Econometric Theory, **16**, 2000, 80–101. Printed in the United States of America.

BAYESIAN REGRESSION ANALYSIS WITH SCALE MIXTURES OF NORMALS

CARMEN FERNÁNDEZ
University of Bristol

MARK F.J. STEEL
University of Edinburgh

This paper considers a Bayesian analysis of the linear regression model under independent sampling from general scale mixtures of normals. Using a common reference prior, we investigate the validity of Bayesian inference and the existence of posterior moments of the regression and scale parameters. We find that whereas existence of the posterior distribution does not depend on the choice of the design matrix or the mixing distribution, both of them can crucially intervene in the existence of posterior moments. We identify some useful characteristics that allow for an easy verification of the existence of a wide range of moments. In addition, we provide full characterizations under sampling from finite mixtures of normals, Pearson VII, or certain modulated normal distributions. For empirical applications, a numerical implementation based on the Gibbs sampler is recommended.

1. INTRODUCTION

The present paper focuses on Bayesian inference in the context of the linear regression model with independent errors distributed as scale mixtures of normals, to allow for flexible tails. More explicitly, we shall analyze the existence of the posterior distribution and of its moments under a commonly used improper prior and comment on numerical techniques for evaluating posterior quantities of interest. Whereas a growing number of Bayesian studies have used this model, the theoretical foundations have, so far, not been established. This paper aims to fill that gap.

It has long been recognized that the usually convenient assumption of normal sampling might be overly restrictive for many practical modeling situations. In particular, the thin tails of a normal distribution are often not a natural choice. An early contribution to this literature was Jeffreys (1961), whereas Maronna (1976) and Lange, Little, and Taylor (1989) discussed maximum likelihood estimation

We thank Jacek Osiewalski and three anonymous referees for helpful comments. Both authors were affiliated with CentER and the Department of Econometrics, Tilburg University, The Netherlands, during much of the work on this paper. Address correspondence to: Mark F.J. Steel, Department of Economics, University of Edinburgh, 50 George Square, Edinburgh EH8 9JY, United Kingdom; e-mail:Mark.Steel@ed.ac.uk.

  0266-4666/00 $9.50

for models with heavier-tailed Student-*t* disturbances. Bayesian results for outlier problems are provided in West (1984) for the wider class of scale mixtures of normals; however, he did not address the issue of existence of the posterior distribution and its moments under an improper prior distribution. Geweke (1993) considered the same Bayesian model as treated here for the case of Student-*t* sampling, but an unfortunate error in his main proof invalidates his results on posterior propriety and the existence of moments. The present analysis is thus required to validate the interesting numerical results obtained in Geweke (1993) on the basis of the Gibbs sampler and, more generally, to establish a basis for feasible Bayesian inference. In addition, we cover the entire class of scale mixtures of normals.

The class of scale mixtures of normals is generated by allocating to the disturbance of the *i*th observation, say, ε_i, the following distribution:

$$\varepsilon_i \stackrel{d}{=} z_i/\lambda_i^{1/2}, \tag{1.1}$$

where z_i is a normal$(0,1)$ random variable and λ_i an independent random variable on $(0,\infty)$. By assuming different probability distributions P_{λ_i} for λ_i, we map the entire class of scale mixtures of normals. Table 1 groups some known distributions of ε_i implied by (1.1) together with the corresponding distributions for λ_i. It is clear from this table that quite a rich class of continuous symmetric and unimodal distributions can be described by scale mixtures of normals, so that processes with thicker-than-normal tails will often be adequately modeled by choosing a distribution from this class. A formal characterization of the extent of this class is given in, e.g., Kelker (1970, Theorem 10) or Fang, Kotz, and Ng (1990, Theorem 2.21). Viewed in a multivariate spherical context, scale mixtures of normals are the only spherical distributions that can coherently be extended in dimension indefinitely. In other words, they can always be interpreted as the marginals of higher-dimensional spherical distributions.

We can cite a number of examples that testify to the growing impact of scale mixtures of normals in applied statistical practice. Modeling distributions of high-frequency financial data with the help of scale mixtures of normals is recently becoming more and more popular. In the context of stochastic volatility models, Harvey, Ruiz, and Shephard (1994) and Jacquier, Polson, and Rossi (1995) used a Student-*t*, and in Shephard (1994a, 1994b) we find an exponential power distribution and a finite mixture of normals. Bauwens and Lubrano (1998) considered GARCH models with Student-*t* disturbances. Lange et al. (1989) reported a number of examples from statistical practice where Student-*t* models provide a better fit to the data than their normal counterparts. For modeling macroeconomic time series, Geweke (1993) found relatively high posterior odds in favor of Student-*t* sampling as opposed to normal sampling.

We shall use a linear regression model under independent sampling from a scale mixture of normals with known mixing distribution P_{λ_i}. We complete the Bayesian model with a commonly used improper Jeffreys' prior on the param-

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

TABLE 1. Classes of scale mixtures of normals

Distribution of ε_i	Mixing distribution on λ_i	Reference
1. Finite mixture of normals	Discrete with finite support	
a. Normal	Dirac	
b. Contaminated normal	Most mass in one point	Johnson and Kotz (1970)
2. Generalized hyperbolic	Generalized inverse Gaussian	Barndorff-Nielsen et al. (1982)
a. Hyperbolic	$h \propto \lambda_i^{-2} \exp\left\{-\frac{1}{2}\left(\frac{\kappa}{\lambda_i} + \delta\lambda_i\right)\right\}, \delta \geq 0, \kappa > 0$	Barndorff-Nielsen et al. (1982)
(i) Laplace	$\delta = 0, \kappa = 1$	Andrews and Mallows (1974)
b. Pearson type VII	Gamma$(\nu/2, \mu/2) \quad \nu, \mu > 0$	Johnson and Kotz (1970)
(i) Student-t	$\nu = \mu$	
1. Cauchy	$\nu = \mu = 1$	
3. Symmetric z-distribution	$h = \lambda_i^{-2} \sum_{k=0}^{\infty} \binom{-2\delta}{k} \frac{\delta+k}{B(\delta,\delta)} \exp\left\{-\frac{(\delta+k)^2}{2\lambda_i}\right\}, \delta > 0$	Barndorff-Nielsen et al. (1982)
a. Generalized logistic	$\delta = 1, 2, \ldots$	Barndorff-Nielsen et al. (1982)
(i) Logistic	$\delta = 1$	Andrews and Mallows (1974)
b. Hyperbolic cosine	$\delta = \frac{1}{2}$	Barndorff-Nielsen et al. (1982)

4. Symmetric stable(α), $0 < \alpha < 2$	λ_i^{-1} is positive stable$\left(\frac{\alpha}{2}\right)$	Feller (1971)
a. Cauchy	$\alpha = 1$	
5. Exp. power(α), $1 \leq \alpha < 2$	$h \propto \lambda_i^{-1/2} \times$ p.d.f. of positive stable$\left(\frac{\alpha}{2}\right)$	West (1987)
a. Laplace	$\alpha = 1$	
6. Modulated normal type I	Pareto$(1, \nu/2)$ on $(1,\infty)$, $\nu > 0$	Romanowski (1979)
7. Modulated normal type II	Beta$(\nu/2,1)$ on $(0,1)$, $\nu > 0$	Rogers and Tukey (1972)
a. Slash	$\nu = 1$	Rogers and Tukey (1972)
b. Q-distribution	$\nu = 2$	Rogers and Tukey (1972)

*h indicates the p.d.f. of λ_i.

eters (under "independence"). The latter prior was shown by Fernández and Steel (1999a) to also have the interpretation of the "reference prior," based on formal information theory arguments (see Berger and Bernardo, 1992). Under independent sampling, Jeffreys' prior is a popular choice in the absence of compelling prior information. For time series models, the application of Jeffreys' principle is more contentious, as is evidenced by the discussion in Phillips (1991).

The explicit aim of this paper is the study of existence of the posterior distribution and the posterior moments of the parameters. Especially in view of the added complexity of sampling from scale mixtures of normals, numerical methods will typically be required, and usually the Gibbs sampler (proposed by Geweke, 1993, for Student-*t* sampling) provides an attractive approach, as illustrated in Section 5.2. This carries some inherent dangers, however, quite beyond numerical accuracy. The Gibbs sampler essentially approximates drawings from a joint distribution by a Markov chain of drawings from the full conditional distributions (see, e.g., Gelfand and Smith, 1990; Tierney, 1994). As, e.g., Casella and George (1992) have illustrated in an example, all the full conditionals may well be proper distributions, without existence of the joint distribution. Hobert and Casella (1996) pointed out the pitfalls of careless use of Markov chain Monte Carlo methods in cases where no posterior distribution exists (see also Fernández, Osiewalski, and Steel, 1997). Thus, under an improper prior distribution, it becomes crucial to verify propriety of the posterior to validate Bayesian inference. This argument also carries over to the existence of posterior moments of the parameters: the mere fact that the full conditional posterior distribution of a parameter allows for a finite moment of a certain order does not guarantee existence of this moment in the marginal posterior distribution. The problem of higher-order moments can even be more severe as it does not disappear by using a proper prior distribution. Our explicit focus on the existence of the posterior distribution and its moments is, thus, meant to indicate whether Bayesian inference is at all possible and, if so, which moments we can meaningfully try to calculate. We do not deal here with the issue of how precise this inference will be in particular empirical contexts.

This paper will be concerned with n independent and identically distributed (i.i.d.) univariate disturbances $\varepsilon_i, i = 1, \ldots, n$, as in (1.1), in contrast to the literature on multivariate scale mixtures of normals, where we only obtain one n-dimensional vector observation (see, e.g., Osiewalski,1991; and, for the special case of multivariate Student-*t*, Zellner, 1976). In the latter case, $(\varepsilon_1, \ldots, \varepsilon_n)'$ is distributed as a standard n-variate normal $(z_1, \ldots, z_n)'$ divided by a single scalar, say, λ_1 with some distribution P_{λ_1}. As this multivariate scale mixture of normals is in the class of n-variate spherical distributions, we know from Kelker (1970, Lemma 5) that the only intersection between our i.i.d. sampling case and this multivariate case is that of normality. In the course of the paper, we shall briefly compare both sampling schemes with respect to the existence of posterior moments.

The next section of the paper introduces the Bayesian model and treats propriety of the posterior. Conditions for the existence of moments of the regression

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

coefficients are analyzed in Section 3, whereas Section 4 focuses on the moments of the scale parameter. The following section deals with some practical approaches to conducting Bayesian inference in the context of scale mixtures of normals. Section 6 groups some concluding remarks. Throughout the paper, the notation for distributions and probability density functions follows DeGroot (1970). All proofs are referred to the Appendix, without explicit mention in the main text.

2. THE BAYESIAN MODEL

In this section we shall examine the linear regression model corresponding to (1.1). In particular, we assume the observations $y_i \in \Re$ $(i = 1,\ldots,n)$ to be generated from

$$y_i = x_i'\beta + \sigma\varepsilon_i, \tag{2.1}$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. random variables distributed as a scale mixture of normals. Thus, $\varepsilon_i \overset{d}{=} z_i/\lambda_i^{1/2}$ where z_i is a standard normal random variable and λ_i an independent random variable with some known probability distribution P_{λ_i} on $(0,\infty)$. The k-dimensional vector x_i groups the explanatory variables; we interpret (2.1) as modeling y_i given x_i, but we shall not make explicit the fact that we condition on x_i in the sequel. The parameters introduced in (2.1) are the regression coefficients $\beta \in \Re^k$ and the scale $\sigma > 0$.

The sampling model is thus characterized by the density function

$$p(y_i|\beta,\sigma) = \int_0^\infty \frac{\lambda_i^{1/2}}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{\lambda_i}{2\sigma^2}(y_i - x_i'\beta)^2\right\} dP_{\lambda_i}. \tag{2.2}$$

Independent replications from (2.2) will constitute the sampling information regarding the common regression and scale parameters. Let us now group the explanatory variables into an $n \times k$ matrix $X = (x_1,\ldots,x_n)'$ that is assumed to be of full column rank (and thus $n \geq k$). In addition, we define $y = (y_1,\ldots,y_n)'$ as the vector of observations.

A common choice for a noninformative prior distribution is the "independence" Jeffreys' prior given by

$$p(\beta,\sigma) \propto \sigma^{-1}. \tag{2.3}$$

The prior in (2.3) is also the reference prior in the sense of Berger and Bernardo (1992) for any regular distribution on ε_i as shown in Fernández and Steel (1999a).

Because the prior distribution in (2.3) is not proper, the existence of the posterior distribution (defined as the conditional distribution of the parameters given the observables) is not guaranteed. The results in Mouchart (1976) and Florens, Mouchart, and Rolin (1990) imply that such a conditional distribution exists only when the predictive distribution is σ-finite, i.e., $p(y) \equiv \int p(y|\beta,\sigma)p(\beta,\sigma)\,d\beta\,d\sigma < \infty$ except possibly on a set of y's of Lebesgue measure zero in $\Re^n$. In the context of

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

the model (2.2) and (2.3), we can obtain the following result concerning the feasibility of Bayesian inference.

THEOREM 1. (Propriety of posterior) *Under the prior in* (2.3) *and with n independent observations from* (2.2), *the conditional distribution of* (β,σ) *given y exists if and only if* $n \geq k+1$, *for any choice of the mixing distribution* P_{λ_i}.

Note that the condition $n \geq k+1$ is both necessary and sufficient and does not involve any properties of the mixing distribution. Surprisingly, the wide range of tails accommodated within the class of scale mixtures of normals has no influence whatsoever on the existence of the posterior.

Before proceeding with the remainder of the paper, a remark is in order. Note that Theorem 1 is concerned with establishing the existence of the conditional distribution of the parameters given the observables, which puts us on equal footing with the case where a proper prior is used. This, however, does not rule out the possibility that $p(y)$, the denominator in the usual Bayes' formula, becomes infinite in a set of y's that has Lebesgue measure zero in $\Re^n$. Whereas any such sample has, by definition, zero probability of being observed under our assumed sampling model, the rounding implicit in any data set means that, in practice, there could be a positive probability of observing an "offending" value of y. The same type of comment applies to the existence of posterior moments, examined in the following two sections. We stress, however, that these problems are inherent to any statistical analysis using continuous sampling distributions and are by no means restricted to the use of improper priors or Bayesian methods. A detailed discussion of these issues together with a general solution within a Bayesian framework can be found in Fernández and Steel (1999b).

In the sequel, we assume $n \geq k+1$ so that Theorem 1 applies and turn to the question of existence of moments. To facilitate the discussion in the remainder of the paper, we shall introduce the following definitions of characteristics of the design matrix X and the mixing distribution P_{λ_i}.

DEFINITION 1. (Singularity index for column j) *Given an* $n \times k$ *full column-rank matrix X, we define the singularity index for column* $j = 1,\ldots,k$ *as the largest number* p_j *such that there exists a* $(k-1+p_j) \times k$ *submatrix of X of rank* $k-1$ *that retains rank* $k-1$ *after removing its jth column.*

From the definition, $k-1+p_j$ gives the largest number of observations in the sample for which β_j, the jth component of β, is not identified. Clearly, $0 \leq p_j \leq n-k$ because X is of full column rank. A simple way of computing p_j is as follows: consider all sets of $k-1$ rows of X such that the rank of the corresponding submatrix without column j is $k-1$. Then p_j is the maximum number of rows that can be added to any such set without increasing the rank. If X contains rows of zeros, then p_j is at least equal to the number of such zero 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 X is nonsingular.

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

DEFINITION 2. (Moment set and moment index) *Let P_{λ_i} be the probability distribution of a random variable λ_i in $\Re_+$. We define*:

(i) *Moment set of* P_{λ_i}: $\mathcal{M} = \{s \in \Re : E(\lambda_i^{s/2}) \equiv \int_0^\infty \lambda_i^{s/2}\, dP_{\lambda_i} < \infty\}$.
(ii) *Moment index of* P_{λ_i}: $m = \sup\{s \geq 0 : s \in \mathcal{M} \text{ and } -s \in \mathcal{M}\}$.

Clearly, $0 \in \mathcal{M}$ because P_{λ_i} is a probability distribution, and thus $m \geq 0$ is always defined.

3. POSTERIOR MOMENTS OF REGRESSION COEFFICIENTS

In this section, we denote by $(r_1, \ldots, r_k)$ the order of the moment of $\beta = (\beta_1, \ldots, \beta_k)'$ and define $r = \sum_{j=1}^k r_j$. Our most general result is stated in the following theorem.

THEOREM 2. (Posterior moments of β) *Consider the Bayesian model in* (2.2) *and* (2.3) *and any choice of $r_j \geq 0$ for $j = 1, .., k$ such that $r > 0$. We obtain that*

(i) *necessity*: *if $r \geq n - k$, then $E(\Pi_{j=1}^k |\beta_j|^{r_j} | y) = \infty$;*
(ii) *sufficiency*: *if $r < \min\{n - k, n - k - p(r_1, \ldots, r_k) + m\}$, where m is the moment index of the mixing distribution P_{λ_i} and $p(r_1, \ldots, r_k) = \max\{p_j : r_j > 0\}$ with p_j the singularity index for column j of the design matrix X, then $E(\Pi_{j=1}^k |\beta_j|^{r_j} | y) < \infty$.*

Theorem 2 only addresses the situation of nonnegative moments. Using the fact that the first negative moment of a normally distributed random variable does not exist, it is straightforward to prove that the moment in Theorem 2 is always infinite if any $r_j \leq -1$. Theorem 2(i) tells us that there is never any hope for the existence of moments for which $r \geq n - k$, regardless of the characteristics of the design matrix or the mixing distribution. Such lack of existence of moments is, therefore, due to the uncertainty about β and σ rather than to the scale mixing. On the other hand, both X and P_{λ_i} intervene (through $p(r_1, \ldots, r_k)$ and m, respectively) in the sufficient condition for existence of moments with $r < n - k$.

Theorem 2 fully characterizes the existence of positive posterior moments of β whenever X and P_{λ_i} fulfill the following property.

COROLLARY 1. *If, in the context of Theorem* 2, *the design matrix X and the mixing distribution P_{λ_i} are such that* $\max\{p_j : j = 1, \ldots, k\} \leq m$, *then*

$$E\left(\prod_{j=1}^k |\beta_j|^{r_j} | y\right) < \infty \quad \text{if and only if } r < n - k.$$

Thus, under the condition of Corollary 1 the same posterior moments exist as under normal sampling. We mention two important special cases where Corollary 1 applies:

(i) Every $k \times k$ submatrix of X is nonsingular.

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

In this case $\max\{p_j : j = 1, \ldots, k\} = 0 \leq m$, and all posterior moments of β with $r < n - k$ exist for any P_{λ_i}. Therefore, the mixing distribution is entirely irrelevant for the issue of existence of posterior moments. As examples of this situation, we can mention the location-scale model (corresponding to $k = 1$ and $x_i = 1$, $i = 1, \ldots, n$) and models with the x_i's independently drawn from continuous k-variate distributions. We wish to remind the reader that the finite posterior moments of β can (and typically will) take different values for different mixing distributions. The order up to which moments of β are finite, however, is robust with respect to the choice of P_{λ_i}, i.e., in the entire class of scale mixtures of normals.

(ii) The moment index of P_{λ_i} verifies $m \geq n - k$.

Again, Corollary 1 applies, regardless of the form of the matrix X (of rank k). The design matrix, however, will typically influence the actual values of such moments. Sampling from finite mixtures of normals leads to $m = \infty$, thus providing an example of this situation.

In many situations not covered by Corollary 1, Theorem 2 can still provide an answer. If, given a particular order $(r_1, \ldots, r_k)$, the inequality $p(r_1, \ldots, r_k) \leq m$ is verified, then Theorem 2 shows that such a posterior moment of β exists if and only if $r < n - k$. However, when $p(r_1, \ldots, r_k) > m$, the necessary condition ($r < n - k$) and the sufficient condition ($r < n - k - p(r_1, \ldots, r_k) + m$) do not coincide, and Theorem 2 remains inconclusive if $r \in [n - k - p(r_1, \ldots, r_k) + m, n - k)$. By further specifying P_{λ_i}, we can refine Theorem 2, as evidenced by the following theorem concerning marginal posterior moments of the components of β.

THEOREM 3. (Finite mixtures of normals, Pearson VII, and modulated normal sampling) *For the Bayesian model in* (2.2) *and* (2.3), *we obtain for any value of* $r > 0$

$E(|\beta_j|^r | y) < \infty$ *if and only if*

(i) $r < n - k$ *for a discrete mixing distribution with finite support* (*sampling from finite mixtures of normals*) *or a Pareto* $(1, \nu/2)$ *mixing distribution with* $\nu \geq 1$ (*modulated normal type I sampling*),

(ii) $r < \min\{n - k, n - k - p_j + \nu(n - k - p_j + 1)\}$ *for a gamma* $(\nu/2, \mu/2)$ *mixing distribution* (*Pearson VII sampling*) *or a beta* $(\nu/2, 1)$ *mixing distribution* (*modulated normal type II sampling*).

Whereas the characterization for finite mixtures of normals is a direct consequence of Theorem 2 (because $m = \infty$), for the cases of Pearson VII and modulated normal type I and II sampling we have improved upon Theorem 2, which leads to $r < n - k$ and $r < \min\{n - k, n - k - p_j + \nu\}$ as necessary and sufficient conditions, respectively; taking the added information on the mixing distribution into account allows us to characterize the range of positive orders for which posterior moments of β_j exist. Thus, the sufficient condition in Theorem 2 is too conservative if and only if $\nu < p_j$ under modulated normal type I sampling with $\nu \geq 1$ and if and only if $\nu < p_j < n - k$ for Pearson VII and modulated normal type

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

II models. In the former case, Theorem 3(i) states the existence of $p_j - \nu$ additional moments over those guaranteed by Theorem 2, whereas, by Theorem 3(ii), the gain in the latter case is $p_j - \nu$ moments if $\nu \geq p_j/(n-k-p_j+1)$ and $\nu(n-k-p_j)$ moments otherwise. Theorem 3 also illustrates the fact that, in general, both the design matrix and the mixing distribution intervene in the issue of existence of moments, because neither can be neglected in the full characterization provided in (ii). In addition, Theorem 3(ii) shows that different components of β can possess marginal posterior moments up to different orders.

Observe that under Pearson VII sampling the parameter μ of the gamma mixing distribution does not intervene in the issue of existence of marginal posterior moments. Obviously, when $\mu = \nu$ it specializes to the important case of Student-*t* sampling. Furthermore, as ν tends to infinity for this Student sampling and for both types of modulated normal distributions, the sampling distribution converges to normality (which is also a special case of finite mixtures of normals with P_{λ_i} a Dirac distribution).

Finally, we can compare the i.i.d. sampling case treated in this paper with sampling one single vector *y* from an *n*-dimensional distribution. If we specialize the results obtained by Osiewalski and Steel (1992) to the class of scale mixtures of multivariate normals, we note that the posterior distribution of β, and, therefore, its existence of moments, is entirely unaffected by departures from normality within this class (see also Zellner, 1976, for the special case of the multivariate Student-*t* model and for analogous results in a maximum likelihood framework). Thus, in this multivariate context, posterior moments of β always exist as long as $r < n-k$, irrespective of the (full column-rank) matrix X or the mixing distribution. We conclude that the present case of independent sampling, generally, requires stronger conditions for the existence of posterior moments of β as shown by Theorem 3(ii).

4. POSTERIOR MOMENTS OF SCALE PARAMETER

In this section, we shall focus on the existence of moments of the scale σ of any order $r \in \Re$.

THEOREM 4. (Posterior moments of scale) *The Bayesian model in* (2.2) *and* (2.3) *leads to*

(i) *necessity*: *if* $r \geq n-k$, *then* $E(\sigma^r|y) = \infty$;
(ii) *sufficiency*: *if* $r \in (-\infty, n-k) \cap \mathcal{M}$, *where* $\mathcal{M}$ *is the moment set of* P_{λ_i}, *then* $E(\sigma^r|y) < \infty$.

As was the case with the regression coefficients, posterior moments of σ of order $r \geq n-k$ never exist, whatever the choice of the design matrix X or the mixing distribution P_{λ_i}. For values of $r < n-k$, Theorem 4(ii) provides a sufficient condition for existence of the rth moment that relies on the existence of moments of P_{λ_i}, namely, that $r \in \mathcal{M}$. Thus, for $r = 0$ we can deduce Theorem 1

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

for propriety of the posterior distribution, using that $0 \in \mathcal{M}$, as remarked after Definition 2.

The necessary and sufficient conditions presented in Theorem 4 do not coincide in general. The following theorem provides a full characterization of the existence of posterior moments of σ for some distributions of practical interest.

THEOREM 5. (Finite mixtures of normals, Pearson VII, and modulated normal sampling) *For the sampling model in* (2.2) *and the prior in* (2.3) *we obtain that*

$E(\sigma^r|y) < \infty$ *if and only if*

(i) $r < n - k$ *for a discrete mixing distribution with finite support* (*sampling from finite mixtures of normals*) *or a Pareto*$(1, \nu/2)$ *mixing distribution with* $\nu \geq 1$ (*modulated normal type I sampling*),

(ii) $-(n-k)\nu < r < n - k$ *for a gamma*$(\nu/2, \mu/2)$ *mixing distribution* (*Pearson VII sampling*) *or a beta*$(\nu/2, 1)$ *mixing distribution* (*modulated normal type II sampling*).

Clearly, sampling from finite mixtures of normals leads to the moment set $\mathcal{M} = \Re$, from which we can immediately conclude, through Theorem 4, that in this case the rth posterior moment of σ is finite if and only if $r < n - k$. On the other hand, the characterizations for the cases of Pearson VII and both types of modulated normal sampling do not follow from Theorem 4 but are obtained through exploiting the properties of the corresponding mixing distributions.

Under modulated normal type I sampling with $\nu \geq 1$ all moments of order smaller than $n - k$ exist from Theorem 5. The theorem also shows that when sampling within the Pearson VII or the modulated normal type II classes, moments of order $r \in [0, n - k)$ are always assured whereas existence of negative order moments is entirely determined by the parameter ν of the mixing distribution. Under Pearson VII sampling, the value of the parameter μ in the gamma mixing distribution does not intervene as was the case in Theorem 3 for marginal posterior moments of β. Choosing $\mu = \nu$ we obtain the important special case of Student-t sampling, and as ν then tends to infinity, we converge to the normal case where the rth moment of σ is finite if and only if $r < n - k$.

The case of sampling one single vector observation from a scale mixture of multivariate normals was shown in Osiewalski and Steel (1996) to lead to a necessary and sufficient condition for existence of moments of σ that corresponds to (ii) in Theorem 4, where n now represents the dimension of the vector observation instead of sample size. Thus, the latter case, which only intersects with the model analyzed here under normality, generally requires a more stringent condition. This is clearly shown by Theorem 5, where, e.g., under Pearson VII or modulated normal type II sampling the condition of Theorem 4(ii) is not satisfied for $r \in (-(n-k)\nu, -\nu]$, but the rth posterior moment of σ is, nevertheless, finite.

As a final remark, we note that the results in Theorems 4 and 5 can alternatively be used to assess the propriety of the posterior distribution under the more general

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

prior

$$p(\beta,\sigma) \propto \sigma^{r-1}. \tag{4.1}$$

Although the independence Jeffreys' prior in (2.3) is widely used in a noninformative context, the more general prior in (4.1) could be of interest in some cases. Existence of the gth-order posterior moment of σ under the prior in (4.1) can also be examined by means of Theorems 4 and 5, replacing r by $r + g$. On the other hand, existence of posterior moments of the regression coefficients is much more difficult to establish. Note that this problem is equivalent to analyzing the existence of cross moments for β and σ with our present prior in (2.3), for which we have been unable to find simple and useful results. This, however, does not preclude posterior inference on β under the prior in (4.1). Once the propriety of the posterior distribution has been established, we could simply report quantities that are always known to exist, such as quantiles or highest posterior density regions, instead of posterior moments of β.

5. NUMERICAL ASPECTS

Once we have made sure that a Bayesian analysis can meaningfully be conducted (Section 2) and the moments we are interested in actually exist (Sections 3 and 4), we will generally need numerical tools to conduct the necessary analysis. This section gives a generic description of two distinct numerical strategies that could be employed. Both start from the simple observation that given λ_i the sampling model in (2.2) is merely the normal linear regression model. Thus, the posterior analysis of β and σ, using the reference prior in (2.3) and conditioning on $\lambda = (\lambda_1,\ldots,\lambda_n)'$, is entirely standard and described by the following normal-gamma density function on (β,σ^{-2}):

$$p(\beta,\sigma^{-2}|\lambda,y) = f_N^k(\beta|b(\lambda),\sigma^2(X'\Lambda X)^{-1}) f_G\left(\sigma^{-2}\middle|\frac{n-k}{2},\frac{s(\lambda)}{2}\right), \tag{5.1}$$

where $\Lambda = \mathrm{Diag}(\lambda_i)$, $b(\lambda) = (X'\Lambda X)^{-1}X'\Lambda y$, and $s(\lambda) = y'\{\Lambda - \Lambda X(X'\Lambda X)^{-1}X'\Lambda\}y$.

The treatment of the λ_i's will constitute the nonstandard part of the analysis of our Bayesian model. We distinguish the following two approaches.

5.1. Independent Monte Carlo

Here we generate independent drawings from the distribution of (β,σ,λ) given y by drawing consecutively from (5.1) and from the distribution of λ given y, which is proportional to

$$g(\lambda)\prod_{i=1}^{n} P_{\lambda_i}, \tag{5.2}$$

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

where we have defined $g(\lambda) = \{\mathrm{Det}(X'\Lambda X)\}^{-1/2} s(\lambda)^{-(n-k)/2} \prod_{i=1}^{n} \lambda_i^{1/2}$. The implicit assumption underlying the notation in (5.2) is that the probability distribution of λ given y is absolutely continuous with respect to $\prod_{i=1}^{n} P_{\lambda_i}$ with Radon–Nikodym derivative proportional to $g(\lambda)$. Note that $g(\lambda)$ in (5.2) is simply the result of integrating the data density given (β, σ, λ) with the prior of (β, σ) and can be found immediately from (A.2) in the Appendix with $r_j = 0, j = 1, \ldots, k$. Thus, $g(\lambda)$ corresponds to the integrand in (A.6) with $l = 0$ or, equivalently, to that in (A.16) with $r = 0$.

As a result of integrating out β and σ, the components $\lambda_1, \ldots, \lambda_n$ of λ do not preserve independence conditionally upon y, which seriously complicates drawing from (5.2). From the proof of Theorem 2 (see (A.8) in the Appendix), we know that $g(\lambda)$ is a bounded function of λ. Thus, in general, we can use rejection sampling (see, e.g., Devroye, 1986) to draw from (5.2), generating drawings from $\prod_{i=1}^{n} P_{\lambda_i}$ and accepting with a probability proportional to $g(\lambda)$. Especially for large sample size, n, this can, however, prove to be very inefficient. An alternative procedure for generating drawings from (5.2) is importance sampling, as described in, e.g., Geweke (1989). We then need to choose a convenient probability distribution (importance function) on $\Re_+^n$ from which to draw λ that, ideally, closely resembles (5.2) and dominates it in the tails. Again, numerical problems could occur for moderate or high values of n.

In the special case where P_{λ_i} is a discrete distribution with support on, say, q points (sampling from a finite mixture of normals), we can use (5.2) to evaluate the probability mass attached to each of the q^n possible values for $\lambda = (\lambda_1, \ldots, \lambda_n)$ given y. If q^n is not prohibitively large, we can immediately evaluate quantities of interest from (5.1), without recourse to numerical methods. Clearly, if $q = 1$, we have the standard normal regression model.

Generally, drawing from the n-variate distribution in (5.2) will be cumbersome, and, therefore, the following alternative strategy is outlined.

5.2. Gibbs Sampling

This Markov chain Monte Carlo method is based on the full conditional distributions (see, e.g., Gelfand and Smith, 1990; Tierney, 1994). For (β, σ) given λ the posterior distribution is described by (5.1). To complete the Gibbs sampler we need the distribution of λ given (β, σ, y), which is proportional to

$$\prod_{i=1}^{n} g_i(\lambda_i) P_{\lambda_i}, \quad \text{where } g_i(\lambda_i) = f_G\left(\lambda_i \middle| \frac{3}{2}, \frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right). \tag{5.3}$$

Each pass through the sampler thus requires only two steps: one drawing from (5.1) and one from the probability distribution proportional to (5.3). 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).

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

As opposed to the situation in Section 5.1, the λ_i's are independent given y and (β,σ), which greatly facilitates drawing the vector $(\lambda_1,\ldots,\lambda_n)$. A general rejection sampling strategy can be used where each λ_i is drawn from P_{λ_i} and accepted with probability $\{\lambda_i\sigma^{-2}(y_i - x_i'\beta)^2\}^{1/2}\exp[\frac{1}{2}\{1 - \lambda_i\sigma^{-2}(y_i - x_i'\beta)^2\}]$, which corresponds to $g_i(\lambda_i)$ divided by its maximum value. If required, more carefully tailored rejection samplers can, of course, be designed. Alternatively, we could use, e.g., the Metropolis–Hastings algorithm (see, e.g., Tierney, 1994) to draw from (5.3) within the Gibbs sampler. Most importantly, the overall performance of the rejection or Metropolis step is not adversely affected by the necessity to draw in n dimensions: we just require n one-dimensional sampling schemes. In most practical situations, this will more than offset the inherent efficiency loss (with respect to independent Monte Carlo) due to the serial correlation between Gibbs drawings.

The Gibbs sampler simplifies considerably in a number of special cases.

If P_{λ_i} is a gamma$(\nu/2,\mu/2)$ distribution, giving rise to a Pearson type VII sampling distribution in (2.2), we retain a gamma distribution for the full conditional of each λ_i, i.e., each of the n factors in (5.3) is described by the density function

$$p(\lambda_i|\beta,\sigma,y_i) = f_G\left(\lambda_i\middle|\frac{\nu+1}{2},\frac{\mu+\sigma^{-2}(y_i - x_i'\beta)^2}{2}\right), \tag{5.4}$$

which we can draw from easily. For the Student-t case, Geweke (1993) uses a similar Gibbs sampler, and Lange et al. (1989) also mention the conditional distribution in (5.4).

In the modulated normal type II class, introduced by Rogers and Tukey (1972), where the mixing distribution P_{λ_i} is a beta$(\nu/2,1)$ distribution, we obtain

$$p(\lambda_i|\beta,\sigma,y_i) \propto f_G\left(\lambda_i\middle|\frac{\nu+1}{2},\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right)I_{[0,1]}(\lambda_i), \tag{5.5}$$

i.e., a truncated gamma distribution.

Sampling from a generalized hyperbolic distribution corresponds to a generalized inverse Gaussian mixing distribution (see Barndorff-Nielsen, Kent, and Sørensen, 1982). Then, the factors in (5.3) will still be generalized inverse Gaussian distributions with density function

$$p(\lambda_i|\beta,\sigma,y_i) \propto \lambda_i^{-(\gamma+(1/2))}\exp -\frac{1}{2}\left[\frac{\kappa}{\lambda_i} + \lambda_i\left\{\delta + \frac{(y_i - x_i'\beta)^2}{\sigma^2}\right\}\right], \tag{5.6}$$

where $\gamma \in \Re$ and δ and κ take strictly positive values. In addition, for negative γ, κ can be 0, and for positive values of γ the same holds for δ. As can be verified from Table 1, choosing $\gamma = 1$ corresponds to sampling from a hyperbolic distribution, whereas the sampling distribution becomes Laplace if we also take $\delta = 0$ and $\kappa = 1$. Drawing from (5.6) can be implemented in quite an efficient manner,

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

as explained in Devroye (1986, pp. 479–480). In addition, the Pearson type VII family, discussed earlier, is also a subclass of the class of generalized hyperbolic distributions. In particular, when we take $\gamma = -\nu/2$, $\delta = \mu$, and $\kappa = 0$ we obtain the gamma$(\nu/2, \mu/2)$ mixing distribution, and (5.6) reduces to (5.4).

Finally, for finite mixtures of normals with q possible values for each λ_i, the Gibbs sampler provides an alternative in those cases where direct evaluation (using (5.1) and (5.2)) proves very difficult as a result of a large value for q^n. In contrast to the situation using independent Monte Carlo (Section 5.1), it will now typically be feasible to draw values for λ even when q^n is too large for a direct analysis. All we need is to draw from the n independent discrete distributions in (5.3), which is often straightforward even for relatively large q and n.

6. CONCLUDING REMARKS

In this paper, we have treated the linear regression model under independent sampling from scale mixtures of normals. From Table 1, which groups some members of this class of sampling distributions, it is clear that this covers a rather wide variety of behavior. Completing this sampling model with a common noninformative ("reference") prior, we have investigated conditions for the validity of Bayesian inference and the existence of the posterior moments of the regression coefficients and the scale parameter.

There are three characteristics that can influence this existence of moments:

(1) the quantity $n - k$, i.e., the sample size minus the number of regressors in the model,
(2) the structure of the design matrix X, always of full column rank,
(3) the mixing distribution P_{λ_i}.

Throughout, existence of moments will be influenced by (1), whereas (2) and (3) do not always intervene. Our main theoretical results are presented in Theorems 1–5.

To implement a Bayesian analysis of the models treated here, and to actually evaluate the moments that can be shown to be finite, we typically require numerical methods. We mention two distinct strategies in Section 5 and conclude that, especially for moderate or large sample size n, Gibbs sampling seems preferable to independent Monte Carlo.

The assumption of i.i.d. error terms was made here as it corresponds to many empirical modeling situations, but it is by no means crucial for our results; most of the techniques used in our proofs can be used and the analysis can be extended straightforwardly to the case where P_{λ_i} varies across the observations $i = 1, \ldots, n$. In addition, we could even handle the case where the λ_i's are not independent but $\lambda = (\lambda_1, \ldots, \lambda_n)'$ follows some joint distribution on $\Re_+^n$. This situation would arise naturally if each P_{λ_i} depended on a common unknown parameter, for which a prior distribution was assumed. Markov chain Monte Carlo methods can easily be adapted to handle such extensions, as demonstrated in Fernández and Steel (1998) for the case of (skewed) Student sampling with unknown degrees of free-

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

dom. Of course, as we allow for more flexibility on the distribution of λ, our theoretical results will inevitably become less conclusive. Finally, an issue of importance that arises in this more general context of an unknown mixing distribution is how much we can expect to learn about it from the data. Intuitively, one would expect that a large number of observations is required, as the parameters of the mixing density are often largely determined by more extreme observations and not by the bulk of the data. Whereas this topic falls outside the scope of the present paper, which deals exclusively with the case of known mixing distribution, it is quite relevant for empirical work and certainly deserves further investigation.

REFERENCES

Andrews, D.F. & C.L. Mallows (1974) Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B* 36, 99–102.

Barndorff-Nielsen, O., J. Kent, & M. Sørensen (1982) Normal variance-mean mixtures and *z*-distributions. *International Statistical Review* 50, 145–159.

Bauwens, L. & M. Lubrano (1998) Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal* 1, C23–C46.

Berger, J.O. & J.M. Bernardo (1992) On the development of the reference prior method (with discussion). In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (eds.), *Bayesian Statistics 4*, pp. 35–60. Oxford: Oxford University Press.

Casella, G. & E. George (1992) Explaining the Gibbs sampler. *American Statistician* 46, 167–174.

DeGroot, M.H. (1970) *Optimal Statistical Decisions.* New York: McGraw-Hill.

Devroye, L. (1986) *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

Fang, K.-T., S. Kotz, & K.-W. Ng (1990) *Symmetric Multivariate and Related Distributions.* London: Chapman and Hall.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, vol. II, 2nd ed. New York: Wiley.

Fernández, C., J. Osiewalski, & M.F.J. Steel (1997) On the use of panel data in stochastic frontier models with improper priors. *Journal of Econometrics* 79, 169–193.

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

Fernández, C. & M.F.J. Steel (1999a) Reference priors for the general location-scale model. *Statistics and Probability Letters* 43, 377–384.

Fernández, C. & M.F.J. Steel (1999b) On the dangers of modelling through continuous distributions: A Bayesian perspective (with discussion). In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (eds.), *Bayesian Statistics 6*, pp. 213–238. Oxford: Oxford University Press.

Florens, J.P., M. Mouchart, & J.M. Rolin (1990) Invariance arguments in Bayesian statistics. In J. Gabszewicz, J.F. Richard, & L.A. Wolsey (eds.), *Economic Decision Making: Games, Econometrics and Optimisation*, pp. 387–403. Amsterdam: North-Holland.

Gantmacher, F.R. (1959) *The Theory of Matrices*, vol. 1. New York: Chelsea.

Gelfand, A.E. & A.F.M. Smith (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.

Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57, 1317–1340.

Geweke, J. (1993) Bayesian treatment of the independent Student-*t* linear model. *Journal of Applied Econometrics* 8, S19–S40.

Harvey, A.C., E. Ruiz, & N.G. Shephard (1994) Multivariate stochastic variance models. *Review of Economic Studies* 61, 247–264.

Hobert, J.P. & G. Casella (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 91, 1461–1473.
Jacquier, E., N.G. Polson, & P.E. Rossi (1995) Stochastic Volatility: Univariate and Multivariate Extensions. Mimeo, University of Chicago.
Jeffreys, H. (1961) *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.
Johnson, N.L. & S. Kotz (1970) *Distributions in Statistics: Continuous Univariate Distributions—1*. New York: Houghton Mifflin.
Kelker, D. (1970) Distribution theory of spherical distributions and a location-scale parameter generalization, *Sankyā A*32, 419–430.
Lange, K.L., R.J.A. Little, & J.M.G. Taylor (1989) Robust statistical modeling using the *t*-distribution. *Journal of the American Statistical Association* 84, 881–896.
Maronna, R. (1976) Robust M-estimators of multivariate location and scatter. *Annals of Statistics* 4, 51–67.
Mouchart, M. (1976) A note on Bayes theorem. *Statistica* 36, 349–357.
Osiewalski, J. (1991) A note on Bayesian inference in a regression model with elliptical errors. *Journal of Econometrics* 48, 183–193.
Osiewalski, J. & M.F.J. Steel (1992) Robust Bayesian inference in elliptical regression models. *Journal of Econometrics* 57, 345–363.
Osiewalski, J. & M.F.J. Steel (1996) Posterior moments of scale parameters in elliptical regression models. In D. Berry, K. Chaloner, & J. Geweke (eds.), *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* pp. 323–335. New York: Wiley.
Phillips, P.C.B. (1991) To criticize the critics: An objective Bayesian analysis of stochastic trends (with discussion). *Journal of Applied Econometrics* 6, 333–364.
Roberts, G.O. & A.F.M. Smith (1994) Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and Their Applications* 49, 207–216.
Rogers, W.H. & J.W. Tukey (1972) Understanding some long-tailed symmetric distributions. *Statistica Neerlandica* 26, 211–226.
Romanowski, M. (1979) *Random Errors in Observation and the Influence of Modulation on Their Distribution.* Stuttgart: Verlag Konrad Wittwer.
Shephard, N.G. (1994a) Local scale models: State space alternative to integrated GARCH processes. *Journal of Econometrics* 60, 181–202.
Shephard, N.G. (1994b) Partial non-Gaussian state space. *Biometrika* 81, 115–131.
Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 22, 1701–1762.
West, M. (1984) Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B* 46, 431–439.
West, M. (1987) On scale mixtures of normal distributions. *Biometrika* 74, 646–648.
Zellner, A. (1976) Bayesian and non-Bayesian analysis of the regression model with multivariate Student-*t* error terms. *Journal of the American Statistical Association* 71, 400–405.

APPENDIX: PROOFS

We first introduce some definitions and lemmas that will facilitate the proofs of the theorems. The notation used in the Appendix is consistent with that used in the body of the paper; thus, $y = (y_1, \ldots, y_n)'$ is the vector of observations, $X = (x_1, \ldots, x_n)'$ is the $n \times k$ design matrix of rank k, $\Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_n)$, and $\lambda = (\lambda_1, \ldots, \lambda_n)'$.

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

DEFINITION 3. *For $\lambda \in \Re_+^n$, we define $\lambda_{(1)} \leq \cdots \leq \lambda_{(n)}$ to be the ordered λ_i's.*

DEFINITION 4. *For $\lambda \in \Re_+^n$, we define $\{\lambda_{m_1}, \ldots, \lambda_{m_k}\}$ as the set of λ_i's that verifies $\Pi_{i=1}^k \lambda_{m_i} = \max\{\Pi_{i=1}^k \lambda_{s_i} : 1 \leq s_1 < \cdots < s_k \leq n$ and $\mathrm{Det}(x_{s_1}, \ldots, x_{s_k}) \neq 0\}$, where $\mathrm{Det}(x_{s_1}, \ldots, x_{s_k})$ denotes the determinant of the submatrix of X corresponding to the observations $y_{s_1}, \ldots, y_{s_k}$.*

The following lemmas provide bounds on functions of λ that will repeatedly appear in the proofs of the theorems. These bounds are given up to proportionality constants, which can depend on the fixed values of X and y (see also the remark after Lemma 1, which follows).

LEMMA 1. *$\mathrm{Det}(X'\Lambda X)$ has upper and lower bounds that are both proportional to $\Pi_{i=1}^k \lambda_{m_i}$.*

Remark. Lemma 1 means that there exist positive finite constants, $0 < C_1(X) \leq C_2(X) < \infty$, such that $C_1(X)\Pi_{i=1}^k \lambda_{m_i} \leq \mathrm{Det}(X'\Lambda X) \leq C_2(X)\Pi_{i=1}^k \lambda_{m_i}$. The remaining lemmas should be interpreted similarly, with constants possibly depending on X and/or y.

Proof of Lemma 1. Direct application of the Binet–Cauchy formula (Gantmacher, 1959, p. 9) leads to $\mathrm{Det}(X'\Lambda X) = \sum_{1 \leq s_1 < \cdots < s_k \leq n} (\Pi_{i=1}^k \lambda_{s_i}) \mathrm{Det}^2(x_{s_1}, \ldots, x_{s_k})$. For any functions $a_i(\lambda) > 0$, and constants $b_i > 0$, it is immediate that $a_{\max}(\lambda)\min_i\{b_i\} \leq \sum_i a_i(\lambda) b_i \leq a_{\max}(\lambda)\sum_i b_i$, where $a_{\max}(\lambda) = \max_i\{a_i(\lambda)\}$. Thus, $\sum_i a_i(\lambda) b_i$ has upper and lower bounds proportional to $a_{\max}(\lambda)$. Applying this idea to $\mathrm{Det}(X'\Lambda X)$, in combination with Definition 4, Lemma 1 follows. ■

LEMMA 2. *The Euclidean norm of $b(\lambda) = (X'\Lambda X)^{-1}X'\Lambda y$ is bounded above by a finite constant $C(X, y)$.*

Proof. Direct application of Cramer's rule to the linear system $(X'\Lambda X)b = X'\Lambda y$ leads to the following expression for the elements of $b(\lambda)$: $b(\lambda)_j = \mathrm{Det}(X'\Lambda X)^{-1}\mathrm{Det}(M_j)$, $j = 1, \ldots, k$, where the matrix M_j is obtained from $X'\Lambda X$ substituting the jth column by the vector $X'\Lambda y$. Applying the Binet–Cauchy formula to both determinants leads to

$$b(\lambda)_j = \frac{\displaystyle\sum_{1 \leq s_1 < \cdots < s_k \leq n} \left(\prod_{i=1}^k \lambda_{s_i}\right) \mathrm{Det}(x_{s_1}, \ldots, x_{s_k}) \mathrm{Det}(\tilde{x}_{s_1}, \ldots, \tilde{x}_{s_k})}{\displaystyle\sum_{1 \leq s_1 < \cdots < s_k \leq n} \left(\prod_{i=1}^k \lambda_{s_i}\right) \{\mathrm{Det}(x_{s_1}, \ldots, x_{s_k})\}^2},$$

where $\tilde{x}_{s_i}$, $i = 1, \ldots, k$ denotes the vector x_{s_i} with its jth component replaced by y_{s_i}. The result follows from a similar reasoning to the proof of Lemma 1, applied to both the numerator and denominator of $|b(\lambda)_j|$, after use of the bound $|\sum_i a_i| \leq \sum_i |a_i|$ for the numerator. ■

LEMMA 3. *For all $y \in \Re^n$ barring a set of Lebesgue measure zero, the expression $s(\lambda) = y'\Lambda y - y'\Lambda X(X'\Lambda X)^{-1}X'\Lambda y$ has upper and lower bounds proportional to $\lambda_b = \max\{\lambda_i : i \neq m_1, \ldots, m_k\}$.*

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

Proof. Defining the $n \times (k+1)$ matrix $L = (X : y)$, and subsequently applying the Binet–Cauchy formula, we obtain

$$s(\lambda) = \frac{\mathrm{Det}(L'\Lambda L)}{\mathrm{Det}(X'\Lambda X)} = \frac{1}{\mathrm{Det}(X'\Lambda X)} \sum_{1 \le s_1 < \cdots < s_{k+1} \le n} \left(\prod_{i=1}^{k+1} \lambda_{s_i} \right) \mathrm{Det}^2 \begin{pmatrix} x_{s_1} & \cdots & x_{s_{k+1}} \\ y_{s_1} & \cdots & y_{s_{k+1}} \end{pmatrix},$$

which, in combination with Lemma 1, proves Lemma 3. ■

LEMMA 4. *The jth diagonal element of $(X'\Lambda X)^{-1}$ has upper and lower bounds proportional to $1/\lambda_{\theta_j}$, where $\lambda_{\theta_j} = \min\{\lambda_\theta : \theta \in \{m_1, \ldots, m_k\}$ and $\mathrm{Det}({}^j x_{m_i} : m_i \neq \theta) \neq 0\}$, ${}^j x_i$ denotes the vector x_i without its jth element and $({}^j x_{m_i} : m_i \neq \theta)$ is the $(k-1) \times (k-1)$ matrix obtained from $({}^j x_{m_1}, \ldots, {}^j x_{m_k})$ after removing ${}^j x_\theta$.*

Proof. The term $(X'\Lambda X)^{-1}_{jj}$, the jth diagonal element of $(X'\Lambda X)^{-1}$, is computed as $(X'\Lambda X)^{-1}_{jj} = \mathrm{Det}(M_{jj})/\mathrm{Det}(X'\Lambda X)$, where M_{jj} is the matrix obtained from $X'\Lambda X$ by removing both the jth row and column. Applying the Binet–Cauchy formula, $\mathrm{Det}(M_{jj})$ is seen to be equal to $\sum_{1 \le s_1 < \cdots < s_{k-1} \le n} (\prod_{i=1}^{k-1} \lambda_{s_i}) \mathrm{Det}^2({}^j x_{s_1}, \ldots, {}^j x_{s_{k-1}})$, which, in combination with Lemma 1, leads to Lemma 4. ■

Proof of Theorem 1. This follows from either of the proofs of Theorems 2 and 4. ■

Proof of Theorem 2. Existence of the $(r_1, \ldots, r_k)$th order posterior moment of $\beta = (\beta_1, \ldots, \beta_k)'$ is equivalent to the following integral being finite:

$$\int_{\Re^k \times \Re_+} \left(\prod_{j=1}^{k} |\beta_j|^{r_j} \right) \left\{ \prod_{i=1}^{n} p(y_i|\beta,\sigma) \right\} p(\beta,\sigma)\, d\beta\, d\sigma, \tag{A.1}$$

where $p(y_i|\beta,\sigma)$ is the sampling density in (2.2) and $p(\beta,\sigma)$ the prior in (2.3). Straightforward calculations and the use of Fubini's theorem show that (A.1) is proportional to

$$\int_{\Re_+^n} \int_{\Re_+} \int_{\Re^k} \left(\prod_{j=1}^{k} |\beta_j|^{r_j} \right) f_N^k(\beta|b(\lambda), \sigma^2 (X'\Lambda X)^{-1})\, d\beta\, \sigma^{-(n-k+1)} \exp\left(-\frac{s(\lambda)}{2\sigma^2} \right) d\sigma$$
$$\times \left(\prod_{i=1}^{n} \lambda_i^{1/2} \right) \{\mathrm{Det}(X'\Lambda X)\}^{-1/2}\, dP_{\lambda_1} \ldots dP_{\lambda_n}, \tag{A.2}$$

where $b(\lambda) = (X'\Lambda X)^{-1} X'\Lambda y$ and $s(\lambda) = y'\Lambda y - y'\Lambda X (X'\Lambda X)^{-1} X'\Lambda y$. Observe that $s(\lambda)$ is strictly positive unless y is in the column space of X, which is an event of measure zero provided $n > k$. To first solve the integral on β, which we denote by I_1, we make a variable transformation from β to $q = \beta - b(\lambda)$; thus

$$I_1 = \int_{\Re^k} \left(\prod_{j=1}^{k} |q_j + b(\lambda)_j|^{r_j} \right) f_N^k(q|0, \sigma^2 (X'\Lambda X)^{-1})\, dq. \tag{A.3}$$

We now find a lower and an upper bound for I_1 (which, of course, lead to bounds on the integral in (A.2)). We shall use the lower bound to prove Theorem 2(i) and the upper bound to prove Theorem 2(ii).

Part (i): $r \ge n - k$. We consider the lower bound $|q_j + b(\lambda)_j|^{r_j} \ge |q_j|^{r_j} I_{(0,\infty)}(q_j b(\lambda)_j)$, where $I_A(v)$ takes the value one if $v \in A$ and zero otherwise. Applying this bound to the integral in (A.3) and defining the variable $t = \sigma^{-1} q$, we see that

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

$$I_1 \geq \sigma^r \int_{\Re^k} \left\{ \prod_{j=1}^{k} |t_j|^{r_j} I_{(0,\infty)}(t_j b(\lambda)_j) \right\} f_N^k(t|0,(X'\Lambda X)^{-1})\, dt, \tag{A.4}$$

where $r = \sum_{j=1}^{k} r_j$. Next, we look at the integral with respect to σ in (A.2):

$$\int_{\Re_+} \sigma^{-(n-k-r+1)} \exp\left(-\frac{s(\lambda)}{2\sigma^2}\right) d\sigma \propto \int_{\Re_+} h^{[(n-k-r)/2]-1} \exp\left(-\frac{s(\lambda)h}{2}\right) dh,$$

which requires $n-k-r>0$ for being finite. Thus Theorem 2(i) follows.

Part (ii): $r < n-k$. Because

$$\prod_{j=1}^{k} |q_j + b(\lambda)_j|^{r_j} = \prod_{j=1}^{k} (|q_j + b(\lambda)_j|^r)^{r_j/r} \leq \sum_{j:r_j>0} \frac{r_j}{r} |q_j + b(\lambda)_j|^r,$$

where the last inequality follows directly from the theorem of arithmetic and geometric means, we shall focus on marginal moments for β_j of order r, for those j such that $r_j > 0$. From Lemma 2 we know that $|b(\lambda)_j| \leq C(X,y)$, $j = 1,\ldots,k$, for some positive quantity $C(X,y)$. We then obtain

$$|q_j + b(\lambda)_j|^r \leq (|q_j| + |b(\lambda)_j|)^r \leq \{|q_j| + C(X,y)\}^r \leq 2^r\{C(X,y)\}^r + 2^r|q_j|^r,$$

and, thus, if the integral

$$\int_{\Re_+^n} \int_{\Re_+} \int_{\Re} |q_j|^l f_N^1(q_j|0,\sigma^2(X'\Lambda X)_{jj}^{-1})\, dq_j\, \sigma^{-(n-k+1)} \exp\left(-\frac{s(\lambda)}{2\sigma^2}\right) d\sigma$$
$$\times \left(\prod_{i=1}^{n} \lambda_i^{1/2}\right) \{\mathrm{Det}(X'\Lambda X)\}^{-1/2}\, dP_{\lambda_1} \ldots dP_{\lambda_n} \tag{A.5}$$

is finite for $l=0$ and $l=r$ for all j corresponding to $r_j>0$, the integral in (A.2) will also be finite, and the $(r_1,\ldots,r_k)$th posterior moment of β will exist. Note that propriety of the posterior distribution is equivalent to a finite integral in (A.5) for $l=0$, and thus, the present proof also covers the proof of Theorem 1.

After integrating out q_j and σ, we are left with the integral

$$\int_{\Re_+^n} \{(X'\Lambda X)_{jj}^{-1}\}^{l/2} \left(\prod_{i=1}^{n} \lambda_i^{1/2}\right) \{\mathrm{Det}(X'\Lambda X)\}^{-1/2} s(\lambda)^{-(n-k-l)/2}\, dP_{\lambda_1} \ldots dP_{\lambda_n}. \tag{A.6}$$

We decompose the domain of integration $\Re_+^n$ into the $n!$ possible orderings of $\{\lambda_1,\ldots,\lambda_n\}$. In each of these regions we identify $\lambda_{m_1},\ldots,\lambda_{m_k}$ (Definition 4), λ_b (Lemma 3), and λ_{θ_j} (Lemma 4). Given one of these orderings and applying the previous lemmas we obtain upper and lower bounds of the integrand in (A.6) proportional to

$$F_1(\lambda) = \frac{\prod_{i \neq m_1,\ldots,m_k} \lambda_i^{1/2}}{\lambda_{\theta_j}^{l/2} \lambda_b^{(n-k-l)/2}}. \tag{A.7}$$

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

The definition of singularity index tells us that the largest possible submatrix of X of rank $k-1$ that keeps the same rank after removing column j consists of $k-1+p_j$ vectors x_i. As a consequence, for each possible ordering of $\lambda_1,\ldots,\lambda_n$, there are at most $k-1+p_j$ λ_i's that are larger than λ_{θ_j} (and this includes the $k-1$ λ_{m_i}'s different from λ_{θ_j}). Equivalently, there are at least $n-k-p_j$ λ_i's outside the set $\{\lambda_{m_1},\ldots,\lambda_{m_k}\}$, which are smaller than λ_{θ_j}. This implies that for $l=0$ and $l=r\leq n-k-p_j$,

$$F_1(\lambda)\leq \frac{\lambda_{\theta_j}^{l/2}\lambda_b^{(n-k-l)/2}}{\lambda_{\theta_j}^{l/2}\lambda_b^{(n-k-l)/2}}=1, \tag{A.8}$$

which is integrable with respect to P_{λ_i}. On the other hand, if $l=r>n-k-p_j$, we use the bound

$$F_1(\lambda)\leq \frac{\lambda_{\theta_j}^{(n-k-p_j)/2}\lambda_b^{p_j/2}}{\lambda_{\theta_j}^{l/2}\lambda_b^{(n-k-l)/2}}=\left(\frac{\lambda_{\theta_j}}{\lambda_b}\right)^{(n-k-p_j-l)/2}. \tag{A.9}$$

Clearly, if both $E(\lambda_i^{(n-k-p_j-r)/2})$ and $E(\lambda_i^{(r+p_j+k-n)/2})$ are finite, $F_1(\lambda)$ will be integrable. Using Definition 2, if $r-(n-k-p_j)<m$, both expectations are finite. Thus, $r<n-k-p(r_1,\ldots,r_k)+m$, where $p(r_1,\ldots,r_k)=\max\{p_j: r_j>0\}$, leads to integrability for all j such that $r_j>0$ and Theorem 2(ii) follows. ■

Proof of Theorem 3.

Pareto(1,ν/2) Mixing Distribution with $\nu\geq 1$. The existence of the rth order marginal posterior moment of β_j is equivalent to the integral in (A.6) being finite for $l=r$. Following the same reasoning as in the proof of Theorem 2(ii), we need to integrate $F_1(\lambda)$ in (A.7) (with $l=r$) over all possible orderings of $\{\lambda_1,\ldots,\lambda_n\}$. Because a Pareto(1,$\nu$/2) distribution has support on $(1,\infty)$, we obtain $F_1(\lambda)\leq \lambda_b^{-(n-k-r)/2}\Pi_{i\neq m_1,\ldots,m_k}\lambda_i^{1/2}$ and, thus, the integral of $F_1(\lambda)$ over any ordering of the λ_i's is bounded above by

$$\int_{1<\eta_1<\cdots<\eta_{n-k}<\infty}\eta_{n-k}^{-(n-k-r)/2}\left(\prod_{i=1}^{n-k}\eta_i^{1/2}\right)p(\eta_1)\ldots p(\eta_{n-k})\,d\eta_1\ldots d\eta_{n-k}, \tag{A.10}$$

where $p(\eta_i)\propto \eta_i^{-(\nu/2)-1}$ for $i=1,\ldots,n-k$. Using Fubini's theorem to compute (A.10) in an iterative fashion in the order $\eta_{n-k},\ldots,\eta_1$ leads to a finite value for any $r<n-k$.

Gamma(ν/2,μ/2) and Beta(ν/2,1) Mixing. As is clear from the comments following (A.7), the largest value of $F_1(\lambda)$ corresponds to any ordering of the λ_i's for which $\lambda_{\theta_j}=\lambda_{(n-k-p_j+1)}$. Thus, it is enough to establish the integrability of (A.7) for any such ordering. We again compute the integral iteratively, using Fubini's theorem.

For gamma(ν/2,μ/2) mixing, which corresponds to $p(\lambda_i)\propto \lambda_i^{(\nu/2)-1}\exp(-\mu\lambda_i/2)$ for $\lambda_i>0$, we use the following bounds in each of the n steps of the integration process:

$$\frac{\lambda_\xi^\eta}{\eta}\exp(-\mu\lambda_\xi/2)\leq\int_0^{\lambda_\xi}\lambda_i^{\eta-1}\exp(-\mu\lambda_i/2)\,d\lambda_i\leq\frac{\lambda_\xi^\eta}{\eta},\quad \text{for any } \eta,\mu>0. \tag{A.11}$$

It is easy to see that after the first $n-k-p_j$ steps, we are left with

$$\int_0^{\lambda_{(n-k-p_j+2)}}\lambda_{\theta_j}^{\{(n-k-p_j+1)\nu+n-k-p_j-r\}/2-1}\exp(-\mu\lambda_{\theta_j}/2)\,d\lambda_{\theta_j}, \tag{A.12}$$

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press

which is finite if and only if $r < n - k - p_j + \nu(n - k - p_j + 1)$. Once this condition is imposed, the remaining steps always lead to finite integrals.

The proof for beta mixing is similar throughout, now using that

$$\int_0^{\lambda_\zeta} \lambda_i^{\eta-1}\, d\lambda_i \propto \lambda_\zeta^\eta, \quad \text{for any } \eta > 0, \tag{A.13}$$

instead of the bounds in (A.11). ■

Proof of Theorem 4. The rth posterior moment of σ is finite if and only if the integral

$$\int_{\Re^k \times \Re_+} \sigma^r \left\{ \prod_{i=1}^n p(y_i|\beta,\sigma) \right\} p(\beta,\sigma)\, d\beta\, d\sigma \tag{A.14}$$

is finite, where $p(y_i|\beta,\sigma)$ and $p(\beta,\sigma)$ are given in (2.2) and (2.3), respectively. After integrating out β, using the fact that its conditional distribution given (σ,λ) is a k-variate normal, we are left with the following integral proportional to (A.14):

$$\int_{\Re_+^n} \frac{\prod_{i=1}^n \lambda_i^{1/2}}{\{\mathrm{Det}(X'\Lambda X)\}^{1/2}} \int_{\Re_+} \sigma^{-(n-k-r+1)} \exp\left(-\frac{s(\lambda)}{2\sigma^2}\right) d\sigma\, dP_{\lambda_1} \ldots dP_{\lambda_n}, \tag{A.15}$$

where $s(\lambda) = y'\Lambda y - y'\Lambda X(X'\Lambda X)^{-1}X'\Lambda y > 0$ for all y in $\Re^n$ barring a k-dimensional subspace.

Part (i): $r \geq n - k$. To integrate out σ in (A.15) we require $n - k - r > 0$. Hence Theorem 4(i).

Part (ii): $r < n - k$. In this case we can integrate out σ, and the integral in (A.15) is proportional to

$$\int_{\Re_+^n} \left(\prod_{i=1}^n \lambda_i^{1/2} \right) \{\mathrm{Det}(X'\Lambda X)\}^{-1/2} s(\lambda)^{-(n-k-r)/2}\, dP_{\lambda_1} \ldots dP_{\lambda_n}. \tag{A.16}$$

We now decompose $\Re_+^n$ into all possible orderings of $\{\lambda_1, \ldots, \lambda_n\}$. For each of these regions, the previous lemmas lead to upper and lower bounds for the integrand in (A.16) proportional to

$$F_2(\lambda) = \frac{\prod_{i \neq m_1, \ldots, m_k} \lambda_i^{1/2}}{\lambda_b^{(n-k-r)/2}} \leq \lambda_b^{r/2}, \tag{A.17}$$

where $\lambda_b = \max\{\lambda_i, i \neq m_1, \ldots, m_k\}$. Theorem 4(ii) now follows immediately. ■

Proof of Theorem 5. The proof is entirely parallel to that of Theorem 3, substituting $F_1(\lambda)$ in (A.7) by $F_2(\lambda)$ in (A.17). The result is immediate for Pareto mixing, because $F_2(\lambda)$ exactly corresponds to the upper bound for $F_1(\lambda)$ used in the proof of Theorem 3. For gamma and beta mixing, we respectively apply (A.11) and (A.13) to integrate $F_2(\lambda)$. ■

https://doi.org/10.1017/S0266466600161043 Published online by Cambridge University Press