Electronic Journal of Statistics
Vol. 12 (2018) 572–598
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1401

Flexible linear mixed models with improper priors for longitudinal and survival data[*]

F. J. Rubio and M. F. J. Steel

London School of Hygiene & Tropical Medicine, London, UK
e-mail: Francisco.Rubio@lshtm.ac.uk

University of Warwick, Department of Statistics, Coventry, CV4 7AL, UK
e-mail: M.F.Steel@stats.warwick.ac.uk

Abstract: We propose a Bayesian approach using improper priors for hierarchical linear mixed models with flexible random effects and residual error distributions. The error distribution is modelled using scale mixtures of normals, which can capture tails heavier than those of the normal distribution. This generalisation is useful to produce models that are robust to the presence of outliers. The case of asymmetric residual errors is also studied. We present general results for the propriety of the posterior that also cover cases with censored observations, allowing for the use of these models in the contexts of popular longitudinal and survival analyses. We consider the use of copulas with flexible marginals for modelling the dependence between the random effects, but our results cover the use of any random effects distribution. Thus, our paper provides a formal justification for Bayesian inference in a very wide class of models (covering virtually all of the literature) under attractive prior structures that limit the amount of required user elicitation.

MSC 2010 subject classifications: 62F15, 62J05, 62N01.
Keywords and phrases: Bayesian inference, heavy tails, MEAFT models, posterior propriety, skewness, stochastic frontier models.

Received March 2017.

Contents

[*] This research was partially supported by EPSRC. We gratefully acknowledge insightful comment from two referees.

1. Introduction

Longitudinal and survival data appear in many application areas such as medicine, biology, agriculture, epidemiology, economics and small area estimation. This kind of observations typically present within- and among- subject dependence due to grouping and/or repeated measurements. Hierarchical linear mixed models (LMM) are often used to account for parameter variation across groups of observations as well as dependence. The general formulation of this type of model is given by

$$y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{u}_i + \varepsilon_{ij}, \tag{1}$$

where y_{ij} is the observed response for subject i at time t_{ij}, $j = 1, \dots, n_i$ with n_i recording the number of repeated measurements for subject i, and $i = 1, \dots, r$ where r denotes the number of subjects, $\boldsymbol{\beta}$ is a $p \times 1$ vector of *fixed effects*, $\mathbf{u}_i$ are $q \times 1$ mutually independent random vectors (often called *random effects*) and ε_{ij} are *i.i.d.* residual errors. Let $\mathbf{y} = \{y_{ij}\}$ be the $n \times 1$ vector of response variables and $\mathbf{X}$ and $\mathbf{Z}$ denote the known design matrices of dimension $n \times p$ and $n \times qr$, respectively, while $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of residual errors, and $\mathbf{u} = (\mathbf{u}_1^\top, \dots, \mathbf{u}_r^\top)^\top$. We assume that rank$(\mathbf{X}) = p$, and $n > p + qr$ throughout. In matrix notation we can write (1) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}.$$

These models are used in different fields under different names and notation. In Bayesian analysis of variance, the use of these models goes back to Tiao and Tan (1965). In survival analysis, the n survival times grouped in $\mathbf{T}$ are usually transformed to logarithms and modelled through

$$\log(\mathbf{T}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{2}$$

which is often referred to as a Mixed Effects Accelerated Failure Time model (MEAFT, Komárek and Lesaffre, 2007). In this literature, the random effects $\mathbf{u}$ are typically called *frailties*. An additional challenge that often appears in this case is the presence of censored observations. In the context of production and cost frontiers used in economics, the logarithm of the output (or cost) of r firms are modelled using (1) with certain regularity restrictions on the regression coefficients $\boldsymbol{\beta}$ and sign restrictions on the random effects $\mathbf{u}_i$, which have a clear interpretation here in terms of inefficiencies. These models are known as stochastic frontier models (see *e.g.* Fernández et al., 1997).

The random vectors $\boldsymbol{\varepsilon}$ and $\mathbf{u}$ are often assumed to be normally distributed. Given that the normality assumption can be restrictive in practice (and even impossible for stochastic frontier models), alternative distributional assumptions have been explored, such as smooth flexible (semi-nonparametric) distributions for the random effects and normal residual errors (Zhang and Davidian, 2001), skew Student-t distributions for the random effects (Lee and Thompson, 2008), normal random effects with residual errors modelled through a finite mixture of normals (Komárek and Lesaffre, 2007, 2008), and residual errors and random effects jointly distributed as a multivariate scale mixture of skew-normal distributions (Lachos et al., 2010), which makes the residual errors and the random effects dependent. Jara et al. (2008) model both the residual errors and the random effects using multivariate skew–elliptical distributions, and Dunson (2010) and Jara et al. (2009) use Bayesian nonparametric approaches. Although flexible, Bayesian nonparametric approaches tend to be more computationally expensive than simpler flexible parametric models, while leading to similar conclusions as shown in our examples.

In a Bayesian framework with improper priors, Hobert and Casella (1996) and Sun et al. (2001) present conditions for the propriety of the posterior under a certain improper prior structure for the parameters of model (1) with normal assumptions on both the random effects and the residual errors. When using improper prior structures, it becomes essential to check the propriety of the posterior distribution in order to justify their use and interpretation in a probabilistic framework. Hobert and Casella (1996) characterise cases where certain improper prior structures used in practice lead to proper posteriors. The use of such priors was becoming common at that time since they do allow for the construction of a Gibbs sampler. However, Hobert and Casella (1996) show that the posterior may still be improper in some cases and, consequently, their use is not theoretically justified. Rubio (2015) presents extensions of these propriety results to the use of certain classes of flexible random effects distributions. Hobert and Casella (1996) and Rubio (2015) assume independence of the random effects, while Sun et al. (2001) assume a specific structure for the covariance of the random effects. Fernández et al. (1997) proposed an improper prior structure which allows for using arbitrary random effects distributions, as long they are proper. More specifically, they adopt the following stochastic structure

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), \\ \boldsymbol{\varepsilon} \mid \sigma_\varepsilon &\sim \mathrm{N}_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n) \\ \pi(\boldsymbol{\beta}, \sigma_\varepsilon) &\sim \frac{\pi(\boldsymbol{\beta})}{\sigma_\varepsilon^{b+1}}, \quad b \geq 0, \end{aligned} \tag{3}$$

where $p(\mathbf{u})$ is proper, and $\pi(\boldsymbol{\beta})$ is proper or bounded. The condition that $p(\mathbf{u})$ is proper is rather mild as this allows for using any parametric distribution for the random effects $\mathbf{u}$ with proper priors on the "deeper" parameters that index the random effects distribution. The propriety of $p(\mathbf{u})$ is essential since an improper $p(\mathbf{u})$ never leads to a posterior (see Theorem 2 from Fernández et al., 1997). The prior structure on $(\boldsymbol{\beta}, \sigma_\varepsilon)$ is interesting as it includes priors with useful properties such as invariance with respect to reparameterisation under affine transformations. In this paper, we extend (3) by relaxing the assumption of normality of the residual errors and derive conditions for the existence of the corresponding posterior distribution that also allow for censored observations, covering virtually all models in the recent literature. Before presenting our results, we clarify our main contributions relative to our previous work in flexible linear regression models. Rubio and Genton (2016) study linear regression models with skew-symmetric errors and improper priors, covering the case with censored observations. Analogous results are presented in Rubio and Yu (2017) in the context of linear regression models with two-piece residual errors. Both papers focus on the effect of using skewed residual error distributions in terms of the prediction of the residual life of censored individuals. They do not study the propriety of the posterior distribution of linear regression models with random effects, which requires special attention as shown in this paper.

The Appendix includes proofs not mentioned in the text (Appendix A), conditions on the mixing distributions (Appendix B) and a further simulation study with skewed errors and random effects (Appendix C). The R codes used in the real data examples are available at `http://www.rpubs.com/FJRubio`.

2. Multivariate scale mixtures of normals: a warning

In the context of linear regression, a common strategy for relaxing the assumption of normality of the residual errors consists in using multivariate scale mixtures of normals (Lachos et al., 2010). Recall that a p-variate scale mixture of normal distribution ($\mathrm{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta; H)$) with location parameter $\boldsymbol{\mu}$, symmetric positive definite scale matrix $\boldsymbol{\Sigma}$, and parametric mixing distribution $H(\cdot \mid \delta)$ is defined through

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta) = \int_0^\infty \frac{\tau^{p/2}}{\det\,(2\pi\boldsymbol{\Sigma})^{1/2}} \exp\left[-\frac{\tau}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] dH(\tau \mid \delta).$$

For example, when $H(\tau \mid \delta)$ is a Gamma distribution with parameters $(\delta/2, \delta/2)$ (and mean one), we obtain the p-variate Student-t distribution with δ degrees of freedom. Osiewalski and Steel (1993) and Breusch et al. (1997) have pointed out inferential peculiarities using such error distributions, both in Bayesian and classical regression settings. Here we explore the implications of using multivariate scale mixtures of normal in the context of LMM.

Consider the following hierarchical structure for model (1)

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), & (4)\\ \boldsymbol{\varepsilon} \mid \sigma_\varepsilon, \delta_\varepsilon &\sim \mathrm{SMN}_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n, \delta_\varepsilon; H_\varepsilon), & (5)\end{aligned}$$

where $p(\mathbf{u})$ is proper. For our theoretical results, it does not matter whether $p(\mathbf{u})$ is assumed to be a fixed distribution without further recourse to deeper parameters or whether it is a marginal distribution obtained from a proper joint distribution. In practice, the latter is typically the case and it is derived from a proper parametric distribution with a proper prior on its parameters. Throughout Sections 2 and 3, we will only refer to $p(\mathbf{u})$ for the sake of simplicity of notation.

We adopt the following prior structure

$$\pi(\boldsymbol{\beta},\sigma_\varepsilon,\delta_\varepsilon) \quad \sim \quad \frac{\pi(\boldsymbol{\beta})\pi(\delta_\varepsilon)}{\sigma_\varepsilon^{b+1}}, \tag{6}$$

where $b \geq 0$, $\pi(\boldsymbol{\beta})$ is bounded, and $\pi(\delta_\varepsilon)$ is a proper prior with support $\Delta \subset \mathbb{R}$. The following result shows that the marginal likelihood of the data for this hierarchical model can be factorised as the product of the marginal likelihood under normality multiplied by a quantity that does not depend on the data.

Remark 1. *For model (1), (4)-(6) the marginal likelihood of the data can be written as*

$$\tilde{m}(\mathbf{y}) \quad = \quad m(\mathbf{y})\int_\Delta\int_{\mathbb{R}_+}\tau^{-\frac{b}{2}}\pi(\delta_\varepsilon)dH_\varepsilon(\tau\mid\delta_\varepsilon)d\delta_\varepsilon, \tag{7}$$

where $m(\mathbf{y})$ is the marginal likelihood corresponding to model (1), (3).

The factorisation of the marginal likelihood of the data can be used to obtain conditions for the propriety of the posterior distribution as follows.

Corollary 1. *Consider the model (1), (4)-(6) and let $(X : Z)$ denote the entire design matrix, and suppose that $p(\mathbf{u})$ is proper. Consider also the following conditions:*

(a) $\operatorname{rank}(X : Z) < n$,
(b) $b \geq 0$,
(c) $\int_\Delta\int_{\mathbb{R}_+}\tau^{-\frac{b}{2}}\pi(\delta_\varepsilon)dH_\varepsilon(\tau\mid\delta_\varepsilon)d\delta_\varepsilon < \infty$.

Then, conditions (a) and (c) are necessary for the propriety of the posterior and conditions (a), (b), and (c) are sufficient for the propriety of the posterior.

Proof. From Remark 1 we have that the marginal density of the data can be written as in (7). The finiteness of $m(\mathbf{y})$ in this expression, under assumptions (a)–(b), is proved in Theorem 1 from Fernández et al. (1997). The second factor in (7) is finite by assumption (c).

For $b = 0$, condition (c) is satisfied by any mixing distribution. However, for $b > 0$ this condition may rule out the use of some mixing distributions and/or priors on δ_ε, as discussed in the Supplementary Material (Appendix B). Remark 1 and Corollary 1 point out some issues with this extension. If we use Bayes factors to compare two models $\mathcal{M}_1$ and $\mathcal{M}_2$ of the type described in Remark 1, with different distributions for the residuals and corresponding marginal likelihoods $\tilde{m}_1(\mathbf{y})$ and $\tilde{m}_2(\mathbf{y})$, we obtain (using obvious notation)

$$B_{12} = \frac{\tilde{m}_1(\mathbf{y})}{\tilde{m}_2(\mathbf{y})} = \frac{\int_\Delta\int_{\mathbb{R}_+}\tau^{-\frac{b}{2}}\pi_1(\delta_\varepsilon)dH_{\varepsilon,1}(\tau\mid\delta_\varepsilon)d\delta_\varepsilon}{\int_\Delta\int_{\mathbb{R}_+}\tau^{-\frac{b}{2}}\pi_2(\delta_\varepsilon)dH_{\varepsilon,2}(\tau\mid\delta_\varepsilon)d\delta_\varepsilon}.$$

Therefore, the Bayes factor is determined solely by the mixing distributions and their priors, but not by the data (!). Intuitively, these results indicate that the data do not contain information about the additional shape parameter δ_ε. The factorisation (7) indicates that the use of this kind of multivariate extension is vacuous, in the sense that the additional shape parameters do not really help model selection. Similar issues with the use of certain classes of multivariate distributions and priors have been studied in Maruyama and Strawderman (2014) for linear regression models. The next section presents an alternative extension that overcomes such problems.

3. The proposed extension: flexible linear mixed models

The inferential issues pointed out in the previous section are due to the use of a single mixing variable in the construction of the joint distribution of the residual errors. In other words, (5) really implies only a single realisation of the mixing variable, irrespective of the size of n. Thus, we only have a single observation of the mixing variable, and inference on its (precision) parameter is impossible from the sample. One way to avoid these issues consists of using one mixing variable for each individual error. In this case, each observation contributes to the estimation of the parameter of the distribution of the mixing variable (see West, 1984 for a discussion in the context of Bayesian linear regression). Thus, we consider the model in (1) and (4), but replace (5) by

$$\varepsilon_{ij} \mid \sigma_\varepsilon, \delta_\varepsilon, H_\varepsilon \overset{i.i.d.}{\sim} \text{SMN}_1(0, \sigma_\varepsilon, \delta_\varepsilon; H_\varepsilon). \tag{8}$$

This is a rather wide class of symmetric and unimodal distributions, with tails which are either normal or fatter than normal, covering *e.g.* the normal, Student-t, logistic, Laplace, Cauchy and the exponential power family with power $1 \leq q < 2$ (see Fernández and Steel, 2000 for more details). These distributions naturally lead to models that are robust to the presence of outliers (see Rosa et al., 2003, who also adopt SMN_1 residual errors, but restrict their study to the use of proper priors and normal random effects).

Using the same prior structure (6), we derive the following conditions for the propriety of the posterior distribution.

Theorem 1. *For the model (1), (4), (8) with prior (6), and $p(\mathbf{u})$ a proper distribution, consider the following conditions:*

(a) $\text{rank}(\mathbf{X} : \mathbf{Z}) < n$,
(b) $b \geq 0$,
(c) $\int_\Delta \int_{\mathbb{R}_+} \tau^{-\frac{b}{2}} \pi(\delta_\varepsilon) dH_\varepsilon(\tau \mid \delta_\varepsilon) d\delta_\varepsilon < \infty$.
(d) $\mathbf{y}$ *is not an element of the column space of* $(\mathbf{X} : \mathbf{Z})$.

Condition (a) is necessary for the propriety of the posterior. Conditions (a) – (d) are sufficient for the propriety of the posterior.

Condition (a) indicates the need for repeated measurements (otherwise $n = r$ and $\text{rank}(\mathbf{X} : \mathbf{Z}) = n$ for sensible choices of $\mathbf{Z}$), while conditions (b) and (c) characterise the priors and mixing distributions that produce a proper posterior. Given that all the distributional assumptions in Theorem 1 are continuous, it follows that condition (d) is satisfied with probability one. Condition (d) is equivalent to saying that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ has no solution, which can be checked numerically.

The only requirement on the distribution of the random effects $\mathbf{u}$ is that the marginal $p(\mathbf{u})$ is proper, so as stated before, we can use an arbitrary parametric distribution $\mathbf{u} \sim F(\cdot \mid \boldsymbol{\theta})$, with a proper prior on the parameter $\boldsymbol{\theta}$. This includes, for example, the use of finite mixtures of normals, scale mixtures of normals, skewed scale mixtures of normals, truncated distributions, etc. In our examples, we explore the use of distributions that can capture departures from normality in terms of asymmetry and tail behaviour.

3.1. Mixed effects accelerated failure time models with censoring

The propriety result presented in Theorem 1 can be extended to models used in survival analysis. Let $\mathbf{T}$ be a sample of n survival times. The MEAFT in (2) is exactly the same as (1) for the log survival times. However, a common difficulty that arises with survival regression models is the presence of censored observations (Komárek and Lesaffre, 2007; Vallejos and Steel, 2015). If the sample of survival times does not contain censored observations, the existence of the posterior is provided by Theorem 1. If the sample contains both censored and uncensored observations, the propriety of the posterior can be based on the uncensored observations as follows.

Theorem 2. *Let $\mathbf{T}$ be a sample of survival times where n_0 observations are uncensored, with $p + qr < n_0 < n$. Consider the model (2), (4), (8) with prior (6), and $p(\mathbf{u})$ a proper distribution. Then, the posterior is proper if the marginal likelihood of the uncensored observations is finite.*

The following result presents conditions for the case when the sample contains only censored observations.

Theorem 3. *Suppose that n_c survival times T_j, $j = 1, \dots n_c$, are observed as closed intervals I_j, and the rest of the observations exhibit another type of censoring. Thus, let $I_1, \dots, I_{n_c}$ be finite-length intervals on the positive real line, $n_c \leq n$ and let $(\mathbf{X} : \mathbf{Z})_{n_c}$ represent the design matrix associated to the n_c interval–censored observations. Consider the model (2), (4), (8) with prior (6), $p(\mathbf{u})$ assumed to be proper, and the following condition:*

(d′) The set $\mathcal{E} = I_1 \times \cdots \times I_{n_c}$ and the column space of $(\mathbf{X} : \mathbf{Z})_{n_c}$ are disjoint.

Conditions (a)–(c) from Theorem 1 together with (d′) are sufficient for the propriety of the posterior.

Condition (d′) means that there is no –unobserved– true response that can be written as a linear combination of the columns of X and Z, and can be relaxed in terms of the dimensionality as follows.

(d'') There exists a set $\mathcal{E}' = I_{j_1} \times \cdots \times I_{j_{n'}}$, $p + qr < n' \leq n_c$, for some set of indexes $\mathcal{J} = \{j_1, \dots, j_{n'}\} \subseteq \{1, \dots, n\}$, such that $\mathcal{E}'$ and the column space of $(\mathbf{X}', \mathbf{Z}')$ are disjoint, where $\mathbf{X}'$ and $\mathbf{Z}'$ are the design submatrices associated to the indexes $\mathcal{J}$.

Another way of checking condition (d$'$) consists of formulating it as a linear programming problem. Denote $\boldsymbol{\eta} \in \mathbb{R}^{p+qr}$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{n_c}) \in \mathcal{E}$, and $I_j = [l_j, u_j]$, $j = 1, \dots, n_c$. Define the problem:

$$\begin{aligned} \text{Find} \quad & \max_{\boldsymbol{\eta}, \boldsymbol{\xi}} 1, \\ \text{Subject to} \quad & (\mathbf{X} : \mathbf{Z})_{n_c} \boldsymbol{\eta} = \boldsymbol{\xi}, \\ \text{and} \quad & \log(l_j) \leq \xi_j \leq \log(u_j), \;\; j = 1, \dots, n_c. \end{aligned} \tag{9}$$

Then, condition (d$'$) is equivalent to verifying the infeasibility of (9), for which several theoretical and numerical tools are available.

3.2. Stochastic frontier models

The stochastic frontier model with composed error takes the form of (1) where $\mathbf{u} \in \mathcal{U}$ with $\mathcal{U} = \{\mathbf{u} : \mathbf{Z}\mathbf{u} \in \mathbb{R}^n_+\}$. Several specifications of interest in the context of economics are studied in Fernández et al. (1997). The propriety results in Theorem 1 apply to this model, which represents an extension of Theorem 1 in Fernández et al. (1997).

3.3. Asymmetric errors

So far, we have considered symmetric extensions of the standard linear mixed model with normal errors. We now extend to asymmetric errors by using the class of two–piece distributions. Let f be a continuous density with a unique mode at 0, support on $\mathbb{R}$ and shape parameter δ. A random variable W is distributed according to a two-piece f-distribution (denoted $W \sim \text{TP}(\mu, \sigma, \gamma, \delta; f)$) if its density function can be written as $s(w \mid \mu, \sigma, \gamma, \delta) =$

$$\frac{2}{\sigma[a(\gamma) + b(\gamma)]} \left[f\left(\frac{w - \mu}{\sigma b(\gamma)} \,\middle|\, \delta \right) I(w < \mu) + f\left(\frac{w - \mu}{\sigma a(\gamma)} \,\middle|\, \delta \right) I(w \geq \mu) \right]. \tag{10}$$

where $I(\cdot)$ denotes the indicator function, and $\{a(\cdot), b(\cdot)\}$ are positive differentiable functions. This density is continuous, unimodal, with mode at $\mu \in \mathbb{R}$, scale parameter $\sigma \in \mathbb{R}_+$, and skewness parameter $\gamma \in \Gamma \subset \mathbb{R}$. It coincides with the density f when $a(\gamma) = b(\gamma)$, and is asymmetric for $a(\gamma) \neq b(\gamma)$ while retaining the tails of f in each direction. The most common choices for $a(\cdot)$ and $b(\cdot)$ correspond to the *inverse scale factors* parameterisation $\{a(\gamma), b(\gamma)\} = \{\gamma, 1/\gamma\}$, $\gamma \in \mathbb{R}_+$ (Fernández and Steel, 1998), and the *epsilon-skew* parameterisation $\{a(\gamma), b(\gamma)\} = \{1 - \gamma, 1 + \gamma\}$, $\gamma \in (-1, 1)$ (Mudholkar and Hutson, 2000). Rubio and Steel (2014) study other parameterisations as well as some interpretable choices for the prior of the skewness parameter γ.

Suppose now that f in (10) is a univariate scale mixture of normals with mixing distribution H_ε and consider the model (1) with the following error structure

$$\varepsilon_{ij} \mid \sigma_\varepsilon, \gamma_\varepsilon, \delta_\varepsilon \overset{i.i.d.}{\sim} \text{TP}(0, \sigma_\varepsilon, \gamma_\varepsilon, \delta_\varepsilon; f), \tag{11}$$

and prior

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon, \gamma_\varepsilon, \delta_\varepsilon) \quad \sim \quad \frac{\pi(\boldsymbol{\beta})\pi(\gamma_\varepsilon)\pi(\delta_\varepsilon)}{\sigma_\varepsilon^{b+1}}, \tag{12}$$

where $b \geq 0$, $\pi(\boldsymbol{\beta})$ is bounded, and $\pi(\gamma_\varepsilon)$ and $\pi(\delta_\varepsilon)$ are proper priors with $\gamma_\varepsilon \in \Gamma$.

Theorem 4. *Consider the model (1), (4), (11), (12), $p(\mathbf{u})$ a proper distribution, and conditions (a)–(d) from Theorem 1 together with the following condition:*

(e) $\int_\Gamma M(\gamma_\varepsilon)^b \pi(\gamma_\varepsilon) d\gamma_\varepsilon < \infty$, *where* $M(\gamma_\varepsilon) = \max\{a(\gamma_\varepsilon), b(\gamma_\varepsilon)\}$.

Then, conditions (a)–(e) are sufficient for the propriety of the posterior distribution.

Condition (e) is automatically satisfied for $b = 0$ and any parameterisation $\{a(\cdot), b(\cdot)\}$. For the epsilon-skew parameterisation, condition (e) is satisfied for any $b \geq 0$ and any proper prior $\pi(\cdot)$ given that the functions $a(\cdot)$ and $b(\cdot)$ are upper bounded. Thus, the introduction of skewness does not destroy the existence of the posterior distribution while allowing for more flexibility.

4. Numerical examples

4.1. Simulation study I (Student-t errors, skewed random effects)

In this section we conduct a simulation study, inspired by that presented in Zhang and Davidian (2001), in order to assess the effect of different distributional assumptions on the random effects and the residual errors. We study the model:

$$y_{ij} = t_{ij}\beta_1 + w_i\beta_2 + u_i + \varepsilon_{ij}, \tag{13}$$

where $i = 1, \dots, r = 100$, $j = 1, \dots, n_i = 5$, $t_{ij} = j - 3$, $w_i = I(i \leq 50)$, $\beta_1 = 2$, $\beta_2 = 1$. According to Zhang and Davidian (2001), $r = 100$ represents a case where the amount of information available to estimate both the fixed effects and the random effects is modest. Four scenarios are considered for the distribution of the residual errors ϵ_{ij} and the random effects u_i. For the first scenario we simulate from $\epsilon_{ij} \sim \text{t}(0, 0.5, 2)$ and $u_i \sim \text{TPN}(-1.5, 0.5, 0.5)$; where $\text{t}(0, 0.5, 2)$ represents a Student-t distribution with location parameter 0, scale parameter 0.5 and 2 degrees of freedom, and $\text{TPN}(-1.5, 0.5, 0.5)$ represents a two–piece normal with location parameter -1.5, scale parameter 0.5 and skewness parameter 0.5 using the epsilon-skew parameterisation. For the second scenario we use $\epsilon_{ij} \sim \text{N}(0, 0.5)$ and $u_i \sim \text{TPN}(-1.5, 0.5, 0.5)$. The third scenario consists of

$\epsilon_{ij} \sim \mathrm{t}(0, 0.5, 2)$ and $u_i \sim \mathrm{N}(-1.5, 0.5)$. The fourth scenario uses $\epsilon_{ij} \sim \mathrm{N}(0, 0.5)$ and $u_i \sim \mathrm{N}(-1.5, 0.5)$. We simulate 100 data sets under these configurations. For each of these simulated samples, the model (13) was fitted assuming that $\epsilon_{ij} \sim t(0, \sigma_\varepsilon, \delta_\varepsilon)$ and $u_i \sim \mathrm{TPN}(\mu, \sigma, \gamma)$ with the prior structure:

$$\pi(\beta_1, \beta_2, \sigma_\varepsilon, \delta_\varepsilon, \mu, \sigma, \gamma) \propto \frac{\pi(\delta_\varepsilon)\pi(\mu)\pi(\sigma)\pi(\gamma)}{\sigma_\varepsilon}, \tag{14}$$

where $\pi(\delta_\varepsilon)$ is the weakly informative priors for the degrees of freedom of a Student-t distribution proposed in Rubio and Steel (2015), $\pi(\mu)$ is a uniform prior on $[-100, 100]$, $\pi(\sigma)$ is a half-Cauchy density with mode 0 and unit scale parameter, which has been shown to induce a posterior with good frequentist properties in the context of Bayesian hierarchical models (Polson and Scott, 2012). Finally, $\pi(\gamma)$ is a uniform prior on $(-1, 1)$, which represents a weakly informative prior as described in Rubio and Steel (2014). The prior in (14) is consistent with prior (6) and ensures a proper $p(u_i)$. The propriety of the posterior is then guaranteed by Theorem 1. For each of the 100 simulated sets we obtain a posterior sample of size $1,000$ using an adaptive Metropolis within Gibbs algorithm (Roberts and Rosenthal, 2009) using the 'spBayes' R package (Finley et al., 2007). The posterior samples are obtained after a burn-in period of 7500 iterations and thinned every 10 iterations in order to reduce correlation ($17,500$ draws in total). Table 1 presents summary statistics of the posterior samples, the median (over the datasets) Bayes factors in favour of $\gamma = 0$ (symmetric random effects), calculated using the Savage-Dickey density ratio, and the average odds $p/(1-p)$, where $p = \mathbb{P}(\delta_\varepsilon > 10)$, which provides information about the appropriateness of the assumption of normal tails for the residual errors. The Savage-Dickey density ratio (Verdinelli and Wasserman, 1995) is calculated as the ratio of the marginal posterior density function and the prior density of the parameter under consideration (γ) evaluated at the point of interest ($\gamma = 0$). In order to evaluate the posterior density, we employ a kernel density estimator (with Gaussian kernel) of the posterior samples. The use of the (standard) Savage-Dickey density ratio is valid in our setting as we are using independent proper priors for the parameters that differ across models (see Verdinelli and Wasserman, 1995). Posterior medians are presented instead of posterior means as the existence of the posterior mean is not guaranteed in all the scenarios. From this table, we can observe that despite the relatively small sample size, these Bayesian model selection criteria correctly identify the true model in each scenario.

In order to assess the impact of the assumption of normality, we implement the model with normal residual errors and random effects. Table 2 summarizes posterior inference for this model. The misspecification of the distribution of the residual errors leads to an overestimation of σ_ε, which is in line with the fact that this scale parameter has a different interpretation in normal and Student-t cases, while the presence of unmodelled skewness affects the estimation of the location parameter μ. Note also that misspecification dramatically increases the uncertainty about the fixed effects.

Table 1

Simulation study I: Posterior medians and 95% credible intervals of the median estimators using the general model. The median odds are calculated after removing the cases where all of the drawn values are above 10.

Parameter	Scenario 1 $\varepsilon_{ij} \sim t$ $u_i \sim$ TPN	Scenario 2 $\varepsilon_{ij} \sim$ N $u_i \sim$ TPN	Scenario 3 $\varepsilon_{ij} \sim t$ $u_i \sim$ N	Scenario 4 $\varepsilon_{ij} \sim$ N $u_i \sim$ N
β_1	1.99 (1.96,2.04)	2.00 (1.98,2.03)	2.00 (1.96, 2.03)	2.00 (1.98,2.03)
β_2	0.99 (0.74,1.23)	0.99 (0.80,1.20)	1.01 (0.75, 1.21)	1.00 (0.76,1.19)
σ_ε	0.50 (0.42,0.57)	0.48 (0.44,0.52)	0.50 (0.42,0.57)	0.48 (0.44,0.52)
δ_ε	1.97 (1.61,2.71)	31.54 (13.64, 203.5)	1.96 (1.60,2.70)	30.48 (13.73,203.4)
μ	-1.50 (-1.97, -1.20)	-1.50 (-1.92,-1.22)	-1.47 (-1.91,-1.04)	-1.50 (-1.95,-1.11)
σ	0.49 (0.38,0.60)	0.49(0.41,0.57)	0.48 (0.37,0.59)	0.49 (0.41,0.59)
γ	0.50 (0.00,0.79)	0.54 (-0.04,0.79)	0.00 (-0.51,0.58)	0.01 (-0.52,0.41)
Median BF $\gamma = 0$	0.78	0.43	1.98	2.50
Median odds $\delta_\varepsilon > 10$	0	17.54	0	19.02

Table 2

Simulation study I: Monte Carlo medians and 95% credible intervals of the median estimators using the normal model

Parameter	Scenario 1 $\varepsilon_{ij} \sim t$ $u_i \sim$ TPN	Scenario 2 $\varepsilon_{ij} \sim$ N $u_i \sim$ TPN	Scenario 3 $\varepsilon_{ij} \sim t$ $u_i \sim$ N	Scenario 4 $\varepsilon_{ij} \sim$ N $u_i \sim$ N
β_1	2.00 (1.89,2.09)	2.00 (1.98, 2.03)	2.00 (1.89, 2.10)	2.00 (1.97,2.02)
β_2	0.99 (-0.05,2.24)	1.07 (0.13,2.39)	1.08 (-0.05,2.05)	1.01 (0.78,1.20)
σ_ε	1.53 (1.01,3.05)	0.50 (0.47,0.53)	1.53 (1.01,3.05)	0.50 (0.46,0.53)
μ	-3.41 (-4.21,-2.82)	-3.51 (-4.28,-2.97)	-1.53 (-2.18,-0.92)	-1.50 (-1.65,-1.35)
σ	2.57 (2.18,3.00)	5.27 (2.24,2.94)	2.44 (2.07,2.86)	0.50 (0.42,0.59)

4.2. Simulation study II (Student-t errors and random effects with censoring)

In this section, we present a simulation study with censored responses. We consider a model which is used in practice for modelling the evolution of a marker in longitudinal studies (see *e.g.* Vaida and Liu, 2009), in particular

$$y_{ij} = \beta_j + u_i + \varepsilon_{ij},$$

where ε_{ij} and u_i are mutually independent, $i = 1, \dots, 100$, and $j = 1, \dots, 5$. We generate data using two scenarios: (I) $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0, 0.5)$, and $u_i \overset{i.i.d.}{\sim} N(0, 0.25)$; and (II) $\varepsilon_{ij} \overset{i.i.d.}{\sim} t(0, 0.5, 2)$, and $u_i \overset{i.i.d.}{\sim} t(0, 0.25, 2)$. The theoretical values of the regression parameters are $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (2.5, 3.0, 3.5, 4.0, 4.5)$. We simulate 100 data sets under these two configurations and truncate the values that are greater than $\log_{10}(75000)$. This truncation strategy produces samples with 7%–18% censored observations. Then, we fit the following four models:

Table 3
Simulation study II with Scenario I: Posterior medians and 95% credible intervals of the median estimators.

Parameter	Model 1*	Model 2	Model 3	Model 4
β_1	2.49 (2.37,2.61)	2.50 (2.37,2.61)	2.50 (2.38,2.61)	2.49 (2.37,2.61)
β_2	2.99 (2.90,3.08)	2.99 (2.91,3.08)	2.99 (2.91,3.07)	2.98 (2.90,3.07)
β_3	3.51 (3.39,3.60)	3.51 (3.39,3.60)	3.51 (3.39,3.60)	3.51 (3.39,3.60)
β_4	3.99 (3.90,4.14)	3.99 (3.90,4.14)	3.99 (3.89,4.14)	4.00 (3.90,4.14)
β_5	4.49 (4.40,4.62)	4.50 (4.41,4.63)	4.49 (4.40,4.63)	4.49 (4.40,4.62)
σ_ε	0.50 (0.47,0.54)	0.49 (0.46,0.53)	0.50 (0.47,0.54)	0.48 (0.44,0.53)
δ_ε	–	59.51 (29.15,201.12)	–	32.73 (11.28,212.87)
σ	0.25 (0.20,0.31)	0.25 (0.20,0.31)	0.22 (0.15,0.28)	0.22 (0.16,0.28)
δ	–	–	11.33 (4.64,97.28)	10.77 (4.41, 18.78)

Table 4
Simulation study II with Scenario II: Posterior medians and 95% credible intervals of the median estimators.

Parameter	Model 1	Model 2	Model 3	Model 4*
β_1	2.45 (2.06,2.67)	2.49 (2.33,2.68)	2.49 (2.18,2.69)	2.50 (2.35,2.66)
β_2	2.96 (2.55,3.26)	3.00 (2.83,3.22)	3.00 (2.65,3.34)	3.00 (2.85,3.20)
β_3	3.43 (3.08,3.70)	3.48 (3.31,3.68)	3.47 (3.22,3.70)	3.49 (3.34,3.65)
β_4	3.98 (3.68,4.21)	4.00 (3.80,4.20)	4.00 (3.71,4.24)	4.01 (3.82,4.20)
β_5	4.54 (4.18,4.87)	4.50 (4.31,4.67)	4.57(4.20,4.92)	4.50 (4.33,4.64)
σ_ε	1.10 (0.87,2.64)	0.50 (0.41,0.58)	1.09 (0.87,2.47)	0.50 (0.43,0.58)
δ_ε	–	1.91 (1.33,2.84)	–	1.96 (1.52,2.86)
σ	0.52 (0.26,1.52)	0.48 (0.34,0.76)	0.20 (0.06,0.40)	0.28 (0.18,0.41)
δ	–	–	1.65 (0.97,5.07)	2.17 (1.31,6.55)

Model 1, $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0,\sigma_\varepsilon)$, and $u_i \overset{i.i.d.}{\sim} N(0,\sigma)$, Model 2, $\varepsilon_{ij} \overset{i.i.d.}{\sim} t(0,\sigma_\varepsilon,\delta_\varepsilon)$, and $u_i \overset{i.i.d.}{\sim} N(0,\sigma)$, Model 3, $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0,\sigma_\varepsilon)$, and $u_i \overset{i.i.d.}{\sim} t(0,\sigma,\delta)$ and Model 4, $\varepsilon_{ij} \overset{i.i.d.}{\sim} t(0,\sigma_\varepsilon,\delta_\varepsilon)$, $u_i \overset{i.i.d.}{\sim} t(0,\sigma,\delta)$. For the most general model (Model 4) we adopt the prior structure:

$$\pi(\boldsymbol{\beta},\sigma_\varepsilon,\delta_\varepsilon,\sigma,\delta) \propto \frac{\pi(\delta_\varepsilon)\pi(\sigma)\pi(\delta)}{\sigma_\varepsilon}, \tag{15}$$

where $\pi(\delta_\varepsilon)$ and $\pi(\delta)$ are the weakly informative priors for the degrees of freedom of a Student-t distribution proposed in Rubio and Steel (2015), $\pi(\sigma)$ is a half-Cauchy density with mode 0 and unit scale parameter. For Models 1-3 we use the obvious reduction of this prior. The propriety of the posterior is then guaranteed by Theorem 2.

Tables 3 and 4 (with a star indicating the model used to generate the data) give a summary of the posterior results.

Table 3 suggests that using a more complex model than necessary does not affect the inference on the parameters in the simple model (Model 1), which is encouraging, as we do not really seem to lose anything by allowing for flexibility. We merely find out that the assumed Student-t distributions have large degrees of freedom parameters, which makes them close to normal.

From Table 4 we conclude that wrongly assuming normality of the distributions tends to make inference less precise, especially if we get the tails of the residual errors wrong. It also seems that the (more latent) tail parameter of the random effects is harder to estimate than that of the ε_{ij}'s.

4.3. Framingham heart study

We now illustrate the performance of the proposed models with real data. We use the data set reported in Zhang and Davidian (2001), which consists of measurements of the cholesterol level for 200 randomly selected individuals from the Framingham heart study. The measurements were taken at the beginning of the study and then every 2 years for 10 years. We are interested in the relationship between the cholesterol level and the age (at baseline) and gender of the patients. Zhang and Davidian (2001) model this relationship through the LMM:

$$y_{ij} = \beta_1 + \beta_2 \text{age}_i + \beta_3 \text{sex}_i + \beta_4 t_{ij} + u_{1i} + u_{2i} t_{ij} + \varepsilon_{ij}, \tag{16}$$

where y_{ij} represents the cholesterol level divided by 100 at the jth time for subject i and t_{ij} is $(\text{time} - 5)/10$, with time measured in years from baseline. Zhang and Davidian (2001) assume normal residual errors ε_{ij} and the density of the random effects $\mathbf{u}_i = (u_{1i}, u_{2i})^\top$ is represented by a semi-nonparametric truncated series expansion.

Model (16) is simply the LMM in (1) with $p = 4$ and $q = 2$. Here we adopt the following hierarchical structure:

$$\begin{aligned} \varepsilon_{ij} \mid \sigma_\varepsilon, \delta_\varepsilon &\overset{i.i.d.}{\sim} \text{t}(0, \sigma_\varepsilon, \delta_\varepsilon), \\ \mathbf{u}_i \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \rho &\sim \text{GC}(F_1, F_2 \mid \rho), \end{aligned}$$

where GC denotes a bivariate Gaussian copula with marginals F_1 and F_2, and correlation parameter ρ. For these marginal distributions we use a two-piece sinh-arcsinh distribution with the epsilon–skew parameterisation (Rubio et al., 2016; Rubio and Steel, 2015) which contains a scale parameter $\sigma_i > 0$, a skewness parameter $\gamma_i \in (-1, 1)$, and a kurtosis parameter $\delta_i > 0$, $i = 1, 2$. We denote this model as Model 1. We adopt the prior structure:

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon, \delta_\varepsilon, \sigma_1, \sigma_2, \gamma_1, \gamma_2, \delta_1, \delta_2, \rho) \sim \frac{\pi(\delta_\varepsilon)\pi(\rho)}{\sigma_\varepsilon} \prod_{i=1}^{2} \pi(\sigma_i)\pi(\gamma_i)\pi(\delta_i). \tag{17}$$

As a weakly informative prior for the degrees of freedom of the Student-t distribution, δ_ε we employ the prior proposed in Rubio and Steel (2015). For each of the scale parameters (σ_1, σ_2) we adopt a half-Cauchy prior (Polson and Scott, 2012). The shape parameters (δ_1, δ_2) are assigned the prior proposed in Rubio and Steel (2015) for a general class of kurtosis parameters. For each of the skewness parameters (γ_1, γ_2) we adopt a uniform distribution on $(-1, 1)$. In a

Table 5

Framingham data: Bayes factors (against the general Model 1) for several hypotheses.

Model	$\gamma_1 = 0$	$\gamma_2 = 0$	$\delta_1 = 1$	$\delta_2 = 1$	$\rho = 0$	$(\delta_1, \delta_2) = (1,1)$	$(\gamma_2, \delta_1, \delta_2) = (0,1,1)$
BF	0.04	0.65	0.77	1.04	0.02	0.63	0.42

bivariate Gaussian copula the Spearman measure of association, $r_\rho \in (-1, 1)$, can be calculated in closed form as

$$r_\rho = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right),$$

for which we assume a uniform prior $r_\rho \sim U(-1, 1)$, inducing the following prior density on ρ

$$\pi(\rho) \propto \frac{1}{\sqrt{1 - (\rho/2)^2}}.$$

The propriety of the corresponding posterior distribution is guaranteed by Theorem 1. We also consider the following submodels: Model 2 $(\delta_1, \delta_2) = (1, 1)$ (two–piece normal marginal random effects); Model 3 $(\delta_1, \delta_2, \gamma_2) = (1, 1, 0)$ (one marginally normal random effect and one marginally two-piece normal random effect); Model 4 $(\delta_1, \delta_2, \gamma_1, \gamma_2) = (1, 1, 0, 0)$ (normal random effects); and Model 5 $(\delta_1, \delta_2, \gamma_1, \gamma_2, \delta_\varepsilon) = (1, 1, 0, 0, \infty)$ (normal random effects and normal residual errors). Finally, Model 6 assumes normality of all distributions as well as independence between both random effects. A posterior sample of size $5,000$ was obtained after a burn-in period of $15,000$ iterations and thinning every 25 iterations ($140,000$ simulations in total). Table 6 presents a summary of the posterior results.

The Bayes factors, calculated using the Savage-Dickey density ratio, slightly favour the general model. However, the evidence in favour of this model is not decisive (see Table 5). The Bayes factors associated to Models 4–6 (against Model 1) are not shown in Table 5 since they are virtually zero, but the Bayes factors associated to other sub-models are presented in this table for illustration. Bayes factors clearly indicate support for skewness of u_{1t} and dependence of both random effects, which leaves us with Models 1-3. An argument of parsimony could, then, lead to selecting Model 3. Notice that the intervals for the general model are more spread out (*e.g.* for β_4 in Table 6). This indicates that the sample does not contain enough information to accurately estimate all the parameters of this model. The posterior probability of $\{\delta_\varepsilon > 10\}$ in Models 1 and 3 is approximately 0.005 and for Model 2 is approximately 0.015, which suggests that models with heavier tails than normal are favoured.

To assess the predictive performance of the different models we use the log pseudomarginal likelihood (LPML), which is defined as the sum of the log conditional predictive ordinate (CPO) statistic for each observation. The CPO for subject i is defined as the predictive density of the vector of observations $\mathbf{y}_i$ given the rest of the observations. This is, $\text{CPO}_i = \pi(\mathbf{y}_i \mid \mathbf{y} \backslash \{\mathbf{y}_i\})$, where the

Table 6. *Framingham data: Summary of the posterior samples (medians and 95% credible intervals) and LPML.*

Model	1	2	3	4	5	6
β_1	1.494 (1.218,1.794)	1.649 (1.436,1.973)	1.640 (1.366,1.878)	1.621 (1.336,1.957)	1.610 (1.223,1.939)	1.715 (1.399,1.980)
β_2 (age)	0.016 (0.008,0.024)	0.012 (0.005,0.017)	0.013 (0.007,0.020)	0.017 (0.010,0.024)	0.018 (0.010,0.027)	0.015 (0.009, 0.023)
β_3 (sex)	-0.057 (-0.154,0.044)	-0.059 (-0.162,0.040)	-0.061 (-0.159,0.042)	-0.048 (-0.159,0.074)	-0.060 (-0.172, 0.048)	-0.015 (-0.123,0.094)
β_4 (time)	0.130 (-0.011,0.384)	0.121 (-0.002,0.321)	0.290 (0.243,0.336)	0.289 (0.241,0.337)	0.282 (0.233,0.327)	0.282 (0.235,0.329)
σ_ε	0.163 (.146,0.182)	0.163 (0.147,0.181)	0.164 (0.147,0.182)	0.163 (.146,0.182)	0.209 (0.198,0.221)	0.209 (0.198,0.220)
δ_ε	5.094 (3.513,8.095)	5.116 (3.522,8.153)	5.202 (3.586,8.459)	5.115 (3.524,8.235)	–	–
σ_1	0.275 (0.181,0.412)	0.372 (0.336,0.418)	0.372 (0.335,0.416)	0.384 (0.347,0.428)	0.380 (0.344,0.425)	0.376 (0.340,0.419)
σ_2	0.201 (0.075,0.829)	0.186 (0.133,0.244)	0.203 (0.144,0.257)	0.204 (0.147,0.261)	0.193 (0.125,0.254)	0.193 (0.127,0.252)
γ_1	-0.331 (-0.524,-0.115)	-0.345 (-0.521,-0.145)	-0.339 (-0.524,-0.138)	–	–	–
γ_2	-0.546 (-0.972,0.277)	-0.597 (-0.969,0.086)	–	–	–	–
δ_1	0.820 (0.656,1.057)	–	–	–	–	–
δ_2	1.047 (0.630,3.407)	–	–	–	–	–
ρ	0.425 (0.177,0.640)	0.397 (0.158,0.611)	0.383 (0.137,0.620)	0.397 (0.141,0.638)	0.4243 (0.168,0.686)	–
LPML	-22.81	-25.24	-19.12	-26.75	-30.88	-32.98

predictive density is evaluated at the vector of observations corresponding to the i^{th} subject (see Bandyopadhyay et al., 2012). These quantities are approximated using a harmonic mean estimator as described in Bandyopadhyay et al. (2012). Table 6 shows the LPML for the different models, which favour Model 3. A larger sample would likely provide stronger evidence for model selection.

The analysis in Zhang and Davidian (2001) suggests departures from normality in the distribution of random intercepts, while the estimated random slope distribution can be approximated by a normal density. Our conclusions are in line with the conclusions of the parametric, semi-nonparametric and the nonparametric analyses in Zhang and Davidian (2001), Jara et al. (2008), and Jara et al. (2009). However, the models proposed here also allow us to interpret the meaning of the parameters, separating those controlling the shape of the distribution from those controlling the dependence between the random effects.

4.4. HIV-1 viral load after unstructured treatment interruption

Here we revisit the data set analysed in Vaida and Liu (2009), which concerns the study of 72 perinatally HIV-infected children and adolescents. Some of the subjects present unstructured treatment interruption, which is a common phenomenon among HIV-infected people, due mainly to treatment fatigue (Vaida and Liu, 2009). The number of observations from baseline (month 0) to month 24 ranges from 13 to 71. Out of 362 observations, 26 observations (7%) were below the detection limits and were censored at these values. Vaida and Liu (2009) proposed the model:

$$y_{ij} = \beta_j + u_i + \varepsilon_{ij},$$

where y_{ij} is the $\log_{10}$ HIV RNA (HIV Ribonucleic acid, which is used to monitor the status of HIV) for subject i at time t_j, using $t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $t_4 = 6$, $t_5 = 9$, $t_6 = 12$, $t_7 = 18$ and $t_8 = 24$ months.

We allow for Student-t tails in the random effects and the errors (with zero locations), and thus use Model 4 in Subsection 4.2 with the same prior structure as in (15). We also try the simpler models 1-3 as defined in Subsection 4.2.

The propriety of the posterior is ensured by Theorem 2.

Table 7 reveals clear evidence for fat tails in both distributions, especially the residual errors. The predictive criterion LPML also favours the general model and particularly penalizes model with normal residual errors. Clearly, ignoring the heavy tails affects the inference on the fixed effects, especially for β_1.

Finally, the predictive performance as measured by LPML is substantially better for our models 4 and 2 than for the best model in Bandyopadhyay et al. (2012), who obtain LPML $= -366.27$ on the same data using the Skew-contaminated normal model of Lachos et al. (2010), which has normal tails for both random effects and error distributions. Also, their posterior distribution of β_1 is substantially shifted towards smaller values (consistent with our models 1 and 3, which impose normal errors).

Table 7
HIV data: Summary of the posterior samples (sample median and 95% credible intervals) and LPML.

Model	1	2	3	4
β_1	3.62 (3.36,3.87)	4.00 (3.76,4.24)	3.70 (3.45,3.93)	4.16 (3.95,4.38)
β_2	4.19 (3.93,4.43)	4.22 (4.00,4.45)	4.26 (4.02,4.50)	4.37 (4.18,4.56)
β_3	4.26 (4.00,4.52)	4.26 (4.04,4.48)	4.34 (4.09,4.59)	4.41 (4.22,4.60)
β_4	4.38 (4.12,4.64)	4.43 (4.20,4.65)	4.46 (4.21,4.70)	4.58 (4.38,4.77)
β_5	4.61 (4.33,4.89)	4.53 (4.31,4.76)	4.68 (4.41,4.96)	4.68 (4.48,4.87)
β_6	4.60 (4.30,4.89)	4.51 (4.28,4.74)	4.67 (4.38,4.95)	4.65 (4.45,4.85)
β_7	4.70 (4.36,5.02)	4.70 (4.45,4.95)	4.77 (4.46,5.08)	4.84 (4.61,5.06)
β_8	4.81 (4.40,5.22)	4.73 (4.44,5.03)	4.88 (4.47,5.28)	4.88 (4.62,5.14)
σ_ε	0.60 (0.55,0.66)	0.20 (0.16,0.25)	0.61 (0.56,0.66)	0.20 (0.16,0.24)
δ_ε	–	1.27 (0.97,1.66)	–	1.30 (1.01,1.69)
σ	0.90 (0.75,1.10)	0.88 (0.72,1.07)	0.66 (0.46,0.90)	0.56 (0.37,0.79)
δ	–	–	3.86 (1.65,16.81)	2.48 (1.24,7.71)
LPML	-397.4	-291.4	-401.5	-287.6

5. Discussion

This paper focuses on hierarchical linear mixed models which are used frequently for both longitudinal and survival modelling in a number of application areas, such as medicine, epidemiology and economics. These models are often based on simple distributional assumptions, such as normality for both the residual errors and the random effects. There is ample evidence in the literature that such assumptions can seriously affect inference on the random effects and the predictive distribution (see *e.g.* Lee and Thompson, 2008; Zhang and Davidian, 2001) and we add to that evidence in our analysis of simulated data. Misspecification of the distribution of the residual errors and the random effects tends to have a substantial effect on the predictive distribution, which explicitly depends on the aforementioned distributions. The use of normal assumptions when the true generating model is not normal also has a marked effect on the estimation of the variance of the residual errors and the random effects, as well as a decrease in the precision of the estimation of the fixed effects.

We consider Bayesian inference for these models with flexible distributions and sensible prior structures. These priors are intended to convey a lack of strong prior information; as they are based on a combination of formal arguments (such as invariance) and pragmatic common sense, they end up being improper. Our results provide a formal justification for Bayesian inference in these wide classes of models: our models accommodate any proper distribution for the random effects (which could even be dependent, if required) combined with two-piece scale mixtures of normal distributions for the residual errors, as well as potential censoring. Our results thus cover all random effects distributions (both parametric and nonparametric) that were proposed in the literature except for those that are dependent on the residual errors (such as those in Lachos et al., 2010, which where used in Bandyopadhyay et al., 2012). The conditions for the propriety of the posterior distribution are rather mild and easy to check in practice, especially compared to those in previous papers (Hobert and Casella, 1996; Sun et al., 2001; Rubio, 2015), which obtain sets of conditions that combine the sample size, the number of clusters, the design matrix, and the value of the hyperparameters of the priors on the random effects. This also suggests that most of the posterior impropriety issues come from assuming improper priors on the "deeper" parameters of the random effects.

The analysis of simulated datasets leads to quite promising results, which suggest that priors are well-calibrated and sensible, and Bayes factors seem to provide reliable indications. In addition, Appendix C in the Supplementary Material presents results for the case where both the residual errors and the random effects display skewness, which corroborate the results presented in the main text.

Inference on both real datasets strongly indicates departures from normality in that the residual error distributions have much fatter tails than normal. In addition, the Framingham heart data support dependence between the random effect components with normal random slope and skew-normal random intercept. The HIV data present clear evidence for heavy-tailed random effects. By using simple Student-t distributions on both errors and random effects, we obtain considerably better predictive results on the HIV data than those presented by Bandyopadhyay et al. (2012), using their preferred model. We remind the reader that the distributions considered in Bandyopadhyay et al. (2012) are multivariate scale mixtures of skew-normal distributions, so perhaps might be affected by the kind of issues we describe for multivariate scale mixtures of normals in Section 2. For the real data, evidence based on Bayes factors points in the same direction as the predictive performance measured by LPML, which is reassuring.

Here we have employed the R package 'spBayes' in all of our numerical examples, mainly for reasons of efficiency and simplicity. However, the user can also implement the models proposed here in any other "general purpose" MCMC software such as Stan (gradient-based MCMC), PROC MCMC (from SAS, which implements a random walk Metropolis), or any package which includes a Metropolis-within-Gibbs sampler.

Appendix A: Proofs

Proof of Remark 1

The marginal density of the data is given by

$$\begin{aligned}\tilde{m}(\mathbf{y}) &= \int p(\mathbf{y} \mid \mathbf{u}, \boldsymbol{\beta}, \sigma_\varepsilon, \delta_\varepsilon) p(\mathbf{u}) \pi(\boldsymbol{\beta}) \frac{1}{\sigma_\varepsilon^{b+1}} \pi(\delta_\varepsilon)\, d\mathbf{u} d\boldsymbol{\beta} d\sigma_\varepsilon d\delta_\varepsilon \\ &= \int \frac{\tau^{n/2}}{\sigma_\varepsilon^n} \prod_{i=1}^{r} \prod_{j=1}^{n_i} \phi\left[\frac{\tau^{\frac{1}{2}}}{\sigma_\varepsilon}(y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta} - \mathbf{z}_{ij}^\top \mathbf{u}_i)\right] \\ &\times \quad p(\mathbf{u}) \pi(\boldsymbol{\beta}) \frac{1}{\sigma_\varepsilon^{b+1}} \pi(\delta_\varepsilon) d\mathbf{u} d\boldsymbol{\beta} d\sigma_\varepsilon dH_\varepsilon(\tau \mid \delta_\varepsilon) d\delta_\varepsilon.\end{aligned}$$

Consider the change of variable $\tilde{\sigma}_\varepsilon = \sigma_\varepsilon / \tau^{\frac{1}{2}}$, with corresponding Jacobian $|\mathcal{J}| = \tau^{\frac{1}{2}}$. Then,

$$\begin{aligned}\tilde{m}(\mathbf{y}) &= \int \frac{1}{\tilde{\sigma}_\varepsilon^n} \prod_{i=1}^{r} \prod_{j=1}^{n_i} \phi\left[\frac{1}{\tilde{\sigma}_\varepsilon}(y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta} - \mathbf{z}_{ij}^\top \mathbf{u}_i)\right] \\ &\times \quad p(\mathbf{u}) \frac{1}{\tilde{\sigma}_\varepsilon^{b+1}} \pi(\boldsymbol{\beta}) \pi(\delta_\varepsilon) \tau^{-\frac{b}{2}} d\mathbf{u} d\boldsymbol{\beta} d\tilde{\sigma}_\varepsilon dH_\varepsilon(\tau \mid \delta_\varepsilon) d\delta_\varepsilon \\ &= m(\mathbf{y}) \int_\Delta \int_{\mathbb{R}_+} \tau^{-\frac{b}{2}} \pi(\delta_\varepsilon) dH_\varepsilon(\tau \mid \delta_\varepsilon) d\delta_\varepsilon,\end{aligned}$$

where we can identify the first factor as the marginal density of the data associated to the model with normal distributional assumptions on $\boldsymbol{\varepsilon}$.

Proof of Theorem 1

For practical purposes we will work with σ_ε^2 instead of σ_ε. The marginal density of the data is given by

$$\begin{aligned}\tilde{m}(\mathbf{y}) &= \int p(\mathbf{y} \mid \mathbf{u}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \delta_\varepsilon) p(\mathbf{u}) \pi(\boldsymbol{\beta}) \frac{1}{(\sigma_\varepsilon^2)^{\frac{b}{2}+1}} \pi(\delta_\varepsilon)\, d\mathbf{u} d\boldsymbol{\beta} d\sigma_\varepsilon^2 d\delta_\varepsilon \\ &= \int \prod_{i=1}^{r} \prod_{j=1}^{n_i} \frac{\tau_{ij}^{\frac{1}{2}}}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{\tau_{ij}}{2\sigma_\varepsilon^2}(y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta} - \mathbf{z}_{ij}^\top \mathbf{u}_i)^2\right] \\ &\times \quad p(\mathbf{u}) \pi(\boldsymbol{\beta}) \frac{1}{(\sigma_\varepsilon^2)^{\frac{b}{2}+1}} \pi(\delta_\varepsilon) d\mathbf{u} d\boldsymbol{\beta} d\sigma_\varepsilon^2 \left[\prod_{i=1}^{r} \prod_{j=1}^{n_i} dH_\varepsilon(\tau_{ij} \mid \delta_\varepsilon)\right] d\delta_\varepsilon. \quad (18)\end{aligned}$$

Denote $\tau_+ = \max_{i,j} \tau_{ij}$ and $\tau_- = \min_{i,j} \tau_{ij}$. Then, we can obtain the following lower bound

$$\tilde{m}(\mathbf{y}) \geq \int \prod_{i=1}^{r} \prod_{j=1}^{n_i} \frac{\tau_-^{\frac{1}{2}}}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{\tau_+}{2\sigma_\varepsilon^2}(y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta} - \mathbf{z}_{ij}^\top \mathbf{u}_i)^2\right]$$

$$\times \quad p(\mathbf{u})\pi(\boldsymbol{\beta})\frac{1}{(\sigma_\varepsilon^2)^{\frac{b}{2}+1}}\pi(\delta_\varepsilon)d\mathbf{u}d\boldsymbol{\beta}d\sigma_\varepsilon^2\left[\prod_{i=1}^{r}\prod_{j=1}^{n_i}dH_\varepsilon(\tau_{ij}\mid\delta_\varepsilon)\right]d\delta_\varepsilon.$$

Consider the change of variable $\tilde{\sigma}_\varepsilon^2 = \dfrac{\sigma_\varepsilon^2}{\tau_+}$

$$\begin{aligned}
\tilde{m}(\mathbf{y}) &\geq \int\prod_{i=1}^{r}\prod_{j=1}^{n_i}\frac{\tau_-^{\frac{1}{2}}}{\sqrt{2\pi\tilde{\sigma}_\varepsilon^2\tau_+}}\exp\left[-\frac{1}{2\tilde{\sigma}_\varepsilon^2}(y_{ij}-\mathbf{x}_{ij}^\top\boldsymbol{\beta}-\mathbf{z}_{ij}^\top\mathbf{u}_i)^2\right] \\
&\times \quad \tau_+p(\mathbf{u})\pi(\boldsymbol{\beta})\frac{1}{(\sigma_\varepsilon^2\tau_+)^{\frac{b}{2}+1}}\pi(\delta_\varepsilon)d\mathbf{u}d\boldsymbol{\beta}d\sigma_\varepsilon^2\left[\prod_{i=1}^{r}\prod_{j=1}^{n_i}dH_\varepsilon(\tau_{ij}\mid\delta_\varepsilon)\right]d\delta_\varepsilon \\
&= \quad m(\mathbf{y})\int\left(\frac{\tau_-^{n/2}}{\tau_+^{n/2}}\right)\tau_+^{-\frac{b}{2}}\pi(\delta_\varepsilon)\prod_{i=1}^{r}\prod_{j=1}^{n_i}dH_\varepsilon(\tau_{ij}\mid\delta_\varepsilon)d\delta_\varepsilon,
\end{aligned}$$

Using the latter expression together with Theorem 1 from Fernández et al. (1997) it follows that (a) is a necessary condition for the propriety of the posterior.

Now, define $\mathcal{T} = \operatorname{diag}(\tau_1,\dots,\tau_n)$ (the diagonal matrix of mixing variables), $\tilde{\mathbf{y}} = \mathcal{T}^{\frac{1}{2}}\mathbf{y}$, $\tilde{\mathbf{X}} = \mathcal{T}^{\frac{1}{2}}\mathbf{X}$, and $\tilde{\mathbf{Z}} = \mathcal{T}^{\frac{1}{2}}\mathbf{Z}$, where $\mathcal{T}^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\tau_1},\dots,\sqrt{\tau_n})$. Then, by using that $\pi(\boldsymbol{\beta}) \leq K$, we can obtain the following upper bound for (18)

$$\begin{aligned}
\tilde{m}(\mathbf{y}) &\leq K\int\frac{\det(\mathcal{T})^{\frac{1}{2}}}{(2\pi\sigma_\varepsilon^2)^{\frac{n}{2}}}\exp\left[-\frac{1}{2\sigma_\varepsilon^2}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}-\tilde{\mathbf{Z}}\mathbf{u})^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}-\tilde{\mathbf{Z}}\mathbf{u})\right] \\
&\times \quad p(\mathbf{u})\frac{1}{(\sigma_\varepsilon^2)^{\frac{b}{2}+1}}\pi(\delta_\varepsilon)d\mathbf{u}d\boldsymbol{\beta}d\sigma_\varepsilon^2\left[\prod_{i=1}^{r}\prod_{j=1}^{n_i}dH_\varepsilon(\tau_{ij}\mid\delta_\varepsilon)\right]d\delta_\varepsilon.
\end{aligned}$$

Following the proof of Theorem 1 in Fernández et al. (1997) we have

$$(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}-\tilde{\mathbf{Z}}\mathbf{u})^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}-\tilde{\mathbf{Z}}\mathbf{u}) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}}),$$

where $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{Z}}\mathbf{u})$ and $c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}}) = (\tilde{\mathbf{y}}-\tilde{\mathbf{Z}}\mathbf{u})^\top\mathbf{M}_{\tilde{\mathbf{X}}}(\tilde{\mathbf{y}}-\tilde{\mathbf{Z}}\mathbf{u})$, with $\mathbf{M}_{\tilde{\mathbf{X}}} = \mathbf{I}_n - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top$. Then, we get

$$\begin{aligned}
\tilde{m}(\mathbf{y}) &\leq K\int\frac{\det(\mathcal{T})^{\frac{1}{2}}}{(2\pi\sigma_\varepsilon^2)^{\frac{n}{2}}}\exp\left[-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\exp\left[-\frac{c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})}{2\sigma_\varepsilon^2}\right] \\
&\times \quad p(\mathbf{u})\frac{1}{(\sigma_\varepsilon^2)^{\frac{b}{2}+1}}\pi(\delta_\varepsilon)d\mathbf{u}d\boldsymbol{\beta}d\sigma_\varepsilon^2\left[\prod_{i=1}^{r}\prod_{j=1}^{n_i}dH_\varepsilon(\tau_{ij}\mid\delta_\varepsilon)\right]d\delta_\varepsilon.
\end{aligned}$$

By integrating $\boldsymbol{\beta}$ out we obtain

$$\tilde{m}(\mathbf{y}) \quad\leq\quad K\int\frac{\det(\mathcal{T})^{\frac{1}{2}}}{\det(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{\frac{1}{2}}(2\pi\sigma_\varepsilon^2)^{\frac{n-p}{2}}}\exp\left[-\frac{c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})}{2\sigma_\varepsilon^2}\right]$$

$$\times \quad p(\mathbf{u})\frac{1}{(\sigma_\varepsilon^2)^{\frac{b}{2}+1}}\pi(\delta_\varepsilon)d\mathbf{u}d\sigma_\varepsilon^2\left[\prod_{i=1}^{r}\prod_{j=1}^{n_i}dH_\varepsilon(\tau_{ij}\mid\delta_\varepsilon)\right]d\delta_\varepsilon.$$

Now, after integrating σ_ε^2 we obtain

$$\begin{aligned}\tilde{m}(\mathbf{y}) &\leq K_1\int\frac{\det(\mathcal{T})^{\frac{1}{2}}}{\det(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{\frac{1}{2}}}c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})^{-\frac{n+b-p}{2}}p(\mathbf{u})\pi(\delta_\varepsilon)d\mathbf{u}\\ &\times \left[\prod_{i=1}^{r}\prod_{j=1}^{n_i}dH_\varepsilon(\tau_{ij}\mid\delta_\varepsilon)\right]d\delta_\varepsilon, \end{aligned}\tag{19}$$

where $K_1>0$ is a known constant. Note that $\det(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})=\det(\mathbf{X}^\top\mathcal{T}\mathbf{X})$, and

$$\begin{aligned}c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}}) &= (\mathbf{y}-\mathbf{Z}\mathbf{u})^\top\mathcal{T}^{\frac{1}{2}}\mathbf{M}_{\tilde{\mathbf{X}}}\mathcal{T}^{\frac{1}{2}}(\mathbf{y}-\mathbf{Z}\mathbf{u})\\ &= (\mathbf{y}-\mathbf{Z}\mathbf{u})^\top\mathcal{T}(\mathbf{y}-\mathbf{Z}\mathbf{u})+(\mathbf{y}-\mathbf{Z}\mathbf{u})^\top\mathcal{T}\mathbf{X}(\mathbf{X}^\top\mathcal{T}\mathbf{X})^{-1}\mathbf{X}^\top\mathcal{T}(\mathbf{y}-\mathbf{Z}\mathbf{u}).\end{aligned}$$

From Lemma 1 from Fernández and Steel (2000) we know that $\det(\mathbf{X}^\top\mathcal{T}\mathbf{X})$ has upper and lower bounds that are both proportional to $\prod_{i=1}^{p}\tau_{m_i}=\max\{\prod_{i=1}^{p}\tau_{s_i}:1\leq s_1<\cdots<s_p\leq n \text{ and } \det(\mathbf{x}_{s_1},\dots,\mathbf{x}_{s_p})\neq 0\}$, where $\det(\mathbf{x}_{s_1},\dots,\mathbf{x}_{s_p})$ denotes the determinant of the submatrix of X corresponding to the observations $y_{s_1},\dots,y_{s_p}$. Define also $\tau_c=\max\{\tau_i:i\neq m_1,\dots,m_p\}$. Using these results we obtain the following inequality

$$\begin{aligned}\frac{\det(\mathcal{T})^{\frac{1}{2}}}{\det(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{\frac{1}{2}}}c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})^{-\frac{n+b-p}{2}} &\leq K_2\frac{\prod_{j\neq m_1,\dots,m_p}\tau_j^{\frac{1}{2}}}{c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})^{\frac{n+b-p}{2}}}\\ &\leq K_2\frac{\tau_c^{\frac{n-p}{2}}}{c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})^{\frac{n+b-p}{2}}},\end{aligned}$$

where $K_2>0$ is a constant. Now, by defining the $n\times(p+1)$ matrix $\mathbf{L}=(\mathbf{X}:\mathbf{y}-\mathbf{Z}\mathbf{u})$ and subsequently applying the Binet–Cauchy theorem (Fernández and Steel, 2000), we obtain

$$\begin{aligned}c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}}) = \frac{\det(\mathbf{L}^\top\mathcal{T}\mathbf{L})}{\det(\mathbf{X}^\top\mathcal{T}\mathbf{X})} &= \frac{1}{\det(\mathbf{X}^\top\mathcal{T}\mathbf{X})}\\ \times \sum_{1\leq s_1<\cdots<s_{p+1}\leq n}\left(\prod_{j=1}^{p+1}\tau_{s_j}\right)&\det{}^2\begin{pmatrix}\mathbf{x}_{s_1} & \dots & \mathbf{x}_{s_{p+1}}\\ y_{s_1}-\mathbf{z}_{s_1}^\top\mathbf{u} & \dots & y_{s_{p+1}}-\mathbf{z}_{s_{p+1}}^\top\mathbf{u}\end{pmatrix}.\end{aligned}$$

Given that $\mathbf{y}-\mathbf{Z}\mathbf{u}$ is not in the column space of $\mathbf{X}$ we have that $c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})>0$. Consequently

$$c(\tilde{\mathbf{Z}},\tilde{\mathbf{y}})\geq P(\mathbf{u})\tau_c,$$

with

$$P(\mathbf{u})=\min_{1\leq s_1<\cdots<s_{p+1}\leq n}\det{}^2\begin{pmatrix}\mathbf{x}_{s_1} & \dots & \mathbf{x}_{s_{p+1}}\\ y_{s_1}-\mathbf{z}_{s_1}^\top\mathbf{u} & \dots & y_{s_{p+1}}-\mathbf{z}_{s_{p+1}}^\top\mathbf{u}\end{pmatrix},$$

where the minimum is taken over the non–zero determinants. Then,

$$\frac{\det(\mathcal{T})^{\frac{1}{2}}}{\det(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{\frac{1}{2}}} c(\tilde{\mathbf{Z}}, \tilde{\mathbf{y}})^{-\frac{n+b-p}{2}} \quad \leq \quad K_2 \frac{\tau_c^{-\frac{b}{2}}}{P(\mathbf{u})^{\frac{n+b-p}{2}}}.$$

Now, note that $P(\mathbf{u})$, being a quadratic polynomial in q variables, is a continuous coercive function (see Definition 2.4 in Güler, 2010). Then, by Corollary 2.5 in Güler (2010), it follows that $P(\mathbf{u})$ achieves a global minimum. Given that $P(\mathbf{u}) > 0$ by assumption, it follows that there exists $M > 0$ such that $P(\mathbf{u}) > M$ for all $\mathbf{u}$. From this lower bound follows that

$$\frac{\det(\mathcal{T})^{\frac{1}{2}}}{\det(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{\frac{1}{2}}} c(\tilde{\mathbf{Z}}, \tilde{\mathbf{y}})^{-\frac{n+b-p}{2}} \quad \leq \quad K_3 \tau_c^{-\frac{b}{2}}, \tag{20}$$

for some constant $K_3 > 0$. Using this inequality in (19) and splitting the integral into the possible orderings of $\{\tau_1, \dots, \tau_n\}$ we can identify τ_c and the result follows.

Proof of Theorem 2

The result follows by using that the censored observations contribute to the likelihood as factors in $[0, 1]$. Then, we can obtain an upper bound for the marginal likelihood of the data which corresponds, up to a proportionality constant, to the marginal likelihood associated to the uncensored observations. The result then follows from Theorem 1.

Proof of Theorem 3

By condition (d$'$), it follows that inequality (20) is satisfied for each $\mathbf{y} = \mathbf{y}^\star$, where $\mathbf{y}^\star \in \mathcal{E}$. Thus, the marginal likelihood of the data associated to each $\mathbf{y}^\star$ is finite (Theorem 1). Given that $\mathcal{E}$ is compact, we can obtain a finite upper bound for the marginal likelihood of the data by integrating out $\mathbf{y}^\star$ over $\mathcal{E}$, and the result follows.

Proof of Theorem 4

By using the inequality

$$s(y \mid \mu, \sigma, \gamma, \delta) \leq \frac{2M(\gamma)}{\sigma[a(\gamma) + b(\gamma)]} f\left(\frac{y - \mu}{\sigma M(\gamma)} \,\middle|\, \delta\right),$$

and the change of variable $\tilde{\sigma}_\varepsilon = \sigma_\varepsilon M(\gamma_\varepsilon)$, we can obtain an upper bound for the marginal likelihood of the data which is a product of the marginal likelihood of model (1), (4), (8) with prior (6) and the expression in condition (e). Then, the result follows by using the proof of Theorem 1.

Appendix B: Conditions on the mixing distributions

Note that the conditions on the mixing distributions in Corollary 1 and Theorem 1 are related to the existence of their marginal moments of negative order.

Gamma mixing distribution

Consider the mixing distribution $\tau \sim \text{Gamma}\left(\frac{\delta}{2}, \frac{\delta}{2}\right)$ with density f_Γ. This mixing distribution produces a Student-t sampling model. Let $c \geq 0$ and $\delta > 2c$, then we have that

$$I = \int_0^\infty \int_0^\infty \tau^{-c} f_\Gamma(\tau \mid \delta)\pi(\delta) d\tau d\delta = \int_0^\infty \frac{2^c \delta^{-c} \Gamma\left(\frac{\delta}{2} - c\right)}{\Gamma\left(\frac{\delta}{2}\right)} \pi(\delta) d\delta$$

If $c = 0$, it follows that $I < \infty$ given that $\pi(\delta)$ is proper. For $c > 0$, it is necessary to use a proper prior $\pi(\delta)$ with support on $(2c + \epsilon, \infty)$, for any $\epsilon > 0$.

Beta mixing distribution

Consider the mixing distribution $\tau \sim \text{Beta}\left(\frac{\delta}{2}, 1\right)$ with density f_β. Let $c \geq 0$, then

$$I = \int_0^\infty \int_0^\infty \tau^{-c} f_\beta(\tau \mid \delta)\pi(\delta) d\tau d\delta \quad = \quad \int_0^\infty \frac{\delta}{\delta - 2c} \pi(\delta) d\delta.$$

For $c = 0$, $I < \infty$ given that $\pi(\delta)$ is proper. However, for $c > 0$ it is necessary to employ a truncated prior on δ with support on $(2c + \epsilon, \infty)$ for any $\epsilon > 0$.

Birnbaum-Saunders mixing distribution

Consider the mixing distribution $\tau \sim \text{BS}(\delta, \delta)$ with density f_{BS} (Birnbaum and Saunders, 1969). Let $c \geq 0$

$$\begin{aligned} I &= \int_0^\infty \int_0^\infty \tau^{-c} f_{BS}(\tau \mid \delta)\pi(\delta) d\tau d\delta \\ &= \int_0^\infty \frac{e^{\frac{1}{\delta^2}} \delta^{c-1} \left(K_{c-\frac{1}{2}}\left(\frac{1}{\delta^2}\right) + K_{c+\frac{1}{2}}\left(\frac{1}{\delta^2}\right) \right)}{\sqrt{2\pi}} \pi(\delta) d\delta. \end{aligned}$$

where $K_n(z)$ represents the modified Bessel function of the second kind. For $c = 0$ it follows that $I < \infty$, however, for $c > 0$ this condition may rule out the use of certain heavy-tailed priors.

TABLE 8
Simulation study III: Monte Carlo median, 95% credible intervals of the median estimators using the two-piece normal model. BF are Bayes factors in favour of symmetry.

Parameter	Scenario 1 $\varepsilon_{ij} \sim$ N $u_i \sim$ N	Scenario 2 $\varepsilon_{ij} \sim$ TPN $u_i \sim$ N	Scenario 3 $\varepsilon_{ij} \sim$ N $u_i \sim$ TPN	Scenario 4 $\varepsilon_{ij} \sim$ TPN $u_i \sim$ TPN
β_1	2.00 (1.97,2.03)	2.00 (1.97,2.02)	2.00 (1.97,2.03)	2.00 (1.97,2.03)
β_2	0.99 (0.77,1.17)	0.98 (0.79,1.18)	0.99 (0.79,1.18)	0.97 (0.77,1.20)
σ_ε	0.50 (0.46,0.53)	0.50 (0.46,0.53)	0.50 (0.46,0.54)	0.50 (0.46,0.53)
γ_ε	0.02 (-0.15,0.21)	0.53 (0.35,0.78)	0.01 (-0.16,0.20)	0.52 (0.35,0.77)
μ	-1.51 (-1.88,-1.06)	-1.47 (-1.92,-1.13)	-1.46 (-1.84,-1.12)	-1.44 (-1.79,-1.14)
σ	0.48 (0.41,0.59)	0.49 (0.41,0.57)	0.49 (0.43,0.56)	0.49 (0.43,0.57)
γ	-0.04 (-0.43,0.45)	-0.02 (-0.57,0.33)	0.51 (0.10,0.82)	0.52 (0.09,0.763)
Median BF $\gamma_\varepsilon = 0$	7.13	2×10^{-21}	7.04	1×10^{-19}
Median BF $\gamma = 0$	2.45	2.81	0.40	0.29

Appendix C: Simulation study III (skewed errors and random effects)

In this section we study the model (13). Four simulation scenarios are considered for the distribution of the residual errors ϵ_{ij} and the random effects u_i. For the first scenario we simulate from $\epsilon_{ij} \sim \text{N}(0, 0.5)$ and $u_i \sim \text{N}(-1.5, 0.5)$. For the second scenario we use $\epsilon_{ij} \sim \text{TPN}(0, 0.5, 0.5)$ and $u_i \sim \text{N}(-1.5, 0.5)$. The third scenario consists of $\epsilon_{ij} \sim \text{N}(0, 0.5)$ and $u_i \sim \text{TPN}(-1.5, 0.5, 0.5)$. The fourth scenario uses $\epsilon_{ij} \sim \text{TPN}(0, 0.5, 0.5)$ and $u_i \sim \text{TPN}(-1.5, 0.5, 0.5)$. We simulate 100 data sets under these configurations. For each of these simulated samples, the model (13) was fitted assuming that $\epsilon_{ij} \sim \text{TPN}(0, \sigma_\varepsilon, \gamma_\varepsilon)$ and $u_i \sim \text{TPN}(\mu, \sigma, \gamma)$ with the prior structure:

$$\pi(\beta_1, \beta_2, \sigma_\varepsilon, \gamma_\varepsilon, \mu, \sigma, \gamma) \propto \frac{\pi(\gamma_\varepsilon)\pi(\mu)\pi(\sigma)\pi(\gamma)}{\sigma_\varepsilon}, \tag{21}$$

where $\pi(\mu)$ is a uniform prior on $[-100, 100]$, $\pi(\sigma)$ is a half-Cauchy density with mode 0 and unit scale parameter and $\pi(\gamma)$ and $\pi(\gamma_\varepsilon)$ are uniform priors on $(-1, 1)$. The propriety of the posterior is then guaranteed by Theorem 1. For each of the 100 simulated datasets we obtain a posterior sample of size $1,000$ after a burn-in period of 7500 iterations and thinned every 10 iterations (a total of $17,500$ MCMC draws). Table 8 summarizes the posterior samples and presents the median Bayes factors in favour of $\gamma_\varepsilon = 0$ or $\gamma = 0$ (symmetric errors or random effects), calculated using the Savage-Dickey density ratio. From this table, we can observe that despite the relatively small sample size, these Bayesian model selection criteria correctly identify the true model in each scenario.

In order to assess the impact of incorrectly imposing normality, we implement the model with normal residual errors and random effects. Table 9 presents posterior results for this model. The misspecification of the distribution of the

Table 9
Simulation study III: Monte Carlo median, quantile intervals of the median estimators using the Normal model.

Parameter	Scenario 1 $\varepsilon_{ij} \sim$ N $u_i \sim$ N	Scenario 2 $\varepsilon_{ij} \sim$ TPN $u_i \sim$ N	Scenario 3 $\varepsilon_{ij} \sim$ N $u_i \sim$ TPN	Scenario 4 $\varepsilon_{ij} \sim$ TPN $u_i \sim$ TPN
β_1	2.00 (1.97,2.03)	2.00 (1.97,2.02)	2.00 (1.97,2.03)	2.00 (1.97,2.02)
β_2	1.01 (0.78,1.18)	0.99 (0.81,1.20)	0.99 (0.79,1.23)	0.98 (0.78,1.25)
σ_ε	0.50 (0.46,0.54)	0.53 (0.49,0.57)	0.50 (0.46,0.54)	0.53 (0.49,0.57)
μ	-1.51 (-1.65,-1.33)	-1.89 (-2.04,-1.73)	-1.89 (-2.08,-1.76)	-2.28 (-2.45,-2.14)
σ	0.49 (0.41,0.58)	0.50 (0.42,0.58)	0.52 (0.44,0.60)	0.52 (0.44,0.60)

residual errors and the random effects clearly affects the estimation of μ (which can be interpreted as an intercept).

Acknowledgments

We thank two referees, an Associate Editor, and the Editor for helpful comments. This research was supported by the EPSRC grant EP/K007521/1.

References

D. Bandyopadhyay, V. H. Lachos, L. M. Castro, and D. K. Dey. Skew-normal/independent linear mixed models for censored responses with applications to HIV viral loads. *Biometrical Journal*, 54(3):405–425, 2012. MR2931082

Z. W. Birnbaum and S. C. Saunders. A new family of life distributions. *Journal of Applied Probability*, 6:319–327, 1969. MR0253493

T. S. Breusch, J. C. Robertson, and A. H. Welsh. The emperor's new clothes: a critique of the multivariate t regression model. *Statistica Neerlandica*, 51(3):269–286, 1997. MR1484953

D. Dunson. Nonparametric Bayes applications to biostatistics. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*, pages 223–273. Cambridge University Press, Cambridge, UK, 2010. MR2730665

C. Fernández and M. F. J. Steel. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998. MR1614601

C. Fernández and M. F. J. Steel. Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16(1):80–101, 2000. MR1749020

C. Fernández, J. Osiewalski, and M. F. J. Steel. On the use of panel data in stochastic frontier models with improper priors. *Journal of Econometrics*, 79(1):169–193, 1997. MR1457701

A. O. Finley, S. Banerjee, and B. P. Carlin. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1, 2007. MR2701470

O. Güler. *Foundations of Optimization*. Springer Science & Business Media, New York, USA, 2010. MR2680744

J. P. Hobert and G. Casella. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436):1461–1473, 1996. MR1439086

A. Jara, F. Quintana, and E. San-Martín. Linear mixed models with skew-elliptical distributions: A Bayesian approach. *Computational Statistics & Data Analysis*, 52(11):5033–5045, 2008. MR2526212

A. Jara, T. E. Hanson, and E. Lesaffre. Robustifying generalized linear mixed models using a new class of mixtures of multivariate Polya trees. *Journal of Computational and Graphical Statistics*, 18(4):838–860, 2009. MR2598032

A. Komárek and E. Lesaffre. Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. *Statistica Sinica*, (17):549–569, 2007. MR2408680

A. Komárek and E. Lesaffre. Bayesian accelerated failure time model with multivariate doubly interval–censored data and flexible distributional assumptions. *Journal of the American Statistical Association*, 103(482):523–533, 2008. MR2523990

V. H. Lachos, P. Ghosh, and R. B. Arellano-Valle. Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, 20:303–322, 2010. MR2640696

K. J. Lee and S. G. Thompson. Flexible parametric models for random-effects distributions. *Statistics in Medicine*, 27(3):418–434, 2008. MR2418453

Y. Maruyama and W. E. Strawderman. Robust Bayesian variable selection in linear models with spherically symmetric errors. *Biometrika*, 101(4):992–998, 2014. MR3286933

G. S. Mudholkar and A. D. Hutson. The epsilon–skew–normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference*, 83(2):291–309, 2000. MR1748020

J. Osiewalski and M. F. J. Steel. Robust Bayesian inference in elliptical regression models. *Journal of Econometrics*, 57:345–363, 1993. MR1237507

N. G. Polson and J. G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009. MR2749836

G. J. M. Rosa, C. R. Padovani, and D. Gianola. Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal*, 45(5):573–590, 2003. MR1998137

F. J. Rubio. On the propriety of the posterior of hierarchical linear mixed models with flexible random effects distributions. *Statistics & Probability Letters*, 96:154–161, 2015. MR3281760

F. J. Rubio and M. G. Genton. Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis. *Statistics in Medicine*, 35(4):2441–2454, 2016. MR3513523

F. J. Rubio and M. F. J Steel. Inference in two-piece location-scale models with Jeffreys priors. *Bayesian Analysis*, 9(1):1–22, 2014. MR3188293

F. J. Rubio and M. F. J. Steel. Bayesian modelling of skewness and kurtosis with two-piece scale and shape distributions. *Electronic Journal of Statistics*, 9(2):1884–1912, 2015. MR3391123

F. J. Rubio and K. Yu. Flexible objective Bayesian linear regression with applications in survival analysis. *Journal of Applied Statistics*, 44(5):798–810, 2017. MR3627599

F. J. Rubio, E. O. Ogundimu, and J. L. Hutton. On modelling asymmetric data using two-piece sinh–arcsinh distributions. *Brazilian Journal of Probability and Statistics*, 30(3):485–501, 2016. MR3531695

D. Sun, R. K. Tsutakawa, and Z. He. Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica*, 11(1):77–95, 2001. MR1820002

G. C. Tiao and W. Y. Tan. Bayesian analysis of random-effect models in the analysis of variance: I. posterior distribution of variance-components. *Biometrika*, 52(1/2):37–53, 1965. MR0208765

F. Vaida and L. Liu. Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*, 18(4):797–817, 2009. MR2750442

C. A. Vallejos and M. F. J. Steel. Objective Bayesian survival analysis using shape mixtures of log-normal distributions. *Journal of the American Statistical Association*, 110(510):697–710, 2015. MR3367258

I. Verdinelli and L. Wasserman. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995. MR1340514

M. West. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society. Series B*, 46:431–439, 1984. MR0790630

D. Zhang and M. Davidian. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802, 2001.