parameterisation for the class of distributions, FS restricted O to a set $\mathcal{O}$, a subset of the set of all orthogonal matrices. Using an argument similar to the one used in Section 2.3 of FS, we can show that this parameterisation is also valid for the subclass considered here.

2.3 The Skew-Elliptical Class

As mentioned in the Introduction, the other main class of skewed distributions that we consider is the class of skew-elliptical distributions of Azzalini and Dalla Valle (1996) and Branco and Dey (2001).

Members of the skew-elliptical class can be derived by a conditioning process on a single unobserved variable. As a consequence, the skewed distribution is generated from the elliptical distribution by introducing skewness along one single particular direction. This leads to a set of distributions that is somewhat limited in terms of modelling skewness in a real phenomenon. As an example of these

limitations, the skew-elliptical class cannot model adequately the (simplest) case when the univariate components of η are independent, each having a skewed distributions. In order to illustrate this, Figure 1 presents contour plots of the densities of two bivariate skewed distributions, one belonging to the subclass defined in Section 2.2 (a), and the other a skew-elliptical distribution (b). While the contours in Figure 1(a) denote (different amounts of) skewness along both coordinate axes, the contours in Figure 1(b) illustrate that skewness is introduced along one single direction (in this case along the direction $\eta_1 = \eta_2$).

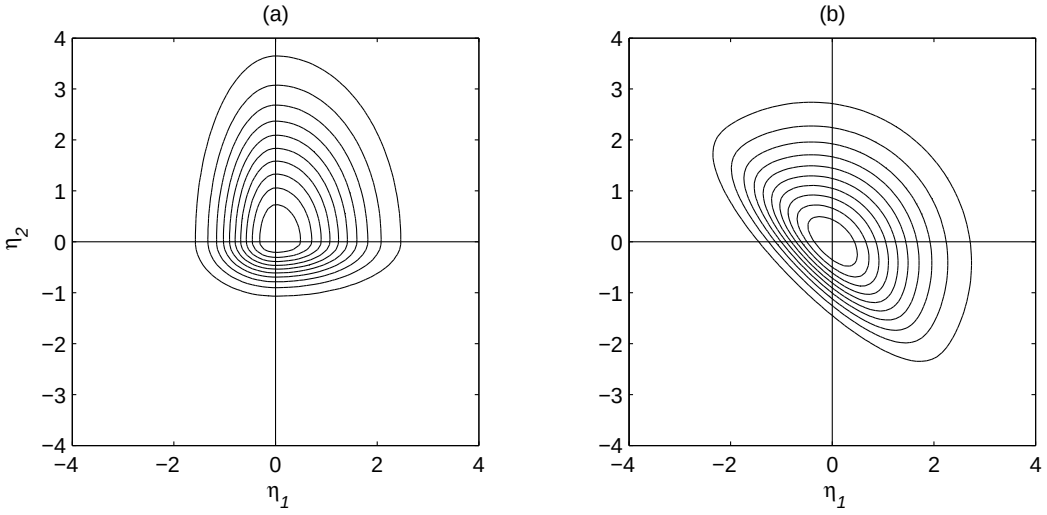

Figure 1: Contour plots of two bivariate skewed distributions with mode at zero, with pdf as in (2) (a) and member of the skew-elliptical class (b).

The reduced flexibility of the skew-elliptical class can also be described using the $\beta_{1,m}$ measure. As an illustration, we compare the range of achievable skewness for versions of the bivariate Normal distribution. For the bivariate skew-Normal distribution of Azzalini and Dalla Valle (1996)

$$\beta_{1,2} \in \left[0, 2\frac{(4-\pi)^2}{(\pi-2)^3}\right).$$

For both the complete class of FS and the subclass that we analyse here, the upper limit for $\beta_{1,2}$ is twice the value of the right limit of the interval above.

Given the characteristics of the skew-elliptical class, a fair comparison with the set of distributions defined in Section 2.2, could only allow for one particular s_j to be skewed, whilst fixing the remaining $m-1$ densities to be symmetric. In this article, we have decided not to do so.

Despite the different characteristics of the two classes of distributions, we point to the fact that, if all components of f are fixed and equal, then both use exactly the same number of parameters.

A recent paper by Gupta *et al.* (2004) introduces a multivariate skewed-Normal distribution by conditioning on an m-dimensional unobserved random variable. Thus, this distribution does not suffer from the limitation mentioned above, and it is also coordinate-free (unlike the class in Sahu *et al.*, 2003). Allowing for conditioning on an unobserved variable of unspecified dimension (possibly even larger than m) leads to a very general class of distributions in Arellano-Valle and Genton (2003), which generalises the approaches of Sahu *et al.* (2003) and Gupta *et al.* (2004). However, these classes are heavily (over)parameterised, which is particularly problematic for prior elicitation and inference. In addition, even with a more restricted parameterisation (for which no clear operational guidelines

are provided), inference on the basis of these types of distributions will be computationally more demanding.

3 Univariate Components

The complete definition of a multivariate skewed distribution, as in Sections 2.1 and 2.2, requires the specification of vectors f and of the transformation that leads to the univariate skewed densities $s_j,\ j = 1, \ldots, m$. In addition, this skewness-inducing transformation will be indexed by the parameter ψ, whereas the densities in f can depend on an extra parameter ν.

The majority of univariate distributions that can model skewness were developed by transforming an originally symmetric distribution. The skewed version then borrows the name of the original distribution, usually with the prefix "skew-". Each element j of vector f is a univariate symmetric density that is transformed to a skewed density s_j using a skewing mechanism parameterised by ψ_j. When all elements of f denote the same univariate density f^*, the multivariate skewed distribution $Sk_m(\mu, D, O, (f^*, \ldots, f^*)', \psi)$ is designated multivariate skew-f^*.

The most important reason for defining distributions that result from *skewing* a symmetric distribution is that in doing so, it is possible to preserve some of properties of the latter. Different skewing mechanisms will preserve different sets of these properties.

Recently, Ferreira and Steel (2004b) introduced a constructive representation of univariate skewed distributions that are generated from symmetric ones. They call a distribution S a skewed version of the unimodal symmetric distribution F, generated by skewing mechanism P, if its density is of the form

$$s(y|F, P) = f(y)p[F(y)], \tag{3}$$

where S and F are distributions in $\Re$ and P is a distribution in $(0, 1)$ and upper (lower) case denote probability distribution (density) functions. By varying the skewing mechanism P it is possible to generate different classes of skewed distributions. In general, the skewing mechanism P is indexed by a parameter ψ, which is specific to each of the methods.

Ferreira and Steel (2004b) review some of the skewing mechanisms that have been proposed in the literature and introduce others. In the sequel, we will mention four distinct skewing mechanisms, all described in greater detail in Ferreira and Steel (2004b).

3.1 Hidden Truncation

The first class of univariate skewed distributions that we review here is based on hidden truncation ideas (see Arnold and Beaver, 2002 for an overview). This is a very well-studied class with the skew-Normal distribution of Azzalini (1985) as its best known member.

The most common versions of univariate skewed distributions generated by hidden truncation have densities that are of the form

$$s(y) = 2f(y)F(\alpha y), \tag{4}$$

where F is a symmetric distribution and α is a real number. Positive (negative) values of α generate right (left) skewed distributions.

Distributions with densities as in (4) are generated using

$$p(x|\alpha) = 2F[\alpha F^{-1}(x)],$$

as the density of the skewing mechanism in (3).

3.2 Inverse Scale Factors

A class of skewed distributions generated by introducing inverse scale factors in the positive and the negative half real lines was proposed in Fernández and Steel (1998). If γ is a scalar in $(0, \infty)$, the distribution of S has density

$$s(y|\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} f\left[y\gamma^{-\text{sign}(y)}\right],$$

with sign$(\cdot)$ the usual sign function in $\Re$.

If $\gamma > 1$ then the distribution is right-skewed, whereas it is left-skewed for $\gamma < 1$. Such distributions can be constructed by using P with density

$$p(x|\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \frac{f[\gamma^{\text{sign}(1/2-x)} F^{-1}(x)]}{f[F^{-1}(x)]},$$

as the skewing mechanism.

3.3 Order Statistics

Another skewing mechanism is defined by the Beta distribution with density given by

$$p(x|\phi_1, \phi_2) = [B(\phi_1, \phi_2)]^{-1} x^{\phi_1 - 1}(1-x)^{\phi_2 - 1}, \tag{5}$$

where $B(\cdot, \cdot)$ is the Beta function. Vector $(\phi_1, \phi_2)' \in \Re_+^2$ parameterises the mechanism. For integer ϕ_1 and ϕ_2, Jones (2004) points out that distributions generated by such mechanism can be thought of as distributions arising from order statistics. Skewed distributions generated using (5) were recently analysed in *e.g.* Jones (2004) and Eugene (2002).

This skewing mechanism will be applied in this paper using a different parameterisation, suggested in Ferreira and Steel (2004c). A parameter $\tau \in \Re_+$ is introduced and (ϕ_1, ϕ_2) set to $\left(\tau, \frac{1}{\tau}\right)$. With this parameterisation, the resulting distribution is always skewed for $\tau \neq 1$. Values of τ larger (smaller) than one correspond to positively (negatively) skewed distributions. An important advantage of this parameterisation is that for certain choices of f, it leads to a one-to-one correspondence between skewness (measured as in Section 3.5) and τ. This greatly facilitates prior elicitation, as explained in Section 5.

3.4 Construct

The last skewing mechanism to be studied here was introduced in Ferreira and Steel (2004b). It was specifically constructed so that it preserves a number of characteristic of the symmetric distribution, such as the mode and the tail behaviour, and so as to generate skewed distributions with pre-defined properties. One of these properties is that right- and left-hand tails are identical.

The Construct mechanism P is defined through a density of the form

$$p(x|\delta) = 1 + l(\delta)[g(x|\delta) - 1],$$

where $\delta \in \Re$ and functions g and l are chosen to ensure that the set of characteristics is met. Further details can be found in the original reference. Positive values of δ induce positive skewness while negative skewness corresponds to negative values of δ. In the following, we always assume that the parameter d, controlling the smoothness of $p(x|\delta)$ as defined in Ferreira and Steel (2004b), equals two.

3.5 Quantifying Univariate Skewness

Measuring the skewness of a distribution is an important problem for which a number of alternatives have been proposed (see Arnold and Groeneveld, 1995 for a review). In order to ensure that the skewness of a unimodal distribution can always be quantified we choose a measure that does not involve moments. Such a measure is thus applicable even for distributions with extremely heavy tails.

In this article, we chose the skewness measure proposed in Arnold and Groeneveld (1995), denoted by AG, and defined as one minus twice the mass to the left of the mode. This measure takes values in $[-1, 1]$ and has an obvious interpretation for unimodal distributions. AG takes a negative (positive) value for left (right) skewed distributions and is equal to zero for symmetric distributions.

3.6 Normalisation of Skewed Distributions

The four versions of a single symmetric distribution, generated by the skewing mechanisms above, are different in several respects. They obviously differ in the way in which the skewness of the distribution is introduced but they also differ in the way location and dispersion is affected by this. In order to compare different univariate distributions with the same skewness we will normalise them with respect to location and scale.

We choose to normalise the distributions using characteristics that do not require existence of the moments of the distributions. For unimodal skewed distributions, the mode is an obvious measure of location. It is always well-defined and can be set at any value in $\Re$ by a shift operation. We use the interquartile range, henceforth IQR, as a measure of dispersion of the distribution. Less (more) dispersed distributions have smaller (larger) IQR values. The application of an appropriate scale transformation can set the IQR of a distribution to any positive value. In the sequel, we will always normalise the distributions so that the mode is at zero and the IQR is equal to one. Besides the obvious advantage that we can now deal with distributions for which the first two moments do not exist, the quantities underlying normalisation are also more robust than moments.

Figure 2 presents plots of the density of the four different normalised skewed versions of the Student-t distribution with two degrees of freedom for three different values of AG skewness: 0.2 (a), 0.5 (b) and 0.8 (c). For the smallest value of AG, the densities are similar. For larger values of AG, the differences between the densities is more evident. Particularly, the version generated by order statistics is very different from the remaining three, both in the central part of the distribution and in the tails. Also, the figure illustrates the identical behaviour in both tails for the construct.

4 Bayesian Regression Modelling

4.1 The Basic Model

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

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

$$y_i = g_i(B) + \lambda_i^{-\frac{1}{2}}\eta_i, \tag{6}$$

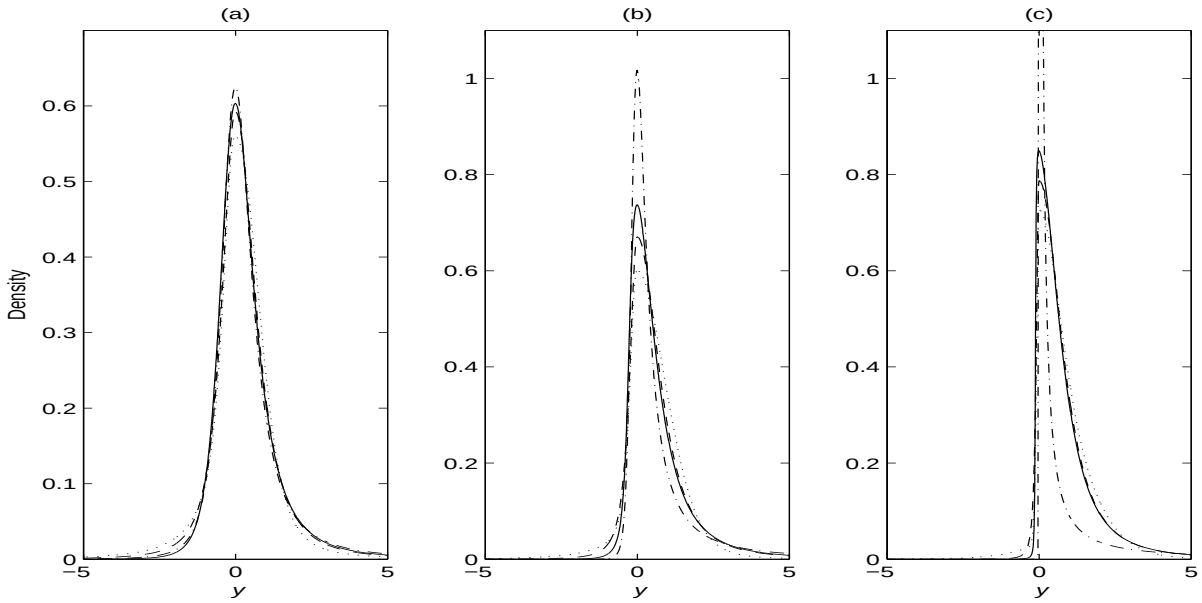

Figure 2: Densities of the hidden truncation (solid), inverse scale factors (dashed), order statistics (dot-dashed) and construct (dotted) versions of the Student-t distribution with 2 degrees of freedom and AG skewness equal to 0.2 (a), 0.5 (b) and 0.8 (c)

where $g_i(\cdot)$ is a known measurable function in $\Re^m$ depending on x_i, B parameterises the location, λ_i are independently drawn from some common underlying distribution P_λ on $\Re_+$ and η_i are independent and identically distributed as $Sk_m(0_m, D, O, f, \psi)$, defined in Section 2.2. In the sequel, we will always assume that s_j denotes the normalised version of the skewed distribution obtained from f_j and the skewness is parameterised by $\psi_j,\ j = 1, \ldots, m$. In addition, f_j is parameterised by ν_j. Since all the skewed densities s_j used here are unimodal, the distribution of η is unimodal by Property 1. Even though this is not necessary, in the sequel we assume that the same skewing mechanism is used in each dimension $j = 1, \ldots, m$.

The model derived from (6) is a general regression model with skewed disturbances. Function $g_i(B)$ represents the mode of the distribution of y_i, given the covariate values, and is chosen in accordance with the problem at hand.

By incorporating λ_i in the analysis, it is possible to model aspects of the sampling distribution that are not captured by the distribution of the disturbance η_i. It enlarges the distribution of the error term $e_i = \lambda_i^{-\frac{1}{2}}\eta_i$, by allowing for a mixture of multivariate skewed distributions. In this article, we will make use of λ_i in the modelling of a common tail behaviour for the elements of e_i.

The effect of D and O on the distribution of the η_is is immediate. Matrix O parameterises the direction of the basic axes and the diagonal elements of D denote the IQR of the distribution along these same axes.

The definition of the Bayesian regression model proceeds with the definition of prior distributions for parameters B, λ, D, O, ν and ψ, where we have defined $\lambda = (\lambda_1, \ldots, \lambda_n)'$ and $\nu = (\nu_1, \ldots, \nu_m)'$. We assume the prior structure given by

$$P_{B,\lambda,D,O,\nu,\psi} = P_B \times P_{D,O} \times P_\lambda \times P_{\psi|\nu} \times P_\nu, \tag{7}$$

where all distributions on the right hand-side will be proper, leading to an overall proper prior. This ensures the existence of a well-defined posterior distribution.

The conditioning of the prior distribution of ψ on ν indicates that the prior on the skewing parameters ψ can change with the distributions that are to be skewed. Distributions $P_{\psi|\nu}$ will be defined in Section 5.

The proposed prior distributions for B, D and O can be specified without taking the remaining parameters into account. They parameterise mode, dispersion and orientation of the distribution regardless of the mixing parameter λ or even the other parameters of the distribution of η.

The prior on B depends on the specification of the regression functions $g_i(B)$ and the particular application at hand, and is left unspecified for now.

The definition of $P_{D,O}$ is based on the relationship between (D, O) and the matrix $\Sigma = O'D^2O$, together with its interpretation with respect to elliptical distributions made in Section 2.2. We define $P_{D,O}$ via a prior on Σ. For the latter, it is common to assume a inverted Wishart prior with parameters, say, Q and v, an $m \times m$ positive definite symmetric matrix and a positive scalar, respectively. The equivalent prior on D and O is then given by density

$$p(D,O) \propto \left| \prod_{l=1}^{m-1} \prod_{j=l+1}^{m} \left(D_{ll}^2 - D_{jj}^2\right) \right| |D|^{-(v+m)} \exp\left\{ -\frac{1}{2} tr\left[\left(O'D^2O\right)^{-1} Q \right] \right\},$$

with tr denoting the trace operation.

If there is no prior information about the direction of the basic axes, choosing $Q = qI$, $q > 0$, leads to $tr\left[\left(O'D^2O\right)^{-1} Q\right] = q\ trD^{-2}$ and thus

$$p(D,O) = p(O)p(D),$$

where $p(O)$ corresponds to the Haar distribution (the invariant distribution on orthogonal matrices) defined on the restricted space $\mathcal{O}$ and

$$p(D) \propto \left| \prod_{l=1}^{m-1} \prod_{j=l+1}^{m} \left(D_{ll}^2 - D_{jj}^2\right) \right| |D|^{-(v+m)} \exp\left\{ -\frac{q}{2} \sum_{j=1}^{m} D_{jj}^{-2} \right\}.$$

In order to specify the prior on λ and ν we first need to introduce the particular choices of skewed distributions that we are going to analyse.

4.2 Choices for Tails and Skewed Distributions

4.2.1 Skew-Normal sampling

The first model that we consider assumes that η_i follows a skew-Normal corresponding to taking all elements of f equal to the standard Normal density and assigning a Dirac prior on $\lambda_i = 1,\ i = 1, \ldots, n$. In this case, ν is void.

4.2.2 Skew-Student sampling

A skew-Student sampling scheme is derived by assuming f and ν as above but assigning a Gamma prior with unitary mean and both parameters equal to $\nu^*/2$ to $\lambda_i,\ i = 1, \ldots, n$. If the distribution of η_i is multivariate Normal then, given ν^*, the distribution of $\lambda_i^{-\frac{1}{2}} \eta_i$ is multivariate Student-t with ν^* degrees of freedom, henceforth Student-t_{ν^*}. Similarly, if η_i has a multivariate skew-Normal distribution, then $\lambda_i^{-\frac{1}{2}} \eta_i$ is said to have a multivariate skew-Student-t_{ν^*}. The skew-Student sampling allows the error

term $\lambda_i^{-\frac{1}{2}}\eta_i$ to have a more flexible distribution. In particular, it allows for tail behaviour, common along all dimensions, which is heavier than the one of the skew-Normal distribution.

We will make use of a proper prior distribution on ν^*, P_{ν^*}.

4.2.3 Skew-IStudent sampling

The final skewed model that we analyse in this article is derived by assuming that f_j denotes the density of the standard univariate Student-t_{ν_j}, $j = 1, \ldots, m$, and setting a Dirac prior on $\lambda_i = 1,\ i = 1, \ldots, n$

By allowing the different elements of f to represent densities of Student-t distributions with varying degrees of freedom, we allow the distribution of η to exhibit different tail behaviours along different basic axes. This class will be denoted by skew-IStudent distributions.

For the prior on $\nu = (\nu_1, \ldots, \nu_m)'$, we assume that $P_\nu = \prod_{j=1}^m P_{\nu_j}$ where $P_{\nu_j} = P_{\nu^*}$, with P_{ν^*} proper.

5 Equivalent Priors on ψ

Bayesian model comparison is known to be sensitive to the choice of prior distributions (Kass and Raftery, 1995). A solution to this problem is to assign common priors whenever possible, and to define priors that are as similar as feasible for the remaining parameters. This is the approach we adopt here.

In Section 3.6, we introduced a normalisation of skewed versions of an underlying symmetric distribution. Figure 2 compared skewed densities with the same amount of skewness but generated by four different mechanisms. By restricting our attention to normalised distributions, common parameters have the same interpretation, irrespective of the particular skewing mechanism. This allows us to choose common prior distributions on all parameters but ψ, *i.e.* the skewness parameters. For the latter, we assume that

$$P_{\psi|\nu} = \prod_{j=1}^{m} P_{\psi_j|\nu_j}$$

and therefore, we will focus on the definition of $P_{\psi_j|\nu_j}$. For notational ease, we shall drop the subscript j in the remainder of this section.

The different mechanisms reviewed in Sections 3.1-3.4 all depend on a single skewness parameter defined over the real line (hidden truncation and construct), or its positive part (inverse scale factors and orders statistics). The choice of prior distributions for these parameters is achieved by fixing a prior on a common characteristic of all the skewed distributions and then by deriving the implied equivalent priors on the original parameters. The common characteristic chosen here is the amount of skewness, measured by AG.

Let P_{AG} be the prior distribution chosen on AG. For a particular value of ψ, the inverse scale factors and the construct skewing mechanisms always introduce the same amount of skewness, irrespective of the symmetric distribution f_j. Thus, the prior on ψ can be specified without taking ν into consideration. In addition, the relationship between the skewness parameter and AG skewness is bijective and invertible. This allows us to specify the prior distribution on $\psi = \gamma$ and δ as

$$p(\gamma|\nu) = p(\gamma) = \frac{4\gamma}{(\gamma^2+1)^2} p_{AG}\left(\frac{\gamma^2-1}{\gamma^2+1}\right)$$

and

$$p(\delta|\nu) = p(\delta) = \frac{48}{5\pi^3} \frac{\arctan^2\left(\frac{2\delta}{5}\right)}{1+\left(\frac{2\delta}{5}\right)^2} p_{AG} \left\{ \left[\frac{2}{\pi} \arctan\left(\frac{2\delta}{5}\right) \right]^3 \right\}.$$

For the two remaining skewing mechanisms, a direct variable transformation from AG to the parameter of the mechanisms is, in general, not possible. Therefore, we have to resort to methods that try to find a prior distribution on α and τ such that the induced prior on AG is *close* to P_{AG}. Another complication is that the amount of skewness, given the values of α and τ, depends on the distribution that is being skewed, and thus on ν.

5.1 Approximation using the Kullback-Leibler Distance

We choose the prior distribution on α and τ, given ν, by selecting a member of a parametric family of distributions G_κ, $\kappa \in \mathcal{K}$, that induces a prior on AG, denoted by $P_{AG|G_\kappa}$, that is closest, with respect to some distance function, to P_{AG}.

As a distance measure, we use the symmetric Kullback-Leibler distance between two discrete distributions. We first create a partition of the space of AG, $S = \{S_1, \ldots, S_L\}$, where the union of the elements of S is $(-1, 1)$. The prior distribution on the parameter of the skewing mechanism is then G_{κ^*}, where

$$\kappa^* = \arg\min_{k\in\mathcal{K}} \sum_{l=1}^{L} P_{AG}(S_l) \ln\left[\frac{P_{AG|G_\kappa}(S_l)}{P_{AG}(S_l)}\right] + P_{AG|G_\kappa}(S_l) \ln\left[\frac{P_{AG}(S_l)}{P_{AG|G_\kappa}(S_l)}\right]. \tag{8}$$

5.2 Prior on AG

If no prior information on AG is directly available, it is reasonable to assume that P_{AG} is a unimodal symmetric distribution with mode at zero. This corresponds to a prior that treats left and right skewness identically and that concentrates prior mass around symmetric distributions. We suggest a Beta prior on AG with both parameters equal to $a > 0$, rescaled to the interval (-1,1), given by density

$$p(AG|a) = 2^{1-2a}[B(a,a)]^{-1}\left[(1+AG)(1-AG)\right]^{a-1}. \tag{9}$$

As the value of a increases, the mass assigned by P_{AG} to heavily skewed distributions decreases.

5.3 Selecting G_k

The selection of a suitable family of distributions G_κ needs to take into account two aspects. First, the parameter space of α and τ will restrict the choice of families G_κ. Also, we need to take into account symmetry properties of the prior distribution P_{AG} that is to be induced.

For skewed distributions generated by hidden truncation, if $\alpha = \alpha^*$ corresponds to a skewness value AG^*, then $-\alpha^*$ corresponds to $-AG^*$. Therefore, if P_{AG} is symmetric then a symmetric family G_κ on α is called for. Here, we use the class of Student-t distributions with mean zero as G_κ. Parameter κ is then a two-component vector, one corresponding to the variance and the other to the degrees of freedom.

The order statistics mechanism is indexed by $\tau \in \Re_+$. If $\tau^* = \exp\{b\}$ leads to skewness AG^*, then $1/\tau^* = \exp\{-b\}$ leads to $-AG^*$, implying that there is a symmetry in the logscale. Thus, we select the prior on $\ln(\tau)$ to be a symmetric Student-t distribution with variance and degrees of freedom to be chosen as in (8).

Figure 3 shows the discrete distribution functions for P_{AG} and the fitted $P_{AG|G_{\kappa^*}}$ when P_{AG} corresponds to (9) with $a = 5$ and κ^* is as in (8), for three different combinations of skewing mechanism and f. The elements S_l of partition S are the intervals $\big(-1+\frac{l-1}{10}, -1+\frac{l}{10}\big]$, $l = 1, \ldots, 20$. For the hidden truncation mechanisms the fit is reasonable, with small deviations around zero. An almost perfect fit is achieved for the order statistics mechanism skewing applied to the t_2 density. We point out that, for this last case, there is an analytical bijective transformation between ψ and AG. However, this is an exceptional case and not the rule for this skewing mechanism.

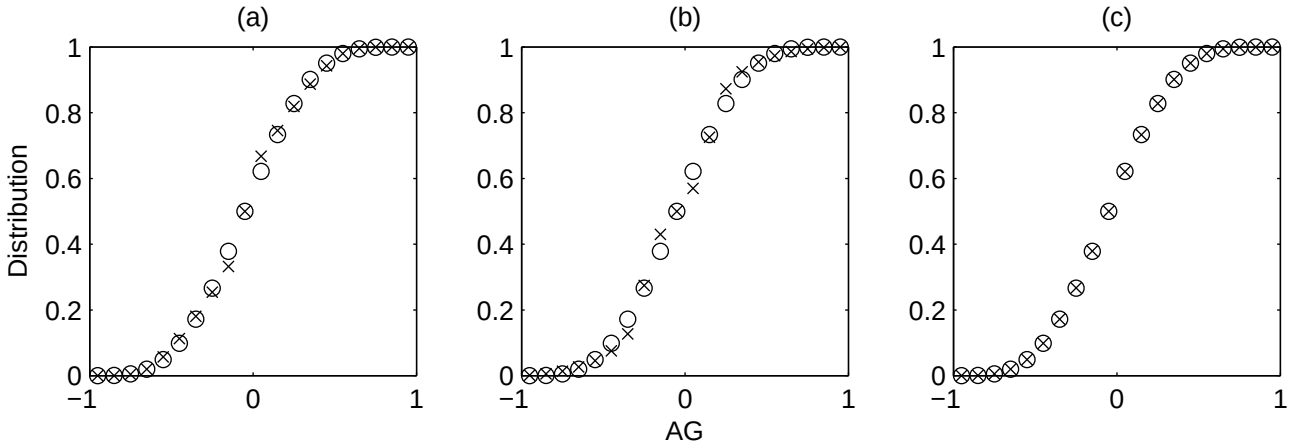

Figure 3: Discrete distribution functions for P_{AG} and $P_{AG|G_{\kappa^*}}$ when P_{AG} is the distribution in (9) with $a = 5$ and κ^* is as in (8) for the hidden truncation mechanism when f is the Normal (a) or the Student-t_2 (b) density, and for the order statistics mechanism with f the Student-t_2 density (c). The circles correspond to P_{AG}, and the crosses represent $P_{AG|G_{\kappa^*}}$.

In the application, we will use the order statistics mechanism only in combination with the Student-t_2 distribution[1], since the relationship between ψ and $AG \in (-1, 1)$ is not bijective for skewed versions of the Normal or Student-t distributions with large degrees of freedom. This is illustrated in Figure 4 for the skew-Normal distribution.

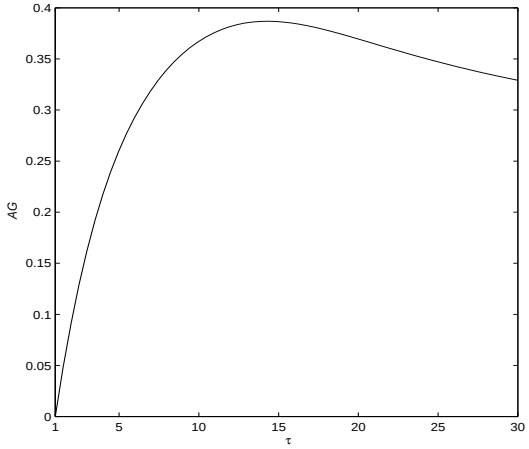

Figure 4: Measure of skewness AG as a function of $\tau > 1$ for the order statistics skew-Normal.

[1]This skewed version of the Student-t_2 was studied in detail in Jones and Faddy (2003).

6 Model Comparison

In order to compare the suitability of the different models we make use of two different measures: Bayes factors and log predictive scores (LPS).

Bayes factors measures the adequacy of two competing models using the ratio between their marginal likelihoods.[2] Estimates of marginal likelihoods are obtained using the p_4 measure in Newton and Raftery (1994).

The log predictive score is a proper scoring rule, introduced in Good (1952) and discussed in further detail in Dawid (1986). It can be interpreted as an approximation to the expected loss with a logarithmic rule. Consider predicting n_p observables, say, in $Y^p \in \Re^{n_p \times m}$, with each row corresponding to one out-of-sample observation, where we condition on the corresponding regressor values and use a model $\mathcal{M}$. Then, LPS is defined as

$$\text{LPS}(Y^p|Y,\mathcal{M}) = -\frac{1}{n_p}\sum_{i=1}^{n_p} \ln p\left(y_i^p|Y,\mathcal{M}\right).$$

Smaller values of $\text{LPS}(Y^p|Y,\mathcal{M})$ indicate that model $\mathcal{M}$ predicts better. For the models that we consider in this article, $\text{LPS}(Y^p|Y,\mathcal{M})$ is not available directly but will be estimated using Markov chain Monte Carlo methods.

In general, a separate sample Y^p is not available and we will take a cross-validation approach. We randomly partition the original sample Y into M disjoint sets $Y^i,\ i = 1,\ldots,M$ of (almost) equal size. We then calculate $\text{LPS}(Y^i|Y^{-i},\mathcal{M})$, where $Y^{-i} = Y \setminus Y^i,\ i = 1,\ldots,M$. The LPS for model $\mathcal{M}$ will be computed as the average over all M partitions.

7 Application to Stochastic Frontiers

7.1 Introduction to the Problem

Stochastic production frontiers describe the possibilities of economic agents to transform inputs into outputs in the most favourable way ("best-practice" production). They are important tools in the study of firm efficiency, as firms will, in practice, often not attain the optimal frontier production and this is typically associated with inefficiency. The usual statistical approach to stochastic frontiers is through a composed error framework (as introduced in Aigner *et al.*, 1977 and Meeusen and van den Broeck, 1977) with two separate error components: a symmetric measurement error and an inefficiency error, which is defined to be one-sided. This standard framework has been examined with Bayesian methods, starting with van den Broeck *et al.* (1994).

An important problem in this area is how to deal with production frontiers corresponding to firms producing multiple outputs. Kumbhakar (1996) and Fernández *et al.* (2000) discuss some of the statistical problems that arise in this context and Fernández *et al.* (2000) propose an analysis based on a parametric production equivalence surface, effectively leading to an aggregate output quantity, which can then be modelled through a univariate frontier, and use a Dirichlet distribution on output shares to complete the multivariate stochastic specification.

Here we follow an alternative approach and use instead a skewed distribution to model the composed error directly. In contrast to the composed error framework mentioned above, this approach

[2] The marginal likelihood is the data density integrated with respect to the prior.

immediately generalises to the analysis of multi-output production. All we need to do is to use a multivariate skewed distribution. In effect, we will model the frontier through a multivariate regression specification with skewed errors, which implies that each output component has its own specific frontier.

There is a direct link between both approaches in the context of the univariate skew-Normal generated by hidden truncation as in Azzalini (1985). The latter distribution can be thought of as arising from summing a Normal and a half-Normal random variable. This fact was mentioned in *e.g.* Arnold and Beaver (2002), and in the discussion to that paper possible links with stochastic frontiers were suggested by Azzalini (2002) and Sarabia (2002). Nakatsuma (2003) uses a univariate skewed Normal effectively based on inverse scale factors for modelling a cost frontier.

7.2 Description of the data

The data that we analyse here was compiled by the Netherlands Agricultural Economics Research Institute and relates to highly specialised dairy farms that were part of the Dutch Farm Accountancy Data Network. Further details on these data can be found in Reinhard *et al.* (1999).

We have 1545 observations of two outputs and three regressors. The outputs are milk (millions of kilograms) and non-milk (millions of 1991 guilders). Non-milk output contains sales of meat, livestock and roughage. The inputs or production factors are family labour (thousands of hours), capital (millions of 1991 guilders) and variable input (thousands of 1991 guilders). Capital includes land, buildings, equipment and livestock. Variable input refers to, *i.a.*, hired labour, concentrates, roughage and fertilizer.

The original data is in the form of an unbalanced panel, with observations of 613 farms for all or some of 1991-1994. Here we discard the temporal information and treat all observations independently. This is clearly an assumption we would like to relax in future but we have used it here as this application is primarily an illustration of the use of multivariate skewed distributions in this context.

7.3 Model Specification

We will entertain ten different multivariate skewed regression models and three symmetric alternatives. The skewed models correspond to Sections 4.2.1-4.2.3, *i.e.* skew-Normal, skew-Student and skew-IStudent sampling, in combination with the hidden truncation, inverse scale factors or construct mechanisms, plus the model defined by the order statistics mechanism applied to skew-IStudent sampling but with all elements of f equal to the Student-t_2 density[3]. The symmetric models are given as in Sections 4.2.1-4.2.3, but with s_j equal to a normalised (as in Section 3.6) symmetric version of $f_j, j = 1, \ldots, m$.

We assume that the bivariate frontier has a simple Cobb-Douglas form. Thus, the regression function for modelling log outputs y_i in (6) is

$$g_i(B) = B'x_i,$$

where B is a $k \times m$ matrix of coefficients and x_i consists of 1 (as the first element) and the log input values for observation i. Economic regularity conditions are easily imposed by constraining all elements of B to be positive, with the exception of the first row. Fernández *et al.* (2002) use the same

[3]This means that P_{ν^*} is Dirac at $\nu^* = 2$. The reason for this limitation is given in Subsection 5.3.

data (with an additional bad output) and also consider a Cobb-Douglas specification for the frontier. A more general translog frontier was used in Reinhard *et al.* (1999) and Fernández *et al.* (2004).

The definition of the Bayesian models is completed by fully specifying the prior distribution described in Sections 4 and 5.

The prior on B is a matricvariate Normal truncated to the regularity region, corresponding to

$$p(B) \propto \exp\left[-\frac{1}{2}\ \text{tr}\ M_1^{-1}B'M_2^{-1}B\right] I_{(B_{i,j}>0,\ i\geq 2)}(B)$$

where we choose $M_1 = 100I_m$ and $M_2 = 100I_k$, and with $I_A(x)$ denoting the indicator function on A. This corresponds to a quite dispersed prior on the matrix B, centred at the zero matrix.

We set $Q = 0.1I_m$ and $v = 4$, leading to a vague prior on D and O. For the prior on AG in (9) we choose $a = 5$. Finally, an Exponential prior with mean and standard deviation equal to 10 was chosen for ν^*.

7.4 Inference

Inference is conducted using Markov chain Monte Carlo methods (MCMC). We use Metropolis-Hastings sampling for all parameters with the exception of λ and ν^*, where we use Gibbs steps. For the Metropolis-Hastings samplers, simple random walk steps are used in updating the elements of B, ν_j, α_j and $\delta_j, j = 1, \ldots, m$. The diagonal elements of D, alongside with γ_j and τ_j are updated in logscale, again using random walks. The sampling of O is more complicated and we refer the reader to FS for further details.

LPS as described in Section 6 is calculated using a partition of the data into $M = 10$ subsets, which are kept constant for the evaluation of the performance of each model.

Inference was conducted using MCMC chains of 125,000 iterations. We retained every 20th sample after a burn-in period of 25,000 draws. Matlab code is available from the authors upon request.

7.5 Results

Table 1 presents the logarithm of the Bayes factors for the different models with respect to the symmetric Normal alternative. A positive value for an entry indicates support in favour of that alternative. The results show that the hidden truncation and, in particular, the inverse scale factors skewed versions find most support in the data. The construct alternatives do not improve on the symmetric counterparts and the order statistics model performs poorly. There is no real evidence in favour of a tails other than Normal, which can be seen from Table 1 and from the fact that the posterior distributions of ν^* and ν_j have most mass on relatively large values. This helps to explain the poor performance of the order statistics method based on the Student-t_2 distribution, as for this model one of the tails of the distribution is necessarily very heavy. This is a consequence of the parameterisation used here, which allows for the prior elicitation described in Section 5. When allowing for tail behaviour different from Normal, the inverse scale factors skewing mechanism appears to be quite robust.

We now concentrate our attention on the posterior distribution of the residuals for the Normal alternatives. Figure 5 presents contour plots for the densities of these distributions. For the four alternatives, the orientation of the density and of the basic axes, is very similar. In plots (b) and (c), corresponding to the hidden truncation and inverse scale factors models, the presence of skewness is

Table 1: Log of Bayes factors for the different models with respect to symmetric Normal model. The entry for the order statistics model corresponds to choosing f_j equal to the density of the Student-t_2 distribution, $j = 1, 2$.

Distribution	Symmetric	Hidden Truncation	Inv. Scale Factors	Construct	Order Statistics
Normal	0	22.3	24.6	-0.4	-
Student	-3.0	13.0	24.6	-2.7	-
IStudent	-1.0	18.0	24.0	-5.5	-227.4

evident. The contours for the two distributions are similar, which is also supported by the results in Table 1. Figure 5(d) illustrates that the construct distribution does not really induce much skewness, in line with Table 1, even though, interestingly, the shapes of the contours are changed with respect to the symmetric ones in Figure 5(a).

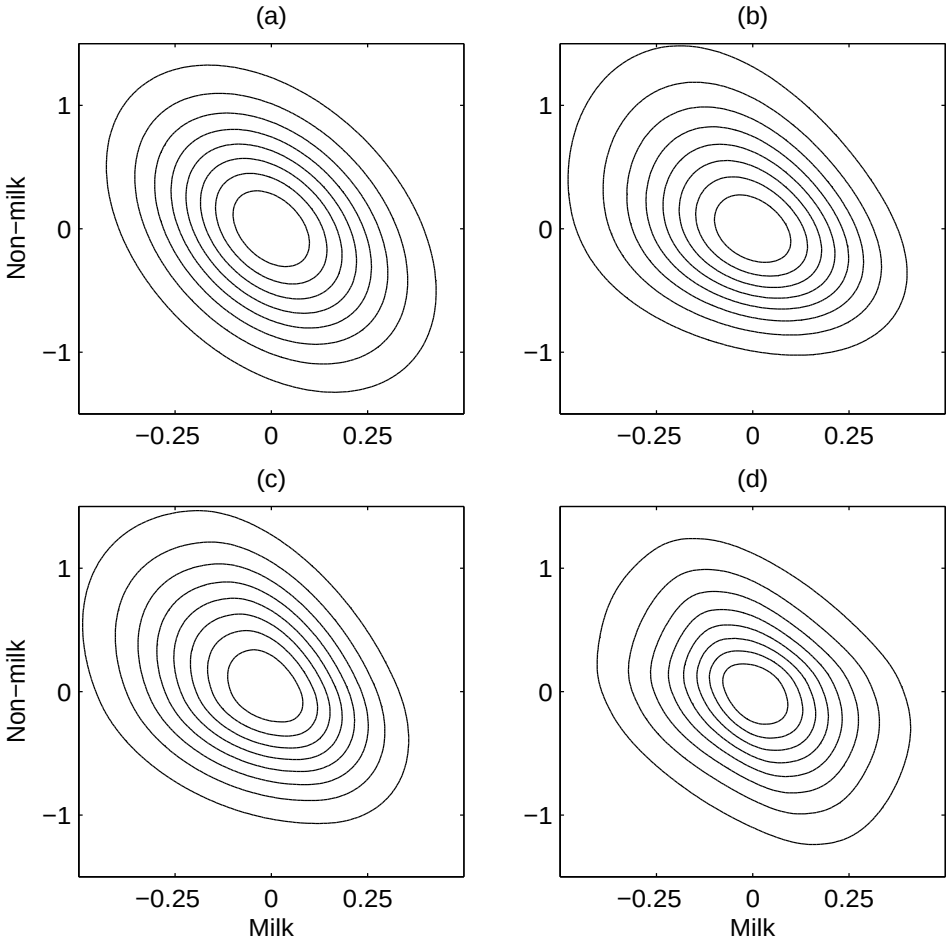

Figure 5: Contour plots of the posterior density of the residuals for the symmetric (a), hidden truncation (b), inverse scale factors (c) and construct (d) versions of the Normal distributions.

In order to further investigate skewness, we can analyse the posterior distribution of AG. Figure 6(b) and (c) present posterior densities of AG for the skewed versions of the Normal distribution, and for the order statistics version of the Student-t_2 distribution, measured along the basic axes. Figure

6(a) presents a grayscale plot of the direction of the basic axes for the inverse scale factors version of the Normal distribution. For the other models, the directions for these axes have a similar posterior distribution. As expected from Figure 5(d), the posterior densities of AG for the construct model are centred at zero. Despite the fact that the order statistics Student-t_2 is not supported by the data, it can still capture some of the skewness present in the application. Finally, even though Figures 5(b) and (c) seem to exhibit similar contours at first sight, the respective posterior densities of AG are quite different. The densities for the hidden truncation model concentrate mass on smaller values of AG than the inverse scale factor version.

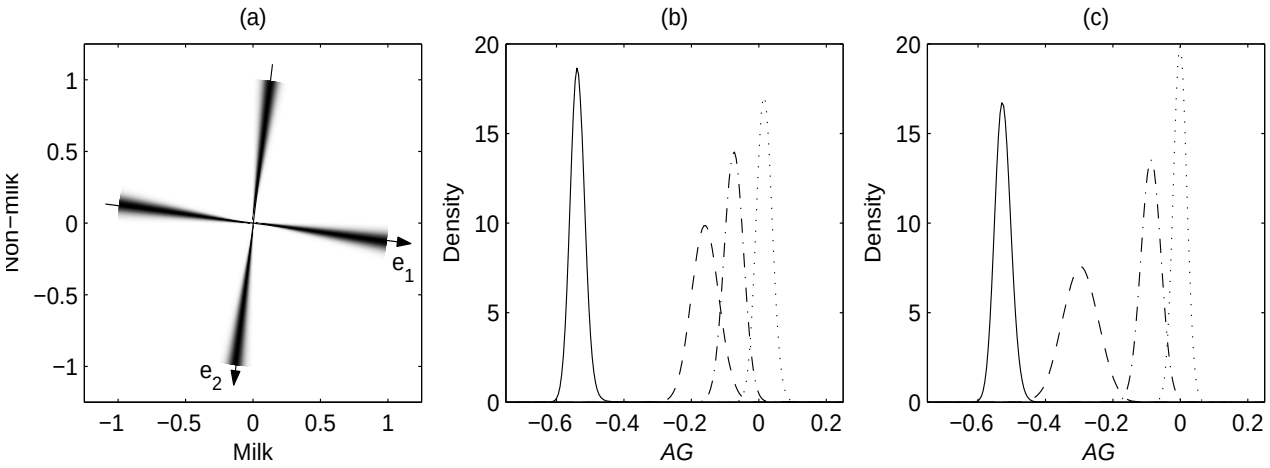

Figure 6: (a) Grayscale plot of the posterior density of the basic axes as defined in Section 2, for the inverse scale factors version of the Normal distribution; marginal posterior densities for the AG measure of the distribution along the first (b) and second (c) basic axes for the hidden truncation (solid), inverse scale factors (dashed) and construct (dotted) versions of the Normal distribution and the order statistics version of the Student-t_2 (dot-dashed).

A link of these distributions with efficiency behaviour is less immediate than in the composed error framework, and will be the focus of future research in this area. As an initial idea, however, we can simply compute the orthant probabilities in terms of the original output coordinates. For a skewed error distribution to behave in line with the composed error framework for production frontiers (where a positive error term is subtracted from a symmetric one), it needs to be negatively skewed. It is important to stress that we are not restricting the direction of skewness in our setting. Thus, our framework implicitly investigates whether the distribution of the data is compatible with the frontier idea. Table 3 presents the median posterior orthant probabilities along the original coordinate axes for the most favoured model. There is a clear negative skewness in the milk dimension, corresponding to a median AG of -0.3, indicating a noticeable level of inefficiency in the sector. This is mostly due to the negative skewness along the first basic axis (e_1 in Figure 6(a)), which is roughly in the same direction. The non-milk dimension, however, indicates skewness in the opposite sense, thus contradicting the interpretation of the regression model in this direction as a frontier. This is perhaps not overly surprising as these farms are specialised in milk production and the behaviour of farms is bound to be more in line with the usual economic rationale in that direction, whereas the non-milk production is much more incidental and less of a planned economic activity. In other words, farms are not really looking to reach the frontier in that direction, so we can not necessarily expect a left-skewed distribution. This is also in line with the findings of Fernández *et al.* (2004), who measure efficiency

(with respect to a common frontier) separately for both outputs and find efficiency to be much lower for non-milk output than for milk.

Table 2: Median posterior orthant probabilities of the residuals for the inverse scale factors version of the Normal.

		Milk		
		Negative	Positive	**Marginal**
Non-milk	Negative	0.18	0.18	0.36
	Positive	0.47	0.17	0.64
	Marginal	0.65	0.35	1

From Table 2, we can also deduce that the correlation between the error terms in terms of the original output directions is negative, indicating that a farm which does particularly well with respect to its frontier in terms of milk is expected to do worse than average in terms of the production of non-milk output. This may reflect the level of specialisation of the farm.

Inference on the regression models for both outputs is summarized in Table 3. As mentioned above, only the one for milk is consistent with the interpretation of frontier in the usual economic sense. The coefficients corresponding to the different inputs are input elasticities and their values for the milk frontier are reasonable, and rather close to those obtained in Fernández *et al.* (2002): the main difference is a somewhat larger capital elasticity. Returns to scale are founds to be increasing, as in Fernández *et al.* (2002) and Reinhard *et al.* (1999).

Table 3: Posterior median and percentiles for regression parameters.

		2.5%	Median	97.5%
Milk	Intercept	-2.62	-2.24	-1.99
	Labor	0.05	0.11	0.14
	Capital	0.62	0.65	0.71
	Variable	0.39	0.42	0.44
	RTS	1.15	1.17	1.21
Non-milk	Intercept	-2.17	-1.48	-0.81
	Labor	0.15	0.26	0.35
	Capital	0.0002	0.008	0.044
	Variable	0.78	0.83	0.88
	RTS	1.02	1.10	1.17

The LPS results on out-of-sample prediction for the different models are summarised in Table 4, which presents the number of times (out of the ten prediction subsets) that each model performed best and worst (including ties), as well as the average of the LPS values over the ten different subsets. The results are in close agreement with the evidence from Bayes factors. The predictive performances of the Normal alternatives are again favoured. Also, the inverse scale factor versions do somewhat better than the hidden truncation ones, whereas the construct does not really improve on the symmetric models.

Throughout, the worst predictions come from the order statistics model (based on the Student-t_2).

Table 4: LPS for the different models.

Version	Distribution	# best	# worst	**Average**
Symmetric	Normal	0	0	.711
	Student	0	0	.713
	IStudent	0	0	.712
Hidden Truncation	Normal	1	0	.696
	Student	0	0	.701
	IStudent	0	0	.700
Inv. Scale Factors	Normal	6	0	.694
	Student	2	0	.695
	IStudent	2	0	.695
Construct	Normal	0	0	.711
	Student	1	0	.712
	IStudent	0	0	.713
Order Statistics	Student-t_2	0	10	.859

An additional visual aid for the assessment of the predictive quality of different models is provided by the plot in Figure 7. There, the LPS values for the ten prediction subsets are presented for the symmetric, hidden truncation and inverse scale factors version of the Normal distribution, with lines connecting LPS values corresponding to the same subset. The plot highlights the predictive gain of the two skewed models with respect to the symmetric one. It also shows the (slight) edge of the inverse scale factors model. As can be expected, there is quite some variation in the LPS values for the different prediction subsets. For example, the range of LPS values for the inverse scale factors Normal model is [0.547, 0.781].

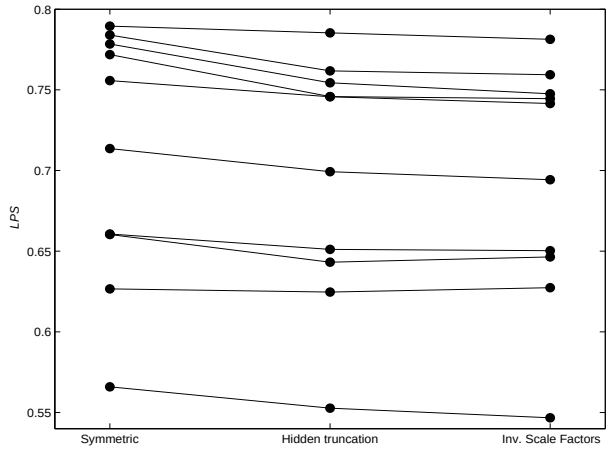

Figure 7: Comparison of the LPS values for the symmetric, hidden truncation and inverse scale factors versions of the Normal distribution. Each line corresponds to a different prediction subset.

8 Conclusion

In this paper we consider model comparison of coordinate-free skewed distribution through Bayesian methods. We discuss various classes of such distributions and use a new class (a subclass of an existing proposal), which is useful for our purposes. The skewed distributions are used as error distributions in a multivariate linear regression framework.

As the coordinate-free class of skewed distributions that we use is based on affine linear transformations of univariate skewed variables, one important issue is how to select appropriate forms for the distributions of these univariate components. We consider a variety of models, differing in the skewing mechanism and in the underlying symmetric distributions, allowing for *e.g.* potentially heavy-tailed distributions.

We propose a Bayesian framework for comparison, where we make sure that the prior distributions on the various models are equivalent. We elicit the priors on model-specific (skewness) parameters through a common implied prior on a skewness measure that can be defined irrespective of moment existence. This elicitation process is exact for some candidate distributions and for other distributions we propose a Kullback-Leibler approximation. We also normalise the distributions of the skewed univariate components, so that the other parameters have the same interpretation across models. Again, the normalisation we use does not rely on moments, but on robust measures of location and spread, so that we do not need to exclude models with very fat tails. These ideas of prior matching and normalisation provide a general framework for comparing multivariate skewed regression models based on the entire class of distributions considered here.

We use two different measures to assess model performance: Bayes factors, to assess within-sample fit, and log predictive scores, to capture out-of-sample predictive ability.

An application to stochastic production frontiers is provided. In this context, we depart from the usual composed error framework with two separate error terms (a symmetric measurement error and a one-sided inefficiency error) in favour of direct modelling through a skewed distribution. This framework carries over immediately to the multi-output case, where we simply use the classes of multivariate skewed distributions mentioned above, and the regression function captures the frontiers. We feel this is an interesting modelling development in the frontier literature, and we intend to pursue further research in this area. The present application is mainly a pilot study, illustrating that such models can be used in practically relevant frontier problems.

The data used consist of 1545 observations on specialised Dutch dairy farms with two separate outputs. In the context of this application both model comparison criteria lead to similar conclusions. The inverse scale factors and the hidden truncation skewing mechanisms are the most adequate and represent a strong improvement over the symmetric alternatives. Results for the main focus of the economic activity of these farms (milk production) are in line with economic theory and with previous results obtained using the composed error framework (where the frontier was shared for both outputs).

This article deals with model comparison based solely on the two model performance criteria mentioned above. In practice, other issues can also be of importance: simplicity and interpretability of the models, and computational ease of the inference. The simplicity of the class of skewed distributions that we analyse does not depend on the particular choices for the distributions of the univariate components. Interpretability favours the inverse scale factors and construct skewing mechanisms, as for these choices skewness only depends on the skewness parameter and not on the properties of the underlying symmetric distribution. In terms of computational ease, the inverse scale factors mechanism

is certainly the easiest to work with. This is due to the fact that, unlike the other methods, it is not necessary to compute values of a probability distribution function.[4] For conducting Bayesian inference with the hidden truncation mechanism, this can be replaced by performing data augmentation in the sampler.

Appendix: Proofs

Proof of Property 1. As the densities s_j are unimodal and have mode at zero, η is a stationary point of $p(\eta|\mu, A, f, \psi)$ if

$$\frac{dp(\eta|\mu, A, f, \psi)}{d\eta} = 0_m \Leftrightarrow (A')^{-1}(\eta - \mu) = 0_m. \tag{10}$$

Since the matrix A is nonsingular, the equation on the left hand-side of (10) is satisfied if and only if $\eta = \mu$.

That $\eta = \mu$ is the unique mode follows directly from the fact that the s_js are unimodal, $j = 1, \ldots, m$. □

Proof of Property 2.

Let η be defined as in (1) and, without loss of generality, let $\mu = 0_m$. Further, let $r_1, \ldots, r_m$ be non-negative integers and a_{ij} denote the element in row i and column j of A. Then,

$$E\left[\prod_{j=1}^{m} \eta_j^{r_j}\right] = E\left[\prod_{j=1}^{m}\left(\sum_{l=1}^{m} a_{lj}\epsilon_l\right)^{r_j}\right]. \tag{11}$$

By the multinomial theorem (11) is equal to

$$E\left[\sum_{j=1}^{m}\sum_{l=1}^{r} d_{lj}\epsilon_j^l\right] = \sum_{j=1}^{m}\sum_{l=1}^{r} d_{lj}E\left[\epsilon_j^l\right], \tag{12}$$

where d_{lj} are constants, $r = \sum_{j=1}^{m} r_j$ and the equality in (12) follows because the ϵ_js are independent, $j = 1, \ldots, m$.

Therefore, if $E\left[\epsilon_j^l\right]$ exists for $j = 1, \ldots, m$ and $l = 1, \ldots, r$ then $E\left[\prod_{j=1}^{m} \eta_j^{r_j}\right]$ also exists, concluding the proof. □

References

Aigner, D., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21-37.

Arellano-Valle, R.B., Genton, M.G., 2003. On fundamental skew distributions. Mimeo, North Carolina State University.

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

[4] For instance, execution times for the Normal models are roughly 1.7 times longer for the hidden truncation method than for the inverse scale method.

Arnold, B.C., Groeneveld, R.A., 1995. Measuring skewness with respect to the mode. The American Statistician 49, 34-38.

Azzalini, A., 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12, 171-178.

Azzalini, A., 2002. Discussion of Arnold, B.C., Beaver, R.J. "Skewed multivariate models related to hidden truncation and/or selective reporting". Test 11, 36.

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

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

van den Broeck, J., Koop, G., Osiewalski J., Steel, M.F.J., 1994. Stochastic frontier models: A Bayesian perspective. Journal of Econometrics 61, 273-303.

Dawid, A.P., 1986. Probability forecasting. In: Kotz, S., Johnson, N.L., Read C.B. (Eds.), Encyclopedia of Statistical Sciences Vol. 7. John Wiley & Sons, New York, pp. 210-218.

Eugene, N., Lee, C., Famoye, F., 2002. Beta-Normal distribution and its applications. Communications in Statistics: Theory & Methods 31, 497-512.

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

Fernández, C., Koop, G., Steel, M.F.J., 2000. A Bayesian analysis of multiple-output production frontiers. Journal of Econometrics 98, 47-49.

Fernández, C., Koop, G., Steel, M.F.J., 2002. Multiple-output production with undesirable outputs: An application to nitrogen surplus in agriculture. Journal of the American Statistical Association 97, 432-442.

Fernández, C., Koop, G., Steel, M.F.J., 2004. Alternative efficiency measures for multiple-output production. Journal of Econometrics, forthcoming.

Fernández, C., Steel, M.F.J., 1998. On Bayesian modeling of fat tails and skewness. Journal of the American Statistical Association 93, 359-371.

Ferreira, J.T.A.S., Steel, M.F.J., 2004a. Bayesian multivariate regression analysis with a new class of skewed distributions. Mimeo, University of Warwick.

Ferreira, J.T.A.S., Steel, M.F.J., 2004b. A constructive representation of univariate skewed distributions. Statistics Research Report 422, University of Warwick.

Ferreira, J.T.A.S., Steel, M.F.J., 2004c. Discussion of M.C. Jones, "Families of distributions arising from distributions of order statistics". Test, forthcoming.

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

Good, I.J., 1952. Rational decisions. Journal of the Royal Statistical Society: Series B 14, 107-114.

Gupta, A.K., González-Farías, G., Domínguez-Molina, J.A., 2004. A multivariate skew normal distribution. Journal of Multivariate Analysis 89, 181-190.

Hoggart, C.J., Walker, S.G., Smith, A.F.M., 2003. Bivariate kurtotic distributions of garment fibre data. Applied Statistics 52, 323-335.

Jones, M.C., 2004. Families of distributions arising from distributions of order statistics. Test, forthcoming.

Jones, M.C., Faddy, M.J., 2003. A skew extension of the t-distribution, with applications. Journal of the Royal Statistical Society: Series B 65, 159-174.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773-795.

Kelker, D., 1970. Distribution theory of spherical distributions and a location-scale parameter generelization. Sankhy$\bar{a}$ 32, 419-430.

Kumbhakar, S. (1996). Efficiency measurement with multiple outputs and multiple inputs. Journal of Productivity Analysis 7, 225-255.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519-530.

Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed errors. International Economic Review 8, 435-444.

Nakatsuma, T., 2003. Bayesian analysis of two-piece Normal regression models. 2003 Proceedings of the Joint Statistical Meetings. American Statistical Association, Alexandria, VA, pp. 2989-2996.

Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). Journal of the Royal Statistical Society: Series B 56 ,3-48.

Reinhard, S., Lovell, C.A.K., Thijssen, G., 1999. Econometric application of technical and environmental efficiency: An application to Dutch dairy farms. American Journal of Agricultural Economics 81, 44-60.

Sahu, S., Dey, D.K., Branco, M.D., 2003. A new class of multivariate skew distributions with applications to Bayesian regression models. The Canadian Journal of Statistics 31, 129-150.

Sarabia, J.M., 2002. Discussion of Arnold, B.C., Beaver, R.J. "Skewed multivariate models related to hidden truncation and/or selective reporting". Test 11, 48-52.