

## On Bayesian Modelling of Fat Tails and Skewness

Fernández, C.; Steel, M.F.J.

*Publication date:*  
1996

[Link to publication](#)

*Citation for published version (APA):*  
Fernández, C., & Steel, M. F. J. (1996). On Bayesian Modelling of Fat Tails and Skewness. (CentER Discussion Paper; Vol. 1996-58). Tilburg: Econometrics.

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

### Take down policy

If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# ON BAYESIAN MODELLING OF FAT TAILS AND SKEWNESS

By Carmen Fernández and Mark F.J. Steel<sup>1</sup>

*CentER for Economic Research and Department of Econometrics  
Tilburg University, 5000 LE Tilburg, The Netherlands*

## Abstract

We consider a Bayesian analysis of linear regression models that can account for skewed error distributions with fat tails. The latter two features are often observed characteristics of empirical data sets, and we will formally incorporate them in the inferential process. A general procedure for introducing skewness into symmetric distributions is first proposed. Even though this allows for a great deal of flexibility in distributional shape, tail behaviour is not affected. In addition, the impact on the existence of posterior moments in a regression model with unknown scale under commonly used improper priors is quite limited. Applying this skewness procedure to a Student- $t$  distribution, we generate a “skewed Student” distribution, which displays both flexible tails and possible skewness, each entirely controlled by a separate scalar parameter. The linear regression model with a skewed Student error term is the main focus of the paper: we first characterize existence of the posterior distribution and its moments, using standard improper priors and allowing for inference on skewness and tail parameters. For posterior inference with this model, a numerical procedure is suggested, using Gibbs sampling with data augmentation. The latter proves very easy to implement and renders the analysis of quite challenging problems a practical possibility. Two examples illustrate the use of this model in empirical data analysis.

KEY WORDS: Gibbs sampling; Improper prior; Linear regression model; Posterior moments; Student- $t$  sampling.

---

<sup>1</sup> Carmen Fernández is Research Fellow, CentER for Economic Research and Assistant Professor, Department of Econometrics, Tilburg University, 5000 LE Tilburg, The Netherlands. Mark Steel is Senior Research Fellow, CentER for Economic Research and Associate Professor, Department of Econometrics, Tilburg University, 5000 LE Tilburg, The Netherlands. We gratefully acknowledge stimulating discussions with James O. Berger and Jean-Francois Richard, and useful comments from an anonymous referee. Both authors benefitted from a travel grant awarded by the Netherlands Organization for Scientific Research (NWO).

# 1. INTRODUCTION

This paper aims at introducing two pervasive features of empirical data into statistical modelling and inference. In particular, we shall introduce a class of sampling models that can simultaneously account for both skewness and fat tails, and conduct Bayesian inference in the context of a regression model with unknown scale. Quite surprisingly, the currently existing toolbox for handling the frequently occurring phenomenon of skewed data with fat tails seems very limited indeed. The solutions that we are aware of, *e.g.* using Stable laws as in Buckle (1995), seem quite complicated to implement numerically and, more importantly, seem to lack the flexibility and ease of interpretation that an applied statistician would typically require.

In a general context, Section 2 introduces skewness into any continuous (with respect to Lebesgue measure in  $\mathbb{R}$ ), unimodal and symmetric distribution in a rather straightforward way: we simply use inverse scaling of the probability density function (p.d.f.) both sides of the mode. This does not affect the unimodality and allows us to control, with a single unidimensional parameter, the amount of probability mass both sides of the mode. Tail behaviour is not affected by this operation, yet a great deal of flexibility in distributional shape is introduced at the expense of a scalar parameter. Clearly, simultaneously capturing thick tails and skewness can now be achieved by applying this method to a symmetric fat-tailed distribution.

Despite the relative simplicity of the latter idea, one can not hope to use analytical methods to perform posterior and predictive inference in such models allowing for skewness. Therefore, numerical methods will have to be employed. A very useful type of Monte Carlo method in this context is based on Markov chains. The recent statistical literature in the area of Markov chain Monte Carlo (MCMC) abounds and it suffices to refer the reader to Tierney (1994) for a general discussion. A particularly useful version of MCMC is Gibbs sampling, for which we mention the seminal paper of Gelfand and Smith (1990) and the very clear exposition in Casella and George (1992). Gibbs sampling approximates drawings from a (complicated) joint distribution by a Markov chain of drawings from all full conditional distributions. Properness of these full conditionals, however, does not imply properness of the joint distribution [an example is provided in Casella and George (1992)]. Thus, if one uses such methods under improper priors, it becomes crucial to verify existence of the posterior before actually conducting the numerical analysis. Furthermore, efficient estimates of marginal moments are often achieved by averaging over the conditional moments, using the Rao-Blackwell argument introduced in Gelfand and Smith (1990). Again, existence of the conditional moment does not imply that the marginal moment from the joint distribution is finite. The problem of existence of moments does not even vanish when proper priors are used. Thus, we should also check whether the posterior moments that we wish to compute actually exist. Therefore, Sections 3 and 4 are devoted to checking for the existence of the posterior distribution and its moments.

Section 3 considers a general regression model with unknown scale under an improper prior distribution, and examines the impact of introducing skewness (following the method outlined above) into the error distribution on the existence of the posterior distribution and of its moments.

Section 4 specifies the model further, by considering a linear regression structure with independent skewed Student error terms and an unknown scale factor. We consider a standard “non-informative” prior on the regression and scale parameters. Furthermore, we do not fix tail behaviour (controlled by the degrees-of-freedom parameter) nor skewness, but leave both subject to inference. This model, which will be the focus of the sequel of the paper, thus allows for both skewness and flexible tail behaviour.

In Section 4 we examine when a Bayesian analysis can be conducted (*i.e.* properness of the posterior) and which moments of regression coefficients and scale parameter can meaningfully be computed. We then design a Gibbs sampler (using data augmentation) to conduct posterior inference using this model. The actual numerical implementation will be shown to result in a very simple sampler, that can easily be run on a PC for the analysis of moderately large data sets. Section 5 presents the details, and illustrates that judgmental user input is restricted to a minimum.

Finally, Section 6 presents two examples: a location-scale model applied to a data set of share price returns, which was used in Buckle (1995) with the Stable distribution as a modelling device. The second example concerns a data set from astronomy (a Hertzsprung-Russell diagram) where a regression model with two explanatory variables is used. In both examples, posterior and predictive inference is conducted for the general model with skewness and fat tails, and also for models that only account for one of both features. In addition, Bayes factors between these models are computed using the methods advocated in Chib (1995) and in Verdinelli and Wasserman (1995). A final section concludes.

In summary, we will argue that the approach proposed here leads to very flexible modelling of both skewness and fat tails, using only two scalar parameters that are clearly interpretable with well-defined modelling purposes. In addition, the numerical requirements are quite modest and the model can easily be used to tackle problems of direct practical relevance.

All proofs will be grouped in the Appendix, without explicit mention in the main text.

## 2. INTRODUCING SKEWNESS

In this Section we present a general method for transforming a symmetric distribution into a skewed distribution. This generalizes the approach followed in Fernández, Osiewalski and Steel (1995), where a skewed version of the Exponential Power distribution was introduced.

Let us consider a univariate p.d.f.  $f(\cdot)$ , which is unimodal and symmetric around zero. More formally, we assume that  $f(s) = f(|s|)$  and that the latter is decreasing in  $|s|$ . We then generate the following class of skewed distributions, indexed by a scalar  $\gamma \in (0, \infty)$ :

$$p(\varepsilon|\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ f\left(\frac{\varepsilon}{\gamma}\right) I_{[0,\infty)}(\varepsilon) + f(\gamma\varepsilon) I_{(-\infty,0)}(\varepsilon) \right\}. \quad (2.1)$$

The basic idea underlying (2.1) is simply the introduction of inverse scale factors in the positive and the negative orthant. Clearly,  $p(\varepsilon|\gamma)$  retains the unique mode at zero, but

loses symmetry whenever  $\gamma \neq 1$ . More formally, we deduce

$$p(\varepsilon|\gamma = 1) = f(\varepsilon), \quad (2.2)$$

$$\frac{P(\varepsilon \geq 0|\gamma)}{P(\varepsilon < 0|\gamma)} = \gamma^2, \quad (2.3)$$

from which it is clear that  $\gamma$  controls the allocation of mass to each side of the mode. Furthermore, the way  $\gamma$  intervenes in (2.1) implies

$$p(\varepsilon|\gamma) = p(-\varepsilon|1/\gamma), \quad (2.4)$$

so that inverting  $\gamma$  produces the mirror image around zero. In addition,  $p(\varepsilon|\gamma)$  will inherit the differentiability properties of  $f(\cdot)$ . By way of illustration, Figure 1 displays a symmetric distribution ( $\gamma = 1$ ) and its skewed counterparts for  $\gamma = 1.5$  and 2.

In order to gain more insight in the properties of (2.1), let us examine how  $\gamma$  affects its moments. Generally, (2.1) leads to a finite  $r^{th}$  order moment ( $r \in \mathbb{R}$ ) if and only if the corresponding moment of  $f(\cdot)$  exists (*i.e.* for  $\gamma = 1$ ). In particular, we obtain

$$E(\varepsilon^r|\gamma) = M_r \frac{\gamma^{r+1} + \frac{(-1)^r}{\gamma^{r+1}}}{\gamma + \frac{1}{\gamma}}, \quad (2.5)$$

where

$$M_r = \int_0^\infty s^r 2f(s) ds, \quad (2.6)$$

*i.e.* the  $r^{th}$  order moment of  $f(\cdot)$  truncated to the positive real line. Of course,  $E(\varepsilon^r|\gamma)$  will only be real-valued for integer  $r$ . In addition, the assumptions on  $f(\cdot)$  imply that  $M_r = \infty$  for  $r \leq -1$ . Let us, therefore, concentrate on positive integer order moments. From (2.5), the following properties can be shown to hold for noncentered moments: for odd  $r$ , the  $r^{th}$  order moment retains the same absolute value but changes sign if we invert  $\gamma$ , takes the value zero only for  $\gamma = 1$ , and is an increasing function of  $\gamma$  with  $\lim_{\gamma \rightarrow \infty} E(\varepsilon^r|\gamma) = \infty$ . Even moments, on the other hand, are entirely unaffected by inverting  $\gamma$  and, again, increase without bounds in  $\gamma$  for  $\gamma > 1$ . As a consequence,  $\min_\gamma E(\varepsilon^r|\gamma) = E(\varepsilon^r|\gamma = 1)$  for even  $r$ .

If we now consider centered moments, we obtain the following expressions (provided  $f(\cdot)$  allows for the existence of these moments):

$$E(\varepsilon|\gamma) = M_1 \left( \gamma - \frac{1}{\gamma} \right), \quad (2.7)$$

$$Var(\varepsilon|\gamma) = (M_2 - M_1^2) \left( \gamma^2 + \frac{1}{\gamma^2} \right) + 2M_1^2 - M_2, \quad (2.8)$$

where  $Var(\varepsilon|\gamma)$  possesses all the properties mentioned above for even noncentered moments.

Skewness, as measured by the standardized third cumulant [see Box and Tiao (1973, p.150)], is given by

$$Sk(\varepsilon|\gamma) = \left(\gamma - \frac{1}{\gamma}\right) \frac{(M_3 + 2M_1^3 - 3M_1M_2) \left(\gamma^2 + \frac{1}{\gamma^2}\right) + 3M_1M_2 - 4M_1^3}{\left\{(M_2 - M_1^2) \left(\gamma^2 + \frac{1}{\gamma^2}\right) + 2M_1^2 - M_2\right\}^{3/2}}. \quad (2.9)$$

As with noncentered odd moments, we find  $Sk(\varepsilon|\gamma) = -Sk(\varepsilon|1/\gamma)$  and  $Sk(\varepsilon|\gamma = 1) = 0$ , but now we have a finite limit as  $\gamma \rightarrow \infty$ , namely the skewness of  $f(\cdot)$  truncated to the positive real line.

Another popular measure of skewness is the Pearson measure, defined through the standardized difference between mean and mode. Since the p.d.f. in (2.1) has zero mode, we obtain:

$$SP(\varepsilon|\gamma) = \frac{M_1 \left(\gamma - \frac{1}{\gamma}\right)}{\left\{(M_2 - M_1^2) \left(\gamma^2 + \frac{1}{\gamma^2}\right) + 2M_1^2 - M_2\right\}^{1/2}}. \quad (2.10)$$

This skewness measure changes sign as a result of inverting  $\gamma$  and is strictly increasing in  $\gamma$ , converging to the Pearson skewness measure of  $2f(s)I_{(0,\infty)}(s)$  as  $\gamma \rightarrow \infty$ .

In the context of the class of unimodal distributions defined in (2.1), a natural measure of skewness is that introduced in Arnold and Groeneveld (1995), defined as one minus two times the probability mass left of the mode, leading to

$$SM(\varepsilon|\gamma) = \frac{\gamma^2 - 1}{\gamma^2 + 1}, \quad (2.11)$$

which is a strictly increasing function of  $\gamma$ , taking values anywhere in  $(-1, 1)$ . The results in Arnold and Groeneveld (1995) imply that the latter skewness measure maintains the convex ordering of distributions introduced by van Zwet (1964) if  $f(\cdot)$  is differentiable. Clearly, we also have  $SM(\varepsilon|\gamma) = -SM(\varepsilon|1/\gamma)$  and  $SM(\varepsilon|\gamma = 1) = 0$ . In contrast to the skewness coefficients in (2.9) and (2.10), (2.11) does not depend on the choice of  $f(\cdot)$ , and the entire range of this skewness measure can be covered by choosing  $\gamma$  appropriately with  $\lim_{\gamma \rightarrow 0} SM(\varepsilon|\gamma) = -1$  (extreme left skewness) and  $\lim_{\gamma \rightarrow \infty} SM(\varepsilon|\gamma) = 1$  (extreme right skewness).

### 3. EFFECT OF SKEWNESS ON THE EXISTENCE OF POSTERIOR MOMENTS

Let us now consider the impact of introducing skewness into the sampling distribution on Bayesian inference in the context of a general regression model. In particular, we examine the issue of existence of the posterior distribution and of its moments.

We shall assume the observables  $y_i \in \mathfrak{R}$ ,  $i = 1, \dots, n$ , to be generated from

$$y_i = g_i(\beta) + \tau^{-1}\varepsilon_i, \quad (3.1)$$

where  $g_i(\cdot)$  is a known measurable function from  $\mathbb{R}^k (k \geq 1)$  to  $\mathbb{R}$ ,  $\beta = (\beta_1, \dots, \beta_k)' \in \mathbb{R}^k$  parameterizes the location and  $\tau \in \mathbb{R}_+$  is a precision parameter. We assume the error terms  $\varepsilon_1, \dots, \varepsilon_n$  to be i.i.d. given a parameter  $\nu \in \mathcal{N}$  (possibly of infinite dimension) and  $\gamma \in \mathbb{R}_+$  with conditional p.d.f.

$$p(\varepsilon_i | \nu, \gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ f_\nu \left( \frac{\varepsilon_i}{\gamma} \right) I_{[0, \infty)}(\varepsilon_i) + f_\nu(\gamma \varepsilon_i) I_{(-\infty, 0)}(\varepsilon_i) \right\}, \quad (3.2)$$

where  $f_\nu(\cdot)$  is unimodal and symmetric around zero. This stochastic assumption introduces two extra parameters into the problem:  $\gamma$ , the skewness parameter, as explained in the previous Section, and  $\nu$  which can describe other properties of the sampling distribution. In particular,  $\nu$  will control the thickness of the tails in the next Section.

We shall adopt the following class of prior distributions:

$$P_{(\beta, \tau, \nu, \gamma)} = P_\beta \times P_\tau \times P_\nu \times P_\gamma, \quad (3.3)$$

with  $P_\tau$  the usual noninformative distribution characterized by the improper density

$$p(\tau) \propto \tau^{-1} \quad (3.4)$$

on  $\mathbb{R}_+$ ,  $P_\beta$  is any  $\sigma$ -finite measure on  $\mathbb{R}^k$ , and  $P_\nu$  and  $P_\gamma$  are proper distributions. An important special case of (3.3) is where  $P_\gamma$  is Dirac on 1, which characterizes symmetry of the error distribution. In the sequel of this Section, we shall examine the influence of allowing for skewness on posterior inference. To this end, we compare posterior results under a general  $P_\gamma$  with those where  $P_\gamma$  is a Dirac distribution on 1. For notational simplicity, we shall denote the latter case by  $\gamma = 1$ .

First of all, since the prior distribution in (3.3) – (3.4) is improper, existence of the posterior distribution needs to be verified. In addition, our interest will be focussed on the location and precision parameters  $\beta$  and  $\tau$ , since  $\nu$  and  $\gamma$  are merely auxiliary parameters to widen the class of sampling distributions. We shall therefore also address the issue of existence of posterior moments of  $\beta$  and  $\tau$ . Needless to say, negative order moments of  $\tau$  correspond to positive order moments of the scale  $\sigma = \tau^{-1}$  and vice-versa.

We now present the main results of this Section for the Bayesian model corresponding to (3.1) – (3.4).

**Theorem 1.** *Given  $(r_1, \dots, r_k) \in \mathbb{R}^k$ , we obtain that for any  $P_\gamma$*

$$E\left(\prod_{j=1}^k |\beta_j|^{r_j} | y_1, \dots, y_n\right) < \infty$$

*if and only if the same holds under  $\gamma = 1$ .      •*

Theorem 1 clearly states that the existence of posterior moments of  $\beta$  is entirely unaffected by the added uncertainty on  $\gamma$ . An important special case is where  $r_j = 0$  for all  $j \in \{1, \dots, k\}$ , which establishes the fact that incorporating skewness in the sampling does not affect properness of the posterior distribution either.

**Theorem 2.**

(i) For  $r \leq 0$  and any  $P_\gamma$ ,

$$E(\tau^r | y_1, \dots, y_n) < \infty$$

if and only if the same moment exists under  $\gamma = 1$ .

(ii) Given  $r > 0$  and  $P_\gamma$ , we obtain the following:

(iia)  $E(\tau^r | y_1, \dots, y_n) < \infty$  requires existence of the same moment under  $\gamma = 1$ ,

(iib) if

$$E(\tau^r | y_1, \dots, y_n) < \infty \text{ for } \gamma = 1, \quad \text{and} \quad \int_0^\infty \left[ \max \left\{ \gamma, \frac{1}{\gamma} \right\} \right]^r dP_\gamma < \infty,$$

then  $E(\tau^r | y_1, \dots, y_n) < \infty$  under  $P_\gamma$ . •

Thus, existence of negative order moments of  $\tau$  (equivalently, positive order moments of  $\sigma$ ) is never affected by skewness, whereas for positive order moments of  $\tau$  Theorem 2 (ii) provides necessary and sufficient conditions that do not coincide in general. However, in certain situations, the sufficient condition in Theorem 2 (iib) also becomes necessary, as stated in the following Theorem:

**Theorem 3.** *If both*

$$P_\beta(\cap_{i=1, \dots, n} \{\beta : g_i(\beta) > y_i\}) \quad \text{and} \quad P_\beta(\cap_{i=1, \dots, n} \{\beta : g_i(\beta) < y_i\})$$

are strictly positive, where  $P_\beta$  is the prior measure of  $\beta$ , then for any  $r > 0$  and  $P_\gamma$ :

$$E(\tau^r | y_1, \dots, y_n) < \infty$$

if and only if the same moment exists for  $\gamma = 1$  and  $\int_0^\infty [\max\{\gamma, 1/\gamma\}]^r dP_\gamma < \infty$ . •

The moment condition on  $P_\gamma$ , which is often necessary from Theorem 3, is quite a strong requirement: indeed, many commonly used distributions on  $\mathbb{R}_+$  fail to satisfy this condition even for moderate values of  $r$  (*e.g.* neither Exponential nor half-Normal  $P_\gamma$  allow for the posterior mean of  $\tau$ ).

Finally, we note that the pure location-scale model, where  $g_i(\beta) = \beta \in \mathbb{R}$ , combined with a prior density  $p(\beta)$  strictly positive in all of  $\mathbb{R}$ , is within the framework of Theorem 3; thus, the influence of  $P_\gamma$  on the existence of posterior moments of precision (or scale) is entirely characterized for this model. As a simple example where Theorem 3 does not apply, consider  $n = 2$  and  $k = 1$  with  $g_1(\beta) = \beta$  and  $g_2(\beta) = -\beta$ . Then the set  $\cap_{i=1,2} \{\beta : g_i(\beta) > y_i\}$  is empty whenever  $y_1 \geq -y_2$ , whereas  $\cap_{i=1,2} \{\beta : g_i(\beta) < y_i\}$  is empty if  $y_1 \leq -y_2$ , which precludes the application of Theorem 3.

**4. INFERENCE UNDER SKEWED STUDENT SAMPLING**

In the previous Section, we assessed the effect of skewing a symmetric unimodal error distribution with p.d.f.  $f_\nu(\cdot)$  on the existence of posterior moments. Now, we shall fully



specify a Bayesian model which accounts for both skewness and fat tails and the sequel of the paper will be devoted to posterior and predictive inference from this model. Whereas the present Section groups results on the properness of the posterior and the existence of its moments, the next Section will provide a numerical framework for conducting inference from this model.

In particular, we consider a special case of the model in (3.1)–(3.4), using the following assumptions:

- (a) we specify a linear regression model in (3.1), *i.e.*  $g_i(\beta) = x_i' \beta$ , where  $x_i \in \mathbb{R}^k$  is a vector of explanatory variables. Throughout, we shall condition on  $x_i$  without explicit mention. The entire design matrix  $X = (x_1, \dots, x_n)'$  will always be assumed to be of full column rank  $k$ , which implies that  $n \geq k$ ;
- (b)  $f_\nu(\cdot)$  is chosen to be the p.d.f. of a standard Student- $t$  distribution with  $\nu$  degrees of freedom. Thus,  $\nu \in \mathbb{R}_+$ ;
- (c) for the prior of  $\beta$  we take the improper uniform distribution on  $\mathbb{R}^k$ . This leads to  $p(\beta, \tau) \propto \tau^{-1}$ , which corresponds to the usual noninformative distribution for regression and precision parameters, and is the reference prior in the sense of Berger and Bernardo (1992) if  $\gamma$  and  $\nu$  are known [see Fernández and Steel (1995)]. Following (3.3),  $P_\gamma$  and  $P_\nu$  are taken to be any probability measures on  $\mathbb{R}_+$ .

In summary, we assume  $n$  independent replications from the sampling density

$$p(y_i | \beta, \tau, \nu, \gamma) = 2 \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{1/2}} \frac{\tau}{\gamma + \frac{1}{\gamma}} \left[ 1 + \frac{\tau^2}{\nu} (y_i - x_i' \beta)^2 \left\{ \frac{1}{\gamma^2} I_{[0, \infty)}(y_i - x_i' \beta) + \gamma^2 I_{(-\infty, 0)}(y_i - x_i' \beta) \right\} \right]^{-\frac{\nu+1}{2}}, \quad (4.1)$$

with prior distribution

$$P_{(\beta, \tau, \nu, \gamma)} = P_\beta \times P_\tau \times P_\nu \times P_\gamma, \quad (4.2)$$

where  $P_\beta \times P_\tau$  has density  $p(\beta, \tau) \propto \tau^{-1}$  and  $P_\gamma$  and  $P_\nu$  are proper.

The sampling distribution in (4.1) will be denoted by “Skewed Student” with location  $x_i' \beta$ , precision  $\tau^2$ ,  $\nu$  degrees of freedom and skewness parameter  $\gamma$ . Let us briefly discuss the interpretation of the parameters in (4.1).  $\beta \in \mathbb{R}^k$  groups the regression coefficients, usually of primary interest, and  $\tau \in \mathbb{R}_+$  is the precision parameter. In addition to these parameters of interest, (4.1) contains two more parameters, each with a clearly defined modelling purpose. The thickness of the tails is entirely determined by  $\nu \in \mathbb{R}_+$ . From our results in Section 2 [see *e.g.* (2.5)], we know that introducing skewness does not affect the existence of moments of the underlying symmetric distribution. Thus, the sampling moments will exist up to  $\nu$  (not including), as under Student sampling. Skewness is controlled by  $\gamma \in \mathbb{R}_+$ , as explained in Section 2. Following (2.3),  $\gamma$  determines the amount of mass both sides of the location:

$$\frac{P(y_i \geq x_i' \beta | \beta, \tau, \nu, \gamma)}{P(y_i < x_i' \beta | \beta, \tau, \nu, \gamma)} = \gamma^2. \quad (4.3)$$

Before discussing existence of posterior moments from the Bayesian model in (4.1) – (4.2), we stress that (from Section 3)  $P_\gamma$  does not affect properness of the posterior distribution, nor the existence of posterior moments of  $\beta$  and of negative order moments of  $\tau$ . Thus, most results presented here would also apply to the case of (symmetric) Student- $t$  sampling. The latter was examined in Fernández and Steel (1996), in the context of general scale mixtures of Normals, but under fixed  $\nu$ . Here we shall explicitly incorporate prior uncertainty on the thickness of the tails, as the latter may be a crucial modelling instrument. Thus, even when  $P_\gamma$  does not affect the results, the uncertainty on  $\nu$  precludes direct application of the analysis in Fernández and Steel (1996).

Since the prior distribution in (4.2) is improper, we first investigate properness of the posterior distribution.

**Theorem 4.** *With  $n$  independent replications from the sampling model in (4.1) under the prior in (4.2), we obtain a proper posterior distribution if and only if  $n > k$ , for any choices of  $P_\nu$  and  $P_\gamma$ .* •

This well-known result under Normal sampling is thus seen to hold in our much more general framework, where skewness and fat tails are both allowed for. Clearly, any Bayesian inference from this model will require at least  $k + 1$  observations. Throughout the sequel of the paper, we shall, therefore, assume  $n \geq k + 1$ .

We now present our findings for marginal posterior moments of the components of  $\beta$ . The following technical Definition concerning the design matrix  $X$  will be required to adequately characterize the existence of these moments.

**Definition 1.** singularity index for column  $j$

*Given an  $n \times k$  full column-rank matrix  $X$ , we define the singularity index for column  $j = 1, \dots, k$  as the largest number  $p_j$  ( $0 \leq p_j \leq n - k$ ) such that there exists a  $(k - 1 + p_j) \times k$  submatrix of  $X$  of rank  $k - 1$  which retains rank  $k - 1$  after removing its  $j^{\text{th}}$  column.* •

Clearly, if  $X$  contains rows of zeros, then  $p_j$  is at least equal to the number of such zero rows for all  $j = 1, \dots, k$ . Furthermore,  $\max\{p_j : j = 1, \dots, k\} = 0$  if and only if every  $k \times k$  submatrix of  $X$  is nonsingular. Intuitively, the higher  $p_j$  is, the less information the design matrix  $X$  contains about  $\beta_j$ .

As mentioned previously,  $P_\gamma$  will not affect the existence of posterior moments of  $\beta$ . If  $\nu$  is assumed fixed at some positive value  $\nu_0$  (i.e.  $P_\nu$  is a Dirac distribution on  $\nu_0$ ), we know from Fernández and Steel (1996) that for  $r > 0$

$$E(|\beta_j|^r | y_1, \dots, y_n) < \infty \text{ if and only if } r < \min\{n - k, n - k - p_j + \nu_0(n - k - p_j + 1)\}. \quad (4.4)$$

We now consider a general  $P_\nu$ . In order to examine its influence, we partition the class of probability distributions on  $\mathfrak{R}_+$  on the basis of the presence of mass arbitrarily close to zero.

**Theorem 5.** *Consider  $n$  observations from the sampling model (4.1) and the prior in (4.2) with  $P_\nu$  verifying  $P_\nu(0, c) > 0$  for all positive  $c$  smaller than some constant  $C$ . Then, for any  $r \geq 0$ :*

$$E(|\beta_j|^r | y_1, \dots, y_n) < \infty \quad \text{if and only if} \quad \begin{cases} r < n - k & \text{if } p_j = 0 \\ r \leq n - k - p_j & \text{if } p_j \geq 1 \end{cases} \quad \bullet$$

In practice, the most common situation where Theorem 5 applies is when  $P_\nu$  is given through a p.d.f. verifying  $p(\nu) > 0$  for all  $\nu \in (0, C)$  where  $C$  is some positive constant. As in the case where  $\nu$  is fixed [see (4.4)], the design matrix affects existence of moments of  $\beta$  only through  $p_j$ , the singularity index of column  $j$ . If  $p_j = 0$  (intuitively, the best type of design matrix for  $\beta_j$ ), we have marginal posterior moments up to  $n - k$ , as under Normal sampling. The other extreme corresponds to  $p_j = n - k$ , which does not allow for any positive order moments of  $\beta_j$ . Note that different elements of  $\beta$  can possess posterior moments up to different orders.

The sampling model in (4.1) has moments up to and not including  $\nu$ . Thus, if we allow  $\nu$  to be arbitrarily close to zero, we can preclude the existence of any positive order sampling moment. If we wish to guarantee finite sampling moments of a certain order  $\nu_0 > 0$ , we need to restrict  $\nu$  to be bigger than  $\nu_0$ , *i.e.* we consider distributions  $P_\nu$  with support on  $(\nu_0, \infty)$ . In this situation, more moments of the regression coefficients can be shown to exist, as the next Theorem explains.

**Theorem 6.** *Combining  $n$  observations from (4.1) with the prior (4.2) where  $P_\nu$  has support on  $(\nu_0, \infty)$ ,  $\nu_0 > 0$ , we obtain:*

(i) *if  $r \geq n - k$ , then*

$$E(|\beta_j|^r | y_1, \dots, y_n) = \infty,$$

(ii) *if  $0 \leq r < \min\{n - k, n - k - p_j + \nu_0\}$ , then*

$$E(|\beta_j|^r | y_1, \dots, y_n) < \infty. \quad \bullet$$

The necessary and the sufficient condition in Theorem 6 only coincide when  $\nu_0 \geq p_j$ , in which case moments exist exactly up to  $n - k$  (not including). Otherwise, we can guarantee moments of order smaller than  $n - k - p_j + \nu_0$  and Theorem 6 does not cover the range  $[n - k - p_j + \nu_0, n - k)$ . Clearly, when  $p_j = 0$ , bounding  $\nu$  away from zero does not affect existence of moments, but for  $p_j \geq 1$ , we gain at least the moments of order  $r \in (n - k - p_j, \min\{n - k, n - k - p_j + \nu_0\})$ .

In contrast to the situation where  $P_\nu$  has mass arbitrarily close to zero, analyzed in Theorem 5, moments of order smaller than  $\min\{n - k, \nu_0\}$  will now exist for any design matrix  $X$ . Thus, the design matrix can no longer destroy the existence of all positive order moments of  $\beta$ .

Finally, in the important special case of the location-scale model, *i.e.* where  $x_i' \beta = \beta \in \mathbb{R}$ ,  $p_1 = 0$  and posterior moments of  $\beta$  exist exactly up to  $n - 1$  (not including), irrespective of the choice of  $P_\nu$  (and  $P_\gamma$ ).

Let us now consider posterior moments of  $\tau$  of order  $r \in \mathbb{R}$ . The influence of  $P_\gamma$  on the existence of these moments was addressed in Theorems 2 and 3. Taking  $\gamma = 1$  and  $\nu$  fixed at  $\nu_0 > 0$ , Fernández and Steel (1996) tells us that the range of finite moments of  $\tau$  is given by  $r \in (-(n - k), (n - k)\nu_0)$ . We now consider general probability distributions on  $\nu$ .

First of all, we treat the case where  $\nu$  is not bounded away from zero:

**Theorem 7.** *Under the assumptions of Theorem 5, we obtain*

$$E(\tau^r | y_1, \dots, y_n) < \infty \text{ if and only if } -(n - k) < r \leq 0. \quad \bullet$$

Theorem 7 entirely characterizes the moment existence for  $\tau$ , under any  $P_\gamma$  and choosing any  $P_\nu$  with mass arbitrarily close to zero. This choice of  $P_\nu$  precludes finite moments of  $\tau$  of positive order.

However, choosing distributions for  $\nu$  which give zero probability to some interval  $(0, \nu_0]$  potentially allows for the existence of some positive order moments of  $\tau$ .

**Theorem 8.** *Under the assumptions of Theorem 6, we can derive*

(i) *for  $r \leq 0$ ,*

$$E(\tau^r | y_1, \dots, y_n) < \infty \text{ if and only if } r > -(n - k);$$

(ii) *taking  $\gamma = 1$ ,*

$$E(\tau^r | y_1, \dots, y_n) < \infty \text{ if } 0 < r < \nu_0. \quad \bullet$$

From Theorem 7 and Theorem 8 (i) we immediately deduce that negative order moments of  $\tau$  (positive order moments of the scale parameter  $\sigma$ ) always exist exactly up to  $-(n - k)$ , irrespective of  $P_\nu$  and  $P_\gamma$ . The sufficient condition in Theorem 8 (ii) indicates that some positive order moments of  $\tau$  exist when  $\gamma = 1$ . However, we know from Theorems 2 (ii) and 3 that  $P_\gamma$  can influence these moments. In particular, with our choice of  $P_\beta$ , Theorem 3 applies in the pure location-scale model ( $x'_i \beta = \beta \in \Re$ ) and, thus, existence of the  $r^{th}$  and  $-r^{th}$  prior moments of  $\gamma$  is also required in that case.

## 5. NUMERICAL IMPLEMENTATION

In order to conduct inference with the Bayesian model in (4.1)-(4.2), numerical methods will be required. In particular, we shall use a Markov chain Monte Carlo method, namely the Gibbs sampler with data augmentation. The data augmentation adopted is motivated by the representation of a Student- $t$  distribution as a scale mixture of Normals [see (A.8) in the Appendix]. Thus, we can, alternatively, express the sampling density in (4.1) as

$$p(y_i | \beta, \tau, \nu, \gamma) = \left( \frac{2}{\pi} \right)^{\frac{1}{2}} \frac{1}{\gamma + \frac{1}{\gamma}} \int_0^\infty \lambda_i^{\frac{1}{2}} \tau \exp \left[ -\frac{\lambda_i \tau^2}{2} (y_i - x'_i \beta)^2 \right. \\ \left. \left\{ \frac{1}{\gamma^2} I_{[0, \infty)}(y_i - x'_i \beta) + \gamma^2 I_{(-\infty, 0)}(y_i - x'_i \beta) \right\} \right] f_G \left( \lambda_i | \frac{\nu}{2}, \frac{\nu}{2} \right) d\lambda_i, \quad (5.1)$$

where  $f_G(\lambda_i | \nu/2, \nu/2)$  denotes the p.d.f. of a Gamma distribution parameterized as in DeGroot (1970, p.60). Thus, each observation  $y_i$ ,  $i = 1, \dots, n$ , has its own mixing parameter  $\lambda_i$  and  $\lambda_1, \dots, \lambda_n$  are i.i.d. given  $\nu$ . Augmenting the parameter set with  $(\lambda_1, \dots, \lambda_n)$  will greatly facilitate the numerical analysis. Therefore, we shall conduct a Gibbs sampler on  $(\beta, \tau, \nu, \gamma, \lambda_1, \dots, \lambda_n | y_1, \dots, y_n)$ . Essentially, the Gibbs sampler approximates drawings from the joint distribution by a Markov chain of drawings from the full conditional distributions, which are described subsequently.

### 5.1 Conditional of $\beta$

We will analyze each element of  $\beta$  in a separate Gibbs step. From (5.1) and (4.2), the conditional posterior p.d.f. of  $\beta_j$ ,  $j \in \{1, \dots, k\}$ , is defined by

$$p(\beta_j | \{\beta_s : s \neq j\}, \tau, \nu, \gamma, \lambda_1, \dots, \lambda_n, y_1, \dots, y_n) \propto \exp \left[ -\frac{\tau^2}{2} \sum_{i=1}^n \lambda_i (y_i - x'_i \beta)^2 \left\{ \frac{1}{\gamma^2} I_{[0, \infty)}(y_i - x'_i \beta) + \gamma^2 I_{(-\infty, 0)}(y_i - x'_i \beta) \right\} \right], \quad (5.2)$$

which will now be rewritten in a form that immediately suggests a simple algorithm for generating random drawings. Clearly, those observations for which  $x_{ij}$ , the  $j^{th}$  element of  $x_i$ , is zero do not contribute to the conditional distribution of  $\beta_j$  in (5.2). For the  $m$  remaining observations, we compute

$$w_i^{(j)} = \frac{y_i - x'_i \beta + x_{ij} \beta_j}{x_{ij}}, \quad (5.3)$$

noting that the full column rank assumption on  $X$  implies that  $m \geq 1$ . Then, we order the observations such that  $w_1^{(j)} < w_2^{(j)} \dots < w_m^{(j)}$  and partition  $\Re$ , the domain of  $\beta_j$ , into the sets  $S_0^{(j)} = (-\infty, w_1^{(j)}]$ ,  $S_h^{(j)} = (w_h^{(j)}, w_{h+1}^{(j)}]$  for  $h = 1, \dots, m-1$  and  $S_m^{(j)} = (w_m^{(j)}, \infty)$ . Ultimately, we can express the conditional posterior of  $\beta_j$  as:

$$p(\beta_j | \{\beta_s : s \neq j\}, \tau, \nu, \gamma, \lambda_1, \dots, \lambda_n, y_1, \dots, y_n) \propto \sum_{h=0}^m \{p_h^{(j)}\}^{-\frac{1}{2}} \exp \left( -\frac{\tau^2 l_h^{(j)}}{2} \right) f_N^1 \left( \beta_j | \mu_h^{(j)}, \frac{1}{\tau^2 p_h^{(j)}} \right) I_{S_h^{(j)}}(\beta_j), \quad (5.4)$$

with  $f_N^1(\cdot | t, v)$  the p.d.f. of a univariate Normal distribution with mean  $t$  and variance  $v$ , and

$$\begin{aligned} p_h^{(j)} &= \sum_{i=1}^h \rho_{i1}^{(j)} + \sum_{i=h+1}^m \rho_{i2}^{(j)}, \\ p_h^{(j)} \mu_h^{(j)} &= \sum_{i=1}^h \rho_{i1}^{(j)} w_i^{(j)} + \sum_{i=h+1}^m \rho_{i2}^{(j)} w_i^{(j)}, \\ l_h^{(j)} &= \sum_{i=1}^h \rho_{i1}^{(j)} \{w_i^{(j)}\}^2 + \sum_{i=h+1}^m \rho_{i2}^{(j)} \{w_i^{(j)}\}^2 - p_h^{(j)} \{\mu_h^{(j)}\}^2, \end{aligned} \quad (5.5)$$

where we have defined

$$\begin{aligned} \rho_{i1}^{(j)} &= \lambda_i x_{ij}^2 \left\{ \frac{1}{\gamma^2} I_{(-\infty, 0)}(x_{ij}) + \gamma^2 I_{(0, \infty)}(x_{ij}) \right\}, \\ \rho_{i2}^{(j)} &= \lambda_i x_{ij}^2 \left\{ \gamma^2 I_{(-\infty, 0)}(x_{ij}) + \frac{1}{\gamma^2} I_{(0, \infty)}(x_{ij}) \right\}. \end{aligned} \quad (5.6)$$

The expression in (5.4) is now straightforward to draw from. First, we compute the probabilities attached to each of the sets  $S_h^{(j)}$  forming the partition of  $\mathfrak{R}$ , then we choose one set at random according to those probabilities, and finally we draw the corresponding truncated Normal, using the mixed rejection algorithm of Geweke (1991).

## 5.2 Conditional of $\tau$

It is immediate from (5.2) and (4.2) that

$$p(\tau^2 | \beta, \nu, \gamma, \lambda_1, \dots, \lambda_n, y_1, \dots, y_n) = f_G \left( \tau^2 \middle| \frac{n}{2}, \frac{1}{2} \sum_{i=1}^n \lambda_i (y_i - x'_i \beta)^2 \left\{ \frac{1}{\gamma^2} I_{[0, \infty)}(y_i - x'_i \beta) + \gamma^2 I_{(-\infty, 0)}(y_i - x'_i \beta) \right\} \right), \quad (5.7)$$

from which random drawings can immediately be generated; in particular, we shall use Cheng's (1977) GB algorithm.

## 5.3 Conditional of $\nu$

Generally, the full conditional distribution of  $\nu$  given  $(\beta, \tau, \gamma, \lambda_1, \dots, \lambda_n, y_1, \dots, y_n)$  is proportional to

$$\left(\frac{\nu}{2}\right)^{n\nu/2} \left\{ \Gamma\left(\frac{\nu}{2}\right) \right\}^{-n} \exp \left\{ -\frac{\nu}{2} \sum_{i=1}^n (\lambda_i - \log \lambda_i) \right\} P_\nu, \quad (5.8)$$

*i.e.* the conditional posterior distribution of  $\nu$  is absolutely continuous with respect to the prior  $P_\nu$  with Radon-Nikodym derivative proportional to the first three factors in (5.8). Clearly, the distribution in (5.8) does not directly lend itself to random number generation, but as  $\nu$  is a scalar, many numerical methods should work efficiently.

In our empirical Section, we shall not bound  $\nu$  away from zero and we take  $P_\nu$  to be an Exponential distribution with p.d.f.

$$p(\nu) = d \exp(-d\nu), \quad (5.9)$$

leading to

$$p(\nu | \beta, \tau, \gamma, \lambda_1, \dots, \lambda_n, y_1, \dots, y_n) \propto \left(\frac{\nu}{2}\right)^{n\nu/2} \left\{ \Gamma\left(\frac{\nu}{2}\right) \right\}^{-n} \exp \left[ -\nu \left\{ d + \frac{1}{2} \sum_{i=1}^n (\lambda_i - \log \lambda_i) \right\} \right]. \quad (5.10)$$

Drawings from (5.10) will be generated through rejection sampling [see e.g. Devroye (1986)] using an Exponential source density, with its parameter chosen so as to maximize the overall acceptance probability, as described in Geweke (1994). In particular, we employ the following strategy:

1. Draw  $\nu$  from a distribution with p.d.f.  $(\nu^*)^{-1} \exp(-\nu/\nu^*)$  with  $\nu^*$  the unique solution to

$$\frac{n}{2} \left\{ \log \left( \frac{\nu}{2} \right) - \Psi \left( \frac{\nu}{2} \right) \right\} + \frac{1}{\nu} + \frac{n - \sum_{i=1}^n (\lambda_i - \log \lambda_i)}{2} - d = 0, \quad (5.11)$$

where  $\Psi(\cdot)$  is the digamma function.

2. Accept the drawn value  $\nu$  with probability

$$\left(\frac{\nu}{2}\right)^{\frac{n\nu}{2}} \left(\frac{\nu^*}{2}\right)^{-\frac{n\nu^*}{2}} \left\{\Gamma\left(\frac{\nu}{2}\right)\right\}^{-n} \left\{\Gamma\left(\frac{\nu^*}{2}\right)\right\}^n \exp \left[ (\nu - \nu^*) \left\{ \frac{1}{\nu^*} - \frac{\sum_{i=1}^n (\lambda_i - \log \lambda_i)}{2} - d \right\} \right]. \quad (5.12)$$

For a grid of values of  $n$  (ranging from 50 to 500) and plausible values for  $d + \frac{1}{2} \sum_{i=1}^n (\lambda_i - \log \lambda_i)$  (ranging from slightly larger than  $n/2$  to  $2n$ ) empirical acceptance probabilities are typically in the order of 0.10 and always above 0.05. See also Table A.2 in Geweke (1994).

#### 5.4. Conditional of $\gamma$

With general  $P_\gamma$ , the conditional distribution of  $\gamma$  given  $(\beta, \tau, \nu, \lambda_1, \dots, \lambda_n, y_1, \dots, y_n)$  is proportional to

$$\left(\gamma + \frac{1}{\gamma}\right)^{-n} \exp \left[ -\frac{\tau^2}{2} \sum_{i=1}^n \lambda_i (y_i - x'_i \beta)^2 \left\{ \frac{1}{\gamma^2} I_{[0, \infty)}(y_i - x'_i \beta) + \gamma^2 I_{(-\infty, 0)}(y_i - x'_i \beta) \right\} \right] P_\gamma, \quad (5.13)$$

using the same notation as in (5.8). In our empirical Section, we shall use a Gamma( $a, b$ ) prior on  $\varphi \equiv \gamma^2$ , leading to

$$p(\varphi | \beta, \tau, \nu, \lambda_1, \dots, \lambda_n, y_1, \dots, y_n) \propto \varphi^{\frac{n}{2} + a - 1} (\varphi + 1)^{-n} \exp \left\{ -\left( \frac{\vartheta}{\varphi} + \kappa \varphi \right) \right\}, \quad (5.14)$$

where we have defined

$$\begin{aligned} \vartheta &= \frac{\tau^2}{2} \sum_{i=1}^n \lambda_i (y_i - x'_i \beta)^2 I_{[0, \infty)}(y_i - x'_i \beta) \geq 0, \\ \kappa &= b + \frac{\tau^2}{2} \sum_{i=1}^n \lambda_i (y_i - x'_i \beta)^2 I_{(-\infty, 0)}(y_i - x'_i \beta) > 0. \end{aligned} \quad (5.15)$$

The distribution in (5.14) is not of any standard form, for which random number generators are readily available. However, the density function is bell-shaped and has subquadratic tails, so that the Ratio-of-Uniforms method of Kinderman and Monahan (1977) can be applied. Generally, as explained in Devroye (1986), using this method to draw a scalar variate with p.d.f. proportional to an integrable function  $g(\cdot)$ , consists in:

1. draw a Uniform distribution on the set  $A = \{(u, v) : 0 \leq u \leq \{g(v/u)\}^{1/2}\}$ ;
2. the ratio  $v/u$  is a drawing from the required distribution.

In order to draw from the Uniform distribution on  $A$ , it is convenient to draw a Uniform on a rectangle enclosing  $A$ , accepting the drawing only if it falls in  $A$ . The most efficient implementation of this algorithm corresponds to choosing the smallest possible rectangle enclosing  $A$ , which is generally given by  $[0, \sup_x \{g(x)\}^{1/2}] \times [\inf_x x \{g(x)\}^{1/2}, \sup_x x \{g(x)\}^{1/2}]$ .

Taking  $g(\cdot)$  to be the kernel in (5.14), it is immediate that  $\inf_{\varphi} \varphi^2 g(\varphi) = 0$ , whereas the unique positive solution of  $-\kappa\varphi^3 + (q - \kappa - n)\varphi^2 + (q + \vartheta)\varphi + \vartheta = 0$  maximizes  $g(\varphi)$  for  $q = (n/2) + a - 1$  and  $\varphi^2 g(\varphi)$  for  $q = (n/2) + a + 1$ .

Choosing a wide range of values for  $n$  (from 50 to 500) and a range of empirically plausible values for  $a, \vartheta$  and  $\kappa$  we estimate acceptance rates to be typically around 15% and always exceeding 10%.

### 5.5. Conditional of $\lambda_1, \dots, \lambda_n$

Drawing from the conditional distribution of the mixing parameters is straightforward as they are independent with p.d.f.

$$p(\lambda_1, \dots, \lambda_n | \beta, \tau, \nu, \gamma, y_1, \dots, y_n) = \prod_{i=1}^n f_G \left( \lambda_i \middle| \frac{\nu+1}{2}, \frac{\nu}{2} + \frac{\tau^2(y_i - x'_i \beta)^2}{2} \left\{ \frac{1}{\gamma^2} I_{[0, \infty)}(y_i - x'_i \beta) + \gamma^2 I_{(-\infty, 0)}(y_i - x'_i \beta) \right\} \right). \quad (5.16)$$

The full conditional distributions in (5.4), (5.7), (5.10), (5.14) and (5.16) define a Gibbs sampler with  $k + 4$  steps in  $n + k + 3$  dimensions. Convergence of the induced Markov chain to the posterior distribution is ensured, since the parameter space has a Cartesian product structure [see Roberts and Smith (1994)].

## 6. EMPIRICAL EXAMPLES

### 6.1 Preliminaries

In this Section, we will use the Bayesian model described in Section 4 for the analysis of some examples, following the numerical implementation outlined in the previous Section.

We remind the reader that we adopted the prior distribution in (4.2) with an Exponential distribution on  $\nu$  as in (5.9), and a  $\text{Gamma}(a, b)$  prior for  $\varphi = \gamma^2$ . Thus, a full description of our prior distribution still requires a choice for  $d$  in (5.9) and for  $a$  and  $b$ . In the elicitation of these hyperparameters we shall try to avoid introducing strong prior information. To this end, we choose  $d = 0.1$ , thus obtaining a prior mean of  $\nu$  equal to 10 and a prior variance of 100, essentially allocating substantial prior mass to very thick tails as well as almost Normal tails. For the skewness parameter,  $\gamma$ , we specify a prior with mean one, which centers the prior around the case of symmetric sampling. The latter is equivalent to choosing

$$b = \left\{ \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} \right\}^2, \quad (6.1)$$

and we shall elicit  $a$  using both the prior variance of  $\gamma$  and the prior mass on the interval  $(0, 1)$ . The variance of  $\gamma$  is the following decreasing function of  $a$ :

$$\text{Var}(\gamma) = a \left\{ \frac{\Gamma(a)}{\Gamma(a + \frac{1}{2})} \right\}^2 - 1. \quad (6.2)$$



The expression in (6.2) would seem to suggest that a very small value of  $a$  adequately conveys a lack of prior information: *e.g.*  $a = 0.01$  corresponds to  $\text{Var}(\gamma) = 31.7$ . However, the prior probability that  $\gamma \in (0, 1)$  also decreases in  $a$ , and for  $a = 0.01$  we obtain  $P(\gamma < 1) = 0.93$ . Since we prefer a prior that gives approximately equal weights to left skewness (*i.e.*  $\gamma < 1$ ) and right skewness (*i.e.*  $\gamma > 1$ ), a compromise is in order. We feel that the value  $a = 0.5$ , leading to  $\text{Var}(\gamma) = 0.57$  and  $P(\gamma < 1) = 0.58$  is quite reasonable. This particular value leads exactly to a half-Normal prior for  $\gamma$ . We shall adopt these prior choices in both of the examples subsequently analyzed.

Besides the general model allowing for both skewness and fat tails simultaneously, we shall also consider simpler versions, which incorporate only one of these features at a time. Thus, we examine three possible sampling models, namely the skewed Student in (4.1), the skewed Normal [the limiting case of (4.1) as  $\nu \rightarrow \infty$ ] and the Student- $t$  model [(4.1) with  $\gamma = 1$ ]. Priors for parameters present in the models will always be as described above.

In the sequel, we present posterior inference on model parameters and predictive inference in the context of each model. The latter will be conducted through averaging the sampling density, using the Rao-Blackwell argument suggested in Gelfand and Smith (1990).

Model comparison will formally be done through the use of Bayes factors. Due to the fact that we have proper priors on model-specific parameters, the latter can meaningfully be computed. In order to conduct the actual computations, two distinct methods are employed: the method of Chib (1995) and the Savage-Dickey density ratio mentioned in Verdinelli and Wasserman (1995), based on Dickey (1971).

Throughout, we used a sequential version of the Gibbs sampler, discarding the first 10,000 realizations (the “burn-in”) and basing our results on the following 250,000 drawings. However, much smaller runs already lead to reliable results. All density plots are presented without smoothing and are based on 50 bins.

As a final, but important, note, we stress that the numerical implementation described in Section 5 leads to very efficient algorithms. Using Gauss-386i VM version 3.2, the most complicated models for both examples treated here executed at a rate of over 30,000 Gibbs draws per hour on a PC equipped with a Pentium-100 processor. Thus, the analysis of much more challenging data sets is entirely within reach, even with modest computing facilities.

## 6.2. Share Price Returns

In our first example we use a simple location-scale structure (*i.e.*  $k = 1$  and  $x_i = 1$ ,  $i = 1, \dots, n$ ) to model daily share price returns. The particular data set we use concerns Abbey National shares between July 31 and October 8, 1991, and was used in Buckle (1995). Table 1 in Buckle (1995) lists the price data,  $p_i$ ,  $i = 0, \dots, 49$ , from which we construct the observations  $y_i = (p_i - p_{i-1})/p_{i-1}$ ,  $i = 1, \dots, 49$ .

Buckle (1995) proposed Stable distributions as a way of dealing with skewness and fat tails. Before discussing our results, let us briefly contrast this approach with the approach proposed in the present paper. We feel the main advantages of using the model introduced in Section 4 are model flexibility, interpretability of the parameters and computational simplicity.

In particular, whereas we can account for a smooth transition of very fat to Normal tails, since the sampling density in (4.1) behaves in the tails as a Student distribution with  $\nu$  degrees of freedom, Stable distributions display an inherent discontinuity in tail behaviour, since they either do not possess a finite variance or are Normal. In addition, skewness is only allowed for when the variance does not exist.

A related point is that the skewness and tail parameters are inextricably linked for Stable laws, therefore complicating both the issue of prior elicitation and interpretation of the parameters. In sharp contrast, our approach entirely separates the effect of the skewness parameter  $\gamma$  and the tail parameter  $\nu$ , facilitating their interpretation and making prior independence between the two a plausible assumption.

In addition, the Gibbs sampler used in Buckle (1995) requires far more numerical effort than ours, as it involves four Metropolis-Hastings steps and  $n$  univariate rejection sampling steps for the augmentation variables. Since the p.d.f. of a Stable distribution does not possess a closed form expression, predictive distributions are also much more difficult to evaluate than in our case.

Before our discussion of posterior results, a technical issue still needs to be addressed. Since  $n > 1$ , Theorem 4 assures us of the existence of the posterior distribution. However, this obviously does not prevent the predictive density  $p(y_1, \dots, y_n)$  from being infinite in a set of Lebesgue measure zero in  $\mathbb{R}^n$ . For the location-scale model, the latter set consists of all the samples  $(y_1, \dots, y_n)$  for which  $P_\nu(0, \{s-1\}/\{n-s\}) > 0$ , where  $s$  is the largest number of identical observations. Thus, when  $P_\nu$  has mass arbitrarily close to zero (as is the case with the exponential prior considered here), any sample that contains at least two identical observations leads to  $p(y_1, \dots, y_n) = \infty$ . Whereas theoretically a set of Lebesgue measure zero poses no problem, the censoring and rounding mechanisms underlying many empirical observations may lead to repeated data points, as is the case in our particular data set. One obvious solution would be to restrict  $\nu$  to be bigger than  $(s-1)/(n-s)$ . In practice, this restriction is relatively harmless; *e.g.* in our example, 7 of the 49 observations are repeated ( $y_i = 0$ ), yet  $\nu > 1/7$  is sufficient. In the interest of a fair comparison with the results in Buckle (1995), we have chosen not to restrict the support of  $P_\nu$ , but instead we have slightly perturbed the  $y_i$ 's. The empirical impact of this minor perturbation is, however, quite negligible, since we never obtained any empirical evidence of posterior mass for  $\nu < 1/2$ . Explicitly incorporating the censoring mechanism into the model is, naturally, a very appealing solution. However, this is outside the scope of the present paper, and is the object of ongoing research.

Posterior results using the general sampling model in (4.1) with the prior as explained in Subsection 6.1 are summarized in Table 1 and Figures 2-5. Besides the general skewed Student sampling model, we have also used the Student- $t$  model, which only allows for thick tails, and the skewed Normal, with only skewness accounted for. From our theoretical results in Section 4 we know that positive order posterior moments of  $\beta$  and  $\sigma = \tau^{-1}$  exist up to order  $n - k = 48$  (not including) in all three models, whereas positive order moments of  $\tau$  are precluded under Student or skewed Student sampling. Table 1 reports posterior means and standard deviations of  $\beta$  and  $\sigma$ . The latter vary substantially across models.

**Table 1**

	skewed Student	Student	skewed Normal
mean $\beta$	-0.0068	-0.0012	-0.0064
st. dev. $\beta$	(0.0028)	(0.0018)	(0.0031)
mean $\sigma$	0.0091	0.0103	0.0117
st. dev. $\sigma$	(0.0018)	(0.0018)	(0.0014)

Figure 4 clearly indicates right skewness in the data; thus, if our model does not account for this skewness, the location will be shifted to the right, as occurs for the Student- $t$  model. As Figure 5 indicates,  $\nu$  has substantial posterior mass in regions corresponding to thick tails. Thus, the skewed Normal model, which has Normal tail behaviour, needs to decrease the precision  $\tau$  in order to capture observations in the tails. Figure 3 indicates that precision increases if we account for fat tails and even more if we allow for skewness as well (see also Table 1).

An interesting feature is that inference on skewness is little affected by allowing for thick tails. Indeed, the skewed Student and the skewed Normal lead to similar posterior distributions for  $\gamma$  (Figure 4). Even more striking is the similarity of the posterior distributions for  $\nu$  under Student and skewed Student sampling (Figure 5). Whether we allow for skewness or not has virtually no impact on inference on the degrees of freedom parameter  $\nu$ . In summary, inference on skewness and thickness of tails seems well separated in our model. However, the present data set is not very informative on the thickness of the tails, as we have empirically noticed some sensitivity of posterior inference on  $\nu$  with respect to the choice of  $d$  in (5.9).

Figure 7 displays the post-sample predictive density functions under each of the three models. Note that the predictive from the skewed Student model closely resembles the data histogram in Figure 6. The Student model obviously leads to a symmetric predictive, which seems at odds with the data, whereas the skewed Normal sampling model clearly induces more dispersion in the predictive.

A formal comparison of the three models is now conducted using Bayes factors. We have used the method based on the “Basic Marginal Likelihood Identity” (BMI) developed in Chib (1995). This method estimates the marginal likelihood of the observed sample using Gibbs sampling in combination with the integrating constants of the required full conditionals. Wherever the latter integrating constants were not available analytically (*i.e.* for  $\nu$  and  $\gamma$ ), we have estimated them empirically by normalizing the histograms. All results were based on 75,000 draws after a burn-in of 5,000 draws for each additional Gibbs sampler involved. Table 2 presents the resulting Bayes factors. Entry  $(i, j)$  in the Table indicates the Bayes factor in favour of model  $i$  versus model  $j$ . For completeness, the simple Normal model (for which the marginal likelihood is known analytically) is also included. Clearly, there is some evidence for both fat tails and skewness in the data.

As a check, we also assessed the evidence in favour of skewness using the Savage-Dickey density ratio, as explained in Verdinelli and Wasserman (1995). Comparing skewed Student with Student and skewed Normal with Normal led to the same Bayes factors as displayed in Table 2.

Table 2

	skewed Student	Student	skewed Normal	Normal
skewed Student	1	2.3	4.0	11.9
Student		1	1.7	5.1
skewed Normal			1	3.0
Normal				1

Overall, our results are not incompatible with those found in Buckle (1995), who also recorded evidence of right skewness and heavy tails. Only his posterior findings on the location parameter seem in conflict with ours, as he obtains a posterior mean of 0.00053. Note, however, that the location parameter in Buckle is not interpretable as the mode (under asymmetry), whereas our sampling model in (4.1) always locates the mode at  $x'_i\beta$ . Thus, our location parameter  $\beta$  has the unequivocal interpretation of the mode of the sampling distribution in this example. We feel this is an added advantage of using the Bayesian model described in Section 4.

### 6.3. Hertzsprung-Russell Diagram

Our second example concerns explaining the logarithm of the light intensity of stars ( $y_i$ ) by an intercept and the logarithm of the effective surface temperature of the star. Thus, we now have a regression model with  $k = 2$ ,  $x_{i1} = 1$  and  $x_{i2}$  is the log of the temperature of star  $i$ . We have 47 observations for the star cluster CYG OB1 (in the direction of Cygnus), which are taken from Rousseeuw and Leroy (1987, Table 3, p. 27).

The analysis is conducted using the numerical procedures outlined in Section 5, implemented as described in Subsection 6.1. We consider two sampling models, Student- $t$  and skewed Student, with the priors described in Subsection 6.1. The design matrix  $X$  of our data set verifies  $p_1 = p_2 = 4$ . Recalling Definition 1, this can easily be seen as follows: none of the values  $x_{i2}$  are zero and the maximum number of identical values for  $x_{i2}$  is five. This immediately leads to  $p_1 = p_2 = 4$ . Thus, from Theorem 5, positive order posterior moments of  $\beta_1$  and  $\beta_2$  exist up to the order  $n - k - 4 = 41$  (including), under both sampling assumptions. Theorem 7 implies that the range of finite posterior moments of  $\sigma = \tau^{-1}$  is given by  $[0, 45)$  under both sampling schemes.

As in the previous Example, a technical comment is in order. For the model considered here, *i.e.*  $k = 2$  with an intercept, the practically relevant conditions to check for having a finite predictive value are: no zero observations and those observations corresponding to equal rows of  $X$  should be different. The first condition can easily be achieved by adding a constant to all observations and the intercept. In case the second condition is not fulfilled, restricting  $\nu$  to be bigger than some small value will typically solve the problem. Even though the empirical posterior probability for  $\nu < 1$  is zero in our example, we have based our results on a slightly perturbed sample.

Posterior results are summarized in Table 3 and Figures 8-11. Inference on the regression coefficients is somewhat affected by allowing for skewness, and the posterior mean of  $\sigma$  is smaller under skewed Student sampling. There seems to be evidence of left skewness

in the data (see Figure 10) and, as was the case in our previous example, inference on tail behaviour is largely unaffected by allowing for skewness (see Figure 11).

**Table 3**

	mean $\beta_1$	st. dev. $\beta_1$	mean $\beta_2$	st. dev. $\beta_2$	mean $\sigma$	st. dev. $\sigma$
skewed Student	7.53	(1.37)	-0.495	(0.275)	0.428	(0.132)
Student	6.71	(1.42)	-0.391	(0.327)	0.552	(0.065)

The left skewness revealed in Figure 10 is translated into a skewed predictive plot under skewed Student sampling, conditional on mean values of  $X$  and the full observed sample. Of course, the predictive under Student sampling is symmetric (see Figure 12).

The Bayes factor of skewed Student versus Student sampling was computed to be 1.5 using the Savage-Dickey density ratio. The latter result conveys moderate evidence in favour of skewness.

## 7. CONCLUSION

In this paper, we have introduced a general method for transforming symmetric into skewed distributions, at the cost of a single scalar parameter. Using such a skewed distribution for the error terms in a regression model, we establish that the effects of this skewness on the existence of the posterior distribution and its moments is quite limited. We then consider linear regression under independent skewed Student errors with unknown skewness and thickness of tails, in combination with a commonly used improper prior on the regression coefficients and the precision parameter. For this model, which is central to the paper, we obtain that the posterior is well-defined under the same conditions as for Normal sampling (*i.e.* when sample size exceeds the number of regressors); existence of posterior moments of regression coefficients and precision are examined in detail. A numerical analysis based on the Gibbs sampler is outlined and applied to a number of examples.

We feel that the approach proposed here has a number of attractive features:

- (a) It allows for very flexible modelling of the skewness and fat tail features of the data. Skewness covers the entire range of *e.g.* the skewness measure in Arnold and Groeneveld (1995), which implies that mass can be allocated to the regions both sides of the mode in any proportion, irrespective of the underlying symmetric distribution. Within the skewed Student setup, we can allow for any Student tail behaviour, thus ranging from very fat tails to limiting Normality.
- (b) The extra parameters introduced into the analysis have very clearly defined modelling purposes. The skewness parameter alone controls the allocation of mass with respect to the mode, whereas the degrees of freedom parameter entirely accounts for tail behaviour. The two parameters are, thus, clearly interpretable. Prior independence is typically a very plausible assumption, which drastically simplifies the process

of choosing prior distributions: prior elicitation for each of them can simply be conducted independently. From our empirical examples, it seems that prior independence between these parameters is not substantially altered by the data information.

- (c) The empirical analysis is very feasible indeed. The Gibbs sampler we construct uses either standard algorithms or simple rejection methods that prove to work very efficiently. The speed of execution is such that the analysis of quite challenging problems is a real practical possibility, even for users with modest computing facilities.

## APPENDIX: PROOFS

### Proof of Theorem 1

For the Bayesian model in (3.1) – (3.4),  $E(\prod_{j=1}^k |\beta_j|^{r_j} | y_1, \dots, y_n) < \infty$  if and only if the integral

$$I_\beta = \int_{\mathbb{R}^k \times \mathbb{R}_+ \times \mathcal{N} \times \mathbb{R}_+} \left( \prod_{j=1}^k |\beta_j|^{r_j} \right) \tau^{-1} \left\{ \prod_{i=1}^n p(y_i | \beta, \tau, \nu, \gamma) \right\} dP_\beta d\tau dP_\nu dP_\gamma \quad (A.1)$$

is finite. Since  $f_\nu(s) = f_\nu(|s|)$  is decreasing in  $|s|$ , we obtain the following upper and lower bounds for the sampling density  $p(y_i | \beta, \tau, \nu, \gamma)$ :

$$2 \frac{\tau}{\gamma + \frac{1}{\gamma}} f_\nu \left( \frac{\tau |y_i - g_i(\beta)|}{h(\gamma)} \right), \quad (A.2)$$

with

$$h(\gamma) = \begin{cases} \max\{\gamma, \frac{1}{\gamma}\} & \text{for the upper bound,} \\ \min\{\gamma, \frac{1}{\gamma}\} & \text{for the lower bound.} \end{cases} \quad (A.3)$$

We now substitute each of these bounds inside the integral in (A.1). Applying Fubini's Theorem, we first consider the integral with respect to  $\tau$ . Transforming from  $\tau$  to  $\theta = \tau/h(\gamma)$ , immediately leads to the upper and lower bounds for  $I_\beta$ :

$$2^n \int_{\mathbb{R}_+} \left( \frac{h(\gamma)}{\gamma + \frac{1}{\gamma}} \right)^n dP_\gamma \int_{\mathbb{R}^k \times \mathbb{R}_+ \times \mathcal{N}} \left( \prod_{j=1}^k |\beta_j|^{r_j} \right) \theta^{n-1} \left\{ \prod_{i=1}^n f_\nu(\theta |y_i - g_i(\beta)|) \right\} dP_\beta d\theta dP_\nu, \quad (A.4)$$

with  $h(\gamma)$  as defined in (A.3). Clearly, for both choices of  $h(\gamma)$  in (A.3), the value of the first integral in (A.4) lies in the interval  $(0, 1)$ . In addition, the second integral in (A.4) is finite if and only if  $E(\prod_{j=1}^k |\beta_j|^{r_j} | y_1, \dots, y_n) < \infty$  under  $\gamma = 1$ , thus obtaining Theorem 1. •

### Proof of Theorem 2

Having a finite  $r^{th}$  order posterior moment of  $\tau$  is equivalent to a finite integral

$$I_\tau = \int_{\mathbb{R}^k \times \mathbb{R}_+ \times \mathcal{N} \times \mathbb{R}_+} \tau^{r-1} \left\{ \prod_{i=1}^n p(y_i | \beta, \tau, \nu, \gamma) \right\} dP_\beta d\tau dP_\nu dP_\gamma. \quad (A.5)$$

We now substitute the bounds given in (A.2) – (A.3) for the sampling density inside the integrand in (A.5). Considering first the integral with respect to  $\tau$  and transforming to  $\theta = \tau/h(\gamma)$ , leads to upper and lower bounds for  $I_\tau$  of the form

$$2^n \int_{\mathbb{R}_+} \frac{h(\gamma)^{n+r}}{(\gamma + \frac{1}{\gamma})^n} dP_\gamma \int_{\mathbb{R}^k \times \mathbb{R}_+ \times \mathcal{N}} \theta^{n+r-1} \left\{ \prod_{i=1}^n f_\nu(\theta|y_i - g_i(\beta)|) \right\} dP_\beta d\theta dP_\nu, \quad (A.6)$$

with  $h(\gamma)$  as defined in (A.3). Note that the second integral in (A.6) is finite if and only if  $E(\tau^r|y_1, \dots, y_n) < \infty$  under  $\gamma = 1$ .

(A) Since the first integral in (A.6) is strictly positive for  $h(\gamma) = \min\{\gamma, 1/\gamma\}$ , it follows that  $E(\tau^r|y_1, \dots, y_n) < \infty$  under  $P_\gamma$  requires the same moment to be finite under  $\gamma = 1$ .

(B) In order to obtain a sufficient condition, we consider  $h(\gamma) = \max\{\gamma, 1/\gamma\}$ . The first integral in (A.6) is then finite if and only if

$$\int_{\mathbb{R}_+} \left[ \max \left\{ \gamma, \frac{1}{\gamma} \right\} \right]^r dP_\gamma < \infty, \quad (A.7)$$

which is immediately fulfilled for  $r \leq 0$ , but not for  $r > 0$ .

Combining (A) and (B) proves Theorem 2.  $\bullet$

### Proof of Theorem 3

The necessity of a finite  $r^{th}$  order posterior moment of  $\tau$  under  $\gamma = 1$  was already established in Theorem 2. Thus, we just need to prove that, under the assumptions of Theorem 3, (A.7) is also necessary.

$I_\tau$ , defined in (A.5), can be bounded from below as  $I_\tau \geq I_1 + I_2$ , where  $I_1$  restricts the domain of integration to  $\{\beta : g_i(\beta) > y_i \text{ for all } i\}, \tau \in \mathbb{R}_+, \nu \in \mathcal{N}$ , and  $\gamma \leq 1$ , whereas  $I_2$  covers  $\{\beta : g_i(\beta) < y_i \text{ for all } i\}, \tau \in \mathbb{R}_+, \nu \in \mathcal{N}$ , and  $\gamma \geq 1$ . Integrating first with respect to  $\tau$ , transforming to  $\theta = \gamma\tau$  for  $I_1$  and to  $\theta = \tau/\gamma$  for  $I_2$  leads to the result.  $\bullet$

### Remarks

1. In the remainder of the Proofs we shall be using the fact that the Student distribution is in the class of scales mixtures of Normals. In particular, the p.d.f. of a standard Student- $t$  with  $\nu$  degrees of freedom can be written as

$$f_\nu(\varepsilon) = \int_0^\infty \left( \frac{\lambda}{2\pi} \right)^{1/2} \exp \left( -\frac{\lambda}{2} \varepsilon^2 \right) dP_\lambda, \quad (A.8)$$

with  $P_\lambda$  a Gamma distribution with shape and precision parameters both equal to  $\nu/2$  (*i.e* with unitary mean).

Fernández and Steel (1996) examines Bayesian inference in the context of a linear regression model with i.i.d. errors distributed as a known scale mixture of Normals. Thus,  $\lambda_1, \dots, \lambda_n$ , the mixing parameters corresponding to each of the observations, are i.i.d. with some known probability distribution, say  $P_\lambda$ , on  $\mathbb{R}_+$ . Our setup now is slightly different:

$\lambda_1, \dots, \lambda_n$  are i.i.d. given  $\nu$ , with a  $\text{Gamma}(\nu/2, \nu/2)$  distribution, but the prior  $P_\nu$  destroys the independence between the mixing parameters, leading to the p.d.f.

$$p(\lambda_1, \dots, \lambda_n) = \int_0^\infty \left\{ \prod_{i=1}^n f_G\left(\lambda_i \middle| \frac{\nu}{2}, \frac{\nu}{2}\right) \right\} dP_\nu. \quad (\text{A.9})$$

Despite this difference with Fernández and Steel (1996), many of the proofs and results from the latter paper are useful for the proofs of the present paper; thus, we will frequently refer to it in what follows.

**2.** The following result [see Whittaker and Watson (1927), chap. 12] will be used in the sequel to provide bounds on the Gamma function: for  $z > 0$ ,

$$\Gamma(z) = (2\pi)^{1/2} z^{z-\frac{1}{2}} \exp(-z) \exp\{\phi(z)\}, \quad (\text{A.10})$$

with  $0 < \phi(z) < K/z$  for some positive constant  $K$ . •

#### Proof of Theorem 4

Since, from Theorem 1,  $P_\gamma$  does not affect the existence of the posterior distribution, we consider the Bayesian model in (4.1) – (4.2) taking  $\gamma = 1$ . Using the representation in (A.8), the proof now proceed as follows:

- (A) Consider the joint distribution of  $(y_1, \dots, y_n, \beta, \tau, \lambda_1, \dots, \lambda_n)$ .
- (B) Integrate out  $\beta$  as a  $k$ -variate Normal.
- (C) Integrate out  $\tau$  using a Gamma distribution on  $\tau^2$ , which requires  $n > k$ .
- (D) Finally we are left with a function of  $(\lambda_1, \dots, \lambda_n)$ , which can be shown to be bounded [see proof of Theorem 2 (ii) in Fernández and Steel (1996)]. Thus, it is integrable for any probability distribution of  $(\lambda_1, \dots, \lambda_n)$ ; in particular, it is integrable under (A.9).

•

#### Proof of Theorem 5

Again, from Theorem 1, we simply take  $\gamma = 1$ .

- (A) Following the reasoning in the proof of Theorem 2 (i) in Fernández and Steel (1996), it is immediate that  $r < n - k$  is always required, for any choice of  $P_\nu$ , for the  $r^{\text{th}}$  order posterior moment of  $\beta_j$  to exist.
- (B) Furthermore, from the proof of Theorem 2 (ii) in Fernández and Steel (1996) [see (A.14) – (A.16) in that proof], we obtain that combining  $p_j = 0$  with  $r < n - k$  or  $p_j \geq 1$  with  $r \leq n - k - p_j$  leads to an  $r^{\text{th}}$  order posterior moment of  $\beta_j$ , for any choice of  $P_\nu$ .
- (C) Finally we show that when  $p_j \geq 1$ , posterior moments of  $\beta_j$  of order  $r > n - k - p_j$  do not exist:

From Theorem 3 (ii) of Fernández and Steel (1996) we know that if  $r \geq n - k - p_j + \nu(n - k - p_j + 1)$  [or, equivalently,  $\nu \leq \{r - (n - k - p_j)\}/(n - k - p_j + 1)$ ], then  $E(|\beta_j|^r | y_1, \dots, y_n, \nu) = \infty$ . Clearly, if  $r > n - k - p_j$ ,  $P_\nu(0, \{r - (n - k - p_j)\}/\{n - k - p_j + 1\}) > 0$ , which implies  $E(|\beta_j|^r | y_1, \dots, y_n) = \infty$ . •



### Proof of Theorem 6

Again we take  $\gamma = 1$ . From exactly the same argument used in Parts (A) and (B) of the proof of Theorem 5 we know that  $E(|\beta_j|^r | y_1, \dots, y_n) = \infty$  if  $r \geq n - k$ , whereas the latter integral is finite if  $p_j = 0$  and  $r < n - k$  or if  $p_j \geq 1$  and  $r \leq n - k - p_j$ . Thus, we only need to examine the case where  $p_j \geq 1$  and  $r \in (n - k - p_j, n - k)$ .

From the proof of Theorem 2 (ii) in Fernández and Steel (1996) [in particular, expressions (A.14) – (A.16) in that proof], follows that if  $r < n - k$  and

$$\int_{(0, \infty)^n} \left( \frac{\lambda_1}{\lambda_2} \right)^{(n-k-p_j-r)/2} p(\lambda_1, \dots, \lambda_n) d\lambda_1 \dots d\lambda_n < \infty, \quad (\text{A.11})$$

with  $p(\lambda_1, \dots, \lambda_n)$  as defined in (A.9), then  $E(|\beta_j|^r | y_1, \dots, y_n) < \infty$ . Using Fubini's Theorem we compute the integral in (A.11) in two steps: first we condition upon  $\nu$ , which requires  $r \leq n - k - p_j + \nu_0$  for a finite integral. We then obtain a function of  $\nu$ , which can be shown to be bounded by applying (A.10), whenever  $r < n - k - p_j + \nu_0$ ; therefore (A.11) holds for these values of  $r$ . •

### Proof of Theorem 7

We start by considering  $\gamma = 1$ .

- (A) If  $r \leq -(n - k)$  we know from Theorem 4 (i) in Fernández and Steel (1996) that  $E(\tau^r | y_1, \dots, y_n, \nu) = \infty$  for all  $\nu \in \mathbb{R}_+$ . Thus,  $E(\tau^r | y_1, \dots, y_n) = \infty$  for any  $P_\nu$ .
- (B) We now consider  $-(n - k) < r < 0$ . From the proof of Theorem 4 (ii) in Fernández and Steel (1996) [in particular, (A.23) – (A.24) in that proof], and with  $p(\lambda_1, \dots, \lambda_n)$  as defined in (A.9), we can deduce that

$$\int_0^\infty \int_{0 < \lambda_1 \leq \dots \leq \lambda_{n-k} < \infty} \left( \prod_{i=1}^{n-k} \lambda_i^{1/2} \right) \lambda_{n-k}^{-(n-k+r)/2} \left\{ \prod_{i=1}^{n-k} f_G \left( \lambda_i \left| \frac{\nu}{2}, \frac{\nu}{2} \right. \right) \right\} d\lambda_1 \dots d\lambda_{n-k} dP_\nu < \infty, \quad (\text{A.12})$$

implies a finite  $r^{th}$  order posterior moment of  $\tau$ . We now show that the inside integral in (A.12), which shall be denoted by  $I(\nu)$ , is a bounded function of  $\nu$ , and thus integrable, for any  $P_\nu$ . Since  $I(\nu)$  is continuous in  $\nu$ , we only need to prove that it has finite limits as  $\nu$  converges to zero and infinity. To show that each of these limits is finite, we consider two different upper bounds for  $I(\nu)$ .

- (B1) Limit as  $\nu \rightarrow \infty$ :

Since  $\lambda_{n-k} = \max\{\lambda_1, \dots, \lambda_{n-k}\}$ , we have  $(\prod_{i=1}^{n-k} \lambda_i^{1/2}) \lambda_{n-k}^{-(n-k+r)/2} \leq \lambda_{n-k}^{-r/2}$  and

$$I(\nu) \leq \int_0^\infty \lambda_{n-k}^{-r/2} f_G \left( \lambda_{n-k} \left| \frac{\nu}{2}, \frac{\nu}{2} \right. \right) d\lambda_{n-k}. \quad (\text{A.13})$$

The latter integral is proportional to

$$\nu^{r/2} \Gamma \left( \frac{\nu - r}{2} \right) \left\{ \Gamma \left( \frac{\nu}{2} \right) \right\}^{-1}, \quad (\text{A.14})$$

which, by applying (A.10), can be shown to have a finite limit as  $\nu \rightarrow \infty$ .

(B2) Limit as  $\nu \rightarrow 0$ :

We now perform the integral  $I(\nu)$  iteratively. In each of the  $n - k$  steps of the integration we use the upper bound

$$\int_0^\eta \lambda^{\omega-1} \exp(-\mu\lambda) d\lambda \leq \frac{\eta^\omega}{\omega}, \quad \text{for any } \omega, \mu > 0. \quad (\text{A.15})$$

This leads to an upper bound for  $I(\nu)$  proportional to

$$\frac{\nu^{r/2}}{(\nu+1)^{n-k-1}} \Gamma\left(\frac{(n-k)\nu-r}{2}\right) \left\{\Gamma\left(\frac{\nu}{2}\right)\right\}^{-(n-k)}. \quad (\text{A.16})$$

Applying (A.10) leads to an upper bound for (A.16) which has a finite limit as  $\nu \rightarrow 0$ .

(C) Finally we take  $r > 0$ : From Theorem 5 (ii) in Fernández and Steel (1996) we know that if  $r \geq (n-k)\nu$ , then  $E(\tau^r|y_1, \dots, y_n, \nu) = \infty$ . If  $r > 0$ ,  $P_\nu$  assigns positive probability to the interval  $(0, r/\{n-k\})$ , which precludes a finite  $r^{th}$  order posterior moment of  $\tau$ .

Combining (A)-(C) we obtain that, under  $\gamma = 1$ ,  $E(\tau^r|y_1, \dots, y_n) < \infty$  if and only if  $-(n-k) < r \leq 0$ . Applying Theorem 2 concludes the proof.  $\bullet$

### Proof of Theorem 8

Parts (A) and (B) of the proof of Theorem 7, together with Theorem 2, immediately lead to Theorem 8 (i). In order to prove Theorem 8 (ii), we follow the reasoning in Part (B) of the proof of Theorem 7, now considering  $r > 0$ . As was shown there,  $I(\nu)$  has an upper bound proportional to the expression in (A.14), which is bounded for  $\nu > \nu_0$  provided that  $r < \nu_0$ .  $\bullet$

### REFERENCES

- Arnold, B.C., and Groeneveld, R.A. (1995), "Measuring Skewness With Respect to the Mode," *The American Statistician*, 49, 34-38.
- Berger, J.O., and Bernardo, J.M. (1992), "On the Development of Reference Priors" (with discussion), in *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford University Press, pp. 35-60.
- Box, G.E.P., and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.
- Buckle, D.J. (1995), "Bayesian Inference for Stable Distributions," *Journal of the American Statistical Association*, 90, 605-613.
- Casella, G. and George, E. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167-174.
- Cheng, R.C.H. (1977), "The Generation of Gamma Variables With Non-Integral Shape Parameter," *Applied Statistics*, 26, 71-75.

- Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313-1321.
- DeGroot, M.H. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.
- Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer-Verlag.
- Dickey, J. (1971), "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters," *The Annals of Statistics*, 42, 204-223.
- Fernández, C., Osiewalski, J., and Steel, M.F.J. (1995), "Modeling and Inference With  $v$ -Spherical Distributions," *Journal of the American Statistical Association*, 90, 1331-1340.
- Fernández, C., and Steel, M.F.J. (1995), "Reference Priors in Non-Normal Location Problems", CentER Discussion Paper 9591, Tilburg University, The Netherlands.
- Fernández, C., and Steel, M.F.J. (1996), "On Bayesian Inference Under Sampling From Scale Mixtures of Normals," CentER Discussion Paper 9602, Tilburg University, The Netherlands.
- Gelfand, A., and Smith, A.F.M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.
- Geweke, J. (1991), "Efficient Simulation From the Multivariate Normal and Student- $t$  Distributions Subject to Linear Constraints," in *Computing Science and Statistics*, eds. E.M. Keramidas and S.M. Kaufman, Fairfax Station, VA: Interface Foundation, pp. 571-578.
- Geweke, J. (1994), "Priors for Macroeconomic Time Series and Their Application", *Econometric Theory*, 10, 609-632.
- Kinderman, A.J. and Monahan, J.F. (1977), "Computer Generation of Random Variables Using the Ratio of Uniform Deviates," *ACM Transactions on Mathematical Software*, 3, 257-260.
- Roberts, G.O., and Smith, A.F.M. (1994), "Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms," *Stochastic Processes and Their Applications*, 49, 207-216.
- Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.
- Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *Annals of Statistics*, 22, 1701-1762.
- van Zwet, W.R. (1964), "Convex Transformations of Random Variables", in *Mathematical Centre Tracts 7*, *Mathematisch Centrum*, Amsterdam.
- Verdinelli, I., and Wasserman, L. (1995), "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614-618.
- Whittaker, E.T. and Watson, G.N. (1927), *A Course of Modern Analysis: An Introduction to the General Theory of Infinite Processes and of Analytic Functions*, Cambridge: Cambridge University Press.

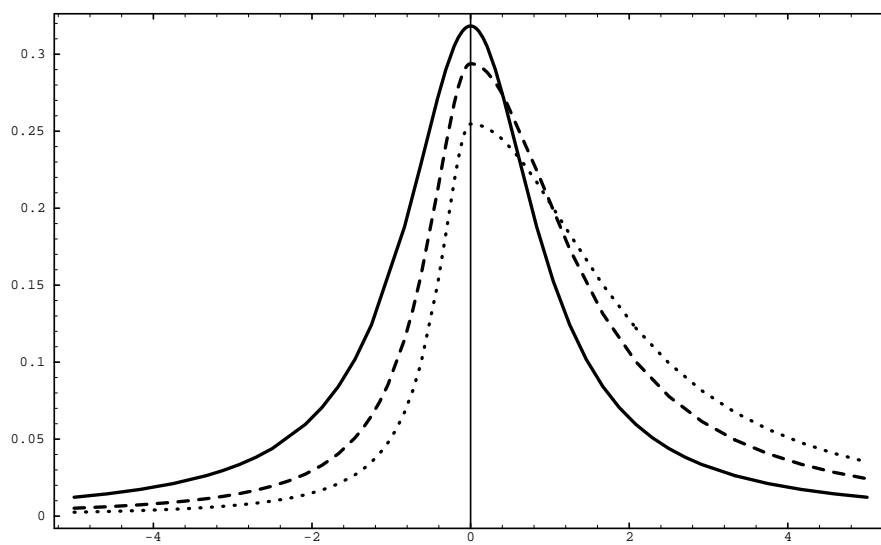


Figure 2: Posterior Density for beta

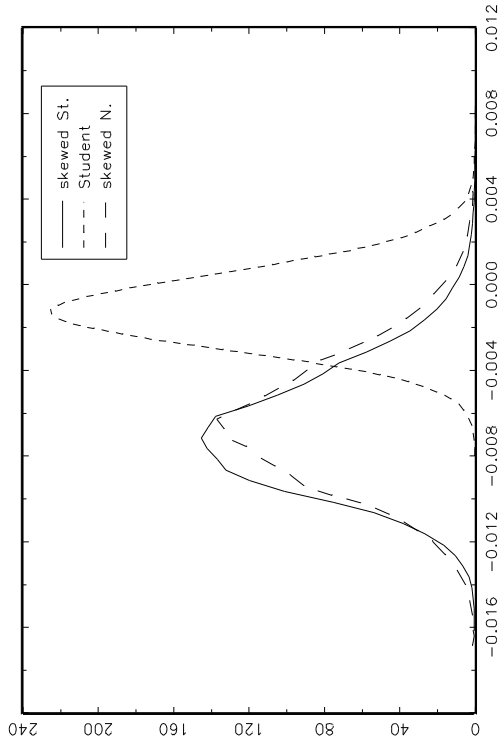


Figure 3: Posterior Density for tau

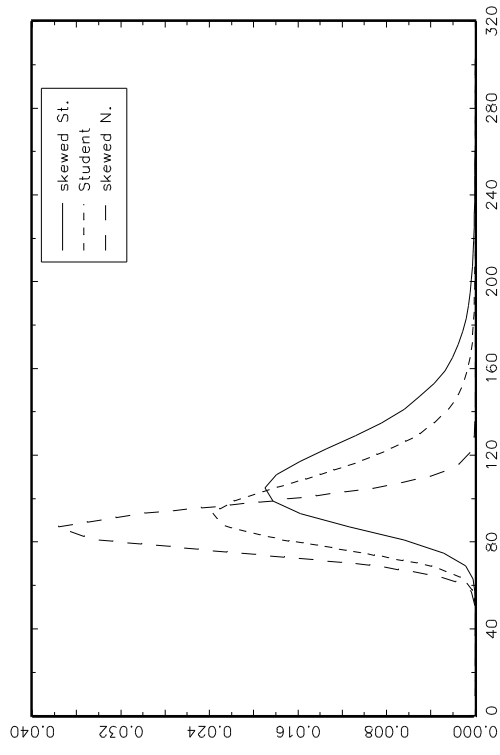


Figure 4: Posterior Density for gamma

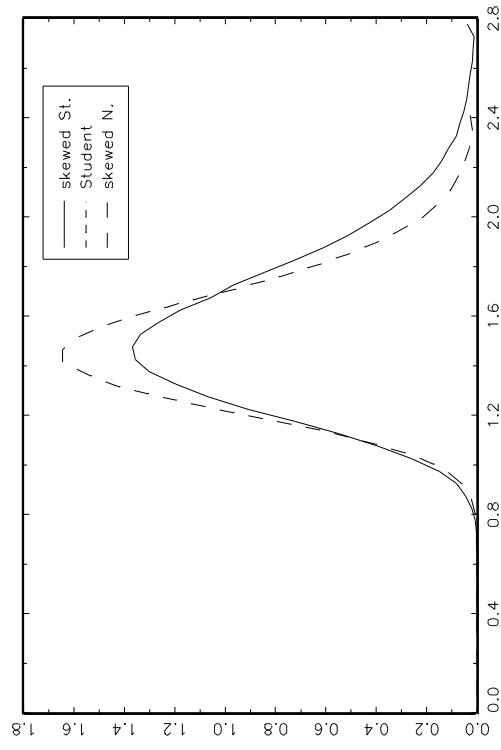


Figure 5: Posterior Density for nu

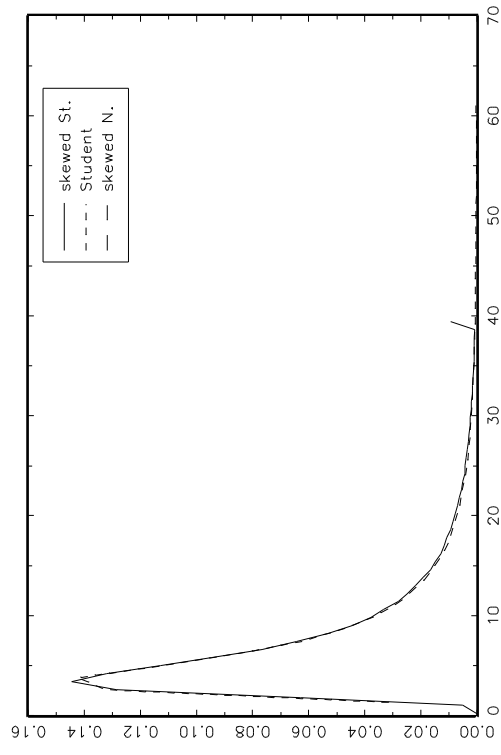


Figure 6: Price return data

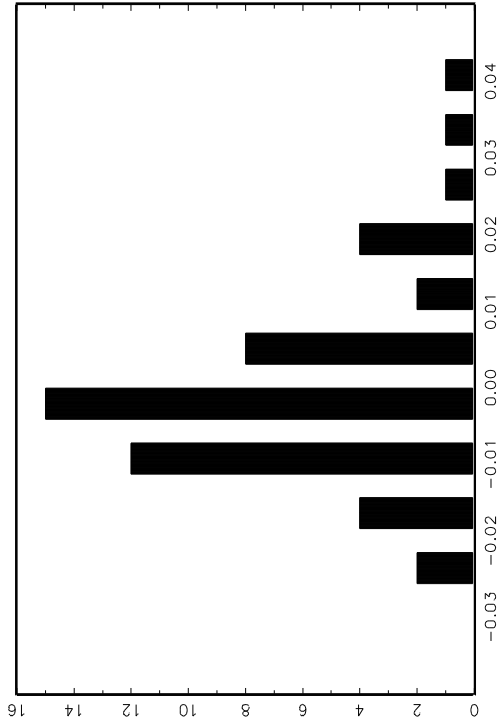


Figure 7: Predictive density

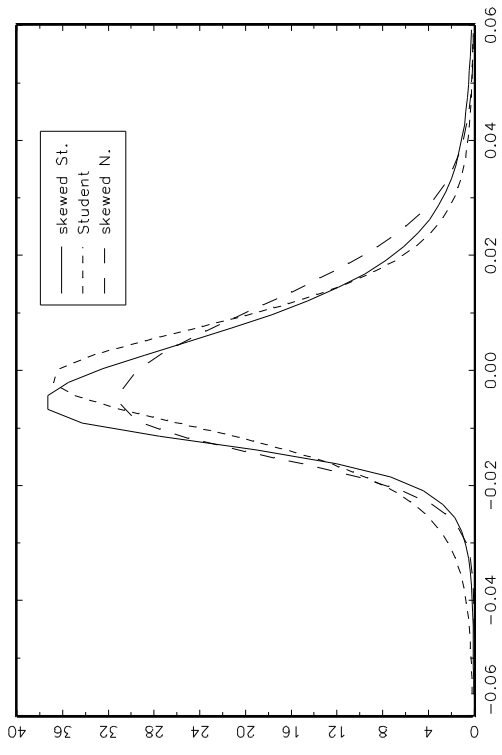


Figure 8: Posterior Density for beta1

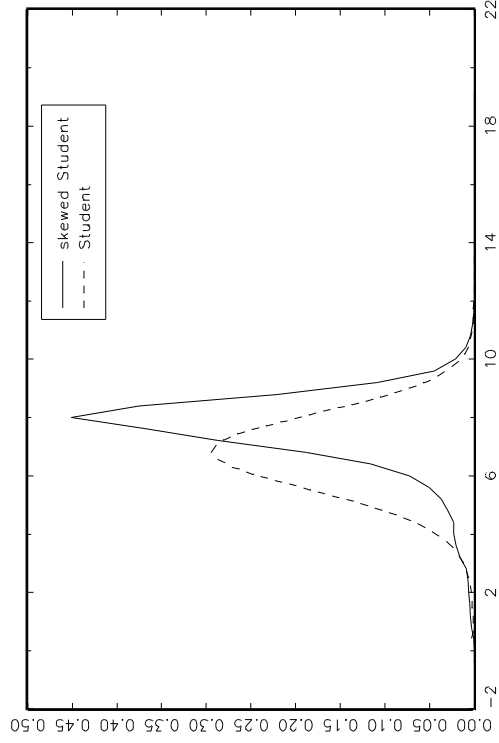


Figure 9: Posterior Density for beta2

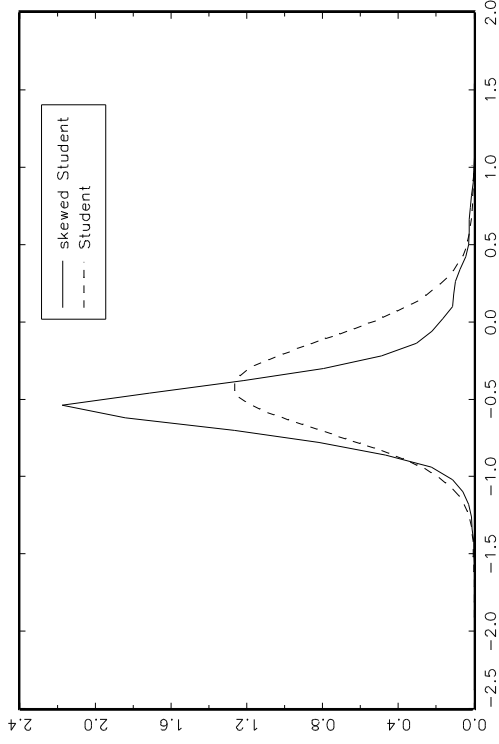


Figure 10: Posterior Density for gamma

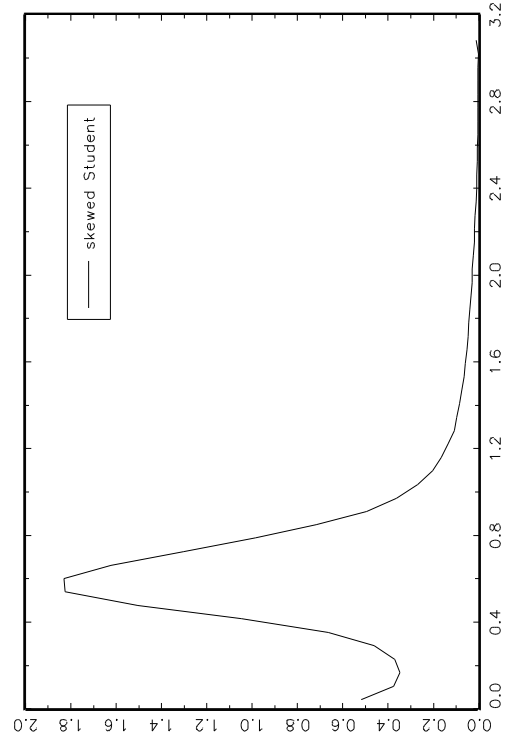


Figure 11: Posterior Density for nu

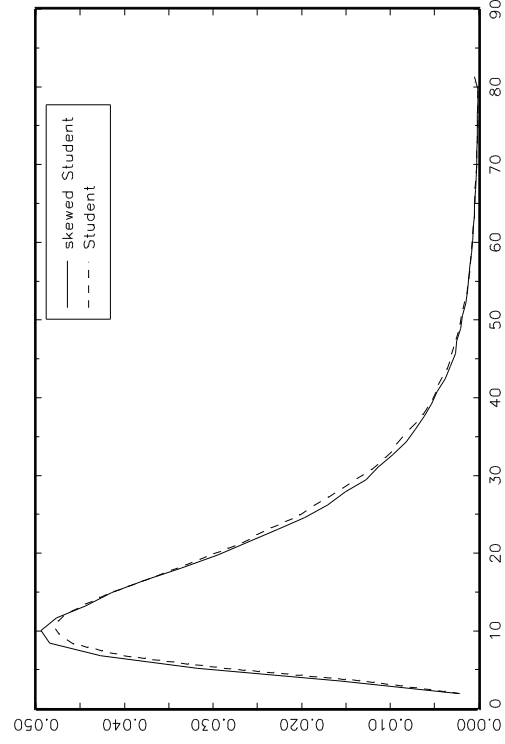


Figure 12: Predictive density for mean values of X

