Statistical Methods and Applications manuscript No.
(will be inserted by the editor)

Objective Bayesian analysis for the multivariate skew-t model

Received: date / Accepted: date

Abstract We propose a novel Bayesian analysis of the p-variate *skew-t* model, providing a new parameterization, a set of non-informative priors and a sampler specifically designed to explore the posterior density of the model parameters. Extensions, such as the multivariate regression model with skewed errors and the stochastic frontiers model, are easily accommodated. A novelty introduced in the paper is given by the extension of the bivariate *skew-normal* model given in Liseo & Parisi (2013) to a more realistic p-variate *skew-t* model. We also introduce the R package `mvst`, which produces a posterior sample for the parameters of a multivariate *skew-t* model.

Keywords Multivariate *skew-t* model · Multivariate *skew-normal* model · Objective Bayes inference · Population Monte Carlo sampler · skewness.

1 Introduction

In the last two decades there has been an explosion of interest around the possibility of constructing models which generalize the Gaussian distributions in terms of skewness and extra-kurtosis. Interest can be partially explained with the empirical observations of phenomena, in different disciplines, which could not be easily represented via Gaussian distributions. See Genton (2004) and Azzalini (2014) for general accounts. In this perspective, different proposals of skew-Student t distributions have been proposed and now they play a prominent role as empirical models for heavy-tailed data, particularly in finance (Rachev et al., 2008).

Among the various proposals we mention the *skew-t* distribution obtained as a scale mixture of skew-normal densities (Azzalini & Capitanio, 2003); the "two-piece" t distributions of Hansen (1994) and Fernandez & Steel (1999); the *skew-t* distribution arising from a conditioning argument (Branco & Dey, 2001; Azzalini & Capitanio, 2003); the *skew-t* distribution of Jones & Faddy (2003), obtained by transforming a beta random variable, and the *skew-t* distribution arising from a *sinh* − *arcsinh* transformation (Rosco et al., 2011). In practice, the most used of these are the Azzalini-type *skew-t* distribution, in the form arising from scale mixing Azzalini's *skew-normal* distribution (Azzalini & Capitanio, 2003) and the "two-piece" *t* distribution.

In the paper we will concentrate on the Azzalini-type *skew-t* distribution. For a Bayesian analysis of the "two-piece" *t* distribution one can refer to Rubio et al. (2015) and Leisen et al. (2016) where a new objective prior is introduced for the degrees of freedom parameter.

Following Azzalini (2014), that version of the multivariate *skew-t* distribution can be obtained as a scale mixture of multivariate *skew-normal* distributions. Let $W_0 \sim SN_p(0, \Omega, \alpha)$, where Ω is the correlation matrix of the multivariate normal density appearing in the density of W_0, and $V \sim \Gamma(\nu/2, \nu/2)$.

Let $W = V^{-\frac{1}{2}} W_0$; integrating out V, one obtains the density of a p-variate *skew-t* random vector as

$$f_W(w; \alpha, \Omega, \nu) = 2 t_p(w; \nu) T_1 \left(\alpha' w \left(\frac{\nu + p}{Q_w + \nu} \right)^{1/2} ; \nu + p \right), \tag{1}$$

where $Q_w = w' \Omega^{-1} w$, $t_p(\cdot, \nu)$ is the density of a p-dimensional Student t random variable with ν degrees of freedom, and $T_1(\cdot, s)$ is the cumulative distribution function of a scalar Student t with s degrees of freedom.

The joint estimation of the skewness vector α and the degrees of freedom parameter ν is hard even in the scalar case. For the symmetric Student's t distribution, it is known that the likelihood function tends to infinity when ν goes to zero (Fernandez & Steel, 1999). Fonseca & al. (2008) gave a condition for the existence of the MLE of ν in that case. For the *skew-t* distribution, an approach based on the deviance function is discussed and implemented in Azzalini & Genton (2008); it is based on the idea of replacing the MLE of (α, ν), by the smallest value (α^*, ν^*) such that $H_0 : (\alpha, \nu) = (\alpha^*, \nu^*)$ is not rejected by a likelihood ratio test statistic based on the χ_{p+1} distribution at a fixed level of significance.

Alternatively, the modified score function approach has been applied to the *skew-t* distribution by Sartori (2006), although no proof of the finiteness of the resulting shape estimator has been provided; besides, this method requires the degrees of freedom parameter ν to be fixed. Branco et al. (2011) provides an objective Bayesian solution to this problem in the scalar case.

In this paper we propose a method which generalizes both the results in Branco et al. (2011) and Liseo & Parisi (2013). In fact we describe a Bayesian analysis of the p-variate *skew-t* (*ST*) model, providing a parameterization, a set of non-informative priors and a sampler specifically designed to explore the posterior density of the parameters of the model. Extensions of the model, such as the multivariate regression model with skewed errors and the stochastic frontiers model, are straightforward.

The main novelty of the present paper is given by the extension of the bivariate *skew-normal* (*SN*) model given in Liseo & Parisi (2013) to a more realistic p-variate *ST* model. Several issues arise in this extension, the most important of which is related to the elicitation of the prior distribution for the shape parameter and the sampling strategy for an additional set of latent variables.

This paper also introduces the R (R Core Team, 2015) package `mvst`, which is available in the CRAN repository.
Several other packages are available for dealing with skew-symmetric distributions; among others, the R packages `sn` (Azzalini, 2015), `EMMIXuskew` (Lee & McLachlan, 2013), `mixsmsn` (Prates et al., 2013) and the Stata (StataCorp., 2015) suite of commands `st0207` by Marchenko & Genton (2010): however, most of them rely upon the frequentist approach.

The rest of the paper is organized as follows: the second section introduces the model and the notation, along with the complete likelihood function and complete maximum likelihood estimators. It finally provides the prior distributions and the proof that the posterior distribution is proper.
The third section introduces the sampler and describes a set of proposal distributions. Results from a simulation study are given in section four.

Throughout the paper, we will switch between three different parameterizations, characterized by the sets of parameters $\theta_\star$, $\theta^\star$ and θ; the former allows us to provide the proofs of our main results, the second one is the most sensible to elicit the prior distributions, while the latter is useful for the sampling strategy.
All the above parameterizations have been already proposed in the literature. The interested reader can see Cancho et al. (2010), Cabral et al. (2012), Ho & Lin (2010) and Lachos et al. (2010).

2 The model

The density of the multivariate *skew-t* random vector has been given in (1). For inferential purposes it is often necessary to introduce location and scale parameters, via the transformation $Y = \xi + \omega W$. We then finally say that a random vector Y is distributed as a p-variate *skew-t* distribution, denoted by $Y \sim ST_p(\xi, \alpha, \Sigma, \nu)$, if its pdf is given by

$$f_Y(y; \xi, \alpha, \Sigma, \nu) = 2\, t_p(y;\nu)\, T_1\left(\alpha'\omega^{-1}(y-\xi)\left(\frac{\nu+p}{Q_y+\nu}\right)^{1/2}; \nu+p\right), \qquad (2)$$

where ξ and α are p-dimensional location and shape parameters, ω is a diagonal matrix with the marginal scale parameters, so that $\Sigma = \omega\Omega\omega$ represents the scale matrix and ν represents the number of degrees of freedom. Moreover,

$$\begin{aligned} Q_y &= (y-\xi)'\Sigma^{-1}(y-\xi), \\ t_p(y;\nu) &= \frac{\Gamma((\nu+p)/2)}{|\Sigma|^{1/2}(\pi\nu)^{p/2}\Gamma(\nu/2)}(1+Q_y/\nu)^{-(\nu+p)/2}. \end{aligned}$$

There exists a useful stochastic representation of the random vector Y which is given in the following proposition.

Proposition 2.0.1 Let

$$\delta = \frac{1}{(1+\alpha'\Omega\alpha)^{1/2}}\Omega\alpha$$

and let $I_A(\cdot)$ be the indicator function of the set A; define

$$\begin{pmatrix} Z \\ X \end{pmatrix} \sim N_{p+1}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \delta^T \\ \delta & \Omega \end{pmatrix}\right], \tag{3}$$

and

$$U = (-1)^{I_{(-\infty,0)}(Z)}X, \qquad V \sim \Gamma(\nu/2, \nu/2),$$

with V independent of U. Then, (a) the random vector

$$Y = \xi + \omega U V^{-1/2} \sim ST_p(\xi, \alpha, \Sigma, \nu)$$

and (b) the joint density of (Y, Z, V) is given by

$$\begin{aligned} f_{p+2}(y,z,v) = f_p(y \mid z, v) f(z) f(v) = N_p\left(\xi + \omega\delta\frac{|z|}{\sqrt{v}}, \frac{1}{v}\omega(\Omega - \delta\delta')\omega\right) \\ \times\, N_1(z, 0, 1) \times \Gamma(v, \nu/2, \nu/2). \end{aligned} \tag{4}$$

Proof: the result is a direct consequence of the definition of the *skew-t* distribution. Details can be found in Appendix A.

2.1 Augmented likelihood function

The above stochastic representation suggests to express the density of a *skew-t* random vector as the marginal density of the augmented vector given in (4).
It is useful to define the parameter vectors $\theta^\star = (\xi, \delta, \Sigma, \nu)$ and $\theta = (\xi, \psi, G, \nu)$, where

$$\begin{aligned} \psi &= \omega\delta, \\ G &= \omega(\Omega - \delta\delta')\omega = \Sigma - \psi\psi'. \end{aligned} \tag{5}$$

Using the new parameterization θ, and in the presence of a sample of n i.i.d. observations y_i from a p-dimensional $ST(\xi, \psi, G, \nu)$ r.v., the augmented likelihood function is

$$\begin{aligned} L(\theta; y, z, v) \propto & \prod_{i=1}^{n}\left\{\phi_p\left(y_i - \xi - \psi\frac{|z_i|}{\sqrt{v_i}}; \frac{1}{v_i}(\Sigma - \psi\psi')\right)\right. \\ & \left.\times\ \phi_1(z_i; 1) \times \Gamma\left(v_i; \frac{\nu}{2}, \frac{\nu}{2}\right)\right\} = \\ & = \frac{\prod_{i=1}^{n} v_i^{p/2}}{|G|^{\frac{n}{2}}} \frac{(\nu/2)^{(n\nu/2)}}{(\Gamma(\nu/2))^n}\left(\prod_{i=1}^{n} v_i\right)^{\nu/2-1} \\ & \times\ \exp\left\{-\nu/2\sum_{i=1}^{n} v_i\right\}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right\} \\ & \times\ \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} v_i\varepsilon_i' G^{-1}\varepsilon_i\right\}, \end{aligned} \tag{6}$$

where $z = (z_1, \ldots, z_n)'$, $v = (v_1, \ldots, v_n)'$, $\varepsilon_i = y_i - \xi - \psi\frac{|z_i|}{\sqrt{v_i}}$.

2.1.1 Complete maximum likelihood estimators

The complete maximum likelihood (CML hereafter) estimators are obtained *as if* we had observed the values of the latent variables Z_i's and V_i's. We will make use of the CML estimatates for the initialization of the sampling strategy, described below. They incorporate an additional piece of information, hence they could also be useful as a benchmark to evaluate and compare different estimators in a simulation experiment.

Given z and v, the likelihood (6) gets transformed into

$$\begin{aligned}L(\theta;y,z,v) \propto\ & |G|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} v_i\, \varepsilon_i'\, G^{-1}\, \varepsilon_i\right\} \\ & \times \frac{(\nu/2)^{(n\nu/2)}}{(\Gamma(\nu/2))^n}\left(\prod_{i=1}^{n} v_i\right)^{\nu/2-1} \exp\left\{-\nu/2\sum_{i=1}^{n} v_i\right\}\end{aligned}$$

After straightforward calculations, the CML estimators are obtained as:

$$\begin{aligned}\hat{\psi}_{CML} &= \frac{1}{(\sum_{i=1}^{n} z_i^2)(\sum_{i=1}^{n} v_i) - (\sum_{i=1}^{n} |z_i|\sqrt{v_i})^2} \\ &\times \left[\left(\sum_{i=1}^{n} v_i\right)\left(\sum_{i=1}^{n} |z_i|\sqrt{v_i} y_i\right) - \left(\sum_{i=1}^{n} |z_i|\sqrt{v_i}\right)\left(\sum_{i=1}^{n} v_i y_i\right)\right], \\ \hat{\xi}_{CML} &= \frac{1}{\sum_{i=1}^{n} v_i}\left[\left(\sum_{i=1}^{n} v_i y_i\right) - \hat{\psi}_{CML}\left(\sum_{i=1}^{n} |z_i|\sqrt{v_i}\right)\right], \\ \hat{G}_{CML} &= \frac{1}{n}\sum_{i=1}^{n} v_i\, \hat{\varepsilon}_i\, \hat{\varepsilon}_i',\end{aligned}$$

where

$$\hat{\varepsilon}_i = y_i - \hat{\xi}_{CML} - \hat{\psi}_{CML}\frac{|z_i|}{\sqrt{v_i}}.$$

The estimator for ν have not a closed form expression: it is the solution of the following equation

$$n \log(\hat{\nu}_{CML}/2) - n\,\psi(\hat{\nu}_{CML}/2) = \sum_{i=1}^{n} v_i - \sum_{i=1}^{n} \log(v_i) - n,$$

where $\psi(\cdot)$ denotes the digamma function.

2.2 Prior distributions

We assume the following prior structure for the parameters

$$\pi(\theta^\star) = \pi(\xi)\pi(\delta,\Sigma)\pi(\nu).$$

As pointed out in Liseo & Parisi (2013), when $p > 1$, even following an objective Bayesian approach, δ and Σ cannot be considered a priori independent of each other. This depends on eq. (5); in fact, if $G = \omega(\Omega - \delta\delta')\omega$, in order to guarantee the positive definiteness of G, one should consider, both in the analytical expression and in the computations, the constraint $\Omega - \delta\delta' \succ 0$.
We further consider the decomposition

$$\pi(\delta, \Sigma) = \pi(\delta|\Sigma)\pi(\Sigma)$$

and we assume a flat prior for ξ and a conjugate Inverse Wishart prior for Σ. This way we adopt the "usual" objective priors for the location and scale parameters as in the multivariate normal model, which is nested in the multivariate ST model, as $\delta = 0$ and $1/\nu \to 0$. In practice, we set

$$\pi(\xi) \propto 1$$
$$\Sigma \sim IW(m, \Lambda)$$

In real applications, we will take $m = 0$ and $\Lambda = 0$. In §2.3, we prove that the use of an improper prior on (ξ, Σ) produces proper posterior distributions, provided that the prior on the degrees of freedom parameter ν is proper and discrete over $\mathbb{N}$. Accordingly, we assume a uniform prior for ν over a set of 20 non-equidistant values ranging from 1 to 100. A recent alternative objective prior for the multivariate symmetric t model is proposed in Villa & Rubio (2017).
Finally, we need to specify $\pi(\delta|\Sigma)$. For each value of Σ, the parameter δ lies in a p-dimensional region whose shape only depends on Σ, or, more specifically, on Ω. In particular, given the expression of δ, it is easy to verify that

$$\delta'\Omega^{-1}\delta < 1, \tag{7}$$

must hold, so the conditional parameter space is an ellipsoid, say Δ_Σ, given by expression (7), centered at the origin and contained in the hyper-cube $(-1, 1)^p$. In any simulation based approach care must be taken that the proposed values actually satisfy (7). For computational convenience we prefer to directly include this constraint on the prior. In the bivariate case, Liseo & Parisi (2013) used an approximation of the Jeffreys' prior, normalized over Δ_Σ. This normalization step, for large p, may become computationally demanding. For this reason, we propose to adopt a uniform prior over Δ_Σ, whose volume can be evaluated in a closed form, so that the normalizing constant is analytically tractable. Then we assume:

$$(\delta|\Sigma) \sim U(\Delta_\Sigma),$$

that is

$$\pi(\delta|\Sigma) = \left(\frac{\pi^{p/2}}{\Gamma(p/2+1)} \sqrt{|\Omega|} \right)^{-1}.$$

In the practical application of the ST model, we will use the θ parameterization for our sampling strategy. Hence, we need to compute the Jacobian of the transformation $\theta^\star \to \theta$, which is given by

$$|J| = \prod_{j=1}^{p} (G_{jj} + \psi_j^2)^{-1/2}.$$

2.3 Posterior propriety

Proposition 2.3.1 The posterior distribution of the model is proper.

Proof: Let $\theta_\star = (\xi, \alpha, \Sigma, \nu)$, using the parameterization in (2),

$$\pi(\theta_\star|y) \propto \pi(\theta_\star)\prod_{i=1}^{n}\Bigg[2t_p(y_i;\nu) \\ \times\ T_1\left(\alpha'\omega^{-1}(y_i-\xi)\left(\frac{\nu+p}{Q_{y_i}+\nu}\right)^{1/2};\nu+p\right)\Bigg].$$

Since the c.d.f. $T_1(\cdot)$ is bounded by 1, one obtains

$$\pi(\theta_\star|y) \leq \bar{\pi}(\theta_\star|y) = \pi(\xi)\pi(\Sigma)\pi(\alpha|\Sigma)\pi(\nu)\prod_{i=1}^{n}[2t_p(y_i;\nu)].$$

Notice that the parameter α only appears in the prior distribution; then it can be integrated out to obtain

$$\bar{\pi}(\xi,\Sigma,\nu|y) \propto \pi(\xi)\pi(\Sigma)\pi(\nu)\prod_{i=1}^{n}[2t_p(y_i;\nu)].$$

The above expression is proportional to the posterior density of the parameters of a multivariate Student-t model, with priors given as in §2.2. Theorem 1 in Fernandez & Steel (1999) then guarantees that the posterior distribution of our *ST* model is proper as soon as the prior on ν is proper and $n \geq p+1$ except, possibly, for a set of Lebesgue measure zero in $\mathbb{R}^{n\times p}$. The finite precision of the data recording process can lead, under some choices for the prior distributions, to improper posterior distributions. However, it is possible to verify this condition for any given dataset, and we refer to the cited article for details.

3 The sampler

In the following, we describe the sampling strategy. We have used a Population Monte Carlo algorithm (PMC hereafter, see Cappé et al., 2004), which improves and generalizes the one used in Liseo & Parisi (2013) for the bivariate *SN* model.
As a Monte Carlo method, the PMC sampler doesn't rely on convergence arguments, hence it can overcome the problem of multimodality of the posterior distribution; moreover, it offers a great flexibility in choosing the proposal density functions. For example, we use (approximations of) the full conditional distributions as proposal densities.

The outline of the algorithm for the *ST* model is as follows:

- At iteration 0, a population of N particles $\eta^{(0)}_{1:N}$, containing the values of $\theta^{(0)}_{1:N}$, $z^{(0)}_{1:N}$ and $v^{(0)}_{1:N}$, is initialized. A possible initialization is described in §3.1.

- At a generic iteration t
  - new values for the particles are proposed following a proposal distribution $q(\eta^{(t)})$, whose parameters possibly depend on the populations of particles in the previous iterations,
  - the importance weights are computed as

$$\begin{aligned}\tilde{\zeta}_j^{(t)} &= \tilde{\pi}(\eta_j^{(t)}|y)/q(\eta_j^{(t)})\\ \zeta_j^{(t)} &= \tilde{\zeta}_j^{(t)}/\sum_{j=1}^{N}\tilde{\zeta}_j^{(t)}\end{aligned}$$

    where $\tilde{\pi}$ and $\tilde{\zeta}$ denote the unnormalized posterior density function and importance weights.
  - A set of quantities are obtained on the basis of the current particles and weights. This set includes the estimates of the parameters $\eta^{(t)}$, a quantity related to the performance of the sampler in the t-th iteration

$$H^{(t)} = -\sum_{j=1}^{N}\zeta_j^{(t)}\log(\zeta_j^{(t)}),$$

    and all the other objects of interest.
  - the particles $\eta_{1:N}^{(t)}$ are multinomially resampled using the weigths $\zeta^{(t)}$.
- After T iterations, the final estimates are obtained as a weighted mean of the estimates $\tilde{\eta}^{(1:T)}$ with (unnormalized) weights given by $H^{(1:T)}$.

The multivariate *skew-t* family of distributions contains, as special cases, the multivariate normal and t family and the multivariate skew-normal family. This implies that, in practical data analysis, one could be often interested in comparing the fit of the above nested families. From this perspective, a quantity of particular interest, which can be easily obtained using the PMC algorithm, is the marginal likelihood of each model. In fact, it can be estimated as

$$\hat{p}(y) \approx \frac{\sum_{t=1}^{T} H^{(t)}\sum_{j=1}^{N}\tilde{\zeta}_j^{(t)}}{N\sum_{t=1}^{T} H^{(t)}}. \tag{8}$$

Since the PMC algorithm produces a sample of particles from the posterior distribution of η, any other Bayesian model adequacy measure can be easily computed from the PMC output. See, for example, the Bayesian chi-square test (Li et al., 2015) or the Bayesian p-value (Zhang, 2014).

In the simulation and applications sections of this paper, we confine ourselves to the analysis of the Bayes factor and the log-pseudo marginal likelihood (Geisser & Eddy, 1979), which is based on the computation Conditional Predictive Ordinate as described in Cancho et al. (2010). The computation of the log-pseudo marginal likelihood is usually performed from a MCMC output; here we adapt the method to our PMC posterior sample.

3.1 Initial values for parameters

The initial points are sampled by mimicking the stochastic representation of the model. Then

1. the values of $\nu_{1:N}^{(0)}$ are sampled from the prior distribution;
2. given $\nu_{1:N}^{(0)}$ the values of the latent variables $z_{1:N}^{(0)}$ and $v_{1:N}^{(0)}$ are sampled by the respective sampling distributions described in Proposition 2.0.1;
3. given $\nu^{(0)}$, $z_{1:N}^{(0)}$ and $v_{1:N}^{(0)}$, the parameters $\xi_{1:N}^{(0)}$, $\psi_{1:N}^{(0)}$ and $G_{1:N}^{(0)}$ are obtained as the CML estimates of the parameters, as described in §2.1.1.

3.2 Proposals

For the common parameters of *SN* and *ST* models, the proposal distributions are similar to those reported in Liseo & Parisi (2013); our versions are given in appendix B. The *ST* model, however, also includes the parameter ν and the latent variables V_i's. The parameter ν assumes values on a finite set, hence it is easy to simulate from its full conditional distribution

$$\pi(\nu|\cdots) \propto \frac{(\nu/2)^{n\nu/2}}{(\Gamma(\nu/2))^n}\left(\prod_{i=1}^{n} v_i\right)^{\frac{\nu}{2}-1} \exp\left\{-\frac{\sum_{i=1}^{n} v_i}{2}\nu\right\}.$$

Instead, to our knowledge, there are no simple methods to draw values from the full conditional distribution of V_i, which is given by

$$\pi(v_i|\cdots) = \frac{1}{k_{v_i}} v_i^{C-1} \exp\{-A_i v_i - B_i\sqrt{v_i}\}, \qquad v_i > 0$$

where

$$\begin{aligned} A_i &= 0.5[\nu + (y_i-\xi)'G^{-1}(y_i-\xi)] \\ B_i &= -(y_i-\xi)'G^{-1}\psi|z_i| \\ C &= (\nu+p)/2 \end{aligned}$$

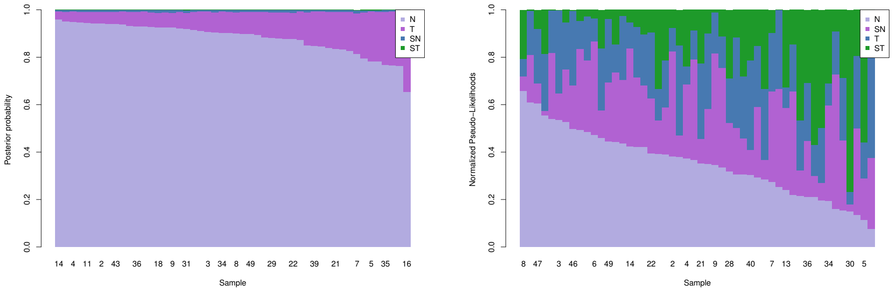
and k_{v_i} is the normalizing constant.
When $B_i = 0$ (for example in the symmetric case, where ψ is a null vector), then the full conditional distribution of v_i is a Gamma distribution. Otherwise, the sign of B_i determines the right tail behavior: when B_i is positive (negative), the right tail of the full conditional distribution is thicker (lighter) than the right tail of a Gamma distribution.
Hence, we cannot propose values from a Gamma distribution, as it could jeopardize the validity of the method when $B_i < 0$. On the other hand, proposing from a distribution with a thick tail could represent a huge loss in the efficiency of the sampler. For these reasons, we propose values using a rejection sampler (see, for example, Robert & Casella, 2004, §2.3) having the full conditional distribution as target density. We will

1. define the distribution of the instrumental variable of the rejection sampler;
2. choose the parameters of this distribution by minimizing the Kullback-Leibler divergence with respect to the target distribution;
3. obtain the constant M required by the rejection sampling algorithm;
4. obtain the normalizing constant k_{v_i}, required by the PMC algorithm.

Details are as follows:

1. define $W = R^2$, with $R \sim \Gamma(\alpha_v, \beta_v)$; the instrumental density function is
$$f(w|\alpha_v, \beta_v) = \frac{\beta_v^{\alpha_v}}{2\Gamma(\alpha_v)} w^{\alpha_v/2-1} \exp(-\beta_v \sqrt{w});$$
this density has a right tail which is thicker than the one of the target distribution;
2. If we set $\alpha_v^\star = 2C$ (see Appendix C), the $KL(f||\pi_{v_i})$ divergence, as a function of β_v, has a minimum (in $\mathbb{R}^+$) in
$$\beta_v^\star = \frac{1}{2}\left(B_i + \sqrt{B_i^2 + 8A_i(2C+1)}\right).$$
Using the parameters $\alpha_v^\star$ and $\beta_v^\star$ we will optimise the efficiency of the rejection sampler.
3. the Rejection Sampling algorithm requires a constant M for which
$$\pi(v_i|\cdots) \leq Mf(v_i).$$
The value of M can be found by defining the ratio $m(v_i) = \pi(v_i|\cdots)/f(v_i)$; given the parameters of the instrumental density, this function has a maximum in
$$v_i^\star = \left(\frac{\beta_v^\star - B_i}{2A_i}\right)^2.$$
The value of M can be finally obtained as $m(v_i^\star)$.
4. To obtain the value of
$$k_{v_i} = \int_{\mathbb{R}^+} v_i^{C-1} \exp\{-A_i v_i - B_i \sqrt{v_i}\}\, dv_i,$$
we use eq. 3.462 (1) in Gradshteyn & Ryzhik (1994, GR hereafter), with $\nu = 2C > 0$, $\beta = A_i > 0$, $\gamma = B_i$,
$$k_v = 2(2A_i)^{-C}\Gamma(2C)\exp\left\{-\frac{B_i^2}{8A_i}\right\} D_{-2C}\left(\frac{B_i}{\sqrt{2A_i}}\right)$$
where $D_p(z)$ is the parabolic cylinder function (GR, eq. 9.240) with $p = -2C$ and $z = B_i/\sqrt{2A_i}$, hence
$$\begin{aligned} D_{-2C}\left(\frac{B_i}{\sqrt{2A_i}}\right) = & \left[\frac{\sqrt{\pi}}{\Gamma\left(\frac{1+2C}{2}\right)} \Upsilon\left(C, \frac{1}{2}; \frac{B_i^2}{4A_i}\right) - \frac{\sqrt{2\pi}\frac{B_i}{\sqrt{2A_i}}}{\Gamma(C)} \Upsilon\left(\frac{1+2C}{2}, \frac{3}{2}; \frac{B_i^2}{4A_i}\right)\right] \\ & \times 2^{-C} \exp\left\{\frac{B_i^2}{8A_i}\right\} \end{aligned} \tag{9}$$
where $\Upsilon(\alpha, \gamma; z)$ denotes the confluent hypergeometric function (GR, eq. 9.210).

4 Simulation study



(a) Posterior probabilities.

(b) Pseudo marginal likelihoods.

Fig. 1 Simulation results for the normally distributed samples.

In this section we use simulated data to evaluate the performance of the proposed approach. Since the multivariate *ST* model may be considered an encompassing model including, as special cases, the multivariate Student-*t* model, the multivariate *SN* model and the multivariate normal one, it is of primary importance to verify the ability of the proposed approach to discriminate among these nested models.

For each of the four models, we have generated 50 samples; for each sample, we compute the posterior probabilities and the the pseudo marginal likelihoods of each candidate model. The posterior probabilities are estimated using equation (8) together with a uniform prior over the model space.

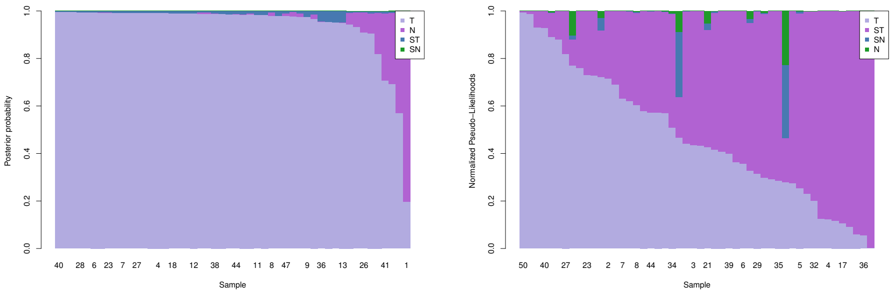
In our simulations, each sample consists of $n = 200$ observations with $p = 4$ and

$$\xi = (5,9,3,10)', \qquad \Sigma = \begin{pmatrix} 7 & 2 & 1 & 1 \\ 2 & 8 & -2 & 3 \\ 1 & -2 & 5 & -2 \\ 1 & 3 & -2 & 8 \end{pmatrix}.$$

Samples from the *SN* and *ST* models have been generated using $\alpha = (4,4,4,4)'$, which corresponds to a significant - although not extreme - amount of skewness. Finally, we set $\nu = 10$ to generate data from the Student-*t* and *ST* models.

For each sample we have run the PMC algorithm using 10000 particles for each of 5 iterations. Results are summarized in the following plots. Barplots in Fig. 1 depicts the results for the Gaussian case. Panel (a) reports the results in terms of the approximate Bayes factor (8) and the induced posterior model probabilities. Panel (b) reports the normalized weights of each model provided by the pseudo marginal likelihood.

Each column in each barplot stacks the weights of the 4 candidate models estimated from a single sample. To improve the readability of the plot, bars have been rearranged in order to have a decreasing weight for the true model.



(a) Posterior probabilities.

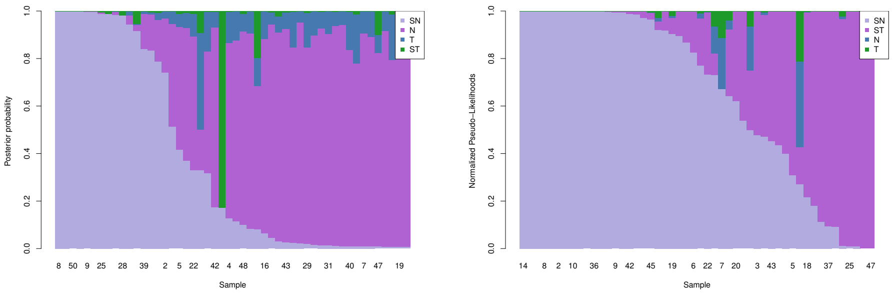
(b) Pseudo marginal likelihoods.

Fig. 2 Simulation results for the Student-*t* distributed samples.

In the first simulation, where data are normally distributed, the Bayes factors always identify the true model. They also assign smaller probabilities to the Student-*t* model and negligible ones for the remaining models.

The situation is even more extreme in the second simulation, where data come from a Student-*t* distribution (Fig. 2). Even if the Bayes factor fails in one sample, it assigns very high weights to the real model in all the other cases. All the other models generally obtain very small probabilities.

In both simulations, the pseudo marginal likelihood recognize the real model for 26 samples. It also assign higher weights to models which are less parsimonious than the real one.



(a) Posterior probabilities.

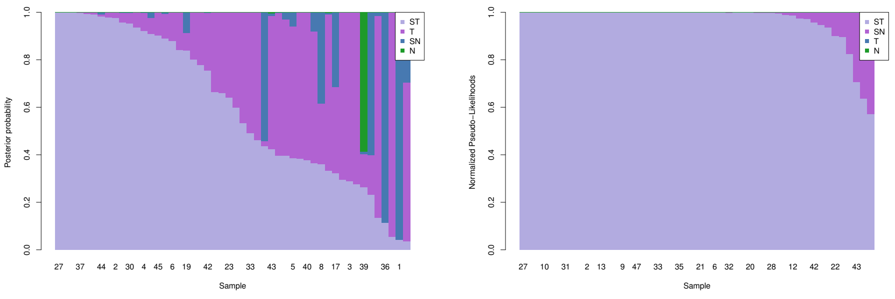
(b) Pseudo marginal likelihoods.

Fig. 3 Simulation results for the *SN*-distributed samples.

The Bayes factors show a lower performance when samples come from a skewed distribution. The worst case happens when the data are generated from a *SN* distribution, which is notoriously the most difficult to deal with because of the multimodality phenomenon, described in Liseo & Parisi (2013). In Fig. 3, it is possible to notice that the procedure detects the correct model in 17 cases, and it most often prefers the multivariate normal model.

Also in the *ST* case (Fig. 4), the true model has been correctly identified in 27 cases. In almost all the other cases, the *T* model has been preferred.

In these simulations, the pseudo marginal likelihood perform better than the Bayes factors. In fact, it recognizes the true model in 33 cases for the *SN* case, giving high weights to the *ST* model in the other cases. It always recognizes the true model in the *ST* case.



(a) Posterior probabilities.

(b) Pseudo marginal likelihoods.

Fig. 4 Simulation results for the *ST*-distributed samples.

As far as the posterior distribution of ν, we report, for each sample, the 95% credible intervals for the estimated Student-*t* and *ST* models. Fig. 5 reports the intervals, divided by generating and estimated model. In order to improve the readability of the plot, intervals are ordered according the lower limit.

When data come from a skewed distribution, the intervals are small and concentrated around the true value. When, instead, data come from symmetric distributions, the intervals are wide and contains all the largest values admitted by the prior distribution.

4.1 The `mvst` package

The simulation results have been obtained in R, using the package `mvst`. It contains functions to estimate the parameters of the *ST* (and nested) models, and to simulate data from them. It uses the model and the proposals described above, even if it allows to define customized prior and proposal distributions.

It makes use of the GNU Scientific Library (see Gough, 2009) to speed up the heaviest parts of the code and, in particular, for the computation of (9). Besides, it requires three R packages: `mvtnorm` (Genz et al., 2015), `MCMCpack` (Martin et al., 2011) and `mnormt` (Azzalini & Genz, 2016). It also makes use of three scripts available in the `RcppGSL` package (Eddelbuettel & Romain, 2013).

5 A real dataset

As a final illustration of the proposed algorithm, we consider the wine data of the Grignolino cultivar, used in §6.2.6 of Azzalini (2014). The dataset contains 71 obser-

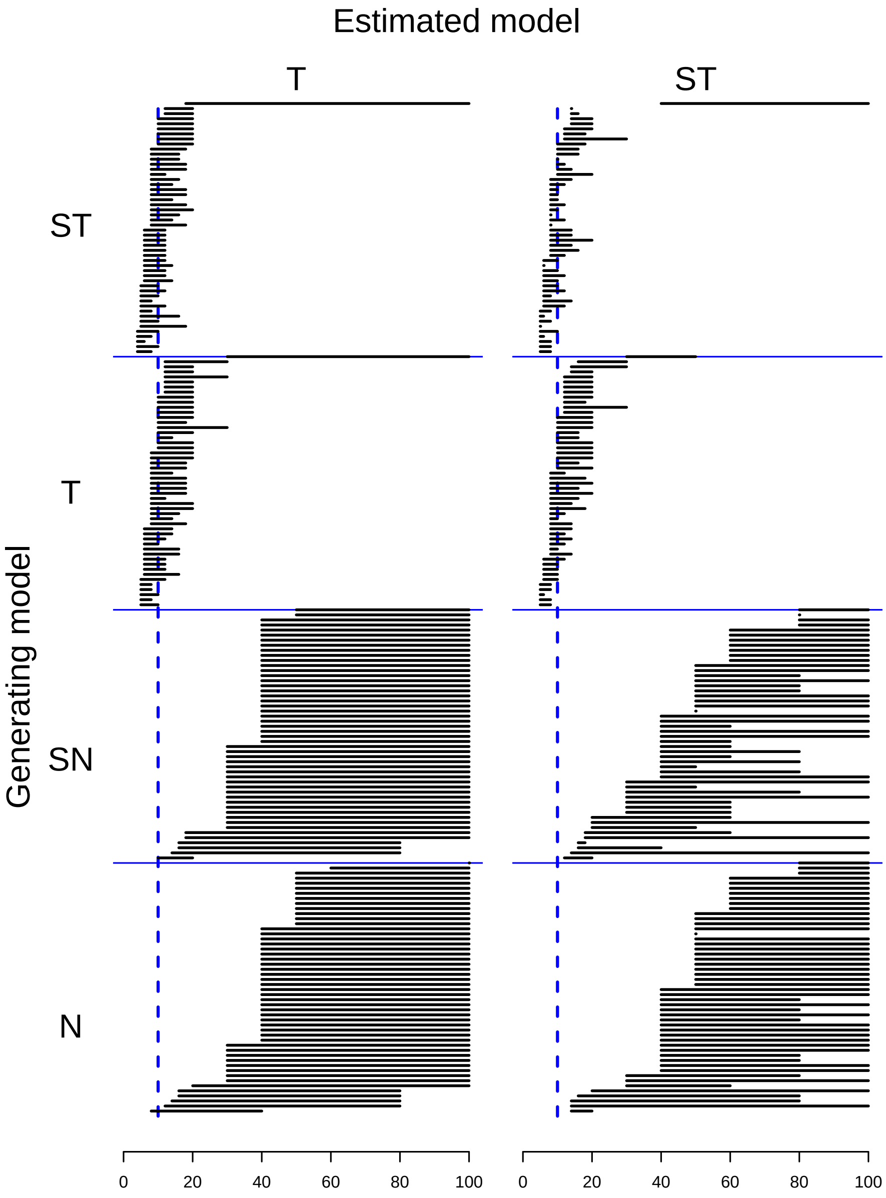


Fig. 5 95% credible intervals for ν.

vations on 3 variables (chloride, glycerol and magnesium). Data are available in the sn package.

We have performed a PMC sampler with 5 iterations, 10000 particles each. The posterior probabilities for the four models are given in Table (1). Models with light tails have negligible probabilities, and the preferred one is the *skew-t* model.

Model	*N*	Student-*t*	*SN*	*ST*
$\hat{\pi}(M\|y)$	8.57e-14	4.20e-01	4.53e-11	5.80e-01

Table 1 Models' posterior probabilities.

Conditionally on this model, the posterior mean of ν is approximately equal to 3.18, while the ML estimate in Azzalini (2014) is equal to 3.4. The 95% credible interval for ν is $[2,4]$.

Figure 6 shows the case deletion diagnostics (see Cancho et al., 2010 for details) for the Grignolino example. The quantities p_i in the histogram take values in the interval $[0.5,1]$; a value of $p_i \gg 0.5$ flags an influential observation.
Admittedly, given the inferred *ST* model, there seems to be two outliers in the dataset; this issue could be addressed by including some adjusting covariates as in Azzalini (2014). We are currently working on a new version of the mvst package which will include the possibility to estimate (possibly multivariate) regression models.

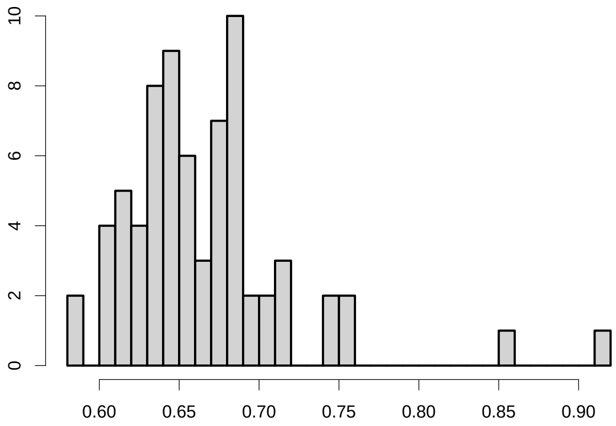


Fig. 6 Case influence diagnostics for the Grignolino example.

A Proof of Proposition 2.0.1

(a): From one of the possible definitions of a multivariate *ST* r.v., it is known that $U \sim SN_p(0,\alpha,\Omega,\nu)$; since Y is a simple transformation of U, its distribution is readily obtained.
(b): Start from $f(y,z,v) = f(v)f(z)f(y \mid z,v)$. By assumption, $f(z)$ is a standard Gaussian density, and

$$(Y \mid Z=z, V=v) = (\xi + \omega U \mid Z=z, V=v) = \begin{cases} \xi + \omega X v^{-1/2} & z \geq 0 \\ \xi - \omega X v^{-1/2} & z < 0 \end{cases}.$$

Then, by using simple results on conditional Gaussian densities, one gets

$$(Y \mid Z = z, V = v) \sim \begin{cases} N_p\left(\xi + \omega\delta\frac{z}{\sqrt{v}}, \frac{1}{v}\omega(\Omega - \delta\delta')\omega\right) & z \geq 0 \\ N_p\left(\xi - \omega\delta\frac{z}{\sqrt{v}}, \frac{1}{v}\omega(\Omega - \delta\delta')\omega\right) & z < 0 \end{cases}$$

Hence the result in (4).

B Proposal distributions

We use the full conditional distributions as proposals for the latent variables Z and ξ: each Z_i has the following full conditional distribution

$$\pi(z_i|\cdots) = \frac{\phi(z_i^+|m_i, v_\theta)}{2(1-\Phi(z_i|m_i, v_\theta))} \tag{10}$$

where

$$\begin{aligned} v_\theta &= (1+\psi' G^{-1}\psi)^{-1} \\ m_i &= v_\theta\sqrt{v_i}(\psi' G^{-1}(y_i - \xi)) \end{aligned}$$

The variables Z_i can be drawn as the product of Z_i^+, a normal r.v. with parameters m_i and v_θ truncated in 0 and the sign S_i, uniform on $\{-1,1\}$. To generate values Z^+ a rejection sampler has been employed (see Robert, 1995).

The parameter ξ has the following full conditional density:

$$(\xi|\cdots) \sim N_p\left(\frac{1}{\sum_{i=1}^n v_i}\left(\sum_{i=1}^n (v_i y_i) - \psi\sum_{i=1}^n \sqrt{v_i}|z_i|\right), \frac{1}{\sum_{i=1}^n v_i} G\right)$$

The parameters ψ and G have untractable full conditional distributions. To obtain a proposal distribution, they are approximated using only the contribution of the likelihood to the full conditional density.
The parameter ψ has the following full conditional distribution

$$\begin{aligned} \pi(\psi \mid \cdots) &\propto \prod_{j=1}^p \left[(G_{jj} + \psi_j^2)^{-1/2}\right] \mathbb{1}_\delta(\Delta_\Sigma) \\ &\times \exp\left\{-\frac{1}{2}\sum_{i=1}^n v_i\left(y_i - \xi - \psi\frac{|z_i|}{\sqrt{v_i}}\right)' G^{-1}\left(y_i - \xi - \psi\frac{|z_i|}{\sqrt{v_i}}\right)\right\}, \end{aligned}$$

where $\mathbb{1}_x(\cdot)$ denotes the indicator function. By ignoring the first two factors, we obtain the following proposal distribution

$$q(\psi) = \phi_p\left(\psi \,\middle|\, \frac{1}{\sum_{i=1}^n z_i^2}\sum_{i=1}^n |z_i|\sqrt{v_i}(y_i - \xi), \frac{1}{\sum_{i=1}^n z_i^2} G\right)$$

The proposal distribution has a positive density on $\mathbb{R}^p$, while the full conditional is bounded on Δ_Σ. This feature improves the ability of the sampler to explore the parameter space; moreover, particles which don't respect the constraint (7) will be automatically discarded, as they have null prior (and posterior) probability density, hence a null importance weight.

The parameter G has the following full conditional density

$$\pi(G|\cdots) \propto \pi(\Sigma)|J||G|^{-n/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(G^{-1}\Xi)\right\}$$

where

$$\Xi = \sum_{i=1}^{n} v_i \left(y_i - \xi - \psi \frac{|z_i|}{\sqrt{v_i}} \right) \left(y_i - \xi - \psi \frac{|z_i|}{\sqrt{v_i}} \right)'.$$

Ignoring the prior term we obtain

$$q(G) = IW(n-p-1, \Xi).$$

C Details about the Rejection Sampler

For a generic latent variable V_i, the Kullback Leibler divergence $KL(f||\pi_v)$ is given by

$$KL(f||\pi_v) = \int_{\mathbb{R}^+} f(v_i) \log \left(\frac{k_v \beta_v^{\alpha_v}}{2\Gamma(\alpha_v)} v_i^{\alpha_v/2-C} \exp\{A_i v_i + (B_i - \beta_v)\sqrt{v_i}\} \right) dv_i$$

which has an analytical solution for $\alpha_v^\star = 2C$:

$$KL(f||\pi_v) = \log \left(\frac{k_v \beta_v^{2C}}{2\Gamma(2C)} \right) + \frac{2C(2C+1)A_i}{\beta_v^2} + 2C \log(\beta_v) + \frac{2CB_i}{\beta} - 2C.$$

This divergence has always one (and only one) minimum in $\mathbb{R}^+$, given by

$$\beta_v^\star = \frac{1}{2}\left(B_i + \sqrt{B_i^2 + 8A_i(2C+1)} \right).$$

References

AZZALINI, A. (2014). *The Skew-Normal and related families*, (with the collaboration of A. Capitanio). Cambridge: Cambridge University Press.

AZZALINI, A. (2015). *The R package* `sn`*: The Skew-Normal and Skew-t distributions (version 1.3-0).* Università di Padova, Italia.

AZZALINI, A. & CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t* distribution. *Journal of the Royal Statistical Society, B*, 65, 367–389.

AZZALINI, A. & GENTON, M. (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76, 106–119

AZZALINI, A. & GENZ, A. (2016). The R package `mnormt`: The multivariate normal and *t* distributions (version 1.5-4). *http://azzalini.stat.unipd.it/SW/Pkg-mnormt*

BRANCO, M. D. & DEY, D. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 77, 1–15.

BRANCO, M.D., GENTON, M.G. & LISEO, B. (2011). Objective Bayesian Analysis of Skew-t Distributions. *Scandinavian Journal of Statistics*, 40 (1), 63–85

CABRAL, C.R.B., LACHOS, V.H. & PRATES, M.O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Comput. Statist. Data Anal.*, 56(1):126–142.

CANCHO, V.G., ORTEGA, E.M.M. & PAULA, G.A.(2010) On estimation and influence diagnostics for log-Birnbaum-Saunders Student-*t* regression models: full Bayesian analysis. *J. Statist. Plann. Inference*, 140(9):2486–2496.

CAPPÉ, O., GUILLIN, A., MARIN, J. M. & ROBERT, C. P. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* **13**, 907–929.

EDDELBUETTEL, D. & ROMAIN, F. (2013). RcppGSL: 'Rcpp' Integration for 'GNU GSL' Vectors and Matrices. http://CRAN.R-project.org/package=RcppGSL.

FERNANDEZ, C. & STEEL, M. F. J. (1999). Multivariate Student-*t* regression models: pitfalls and inference. *Biometrika* **86**, 153–167.

FONSECA, T. C., FERREIRA, M. A. R. & MIGON, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika* **95**, 325–333.

GEISSER, S. & EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.*, 74, 153–160.

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

GENZ, A., BRETZ, F., MIWA, T., MI, X., LEISCH, F., SCHEIPL, F. & HOTHORN, T. (2015). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-3.

GOUGH, B. (2009). *GNU Scientific Library Reference Manual - Third Edition*. Network Theory Ltd., (3rd ed.)

GRADSHTEYN, I. S. & RYZHIK, I. M. (1994). *Table of integrals, series, and products*. Academic Press, Inc., Boston, MA, russian ed. Translation edited and with a preface by Alan Jeffrey.

HANSEN, B.E. (1994). Autoregressive conditional density estimation. *Intern. Econ. Rev.*, 35, 3, 705–730.

HO, H.J. & LIN, T.I. (2010). Robust linear mixed models using the skew *t* distribution with application to schizophrenia data. *Biometrical Journal*, 52(4):449–469.

JONES, M.C. & FADDY, M.J. (2003) A Skew Extension of the t-Distribution, with Applications. *Journal of the Royal Statistical Society, B*, 65, 1, 159–174

LACHOS, V.H., GHOSH, P. & ARELLANO-VALLE, R.B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statist. Sinica*, 20(1), 303–322.

LEE, S. X. & MCLACHLAN, G. J. (2013). EMMIXuskew: An R Package for Fitting Mixtures of Multivariate Skew *t* Distributions via the EM Algorithm. *Journal of Statistical Software* **55**, 1–22.

LEISEN, F., MARIN, J.M. & VILLA, C. (2016). Objective Bayesian modelling of insurance risks with the skewed Student-t distribution. *ArXiv e-prints*, 1607.04796.

LI, Y., LIU, X. & YU, J. (2015). A Bayesian chi-squared test for hypothesis testing. *Journal of Econometrics*, 189(1), 54-69.

LISEO, B. & PARISI, A. (2013). Bayesian inference for the multivariate skew-normal model: a population Monte Carlo approach. *Comput. Statist. Data Anal.* **63**, 125–138.

MARCHENKO, Y. & GENTON, M. (2010). A suite of commands for fitting the skew-normal and skew-t models. *Stata Journal* **10**, 507–539. Cited By 2.

MARTIN, A. D., QUINN, K. M. & PARK, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software* **42**, 22.

PRATES, M. O., CABRAL, C. R. B. & LACHOS, V. H. (2013). mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. *Journal of Statistical Software* **54**, 1–20.

R CORE TEAM (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RACHEV, S.T., HSU, J.S.J., BAGASHEVA, B.S. & FABOZZI, F.J. (2008). *Bayesian Methods in Finance*. Wiley, New York.

ROBERT, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5(2):121–125, June 1995.

ROBERT, C. P. & CASELLA, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, 2nd ed.

ROSCO, J.F., JONES, M.C. & PEWSEY, A. (2011). Skew t distributions via the sinh-arcsinh transformation. *TEST*, 20, 3, 630–652.

RUBIO, F.J. & STEEL, M.F.J. (2015). Bayesian modelling of skewness and kurtosis with Two-Piece Scale and shape distributions. *Electron. J. Statist.*, 9, 2, 1884–1912.

SARTORI, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew-t distributions. *J. Statist. Plan. Inference* 136, 4259–4275.

STATACORP. (2015). Stata Statistical Software: Release 14.

VILLA, C. & RUBIO, F.J. (2017). Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula. *ArXiv e-prints*, 1701.05638.

ZHANG, J.L. (2014). Comparative Investigation of Three Bayesian P Values. *Comput. Statist. Data Anal.*, 79, 277–291.