On Bayesian Modeling of Fat Tails and Skewness
Author(s): Carmen Fernandez and Mark F. J. Steel
Source: *Journal of the American Statistical Association*, Vol. 93, No. 441 (Mar., 1998), pp. 359-371
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2669632
Accessed: 21/04/2010 06:26

On Bayesian Modeling of Fat Tails and Skewness

Carmen FERNÁNDEZ and Mark F. J. STEEL

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 datasets, and we 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 behavior is not affected. 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 article. 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, we suggest a numerical procedure using Gibbs sampling. The latter proves very easy to implement and renders the analysis of quite challenging problems a practical possibility. Some examples illustrate the use of this model in empirical data analysis.

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

1. INTRODUCTION

This article aims at including two pervasive features of empirical data in statistical modeling and inference. In particular, we 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 common phenomenon of skewed data with fat tails seems rather limited. Buckle (1995) provided a Bayesian analysis using stable laws. In addition, we can mention a few classical proposals: the use of g and h distributions by Badrinath and Chatterjee (1991), estimated through matching percentiles; partially adaptive estimation of generalized beta distributions of the second kind of McDonald and Nelson (1993); and (approximate) maximum likelihood estimation of generalized exponential distributions of Lye and Martin (1993). But all of these solutions seem quite complicated to implement numerically and, more important, 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 $\Re$), unimodal and symmetric distribution in a rather straightforward way; we simply use inverse scaling of the probability density function (pdf) on 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 on both sides of the mode. Tail behavior 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.

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 previously) into the error distribution on the existence of the posterior distribution and of its moments. In particular, it is shown that the existence of posterior moments of the regression parameter and of nonnegative order posterior moments of the scale is entirely unaffected by allowing for skewness.

Section 4 specifies the model further, by considering a linear regression structure with independent skewed Student error terms and unknown scale. We consider a standard "noninformative" prior on the regression and scale parameters. Furthermore, we do not fix tail behavior (controlled by the degrees-of-freedom parameter) or skewness, but leave both subject to inference. This model, which will be the main focus of this article thus allows for both skewness and flexible tail behavior.

In Section 4 we show that a Bayesian analysis can be conducted (i.e., the posterior is proper) if and only if the number of observations is larger than the number of regressors in the model, thus extending the well-known result under normal sampling to skewed Student sampling distributions. In addition, we provide results on the existence of posterior moments of regression and scale parameters.

We then design a Gibbs sampler (see Casella and George 1992; Gelfand and Smith 1990) using data augmentation, to conduct posterior inference with this model. The actual numerical implementation is shown to result in a very simple sampler that can easily be run on a PC for the analysis of moderately large datasets. Section 5 presents the details and illustrates that judgmental user input is restricted to a minimum.

Finally, Section 6 presents two empirical examples: a location-scale model applied to a dataset of share price returns, which was used by Buckle (1995) with the stable distribution as a modeling device, and a dataset from astron-

Carmen Fernández is Research Fellow, CentER for Economic Research, and Assistant Professor, Department of Econometrics, and Mark Steel is Senior Research Fellow, CentER for Economic Research, and Associate Professor, Department of Econometrics, Tilburg University, 5000 LE Tilburg, The Netherlands. The authors gratefully acknowledge comments from two anonymous referees, an associate editor, and the editor, as well as stimulating discussions with James O. Berger, David J. Buckle and Jean-François Richard. Both authors benefitted from a travel grant awarded by the Netherlands Organization for Scientific Research (NWO).

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

omy (a Hertzsprung–Russell diagram), where a regression model with two explanatory variables is used. Posterior and predictive inference is conducted for the general model with skewness and fat tails, and also for models that account for only one of these features. In addition, Bayes factors between these models are computed using the methods advocated by Chib (1995) and Verdinelli and Wasserman (1995). Section 6 also illustrates the generality and flexibility of the class of skewed Student distributions by performing inference on the basis of a rather extreme sample. The resulting predictive distribution very closely matches the stable distribution from which the drawings were generated. All proofs are deferred to the Appendix, without explicit mention in the text body.

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

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 by Fernández, Osiewalski, and Steel (1995), who introduced a skewed version of the exponential power distribution.

Let us consider a univariate pdf $f(\cdot)$, which is unimodal and symmetric around 0. 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 $\gamma \in (0, \infty)$:

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

The basic idea underlying (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 0 but loses symmetry whenever $\gamma \neq 1$. More formally, we deduce

$$p(\varepsilon|\gamma = 1) = f(\varepsilon) \quad (2)$$

and

$$\frac{P(\varepsilon \geq 0|\gamma)}{P(\varepsilon < 0|\gamma)} = \gamma^2, \quad (3)$$

from which it is clear that γ controls the allocation of mass to each side of the mode. Furthermore, the way in which γ intervenes in (1) implies

$$p(\varepsilon|\gamma) = p(-\varepsilon|1/\gamma), \quad (4)$$

so that inverting γ produces the mirror image around 0. 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.

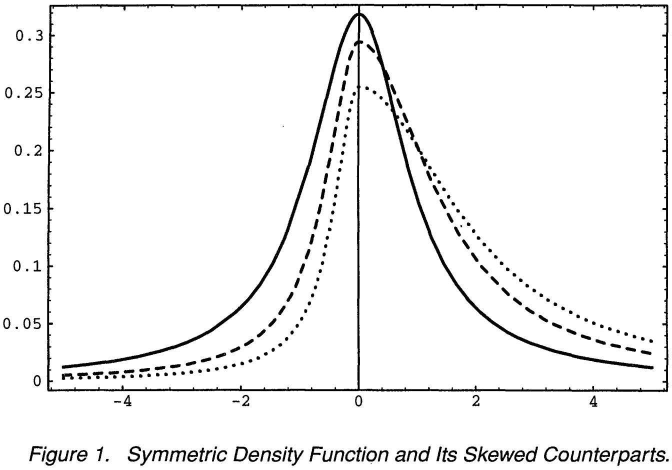

Figure 1. Symmetric Density Function and Its Skewed Counterparts. ——, γ = 1; – – –, γ = 1.5; · · ·, γ = 2.

To gain more insight in the properties of (1), let us examine how γ affects its moments. Generally, (1) leads to a finite rth order moment ($r \in \Re$) if and only if the corresponding moment of $f(\cdot)$ exists (i.e., for $\gamma = 1$). In particular, we obtain

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

where

$$M_r = \int_0^\infty s^r 2f(s)\, ds. \quad (5)$$

Of course, $E(\varepsilon^r|\gamma)$ will be real-valued only for integer r. In addition, the unimodality of $f(\cdot)$ implies that $M_r = \infty$ for $r \leq -1$. Thus, let us concentrate on positive integer order moments. From (5), the following properties can be shown to hold for noncentered moments: For odd r, the rth order moment retains the same absolute value but changes sign if we invert γ, takes the value 0 only for $\gamma = 1$, and is an increasing function of γ with $\lim_{\gamma\to\infty} E(\varepsilon^r|\gamma) = \infty$. Even moments, on the other hand, are entirely unaffected by inverting γ and again increase without bounds in γ for $\gamma > 1$. Consequently, $\min_\gamma E(\varepsilon^r|\gamma) = E(\varepsilon^r|\gamma = 1)$ for even r. Expressions for centered moments are readily available from (5). In particular, the variance possesses all of the properties just mentioned for even noncentered moments.

Skewness, as measured by the third centered moment divided by the cubed standard deviation, is given by

$$\mathrm{Sk}(\varepsilon|\gamma) = \left(\gamma - \frac{1}{\gamma}\right) \times \frac{(M_3 + 2M_1^3 - 3M_1M_2)\left(\gamma^2 + \frac{1}{\gamma^2}\right) + 3M_1M_2 - 4M_1^3}{\left\{(M_2 - M_1^2)\left(\gamma^2 + \frac{1}{\gamma^2}\right) + 2M_1^2 - M_2\right\}^{3/2}}. \quad (6)$$

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

Another popular measure of skewness is the Pearson measure, defined through the difference between mean and mode divided by the standard deviation. Because the pdf in (1) has zero mode, we obtain

$$\mathrm{SP}(\varepsilon|\gamma) = \frac{M_1\left(\gamma - \frac{1}{\gamma}\right)}{\left\{(M_2 - M_1^2)\left(\gamma^2 + \frac{1}{\gamma^2}\right) + 2M_1^2 - M_2\right\}^{1/2}}. \tag{7}$$

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

In the context of the class of unimodal distributions defined in (1), a natural measure of skewness is that introduced by Arnold and Groeneveld (1995), defined as one minus two times the probability mass left of the mode, leading to

$$\mathrm{SM}(\varepsilon|\gamma) = \frac{\gamma^2 - 1}{\gamma^2 + 1}, \tag{8}$$

which is a strictly increasing function of γ, taking values anywhere in $(-1, 1)$. The results of 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 $\mathrm{SM}(\varepsilon|\gamma) = -\mathrm{SM}(\varepsilon|1/\gamma)$ and $\mathrm{SM}(\varepsilon|\gamma = 1) = 0$. In contrast to the skewness coefficients in (6) and (7), (8) does not depend on the choice of $f(\cdot)$, and the entire range of this skewness measure can be covered by choosing γ appropriately with $\lim_{\gamma\to 0}\mathrm{SM}(\varepsilon|\gamma) = -1$ (extreme left skewness) and $\lim_{\gamma\to\infty}\mathrm{SM}(\varepsilon|\gamma) = 1$ (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 assume the observables $y_i \in \Re, i = 1, \ldots, n$, to be generated from

$$y_i = g_i(\boldsymbol{\beta}) + \sigma\varepsilon_i, \tag{9}$$

where $g_i(\cdot)$ is a known measurable function from $\Re^k$ $(k \geq 1)$ to $\Re, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' \in \Re^k$ parameterizes the location, and $\sigma \in \Re_+$ is a scale parameter. We assume the error terms $\varepsilon_1, \ldots, \varepsilon_n$ to be iid given a parameter $\nu \in \mathcal{N}$ (possibly of infinite dimension) and $\gamma \in \Re_+$ with conditional pdf

$$p(\varepsilon_i|\nu, \gamma) = \frac{2}{\gamma + \frac{1}{\gamma}}\left\{f_\nu\left(\frac{\varepsilon_i}{\gamma}\right)I_{[0,\infty)}(\varepsilon_i) + f_\nu(\gamma\varepsilon_i)I_{(-\infty,0)}(\varepsilon_i)\right\}, \tag{10}$$

where $f_\nu(\cdot)$ is unimodal and symmetric around 0. This stochastic assumption introduces two extra parameters into the problem: γ, the skewness parameter (as explained in the previous section), and ν, which can describe other properties of the sampling distribution. In particular, ν will control the thickness of the tails in the next section.

We adopt the following class of prior distributions:

$$P_{(\boldsymbol{\beta},\sigma,\nu,\gamma)} = P_{\boldsymbol{\beta}} \times P_\sigma \times P_\nu \times P_\gamma, \tag{11}$$

where P_σ is the usual noninformative distribution characterized by the improper density

$$p(\sigma) \propto \sigma^{-1} \tag{12}$$

on $\Re_+, P_{\boldsymbol{\beta}}$ is any σ-finite measure on $\Re^k$, and P_ν and P_γ are proper distributions. An important special case of (11) is where P_γ is a point mass at 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_γ to those where P_γ is a Dirac measure at 1. For notational simplicity, we denote the latter case by $\gamma = 1$.

First, because the prior distribution in (11)–(12) is improper, we need to verify existence of the posterior distribution. In addition, our interest is focused on the location and scale parameters $\boldsymbol{\beta}$ and σ, because ν and γ are merely auxiliary parameters to widen the class of sampling distributions. Thus we also address the issue of the existence of posterior moments of $\boldsymbol{\beta}$ and σ.

Theorem 1. Consider n independent replications from the sampling distribution in (9)–(10) and the prior in (11)–(12). Given $(r_1, \ldots, r_k) \in \Re^k$ and $r \geq 0$, we obtain that for any P_γ,

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

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

Theorem 1 clearly states that the existence of posterior moments of $\boldsymbol{\beta}$ is entirely unaffected by the added uncertainty on γ. Furthermore, P_γ is also irrelevant for obtaining a finite posterior moment of σ of nonnegative order, although it can be shown that this result generally does not extend to values of $r < 0$. An important special case of Theorem 1 is where $r = r_j = 0$ for all $j \in \{1, \ldots, k\}$, which establishes the fact that incorporating skewness in the sampling does not affect properness of the posterior distribution.

4. INFERENCE UNDER SKEWED STUDENT SAMPLING

In the previous section we assessed the effect of skewing a symmetric unimodal error distribution with pdf $f_\nu(\cdot)$ on the existence of posterior moments. Here we fully specify a Bayesian model that accounts for both skewness and fat tails, and the sequel of this article will be devoted to posterior and predictive inference from this model. Whereas this section groups results on the properness of the posterior and the existence of its moments, the next section provides a numerical framework for conducting inference from this model.

In particular, we consider the following special case of the model in (9)–(12):

a. We specify a linear regression model in (9); that is, $g_i(\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, where $\mathbf{x}_i \in \Re^k$ is a vector of explanatory variables. Throughout, we condition on $\mathbf{x}_i$ without explicit mention. The entire design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is always assumed to be of full column rank k, which implies that $n \geq k$.
b. $f_\nu(\cdot)$ is chosen to be the pdf of a standard Student t distribution with ν df. Thus $\nu \in \Re_+$.
c. For the prior of $\boldsymbol{\beta}$, we take the improper uniform distribution on $\Re^k$. This leads to $p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1}$, which corresponds to the usual noninformative distribution for regression and scale parameters and is the reference prior in the sense of Berger and Bernardo (1992) if γ and ν are known (see Fernández and Steel 1995). Following (11), P_γ and P_ν are taken to be any probability measures on $\Re_+$.

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

$$p(y_i|\boldsymbol{\beta}, \sigma, \nu, \gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \sigma^{-1} \times \left[1 + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\nu\sigma^2} \left\{\frac{1}{\gamma^2} I_{[0,\infty)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \gamma^2 I_{(-\infty,0)}(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right\}\right]^{-(\nu+1)/2} \tag{13}$$

with prior distribution

$$P_{(\boldsymbol{\beta},\sigma,\nu,\gamma)} = P_{\boldsymbol{\beta}} \times P_\sigma \times P_\nu \times P_\gamma,$$

where

$$P_{\boldsymbol{\beta}} \times P_\sigma \quad \text{has density} \quad p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1}$$

and

$$P_\gamma \text{ and } P_\nu \text{ are proper.} \tag{14}$$

The sampling distribution in (13) is denoted by "skewed Student" with location $\mathbf{x}_i'\boldsymbol{\beta}$, scale σ, ν df, and skewness parameter γ. Now we briefly discuss the interpretation of the parameters in (13). $\boldsymbol{\beta} \in \Re^k$ groups the regression coefficients, usually of primary interest, and $\sigma \in \Re_+$ is the scale parameter. In addition to these parameters of interest, (13) contains two more parameters, each with a clearly defined modeling purpose. The thickness of the tails is determined entirely by $\nu \in \Re_+$. From our results in Section 2 [see, e.g., (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 (but not including) ν, as under Student sampling. Skewness is controlled by $\gamma \in \Re_+$, as explained in Section 2. Following (3), γ determines the amount of mass on both sides of the location:

$$\frac{P(y_i \geq \mathbf{x}_i'\boldsymbol{\beta}|\boldsymbol{\beta}, \sigma, \nu, \gamma)}{P(y_i < \mathbf{x}_i'\boldsymbol{\beta}|\boldsymbol{\beta}, \sigma, \nu, \gamma)} = \gamma^2. \tag{15}$$

Before discussing existence of posterior moments from the Bayesian model in (13)–(14), we stress that (from Sec. 3) P_γ does not affect properness of the posterior distribution, or the existence of posterior moments of $\boldsymbol{\beta}$ and of nonnegative order moments of σ. Thus the results presented here also apply to the case of (symmetric) Student t sampling. We examined the case of symmetric Student t sampling with fixed degrees of freedom ν in earlier work (Fernández and Steel 1996). Here we explicitly incorporate prior uncertainty on the thickness of the tails and on skewness, as both can be crucial modeling instruments.

Because the prior distribution in (14) is improper, we first investigate properness of the posterior distribution.

Theorem 2. With n independent replications from the sampling model in (13) under the prior in (14), we obtain a proper posterior distribution if and only if $n > k$, for any choices of P_ν and P_γ.

This well-known result under normal sampling is thus seen to hold in our much more general framework, which allows for both skewness and fat tails. Clearly, any Bayesian inference from this model will require at least $k+1$ observations. Thus throughout the sequel of the article, we shall assume $n \geq k+1$.

We now present our findings for marginal posterior moments of the components of $\boldsymbol{\beta}$. The following technical definition concerning the design matrix $\mathbf{X}$ is 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 $\mathbf{X}$, we define the singularity index for column $j = 1, \ldots, k$ 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 $\mathbf{X}$ of rank $k-1$ that retains rank $k-1$ after removing its jth column.

Clearly, if $\mathbf{X}$ contains rows of 0s, then p_j is at least equal to the number of such rows for all $j = 1, \ldots, k$. Furthermore, $\max\{p_j : j = 1, \ldots, k\} = 0$ if and only if every $k \times k$ submatrix of $\mathbf{X}$ is nonsingular. The singularity index p_j plays a crucial role in the existence of posterior moments of β_j, as Theorem 3 illustrates.

Theorem 3. Consider n observations from the sampling model (13) and the prior in (14):

a. If $P_\nu(0, c) > 0$ for all positive c, then for any $r \geq 0$,

$$E(|\beta_j|^r|y_1, \ldots, y_n) < \infty$$

$$\text{if and only if } \begin{cases} r < n-k & \text{if } p_j = 0 \\ r \leq n-k-p_j & \text{if } p_j \geq 1. \end{cases}$$

b. If P_ν has support on (ν_0, ∞), for some $\nu_0 > 0$ we obtain:

(1) if $r \geq n-k$, then $E(|\beta_j|^r|y_1,\ldots,y_n) = \infty$,
(2) if $0 \leq r < \min\{n-k, n-k-p_j+\nu_0\}$, then $E(|\beta_j|^r|y_1,\ldots,y_n) < \infty$.

In practice, the most common situation where Theorem 3a applies is when P_ν is given through a pdf verifying $p(\nu) > 0$ for all $\nu \in (0, C)$, where C is some positive constant. Under the assumption of Theorem 3a, the design matrix affects existence of moments of β_j only through p_j, the singularity index of column j, and the order up to which posterior moments of β_j exist decreases with p_j. Thus in some sense, the higher p_j is, the less can be learned about β_j. If $p_j = 0$ (intuitively, the best type of design matrix for β_j), then 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 β_j. Note that different elements of $\boldsymbol{\beta}$ can have posterior moments up to different orders.

The sampling model in (13) has moments up to and not including ν. Therefore, if we wish to guarantee finite sampling moments of a certain order $\nu_0 > 0$, then we need to restrict ν to be larger than ν_0; that is, consider distributions P_ν with support on (ν_0, ∞). In this situation, more moments of the regression coefficients can be shown to exist, as is obvious from Theorem 3b(2). In contrast to the situation where P_ν has mass arbitrarily close to 0, moments of order smaller than $\min\{n-k, \nu_0\}$ will now exist for any design matrix $\mathbf{X}$. Thus the design matrix can no longer destroy the existence of all positive-order moments of $\boldsymbol{\beta}$.

Finally, in the important special case of the location-scale model (i.e., where $\mathbf{x}_i'\boldsymbol{\beta} = \beta \in \Re$), $p_1 = 0$ and posterior moments of β exist exactly up to (and not including) $n-1$, irrespective of the choice of P_ν (and P_γ).

Let us now consider posterior moments of σ of order $r \geq 0$. In this case the order up to which posterior moments are finite does not depend on either P_γ or P_ν, as evidenced by the following theorem.

Theorem 4. For the Bayesian model in (13)–(14), we obtain for $r \geq 0$,

$$E(\sigma^r|y_1,\ldots,y_n) < \infty \quad \text{if and only if } r < n-k.$$

5. NUMERICAL IMPLEMENTATION

To conduct inference with the Bayesian model in (13)–(14), numerical methods are needed. In particular, we 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, which allows us to alternatively express the sampling density in (13) as

$$p(y_i|\boldsymbol{\beta},\sigma,\nu,\gamma) = \left(\frac{2}{\pi}\right)^{1/2} \frac{1}{\gamma+\frac{1}{\gamma}} \int_0^\infty \lambda_i^{1/2}\sigma^{-1} \times \exp\left[-\frac{\lambda_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\left\{\frac{1}{\gamma^2}I_{[0,\infty)}(y_i-\mathbf{x}_i'\boldsymbol{\beta}) + \gamma^2 I_{(-\infty,0)}(y_i-\mathbf{x}_i'\boldsymbol{\beta})\right\}\right] \times f_{\mathrm{G}}\left(\lambda_i\Big|\frac{\nu}{2},\frac{\nu}{2}\right)d\lambda_i, \tag{16}$$

where

$$f_{\mathrm{G}}(\lambda_i|a,b) = b^a\Gamma(a)^{-1}\lambda_i^{a-1}\exp(-b\lambda_i) \tag{17}$$

denotes the pdf of a gamma(a,b) distribution. Thus each observation $y_i, i = 1,\ldots,n$ has its own mixing parameter λ_i and $\lambda_1,\ldots,\lambda_n$ are iid given ν. Augmenting the parameter set with $(\lambda_1,\ldots,\lambda_n)$ will greatly facilitate the numerical analysis. Therefore, we conduct a Gibbs sampler on $(\boldsymbol{\beta},\sigma,\nu,\gamma,\lambda_1,\ldots,\lambda_n|y_1,\ldots,y_n)$. 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 $\boldsymbol{\beta}$

We analyze each element of $\boldsymbol{\beta}$ in a separate Gibbs step. From (16) and (14), the conditional posterior pdf of $\beta_j, j \in \{1,\ldots,k\}$, is defined by

$$p(\beta_j|\{\beta_s\colon s\neq j\},\sigma,\nu,\gamma,\lambda_1,\ldots,\lambda_n,y_1,\ldots,y_n) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n \lambda_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2 \times \left\{\frac{1}{\gamma^2}I_{[0,\infty)}(y_i-\mathbf{x}_i'\boldsymbol{\beta}) + \gamma^2 I_{(-\infty,0)}(y_i-\mathbf{x}_i'\boldsymbol{\beta})\right\}\right], \tag{18}$$

which we now rewrite in a form that immediately suggests a simple algorithm for generating random drawings. Clearly, those observations for which x_{ij}, the jth element of $\mathbf{x}_i$, is 0 do not contribute to the conditional distribution of β_j in (18). For the m remaining observations, we compute $w_i^{(j)} = (y_i - \mathbf{x}_i'\boldsymbol{\beta} + x_{ij}\beta_j)/x_{ij}$, noting that the full column-rank assumption on $\mathbf{X}$ implies that $m \geq 1$. Then we order the observations such that $w_1^{(j)} < w_2^{(j)} \ldots < w_m^{(j)}$ and partition $\Re$, the domain of β_j, into the sets $S_0^{(j)} = (-\infty, w_1^{(j)}], S_h^{(j)} = (w_h^{(j)}, w_{h+1}^{(j)}]$ for $h = 1,\ldots,m-1$ and $S_m^{(j)} = (w_m^{(j)}, \infty)$. Ultimately, we can express the conditional posterior of β_j as

$$p(\beta_j|\{\beta_s\colon s\neq j\},\sigma,\nu,\gamma,\lambda_1,\ldots,\lambda_n,y_1,\ldots,y_n) \propto \sum_{h=0}^m \{p_h^{(j)}\}^{-1/2} \times \exp\left(-\frac{l_h^{(j)}}{2\sigma^2}\right) f_{\mathrm{N}}^1\left(\beta_j\Big|\mu_h^{(j)},\frac{\sigma^2}{p_h^{(j)}}\right) I_{S_h^{(j)}}(\beta_j), \tag{19}$$

with $f_{\mathrm{N}}^1(\cdot|t,v)$ the pdf of a univariate normal distribution with mean t and variance $v, I_S(\cdot)$ is the indicator function of the set S, and

$$p_h^{(j)} = \sum_{i=1}^{h} \rho_{i1}^{(j)} + \sum_{i=h+1}^{m} \rho_{i2}^{(j)},$$

$$p_h^{(j)} \mu_h^{(j)} = \sum_{i=1}^{h} \rho_{i1}^{(j)} w_i^{(j)} + \sum_{i=h+1}^{m} \rho_{i2}^{(j)} w_i^{(j)},$$

and

$$l_h^{(j)} = \sum_{i=1}^{h} \rho_{i1}^{(j)} \{w_i^{(j)}\}^2 + \sum_{i=h+1}^{m} \rho_{i2}^{(j)} \{w_i^{(j)}\}^2 - p_h^{(j)} \{\mu_h^{(j)}\}^2, \quad (20)$$

where we have defined

$$\rho_{i1}^{(j)} = \lambda_i x_{ij}^2 \left\{ \frac{1}{\gamma^2} I_{(-\infty,0)}(x_{ij}) + \gamma^2 I_{(0,\infty)}(x_{ij}) \right\},$$

$$\rho_{i2}^{(j)} = \lambda_i x_{ij}^2 \left\{ \gamma^2 I_{(-\infty,0)}(x_{ij}) + \frac{1}{\gamma^2} I_{(0,\infty)}(x_{ij}) \right\}. \quad (21)$$

The expression in (19) is now straightforward to draw from. Truncated normal random variates are generated through the mixed rejection algorithm of Geweke (1991).

5.2 Conditional of σ

It is immediate from (16) and (14) that

$$p(\sigma^{-2}|\boldsymbol{\beta}, \nu, \gamma, \lambda_1, \ldots, \lambda_n, y_1, \ldots, y_n) = f_{\mathrm{G}}\left(\sigma^{-2} \middle| \frac{n}{2}, \frac{1}{2} \sum_{i=1}^{n} \lambda_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \times \left\{ \frac{1}{\gamma^2} I_{[0,\infty)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \gamma^2 I_{(-\infty,0)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \right\} \right), \quad (22)$$

from which random drawings can immediately be generated.

5.3 Conditional of ν

The general form of the conditional distribution of ν given $(\boldsymbol{\beta}, \sigma, \gamma, \lambda_1, \ldots, \lambda_n, y_1, \ldots, y_n)$ depends on the prior P_ν. In our empirical section, the latter is an exponential distribution with pdf,

$$p(\nu) = d \exp(-d\nu), \quad (23)$$

leading to

$$p(\nu|\boldsymbol{\beta}, \sigma, \gamma, \lambda_1, \ldots, \lambda_n, y_1, \ldots, y_n) \propto \left(\frac{\nu}{2}\right)^{n\nu/2} \left\{ \Gamma\left(\frac{\nu}{2}\right) \right\}^{-n} \times \exp\left[-\nu \left\{ d + \frac{1}{2} \sum_{i=1}^{n} (\lambda_i - \log \lambda_i) \right\} \right]. \quad (24)$$

Drawings from (24) are 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 by Geweke (1992, app. A).

5.4 Conditional of γ

The conditional distribution of γ given $(\boldsymbol{\beta}, \sigma, \nu, \lambda_1, \ldots, \lambda_n, y_1, \ldots, y_n)$ depends on P_γ. In our empirical section, we use a gamma(a, b) (see (17)) prior on $\varphi \equiv \gamma^2$, leading to

$$p(\varphi|\boldsymbol{\beta}, \sigma, \nu, \lambda_1, \ldots, \lambda_n, y_1, \ldots, y_n) \propto \varphi^{(n/2)+a-1} (\varphi + 1)^{-n} \times \exp\left\{ -\left(\frac{\vartheta}{\varphi} + \kappa\varphi \right) \right\}, \quad (25)$$

where we have defined

$$\vartheta = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \lambda_i (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 I_{[0,\infty)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \geq 0$$

and

$$\kappa = b + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \lambda_i (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 I_{(-\infty,0)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) > 0. \quad (26)$$

The distribution in (25) 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 the ratio-of-uniforms method (see Devroye 1986) can be applied.

5.5 Conditional of $\lambda_1, \ldots, \lambda_n$

Drawing from the conditional distribution of the mixing parameters is straightforward, as they are independent with pdf

$$p(\lambda_1, \ldots, \lambda_n|\boldsymbol{\beta}, \sigma, \nu, \gamma, y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{\mathrm{G}}\left(\lambda_i \middle| \frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2} \times \left\{ \frac{1}{\gamma^2} I_{[0,\infty)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \gamma^2 I_{(-\infty,0)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \right\} \right). \quad (27)$$

The full conditional distributions in (19), (22), (24), (25), and (27) 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, because the parameter space has a Cartesian product structure (see Roberts and Smith 1994).

6. EXAMPLES

6.1 Preliminaries

In this section we 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 (14) with an exponential distribution on ν as in (23), and a gamma(a, b) prior (with pdf as in (17)) for $\varphi = \gamma^2$. Thus a full description of our prior distribution still requires a choice for d in (23) and for a and b. In eliciting these hyperparameters we try to avoid introducing strong prior information. To this end, we choose $d = .1$, thus obtaining a prior mean of ν 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, γ, we specify a prior with mean $E(\gamma) = 1$, which centers the prior around the case of symmetric sampling. This allows us to fix b as a function of a. Furthermore, we elicit a using the prior variance of γ and the prior mass on the interval (0, 1). We feel that the value $a = .5$, leading to var$(\gamma) = .57$ and $P(\gamma < 1) = .58$, is quite reasonable. The resulting gamma prior of $\varphi = \gamma^2$ corresponds to a half-normal prior for γ. We adopt these prior choices in all examples subsequently analyzed.

Besides the general model allowing for both skewness and fat tails simultaneously, we also consider simpler versions, which incorporate only one of these features at a time. Thus we examine three possible sampling models: the skewed Student in (13), the skewed normal (the limiting case of (13) as $\nu \to \infty$) and the Student t model ((13) with $\gamma = 1$). Priors for parameters present in the models will always be as described earlier.

In the sequel we will present posterior inference on model parameters and predictive inference in the context of each model. We will conduct the latter through averaging the sampling density, using the Rao–Blackwell argument suggested by 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.

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. But much smaller runs already lead to reliable results.

Before proceeding with the examples, a technical issue still must be addressed. Theorem 2 assures us that whenever $n > k$, where k is the number of regressors in the model, we have a posterior distribution. But this does not prevent the within-sample predictive density $p(y_1, \ldots, y_n)$ from being infinite in a set of Lebesgue measure zero in $\Re^n$. When P_ν has mass arbitrarily close to 0 (as is the case with the exponential prior considered here), we have shown (Fernández and Steel 1997a) that any sample that contains more than k observations that can be fitted exactly as $y_i = \mathbf{x}_i'\boldsymbol{\beta}$ for some fixed value of $\boldsymbol{\beta}$, leads to $p(y_1, \ldots, y_n) = \infty$. Whereas the set of problematic samples clearly has Lebesgue measure zero in $\Re^n$ and thus poses no theoretical problem, the censoring and rounding mechanisms underlying many empirical observations may lead to datasets displaying this problematic behavior (as is the case in Examples 1 and 3, which follow). There are several possible solutions to this, the most appealing of which is to explicitly incorporate the censoring mechanism into the model. This approach, which we have developed in detail elsewhere (Fernández and Steel 1997a,b), is outside the scope of this paper. The solution that we have adopted in Examples 1 and 3 is to slightly perturb the y_i's. In both cases we have checked that the empirical impact of this minor perturbation is quite negligible.

Another closely related issue is that when P_ν has mass arbitrarily close to 0 the postsample predictive density is unbounded in a neighborhood of each observed data point. But this is very unlikely to be noticed when using numerical techniques, because these neighborhoods are extremely small and contain virtually no probability mass. Thus for all practical purposes, the smooth predictive densities plotted in Examples 1 and 2 are very accurate approximations to the actual predictive. (More details on this issue are provided in Fernández and Steel 1997b.)

6.2 Example 1: Share Price Returns

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

Posterior results using the general sampling model in (13) with the prior as explained in Section 6.1 are summarized in Table 1 and in Figures 2–5. Along with the general skewed Student sampling model, we also use 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 β and σ exist up to (but not including) order $n - k = 48$ in all three models. Table 1 reports posterior means and standard deviations of β and σ.

Figure 4 clearly indicates right skewness in the data; thus if our model does not account for this skewness, then the location will be shifted to the right, as occurs for the Student t model (see Fig. 2). As Figure 5 indicates, ν has substantial posterior mass in regions corresponding to thick tails. Thus the skewed normal model, which has normal tail behavior, needs to increase the scale σ to capture observations in the tails. Figure 3 indicates smaller values for σ if we account for fat tails and even more if we allow for skewness as well (see also Table 1).

Table 1. Posterior Moments for Share Price Returns

	Skewed Student	*Student*	*Skewed normal*
Mean β	−.0068	−.0012	−.0064
SD β	(.0028)	(.0018)	(.0031)
Mean σ	.0091	.0103	.0117
SD σ	(.0018)	(.0018)	(.0014)

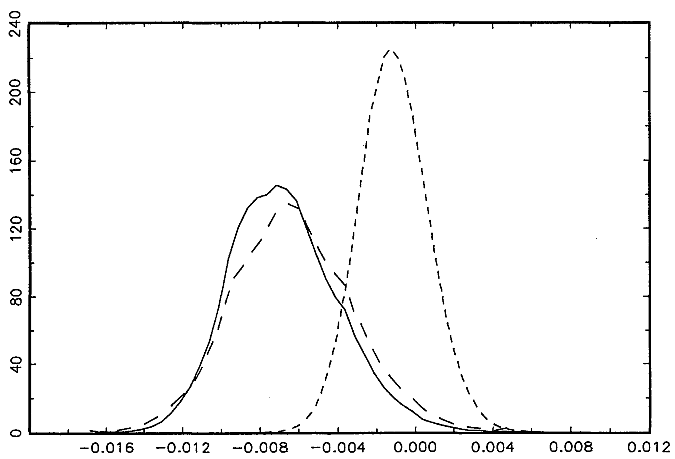

Figure 2. Posterior Density for β. ——, skewed Student; - - -, Student; — — —, skewed normal.

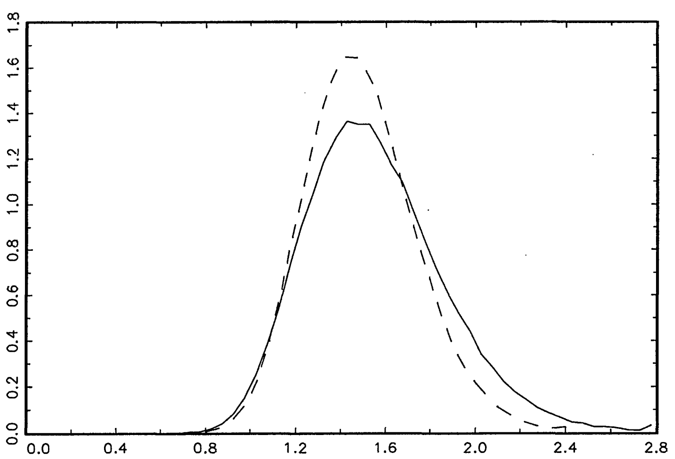

Figure 4. Posterior Density for γ. ——, skewed Student; — — —, skewed normal.

An interesting feature is that the skewed Student and the skewed normal lead to similar posterior distributions for γ (Fig. 4). Thus normal and Student tails lead to similar inference on skewness. Even more striking is the similarity of the posterior distributions for ν under Student and skewed Student sampling (Fig. 5). Whether we allow for skewness or not has virtually no impact on inference on the degrees of freedom parameter ν. In summary, inference on skewness and thickness of tails seems well separated in our model.

Figure 6 displays the postsample predictive density functions under each of the three models. Note that the predictive from the skewed Student model closely resembles the data histogram, which is presented in Figure 7. 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.

We now conduct a formal comparison of the three models using Bayes factors. We have used the method based on the basic marginal likelihood identity (BMI) developed by 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 ν and γ), we 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 favor of model i versus model j. For completeness, the simple normal model (for which the marginal likelihood is known analytically) is also included. Clearly, the data show some evidence for both fat tails and skewness.

As a check, we also assessed the evidence in favor of skewness using the Savage–Dickey density ratio mentioned by Verdinelli and Wasserman (1995), based on Dickey (1971). Comparing skewed Student with Student and skewed normal with normal led to the same Bayes factors as displayed in Table 2.

6.3 Comparison With Stable Model

Buckle (1995) used stable distributions as a way of capturing skewness and fat tails, and presents a Bayesian analysis of a location-scale model under independent sampling from stable distributions. In this section we briefly contrast

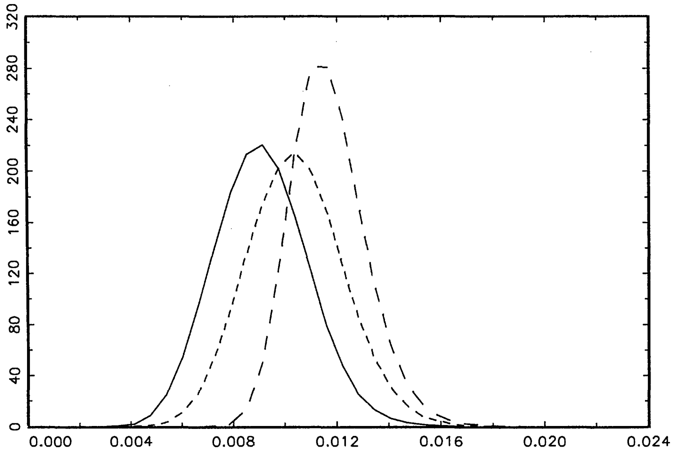

Figure 3. Posterior Density for σ. ——, skewed Student; - - -, Student; — — —, skewed normal.

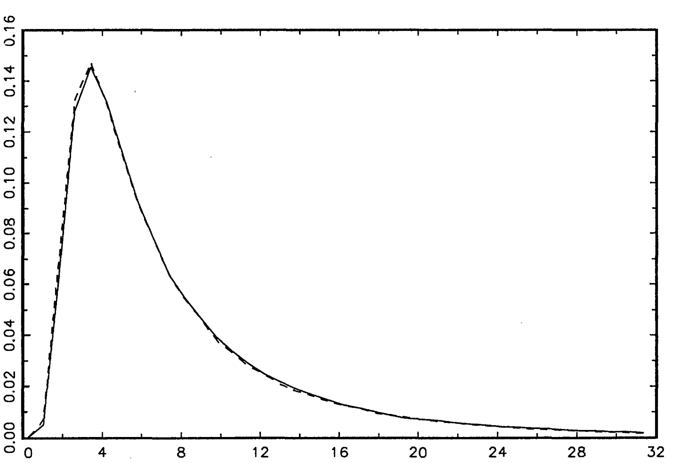

Figure 5. Posterior Density for ν.——, skewed Student; - - -, Student.

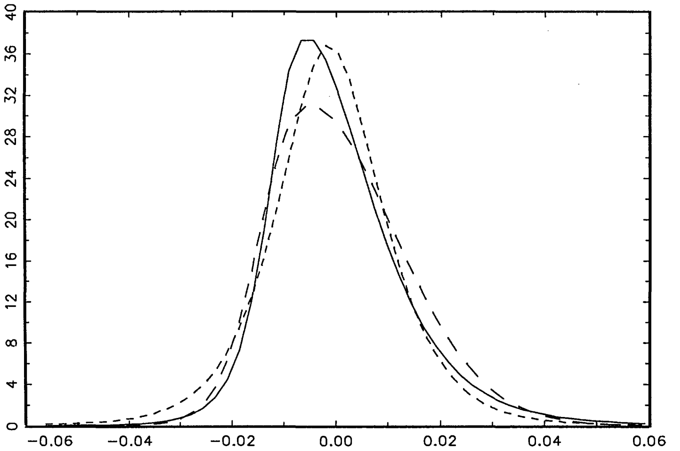

Figure 6. Predictive Densities. ——, *skewed Student;* - - -, *Student;* — — —, *skewed normal.*

this approach with the approach we propose herein. We feel that 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, because the sampling density in (13) behaves in the tails as a Student distribution with ν degrees of freedom, stable distributions display an inherent discontinuity in tail behavior: they either do not possess a finite variance or are normal. In addition, skewness is allowed for only when the variance does not exist.

A related point is that the skewness and tail parameters are inextricably linked for stable laws, thus complicating both the issue of prior elicitation and interpretation of the parameters. In sharp contrast, our approach completely separates the effect of the skewness parameter γ and the tail parameter ν, facilitating their interpretation and making prior independence between the two a plausible assumption.

Table 2. Bayes Factors for Share Price Returns

	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

In addition, the Gibbs sampler used by 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. Because the pdf of a stable distribution does not possess a closed-form expression, predictive distributions are also more difficult to evaluate than in our case.

Buckle (1995) modeled the share price dataset considered in Example 1 using stable distributions. Our results are rather similar to those found by Buckle (1995), who also recorded evidence of right skewness and heavy tails. Only his posterior findings on the location parameter may seem to be in conflict with ours, as he obtained a posterior mean for this parameter of .00053. Note, however, that the location parameter has different interpretations for stable and skewed Student distributions: In the stable case it is interpretable as the mean (if it exists), whereas in the skewed Student case it has the unequivocal interpretation of the mode. For comparison, we plotted the pdf of the stable distribution that Buckle (1995) presented (based on posterior modes of the parameters) in Figure 7 (dashed line), along with the predictive of the skewed Student model (drawn line) and a histogram of the data. Both distributions seem quite in line with the data.

As was suggested by a referee, it is interesting to calibrate our analysis with skewed Student distributions on a very extreme dataset. This is the object of our next example.

6.3.1 Example 2. We generate $n = 250$ observations from a stable distribution with characteristic exponent $\alpha = .5$, skewness parameter $-.5$, location 0, and scale parameter .1, using the same parameterization as used by Buckle (1995). This implies a negatively skewed distribution with extremely fat tails, the pdf of which is very spiked at the mode. Population moments only exist up to (and not including) 1/2, and the actually generated sample contains

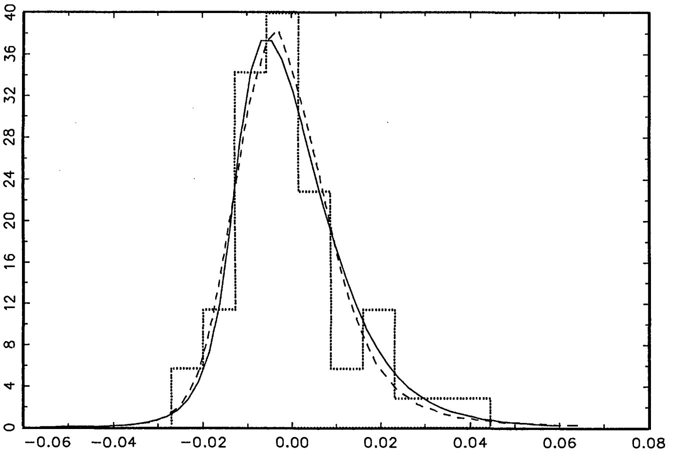

Figure 7. Data, Stable and Predictive. · · ·, *data histogram;* - - -, *stable pdf;* ——, *skewed Student predictive.*

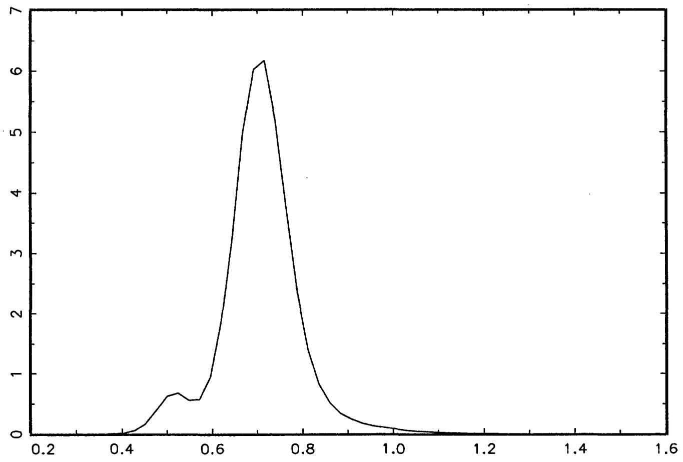

Figure 8. Posterior Density for γ. ——, *skewed Student.*

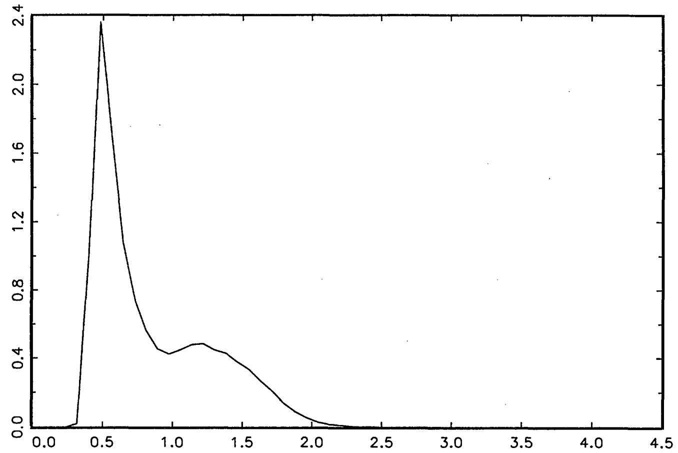

Figure 9. Posterior Density for ν. ——, skewed Student.

values ranging from $-3.53 \cdot 10^8$ to 975.78. Thus this distribution is a very severe test case for our skewed Student t sampling model. Figures 8 and 9 graphically display the posterior results on γ and ν. As expected, the negative skewness induces γ to have virtually all of the posterior mass below 1, and the degrees of freedom, ν, tend to be very small. Interestingly, the posterior mode for ν is very close to $\alpha = .5$, implying that the predictive from the skewed Student model should closely match the tail behavior of the underlying stable distribution. Figure 10 compares this predictive (drawn line) with the actual stable distribution that generated the data (dashed line). Both distributions are remarkably close, even in the central part, testifying to the flexibility of skewed Student modeling. For comparison with the shape of the skewed Student sampling density, Figure 10 also includes the pdf of a skewed Student distribution as in (13) with the parameters equal to the posterior modal values (dotted line). Clearly, it is virtually identical to the predictive.

We stress that the numerical implementation described in Section 5 leads to very efficient algorithms. Using Gauss-386i VM version 3.2, the skewed Student model for Example 1 with $n = 49$ executed at a rate of 40,000 Gibbs draws per hour on a PC equipped with a Pentium 100 processor, whereas for Example 2 (where $n = 250$), computing time is roughly four times as long. Convergence typically occurred after only a few hundred drawings. This compares favorably with the implementation based on the stable distribution of Buckle (1995). Even using a Sun SPARC workstation (with C), his Gibbs sampler seems slower by about a factor of four (based on personal communication). Thus, using skewed Student distributions, the analysis of much more challenging datasets is entirely within reach, even with modest computing facilities.

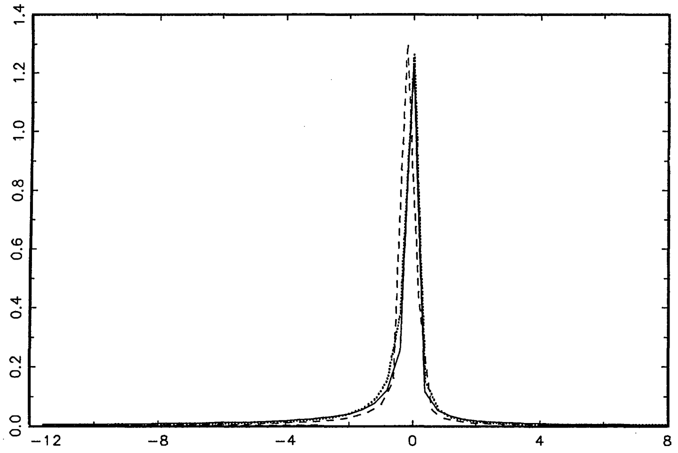

Figure 10. Skewed Student Versus Actual Stable Sampling Density. - - -, actual stable; ——, skewed Student predictive; · · ·, skewed Student pdf.

This article is devoted to the linear regression model with iid errors. We find skewed Student modeling quite appealing in this context. But if we wish to consider dependence between the errors, the approach that we advocate here is less natural. The family of stable distributions is preserved under convolution, which makes it a prime candidate for time-series modeling (see, e.g., Samorodnitsky and Taqqu 1994, chap. 7). Skewed Student distributions do not share this property, rendering the extension to dynamic models less immediate.

6.4 Example 3: Hertzsprung–Russell Diagram

Our third 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 where $k = 2, x_{i1} = 1$, and x_{i2} is the log of the temperature of star i. We have $n = 47$ 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 Section 6.1. We consider two sampling models, Student t and skewed Student, with the priors described in Section 6.1. The design matrix $\mathbf{X}$ of our dataset verifies $p_1 = p_2 = 4$, where p_j denotes the singularity index for column j (see Definition 1). This can easily be seen as follows: None of the values x_{i2} are 0, and the maximum number of identical values for x_{i2} is five. This immediately leads to $p_1 = p_2 = 4$. Thus from Theorem 3, positive-order posterior moments of β_1 and β_2 exist up to the order $n - k - 4 = 41$ (including), under both sampling assumptions. Theorem 4 implies that the range of finite posterior moments of σ of nonnegative order is given by [0, 45) under both models.

Posterior results are summarized in Table 3 and in Figures 11–14. Inference on the regression coefficients is affected somewhat by allowing for skewness, and the posterior mean of σ (see Table 3) is smaller under skewed Student sampling. There seems to be evidence of left skewness in the data (see Fig. 13), and, as was the case in our previous

Table 3. Posterior Moments for Hertzsprung-Russell Diagram

	Mean	*Std. dev.*	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*
	β_1	β_1	β_2	β_2	σ	σ
Skewed Student	7.53	(1.37)	−.495	(.275)	.428	(.132)
Student	6.71	(1.42)	−.391	(.327)	.552	(.065)

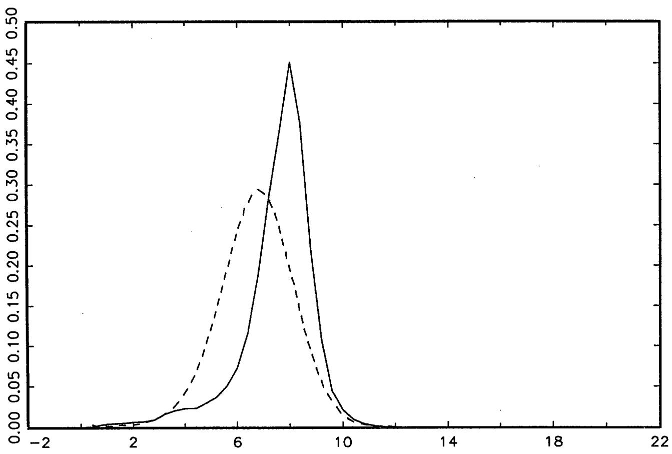

Figure 11. Posterior Density for β_1. ——, skewed Student; - - -, Student.

example, inference on tail behavior is largely unaffected by allowing for skewness (see Fig. 14).

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 favor of skewness.

7. CONCLUSION

In this article we have introduced a general method for transforming symmetric distributions into skewed distributions, at the cost of adding a single scalar parameter. Using such a skewed distribution for the error terms in a regression model, we established that this skewness has no impact on the existence of the posterior distribution and the posterior moments of primary interest. We then considered 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 scale parameter. For this model, which is central to the article, we obtained that the posterior is well defined under the same conditions as those for normal sampling (i.e., when sample size exceeds the number of regressors); we examined the existence of posterior moments of regression coefficients and scale in detail. We outlined a numerical analysis based on the Gibbs sampler and applied it to a number of examples.

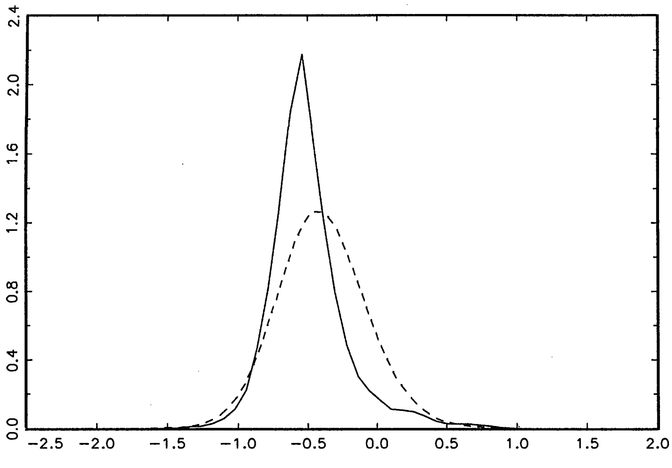

Figure 12. Posterior Density for β_2. ——, skewed Student; - - -, Student.

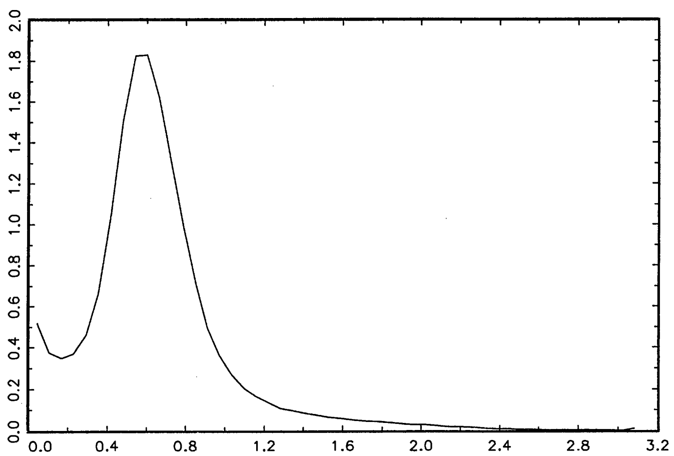

Figure 13. Posterior Density for γ. ——, skewed Student.

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

a. It allows for very flexible modeling of the skewness and fat tail features of the data. We can, for example, cover the entire range of the skewness measure of Arnold and Groeneveld (1995), which implies that mass can be allocated to the regions on both sides of the mode in any proportion. Furthermore, we can allow for any Student tail behavior, thus ranging from very fat tails to limiting normality. Calibration on a quite extreme dataset illustrates the flexibility of the skewed Student model.
b. The two extra parameters introduced into the analysis have very clearly defined modeling purposes. The skewness parameter alone controls the allocation of mass with respect to the mode, whereas the degrees of freedom parameter completely accounts for tail behavior. Thus the additional parameters are clearly interpretable. Prior independence between these two parameters is typically a plausible assumption, which moreover simplifies the process of choosing prior distributions. In our two empirical examples, the prior

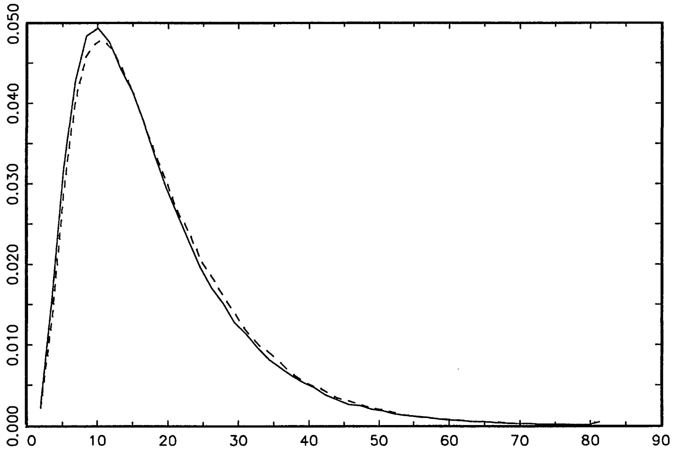

Figure 14. Posterior Density for ν. ——, skewed Student. - - -, Student.

independence between these parameters is not substantially altered in the posterior distribution.

c. The empirical analysis is very feasible indeed. The Gibbs sampler that we constructed 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.

Although we recommend skewed Student modeling for the analysis of independent observations, it is perhaps less suitable for the treatment of dependent data. In contrast to stable distributions, which are closed under addition, (skewed) Student distributions do not naturally extend to time-series modeling.

APPENDIX: PROOFS

Proof of Theorem 1

$E(\sigma^r \prod_{j=1}^k |\beta_j|^{r_j} | y_1, \ldots, y_n) < \infty$ if and only if the integral

$$I = \int_{\Re^k \times \Re_+ \times \mathcal{N} \times \Re_+} \left(\prod_{j=1}^{k} |\beta_j|^{r_j} \right) \sigma^{r-1} \times \left\{ \prod_{i=1}^{n} p(y_i | \boldsymbol{\beta}, \sigma, \nu, \gamma) \right\} dP_{\boldsymbol{\beta}}\, d\sigma\, dP_\nu\, dP_\gamma \quad \text{(A.1)}$$

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

$$\frac{2\sigma^{-1}}{\gamma + \frac{1}{\gamma}} f_\nu \left(\frac{|y_i - g_i(\boldsymbol{\beta})|}{\sigma h(\gamma)} \right),$$

$$\text{with} \quad h(\gamma) = \begin{cases} \max\left\{\gamma, \frac{1}{\gamma}\right\} & \text{for the upper bound} \\ \min\left\{\gamma, \frac{1}{\gamma}\right\} & \text{for the lower bound.} \end{cases} \quad \text{(A.2)}$$

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 σ. Transforming from σ to $\theta = \sigma h(\gamma)$ immediately leads to the upper and lower bounds for I:

$$2^n \int_{\Re_+} \frac{h(\gamma)^{n-r}}{\left(\gamma + \frac{1}{\gamma}\right)^n} dP_\gamma \int_{\Re^k \times \Re_+ \times \mathcal{N}} \left(\prod_{j=1}^{k} |\beta_j|^{r_j} \right) \theta^{r-1-n} \times \left\{ \prod_{i=1}^{n} f_\nu \left(\frac{|y_i - g_i(\boldsymbol{\beta})|}{\theta} \right) \right\} dP_{\boldsymbol{\beta}}\, d\theta\, dP_\nu, \quad \text{(A.3)}$$

with $h(\gamma)$ as defined in (A.2). Clearly, for $h(\gamma) = \min\{\gamma, 1/\gamma\}$, the first integral in (A.3) is strictly positive, whereas for $h(\gamma) = \max\{\gamma, 1/\gamma\}$, this integral is smaller than 1 (because $r \geq 0$ and $\max\{\gamma, 1/\gamma\} \geq 1$). In addition, the second integral in (A.3) is finite if and only if $E(\sigma^r \prod_{j=1}^k |\beta_j|^{r_j} | y_1, \ldots, y_n) < \infty$ under $\gamma = 1$, thus obtaining Theorem 1.

Remarks. 1. In earlier work (Fernández and Steel 1996) we examined Bayesian inference in the context of a linear regression model with iid errors following a symmetric Student t distribution with known degrees of freedom. Even though here we also conduct inference on ν and γ, many of the results in that work are useful for the following proofs, and thus we refer to it often in the remainder of the Appendix.

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 $z > 0$,

$$\Gamma(z) = (2\pi)^{1/2} z^{z-1/2} \exp(-z) \exp\{\phi(z)\}, \quad \text{(A.4)}$$

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

Proof of Theorem 2

Because, from Theorem 1, P_γ does not affect the existence of the posterior distribution, we take $\gamma = 1$. Using the representation in (16), the proof proceeds as follows:

(A) Consider the joint distribution of $(y_1, \ldots, y_n, \boldsymbol{\beta}, \sigma, \lambda_1, \ldots, \lambda_n)$.

(B) Integrate out $\boldsymbol{\beta}$ as a k-variate normal.

For any $(y_1, \ldots, y_n)' \in \Re^n$ barring a set of Lebesgue measure zero, we can perform the following:

(C) Integrate out σ using a gamma distribution on σ^{-2}, which requires that $n > k$.

(D) Finally, we are left with a function of $(\lambda_1, \ldots, \lambda_n)$, which can be shown to be bounded (See the proof of theorem 2 (ii) in Fernández and Steel 1996.) Thus it is integrable for any probability distribution of $(\lambda_1, \ldots, \lambda_n)$; in particular, it is integrable under

$$p(\lambda_1, \ldots, \lambda_n) = \int_0^\infty \left\{ \prod_{i=1}^{n} f_G \left(\lambda_i \middle| \frac{\nu}{2}, \frac{\nu}{2} \right) \right\} dP_\nu. \quad \text{(A.5)}$$

Proof of Theorem 3

Again, from Theorem 1, we take $\gamma = 1$. For any choice of P_ν we obtain the following:

(A) Following the reasoning in the proof of theorem 2 (i) of Fernández and Steel (1996), it is immediate that $r < n - k$ is required for the rth-order posterior moment of β_j to exist.

(B) Furthermore, from the proof of theorem 2 (ii) of Fernández and Steel (1996), 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 rth order posterior moment of β_j.

a. Finally, we show that if $P_\nu(0, c) > 0$ for all c and $p_j \geq 1$, then posterior moments of β_j of order $r > n - k - p_j$ do not exist. From theorem 3 (ii) of Fernández and Steel (1996), we know that if $r \geq n - k - p_j + \nu(n - k - p_j + 1)$, then $E(|\beta_j|^r | y_1, \ldots, y_n, \nu) = \infty$. Thus if $r > n - k - p_j$ and P_ν has mass arbitrarily close to 0, then $E(|\beta_j|^r | y_1, \ldots, y_n) = \infty$.

b. We now examine the case where P_ν has support on (ν_0, ∞), $p_j \geq 1$ and $n - k - p_j < r < n - k$. From the proof of theorem 2 (ii) of Fernández and Steel (1996), it follows that if $r < n - k$ and

$$\int_{(0,\infty)^n} \left(\frac{\lambda_1}{\lambda_2} \right)^{(n-k-p_j-r)/2} \times p(\lambda_1, \ldots, \lambda_n)\, d\lambda_1 \ldots d\lambda_n < \infty, \quad \text{(A.6)}$$

with $p(\lambda_1, \ldots, \lambda_n)$ as defined in (A.5), then $E(|\beta_j|^r | y_1, \ldots, y_n) < \infty$. Using Fubini's theorem, we compute the integral in (A.6) in two steps: First, we condition on ν, which leads to a finite integral if $r \leq n-k-p_j+\nu_0$. We then obtain a function of ν, which can be shown to be bounded by applying (A.4), whenever $r < n-k-p_j+\nu_0$; therefore, (A.6) holds for these values of r.

Proof of Theorem 4

We consider $\gamma = 1$.

(A) If $r \geq n-k$, we know from theorem 4 (i) of Fernández and Steel (1996) that $E(\sigma^r|y_1,\ldots,y_n,\nu) = \infty$ for all $\nu \in \Re_+$. Thus $E(\sigma^r|y_1,\ldots,y_n) = \infty$ for any P_ν.

(B) We now consider $0 < r < n-k$. From the proof of theorem 4 (ii) of Fernández and Steel (1996), and with $p(\lambda_1,\ldots,\lambda_n)$ as defined in (A.5), we can deduce that

$$\int_0^\infty \int_{0<\lambda_1\leq\cdots\leq\lambda_{n-k}<\infty} \times \left(\prod_{i=1}^{n-k} \lambda_i^{1/2}\right) \lambda_{n-k}^{-(n-k-r)/2} \left\{\prod_{i=1}^{n-k} f_{\mathrm{G}}\left(\lambda_i \left| \frac{\nu}{2}, \frac{\nu}{2}\right.\right)\right\} \times d\lambda_1 \ldots d\lambda_{n-k}\, dP_\nu < \infty \tag{A.7}$$

implies a finite rth-order posterior moment of σ. We now show that the inside integral in (A.7), denoted by $I(\nu)$, is a bounded function of ν and thus is integrable for any P_ν. Because $I(\nu)$ is continuous in ν, we need only prove that it has finite limits as ν converges to 0 and infinity.

(B1) Limit as $\nu \to \infty$. Because $\lambda_{n-k} = \max\{\lambda_1,\ldots,\lambda_{n-k}\}$, 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

$$I(\nu) \leq \int_0^\infty \lambda_{n-k}^{r/2} f_{\mathrm{G}}\left(\lambda_{n-k}\left|\frac{\nu}{2},\frac{\nu}{2}\right.\right) \times d\lambda_{n-k} \propto \nu^{-r/2}\Gamma\left(\frac{\nu+r}{2}\right)\left\{\Gamma\left(\frac{\nu}{2}\right)\right\}^{-1},$$

which, by applying (A.4), can be shown to have a finite limit as $\nu \to \infty$.

(B2) Limit as $\nu \to 0$. We now perform the integral $I(\nu)$ iteratively. In each of the $n-k$ steps of the integration, we use the upper bound

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

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

$$\frac{\nu^{-r/2}}{(\nu+1)^{n-k-1}}\Gamma\left(\frac{(n-k)\nu+r}{2}\right) \times \left\{\Gamma\left(\frac{\nu}{2}\right)\right\}^{-(n-k)}. \tag{A.8}$$

Applying (A.4) leads to an upper bound for (A.8) with a finite limit as $\nu \to 0$.

[Received April 1996. Revised January 1997.]

REFERENCES

Arnold, B. C., and Groeneveld, R. A. (1995), "Measuring Skewness With Respect to the Mode," *The American Stastistician*, 49, 34–38.

Badrinath, S. G., and Chatterjee, S. (1991), "A Data-Analytic Look at Skewness and Elongation in Common-Stock-Return Distributions," *Journal of Business and Economic Statistics*, 9, 223–233.

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, U.K.: Oxford University Press, pp. 35–60.

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.

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

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.

——— (1996), "On Bayesian Inference Under Sampling From Scale Mixtures of Normals," CentER Discussion Paper 9602, Tilburg University, The Netherlands.

——— (1997a), "Multivariate Student-t Regression Models: Pitfalls and Inference," CentER Discussion Paper 9708, Tilburg University, The Netherlands.

——— (1997b), "On the Dangers of Modelling Through Continuous Distributions: A Bayesian Perspective," CentER Discussion Paper 9705, 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.

——— (1992), "Priors for Macroeconomic Time Series and Their Application," Discussion Paper 64, Federal Reserve Bank of Minneapolis, Institute for Empirical Macroeconomics.

Lye, J. N., and Martin, V. L. (1993), "Robust Estimation, Nonnormalities, and Generalized Exponential Distributions," *Journal of the American Statistical Association*, 88, 261–267.

McDonald, J. B., and Nelson, R. D. (1993), "Beta Estimation in the Market Model: Skewness and Leptokurtosis," *Communications in Statistics, Part A—Theory and Methods*, 22, 2843–2862.

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.

Samorodnitsky, G., and Taqqu, M. S. (1994), *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance*, New York: Chapman and Hall.

van Zwet, W. R. (1964), "Convex Transformations of Random Variables," in *Mathematical Centre Tracts 7*, Amsterdam: Mathematisch Centrum.

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, U.K.: Cambridge University Press.