Robust Bayesian model selection for heavy-tailed linear regression using finite mixtures

Flávio B. Gonçalves[a], Marcos O. Prates[a] and Victor Hugo Lachos[b]

[a]*Universidade Federal de Minas Gerais*
[b]*University of Connecticut*

Abstract. In this paper, we present a novel methodology to perform Bayesian model selection in linear models with heavy-tailed distributions. We consider a finite mixture of distributions to model a latent variable where each component of the mixture corresponds to one possible model within the symmetrical class of normal independent distributions. Naturally, the Gaussian model is one of the possibilities. This allows for a simultaneous analysis based on the posterior probability of each model. Inference is performed via Markov chain Monte Carlo—a Gibbs sampler with Metropolis–Hastings steps for a class of parameters. Simulated examples highlight the advantages of this approach compared to a segregated analysis based on arbitrarily chosen model selection criteria. Examples with real data are presented and an extension to censored linear regression is introduced and discussed.

1 Introduction

Statistical practitioners generally use model selection criteria in order to select a best model in different applications. However, model selection has been shown not to be an easy task and each criterion performs better under different situations. For more complex models, it is not clear which criterion is preferable (Carlin, 2006; Chen, 2006; Gelman, Hwang and Vehtari, 2014). Recently, Gelman, Hwang and Vehtari (2014) studied and compared different model criteria and concluded that "The current state of the art of measurement of predictive model fit remains unsatisfying". From their study, it is clear that different criteria—Akaike information criterion (AIC), Deviance information criterion (DIC), Widely Applicable Information Criterion (WAIC), fail in selecting the most adequate model under a variety of circumstances (see Appendix B for a description of some criteria). For example, settings with strong prior information or when the posterior distribution is not well summarized by its mean or in a spatial or network setup (for more details, see Gelman, Hwang and Vehtari, 2014). Other authors also comment about the model selection problem, e.g., "In summary, model choice is to Bayesians what multiple comparisons is to frequentists: a really hard problem for which there exist several potential solutions, but no consensus choice" (Carlin, 2006), "we saw that no single measure is dominant in all three cases. The L-measure performed better when the true model becomes more complex and BIC performed better when the true model is more parsimonious." (Chen, 2006).

Under the Bayesian paradigm, a more robust and elegant solution is available, at least in theory, by considering one "full model" that embeds all the individual models of interest. More specifically, this means that a multinomial random variable (r.v.) with each category corresponding to one of the individual models is specified. This way, model selection may be performed based on the posterior distribution of this r.v., that is, the posterior probability of each model. Nevertheless, this approach may be challenging in some cases, specially

when the individual models have different dimensions and distinct parameter. Available solutions may need to rely on complicated reversible jump MCMC (Markov chain Monte Carlo) algorithms and, therefore, other model criteria may be preferred.

A simple and generally efficient solution may be obtained when a mixture distribution can be adopted for one of the model's component (parameter or latent variable) in a way that each mixture component corresponds to one of the individual models (see, for example, Gonçalves, Gamerman and Soares, 2013, George and McCulloch, 1993). This would typically lead to a simple and efficient solution, allowing for model selection to be based on the models' posterior probability.

We consider a model selection problem concerning the specification of the error distribution in linear regression models. In particular, we consider different heavy-tailed distributions and the traditional Gaussian specification. Existing solutions use model selection criteria arbitrarily chosen (see Lachos, Ghosh and Arellano-Valle, 2010; Basso et al., 2010; Cabral, Lachos and Prates, 2012) and, therefore, motivates the development of a more robust methodology.

The distributions of random errors and other random variables are routinely assumed to be Gaussian. However, the normality assumption is doubtful and lacks robustness especially when the data contain outliers or show a significant violation of normality. Thus, previous works have shown the importance of considering more general structures than the Gaussian distribution for this component, such as heavy-tailed distributions (Fernandez and Steel, 1999; Galea, Paula and Uribe-Opazo, 2003; Rosa, Padovani and Gianola, 2003; Galea, Paula and Cysneiros, 2005; Garay et al., 2015). These structures provide appealing robust and adaptable models, for example, the Student-t linear mixed model presented by Pinheiro, Liu and Wu (2001), who showed that it performed well in the presence of outliers. Furthermore, the scale mixtures of normal (SMN) distributions have also been applied into a wide variety of regression models (see Lange and Sinsheimer, 1993; Osorio, Paula and Galea, 2007; Lachos, Angolini and Abanto-Valle, 2011). It is one of the most important subclasses of the elliptical symmetric distributions. The SMN distribution class contains many heavier-than-normal tailed members, such as Student-t, Slash, power exponential and contaminated normal. Recently, Lin and Cao (2013) (see also Lachos, Angolini and Abanto-Valle, 2011) investigated the inference of a measurement error model under the SMN distributions and demonstrated its robustness against outliers through extensive simulations.

As defined by Andrews and Mallows (1974), a continuous random variable Y has a SMN distribution if it can be expressed as follows:

$$Y = \mu + \kappa^{1/2}(U)W,$$

where μ is a location parameter, W is a normal random variable with zero mean and variance σ^2, $\kappa(U)$ is a positive weight function, U is a mixing positive random variable with density $h(\cdot \mid \nu)$ and ν is a scalar or parameter vector indexing the distribution of U. As in Lange and Sinsheimer (1993) and Choy and Chan (2008), we restrict our attention to the case where $\kappa(U) = 1/U$, that is, the normal independent (NI) class of distributions. Thus, $Y \mid U = u \sim \mathcal{N}(\mu, u^{-1}\sigma^2)$ and the marginal probability density function (pdf) of Y is given by

$$f(y \mid \mu, \sigma^2, \nu) = \int_0^\infty \phi((y - \mu)/\sqrt{u^{-1}\sigma^2})h(u \mid \nu)\, du. \tag{1}$$

Note that when $U = 1$, we retrieve the normal distribution. Following the steps of Basso et al. (2010), we have the following properties for the SMN family:

(a) If $E[\kappa^{1/2}(U)] < \infty$, then $E[Y] = \mu$.
(b) If $E[\kappa(U)] < \infty$, then $\text{Var}[Y] = \sigma^2 k_2$.

(c) If $E[\kappa^2(U)] < \infty$, then the excess kurtosis coefficient is given by

$$\gamma_2 = \frac{E[Y - E[Y]]^4}{(\text{Var}[Y])^2} - 3 = \frac{3k_4}{k_2^2} - 3,$$

where $k_m = E[\kappa^{m/2}(U)]$.

Apart from the normal model, we explore two different types of heavy-tailed densities based on the choice of $h(\cdot \mid \boldsymbol{v})$.

- *The Student-t distribution*, $Y \sim \mathcal{T}(\mu, \sigma^2, \nu_t)$.
  The use of the Student-t distribution as an alternative robust model to the normal distribution has frequently been suggested in the literature (Lange, Little and Taylor, 1989). For the Student-t distribution with location μ, scale σ and degrees of freedom ν_t, the pdf can be expressed as

  $$f(y \mid \mu, \sigma, \nu_t) = \int_0^\infty \phi\big((y - \mu)/\sqrt{u^{-1}\sigma^2}\big) f_{\mathcal{G}}\left(u \mid \frac{\nu_t}{2}, \frac{\nu_t}{2}\right) du,$$

  where $f_{\mathcal{G}}(\cdot \mid a, b)$ is the Gamma density function with shape and rate parameters given by a and b, respectively. That is, $Y \sim \mathcal{T}(\mu, \sigma^2, \nu_t)$ is equivalent to the following hierarchical form:

  $$Y \mid \mu, \sigma^2, \nu_t, u \sim \mathcal{N}(\mu, u^{-1}\sigma^2), \qquad U \mid \nu_t \sim \mathcal{G}(\nu_t/2, \nu_t/2).$$

  For the Student-t distribution, we have that

  $$k_m = \left(\frac{\nu_t}{2}\right)^{\frac{m}{2}} \frac{\Gamma(\frac{\nu_t - m}{2})}{\Gamma(\frac{\nu_t}{2})},$$

  therefore, the Student-t has variance $\sigma^2 \frac{\nu_t}{\nu_t - 2}$, for $\nu_t > 2$, and excess kurtosis $\frac{6}{\nu_t - 4}$, for $\nu_t > 4$.
- *The Slash distribution*, $Y \sim \mathcal{S}(\mu, \sigma^2, \nu_s)$.
  This distribution presents heavier tails than those of the normal distribution and it includes the normal case when $\nu_s \uparrow \infty$. Its pdf is given by

  $$f(y \mid \mu, \sigma, \nu_s) = \nu_s \int_0^1 u^{\nu_s - 1} \phi\big((y - \mu)/\sqrt{u^{-1}\sigma^2}\big) du.$$

  Thus, the Slash distribution is equivalent to the following hierarchical form:

  $$Y \mid \mu, \sigma^2, \nu_s, u \sim \mathcal{N}(\mu, u^{-1}\sigma^2), \qquad U \mid \nu_s \sim \mathcal{B}(\nu_s, 1),$$

  where $\mathcal{B}(\cdot, \cdot)$ denotes the beta distribution. For the Slash distribution we have that

  $$k_m = \frac{2\nu_s}{2\nu_s - m},$$

  therefore, the Slash has variance $\sigma^2 \frac{\nu_s}{\nu_s - 1}$, for $\nu_s > 1$, and excess kurtosis $\frac{3}{\nu_s(\nu_s - 2)}$, for $\nu_s > 2$.

The SMN formulation described above is used in a linear regression approach by taking $\mu_i = \mathbf{X}_i \boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is the vector of coefficients and $\mathbf{X}_i$ is the design matrix for the ith individual, for $i = 1, \ldots, n$.

The aim of this paper is to propose a general formulation to perform Bayesian model selection for heavy-tailed linear regression models in a simultaneous setup. That is achieved by specifying a full model which includes the space of all individual models under consideration—specified using the SMN approach described above. This way, the model

selection criterion can be based on the posterior probability of each model. A mixture distribution is adopted to one of the full model's variable, with each component of the mixture referring to one of the individual models. This approach has two main advantages when compared to an ordinary analysis where each model is fitted separately and some model selection criterion is used. Firstly, there is a significant gain in the computational cost since we eliminate the need to fit all the individual models separately. Secondly, the proposed model selection criterion is fully based on the Bayesian Paradigm, meaning that the model choice is based on the posterior probability of each model. This is more robust when compared to some other arbitrarily chosen model selection criteria, such as DIC, expected AIC (EAIC), extended Bayes information criterion (EBIC) (Spiegelhalter et al., 2002), conditional predictive ordinate (CPO) (Geisser and Eddy, 1979) and WAIC (Watanabe, 2010). The examples presented in the paper are meant to provide empirical evidence for this argument. The posterior distribution of the unknown quantities has a significant level of complexity which motivates the derivation of a MCMC algorithm to obtain a sample from this distribution.

This paper is organised as follows: Section 2 presents the general model; Section 3 presents a MCMC algorithm to make inference for the proposed model; a variety of simulated examples are presented in Section 4 and the analysis of two real data sets is shown in Section 5. Finally, Section 6 discusses some extensions of the proposed methodology.

2 Linear regression model with heavy-tailed mixture structured errors

Model selection is an important and complex problem in statistical analysis and the Bayesian approach is particularly appealing to solve it. In particular, the use of mixtures is a nice way to pose and solve the problem, whenever possible. It allows for an analysis where all models are considered and compared in a simultaneous setup without the need of complicated reversible jump MCMC algorithms. Note that, from (1), each model is determined by the distribution of the scale factor u, which suggests that a mixture distribution could be used for this latent variable. We present a general finite mixture model framework capable of capturing different behavior of the response and indicate which individual distribution is preferred.

2.1 The model

Define the n-dimensional response vector $\mathbf{Y}$, the $n \times q$ design matrix $\mathbf{X}$, the q-dimensional coefficient vector $\boldsymbol{\beta}$ and two K-dimensional vectors $\boldsymbol{\gamma} = (\gamma_1 \cdots \gamma_K)'$ and $\mathbf{p} = (p_1 \cdots p_K)'$. Finally, let $\operatorname{diag}(\mathbf{u}^{-1})$ be a n-dimensional diagonal matrix with ith diagonal $u_i^{-1}, i = 1, \ldots, n$. We propose the following general model:

$$(\mathbf{Y}|Z_j = 1, \mathbf{U} = \mathbf{u}) \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \gamma_j \operatorname{diag}(\mathbf{u}^{-1})), \tag{2}$$

$$(U_i|Z_j = 1) \stackrel{\text{iid}}{\sim} F_j(v_j), \qquad i = 1, \ldots, n, \tag{3}$$

$$\mathbf{Z} \sim \operatorname{Mult}(1, p_1, \ldots, p_K), \tag{4}$$

$$\gamma_j = g_j(v_j), \qquad j = 1, \ldots, K, \tag{5}$$

where Mult is the Multinomial distribution and each F_j represents a positive distribution controlled by parameter(s) v_j, which may need to be truncated to guarantee that Y_i has finite variance under each F_j.

The particular structure chosen for the variance in (2) was thought of so that, for each j, the variance of the model is the same—σ^2. This is achieved through specific choices for the functions γ_j and allows us to treat σ^2 as a common parameter to all of the individual models. Otherwise we would need one scale parameter for each model. Therefore, in our approach,

since we have a common μ and σ^2, the models will mainly differ from each other in terms of tail behavior which will favor the model selection procedure.

Note that each component from the mixture distribution of u_i corresponds to one of the models being considered. Model selection is made through the posterior distribution of $\mathbf{Z}$. A subtle but important point here is the fact that there is no i index for Z_j. This means that we assume that all the observations come from the same model, which poses the inference problem in the model selection framework.

Another advantage of the simultaneous approach is that it allows the use of Bayesian model averaging (see Raftery, Madigan and Volinsky, 1995). This is particularly useful in cases where more than one model have a significant posterior probability. Note that the models we consider can be quite similar in some situations—specially for higher values of the degrees of freedom (df) parameters.

2.2 Prior distributions

The Bayesian model is fully specified by (2)–(5) and the prior distribution for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \mathbf{p}, \boldsymbol{\nu})$, for $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$. Due to the complexity of the proposed model, the prior distribution plays an important role on the model identifiability and selection process and, for that reason, needs to be carefully specified.

Prior specification firstly assumes independence among all the components of $\boldsymbol{\theta}$. Secondly, standard priors $\boldsymbol{\beta} \sim \mathcal{N}_q(\boldsymbol{\mu}_0, \tau_0^2 \mathbf{I}_q)$ and $\sigma^2 \sim \mathcal{IG}(a_0, b_0)$ are adopted.

The prior distributions of the tail behavior parameters $\boldsymbol{\nu}$ require special attention. This type of parameter is known to be hard to estimate (see Steel and Fernandez, 1999) and the most promising solutions found in the literature tackle the problem through special choices of prior distributions (see Fonseca, Ferreira and Migon, 2008). Recently, Simpson et al. (2017) proposed a general family of prior distributions for flexibility parameters which includes tail behavior parameters.

In this paper, we adopt the penalised complexity priors (PC priors) from Simpson et al. (2017). In a simple way, the PC priors have as main principle to prefer a simpler model and penalise the more complex one. To do so, the Kullback–Leibler divergence (KLD) (Kullback and Leibler, 1951) is used to define a measure of information loss when a simpler model h is used to approximate a more flexible model $f(\cdot|\nu_j)$. The measure $d(f \parallel h)(\nu_j) = d(\nu_j) = \sqrt{2KLD(f \parallel h)}$ is defined to be a measure of complexity of model $f(\cdot|\nu_j)$ in comparison to h. Further, a density function $\pi(d(\nu_j)) = \lambda \exp(-\lambda d(\nu_j))$ is set for the measure $d(\nu_j)$. Finally, the prior distribution of ν_j is given by

$$\pi(\nu_j) = \lambda \exp\left(-\lambda d(\nu_j)\right) \left| \frac{\partial d(\nu_j)}{\partial \nu_j} \right|, \qquad j = 1, \ldots, K.$$

Note that this is a proper probability distribution (which implies in a proper posterior), since it is a transformation from the proper prior adopted for $d(\nu_j)$. Martins and Rue (2013) showed that in a practical way, for the Student-t regression model, the PC prior can behave very similar to the Jeffrey's priors constructed by Fonseca, Ferreira and Migon (2008). Another interesting practical usage of this prior is that the selection of an appropriate λ is done by allowing the researcher to control the prior tail behavior of the model. For example, for the Student-t distribution the user must select $\nu^\star$ and ξ such that $P(\nu_j < \nu^\star) = \xi$, in other words, how much mass probability ξ is assigned to $\nu_j \in (2, \nu^\star)$ (where j defines the F_j distribution such that the response follows a Student-t distribution). Clearly, the same procedure applies for any other distribution in the NI family that has a flexibility parameter. For more details on the PC priors, see Simpson et al. (2017).

The prior distribution for $\mathbf{p}$ also requires special attention. Note that even in the extreme (unrealistic) case where $\mathbf{Z}$ is observed, it does not provide much information about

p, in fact, it is equivalent to the information contained in a sample of size one from a Mult$(1, p_1, \ldots, p_K)$ distribution. The fact that $\mathbf{Z}$ is unknown aggravates the problem. A simple and practical way to understand the consequences of this is given by the following lemma, which is a generalisation of Lemma 1 from Gonçalves, Gamerman and Soares (2013) where, to the best of our knowledge, this problem was firstly encountered.

Lemma 1. *For a prior distribution* $\mathbf{p} \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$, *the posterior mean of* p_j, $\forall j$, *is restricted to the interval* $(\frac{\alpha_j}{1+\sum_{k=1}^{K} \alpha_k}, \frac{\alpha_j+1}{1+\sum_{k=1}^{K} \alpha_k})$.

Proof. See Appendix A. $\qquad \square$

For example, if $\alpha_j = 1$, $\forall j$, then $\mathbb{E}[p_j|y] \in (1/(K+1), 2/(K+1))$. This result indicates that the estimation of $\mathbf{Z}$ may be compromised by unreasonable choices of the α_j's.

A reasonable solution for this problem is to use a Dirichlet prior distribution with parameters (much) smaller than 1, which makes it sparse. It is important, though, to choose reasonable values for the α_j's, in the light of Lemma 1. Gonçalves, Gamerman and Soares (2013) claim that $\alpha_j = 0.01$, $\forall j$, leads to good results and, in the cases where prior information is available, some of the α_j's may be increased accordingly.

3 Bayesian inference

We derive a MCMC algorithm considering the three most common choices in the NI family—Normal, Student-t, Slash. Nevertheless, based on the formulation presented in Section 2.1, including other possibilities is straightforward. One should be careful, however, as it may lead to serious identifiability issues due to similarities among the individual models. The model is given by:

$$(\mathbf{Y}|Z_j = 1) \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \gamma_j \, \text{diag}(\mathbf{u}^{-1})), \tag{6}$$

$$\mathbf{Z} \sim \text{Mult}(1, p_1, p_2, p_3), \tag{7}$$

$$U_i \overset{\text{iid}}{\sim} \begin{cases} \delta_1, & \text{if } Z_1 = 1, \\ \mathcal{G}(v_t/2, v_t/2), & \text{if } Z_2 = 1, i = 1, \ldots, n, \\ \mathcal{B}(v_s, 1), & \text{if } Z_3 = 1, \end{cases} \tag{8}$$

$$\gamma_j = \begin{cases} 1, & \text{if } j = 1, \\ (v_t - 2)/v_t, & \text{if } j = 1, \\ (v_s - 1)/v_s, & \text{if } j = 1, \end{cases} \tag{9}$$

where δ_1 is a degenerate r.v. at 1 and $\mathcal{G}$ and $\mathcal{B}$ are the Gamma and Beta distributions, respectively. We impose that $v_t > 2$ and $v_s > 1$ so that Y_i has finite variance (σ^2) under each individual model.

Inference is performed via MCMC—a Gibbs sampling with Metropolis Hastings (MH) steps for the df parameters. Details of the algorithm are presented below.

3.1 MCMC

We choose the following blocking scheme for the Gibbs sampler:

$$(\mathbf{p}, \mathbf{Z}, \mathbf{U}), \qquad \boldsymbol{\beta}, \qquad \sigma^2, \qquad (v_t, v_s). \tag{10}$$

This blocking scheme minimises the number of blocks among the algorithms with only one MH step (which is inevitable for the df parameters). The minimum number of blocks reduces

the correlation among the components, which speeds the convergence of the chain. Moreover, the most important and difficult step is the one that samples from $(\mathbf{p}, \mathbf{Z}, \mathbf{U})$ and sampling directly from its full conditional also favors the convergence properties of the chain.

The full conditional densities of (10) are all derived from the joint density of all random components of the model.

$$
\begin{aligned}
\pi(\mathbf{Y}, \boldsymbol{\beta}, \sigma^2, \mathbf{p}, \mathbf{Z}, \mathbf{U}, \boldsymbol{\gamma}, \nu_t, \nu_s | \mathbf{X}) & \\
\propto \pi(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \mathbf{Z}, \mathbf{U}, \boldsymbol{\gamma}, \mathbf{X}) & \pi(\mathbf{U}|\mathbf{Z}, \nu_t, \nu_s) \\
\times \pi(\mathbf{Z}|\mathbf{p}) & \pi(\mathbf{p}) \pi(\nu_t) \pi(\nu_s) \pi(\boldsymbol{\beta}) \pi(\sigma^2).
\end{aligned}
\tag{11}
$$

The first two terms on the right-hand side of (11) are given in Section 1, for each individual model (Z_j). The remaining terms are given in Section 2.

The full conditional distributions of $\boldsymbol{\beta}$ and σ^2 are easily devised and given by:

$$
(\boldsymbol{\beta}|\cdot) \sim \mathcal{N}_q\left(\boldsymbol{\Sigma}_\beta\left((\tau_0^2\mathbf{I}_q)^{-1}\boldsymbol{\mu}_0 + (\sqrt{\mathbf{u}} \odot \mathbf{X})'(\sqrt{\mathbf{u}} \odot \mathbf{y})/(\gamma_j\sigma^2)\right), \boldsymbol{\Sigma}_\beta\right),
$$

$$
(\sigma^2|\cdot) \sim \mathcal{IG}\left(a_0 + n/2, b_0 + \sum_{i=1}^{n} \frac{u_i(y_i - \mathbf{X}_i.\boldsymbol{\beta})^2}{2\gamma_j}\right),
$$

where $\boldsymbol{\Sigma}_\beta = ((\tau_0^2\mathbf{I}_q)^{-1} + (\sqrt{\mathbf{u}} \odot \mathbf{X})'(\sqrt{\mathbf{u}} \odot \mathbf{X})/(\gamma_j\sigma^2))^{-1}$, $\sqrt{\mathbf{u}}$ is the n-dimensional vector with entries $\sqrt{u_i}$, $\odot$ is the Hadamard product which multiplies term by term of matrices with the same dimension and $\mathbf{I}_q$ is the identity matrix with dimension q.

The df parameters are sampled in a MH step with the following transition distribution (at the kth iteration):

$$
q(\nu_t^k, \nu_s^k) = q(\nu_t^k)q(\nu_s^k),
\tag{12}
$$

$$
q(\nu_t^k) = ((1 - Z_2)\mathbb{1}(\nu_t^k = \nu_t^{k-1}) + Z_2 f_\mathcal{N}(\nu_t^k; \nu_t^{k-1}, \tau_t^2)),
\tag{13}
$$

$$
q(\nu_s^k) = ((1 - Z_3)\mathbb{1}(\nu_s^k = \nu_s^{k-1}) + Z_3 f_\mathcal{N}(\nu_s^k; \nu_s^{k-1}, \tau_s^2)),
\tag{14}
$$

where $f_\mathcal{N}(l; a, b)$ is the density of a normal distribution with mean a and variance b evaluated at l. The respective acceptance probability of a move is

$$
\alpha(k - 1 \to k) = \min\left\{1, Z_1 + Z_2\frac{\pi(\nu_t^k|\cdot)}{\pi(\nu_t^{k-1}|\cdot)} + Z_3\frac{\pi(\nu_s^k|\cdot)}{\pi(\nu_s^{k-1}|\cdot)}\right\},
\tag{15}
$$

where

$$
\pi(\nu_t|\cdot) \propto \pi(\mathbf{U}|Z_2 = 1, \nu_t)\pi(\nu_t),
$$

$$
\pi(\nu_s|\cdot) \propto \pi(\mathbf{U}|Z_3 = 1, \nu_s)\pi(\nu_s).
$$

This result is obtained by adopting the following dominating measure for both the numerator and the denominator of the acceptance probability: $\mathbb{L}^2 \otimes \mathbb{L} \otimes m$ if $Z_1 = 0$ and $\mathbb{L}^2 \otimes m^2$ if $Z_1 = 1$, where m is the counting measure and $\mathbb{L}^d$ is the d-dimensional Lebesgue measure. The detailed balance along with the fact that chain is irreducible, makes this a valid MH algorithm (see Tierney, 1998).

Note that, once we have the output of the chain, estimates of the df parameters will be based on samples of $(\nu_t|Z_2 = 1)$ and $(\nu_s|Z_3 = 1)$, which justifies the transition distributions in (12)–(14).

From (11), the full conditional density of $(\mathbf{p}, \mathbf{Z}, \mathbf{U})$ is

$$
\pi(\mathbf{U}, \mathbf{Z}, \mathbf{p}|\cdot) \propto \pi(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{Z}, \mathbf{U}, \boldsymbol{\gamma}, \mathbf{X})\left[\prod_{i=1}^{n} \pi(U_i|\mathbf{Z}, \nu_t, \nu_s)\right]\pi(\mathbf{Z}|p)\pi(\mathbf{p})
$$

$$
\propto \left[\prod_{i=1}^{n} \pi(U_i|\cdot)\right](r_1p_1)^{Z_1}(r_2p_2)^{Z_2}(r_3p_3)^{Z_3}\pi(\mathbf{p}).
$$

Defining $w = \sum_{j=1}^{3} r_j p_j$ and $w_j = r_j p_j / w$, for $j = 1, 2, 3$, we get

$$\pi(\mathbf{U}, \mathbf{Z}, \mathbf{p}|\cdot) \propto \left[\prod_{i=1}^{n} \pi(U_i|\cdot)\right] (w_1)^{Z_1} (w_2)^{Z_2} (w_3)^{Z_3} w \pi(\mathbf{p}). \qquad (16)$$

We can sample from (16) using the following algorithm.

```
1. Simulate p from a density π*(p) ∝ wπ(p);
2. Simulate Z ~ Mult(1, w₁, w₂, w₃);
3. Simulate Uᵢ from the density π(Uᵢ|·), ∀i;
4. OUTPUT (u, z, p).
```

Steps 2 and 3 are straightforward once we have that:

$$r_1 = \prod_{i=1}^{n} \exp\left(-\frac{1}{2\gamma_1\sigma^2}\tilde{y}_i^2\right);$$

$$r_2 = \frac{(\frac{\nu_t - 2}{\nu_t})^{-n/2}(\nu_t/2)^{n\nu_t/2}(\Gamma(\frac{\nu_t+1}{2}))^n}{(\Gamma(\frac{\nu_t}{2}))^n \prod_{i-1}^{n}(\frac{\tilde{y}_i^2}{2\gamma_2\sigma^2} + \frac{\nu_t}{2})^{(\nu_t+1)/2}};$$

$$r_3 = \left(\frac{\nu_s - 1}{\nu_s}\right)^{-n/2} \left(\frac{\Gamma(\nu_s + 1)}{\Gamma(\nu_s)}\Gamma(\nu_s + 1/2)\right)^n \prod_{i=1}^{n}\left[\frac{F_{\mathcal{G}}(1; \nu_s + 1, \frac{\tilde{y}_i^2}{2\gamma_3\sigma^2})}{(\frac{\tilde{y}_i^2}{2\gamma_3\sigma^2})^{\nu_s + 1/2}}\right],$$

where $\tilde{y}_i = y_i - \mathbf{X}_i.\boldsymbol{\beta}$ and $F_{\mathcal{G}}(x; a, b)$ is the distribution function of a Gamma distribution with parameters (a, b) evaluated at x. Moreover,

$$(U_i|Z_1 = 1, \cdot) \sim \delta_1;$$

$$(U_i|Z_2 = 1, \cdot) \sim \mathcal{G}\big((\nu_t + 1)/2, \tilde{y}_i^2/(2\gamma_2\sigma^2) + \nu_t/2\big);$$

$$(U_i|Z_3 = 1, \cdot) \sim \mathcal{G}_{[0,1]}\big(\nu_s + 1, \tilde{y}_i^2/(2\gamma_3\sigma^2)\big),$$

where $\mathcal{G}_{[0,1]}$ is a truncated Gamma distribution in $[0, 1]$.

Step 1 is performed via rejection sampling (RS) proposing from the prior $\pi(\mathbf{p})$ and accepting with probability $\frac{w}{\max_j\{r_j\}}$. Simulated studies indicated that the algorithm is computationally efficient.

Monte Carlo estimates of the posterior distribution of $\mathbf{Z}$ (denoted by $\boldsymbol{\rho}$), that is, the models' posterior probabilities, based on a sample of size M, are given by

$$\hat{\rho}_j = P(\widehat{Z_j = 1}|y) = \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}(Z_j^{(m)} = 1), \qquad j = 1, 2, 3.$$

3.2 Practical implementation

The MCMC algorithm described in the previous section requires special attention to some aspects to guarantee its efficiency.

An indispensable strategy consists of warming up the chain inside each of the heavy-tailed models (Student-t and Slash). It contributes in several ways to the efficiency of the algorithm.

First, it contributes to the mixing of the chain among the different models. If the chain starts at arbitrary values for the df parameters, it may move to high posterior density values for one of them while the other is still at a low posterior density value. This will make moves from the former model to the latter very unlike, jeopardising the convergence. More specifically,

one may take the sample mean of the df parameters from their respective warm-up chains, after discarding a burn-in, as the starting values for the full chain.

Secondly, the warm-up chains will achieve or approach local convergence (inside each model). This will significantly speed the convergence of the full chain, which will have as main purpose the convergence of the $\mathbf{Z}$ coordinate.

Finally, the warm-up chains are a good opportunity to tune the MH steps of the df parameters. Given the unidimensional nature of the step and the random walk structure, the acceptance rates should be around 0.44 (see Roberts, Gelman and Gilks, 1997).

3.3 Prediction

An often common step in any regression analysis is prediction for a new configuration $\mathbf{X}_{n+1}$ of the covariates. This procedure is straightforward in a MCMC context where a sample from the posterior predictive distribution of Y_{n+1} can be obtained by adding two simple steps at each iteration of the Gibbs sampler after the burn-in.

Let $(\mathbf{Z}^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma^{2(m)}, \boldsymbol{\gamma}^{(m)}, v_t^{(m)}, v_s^{(m)})$ be the state of the chain at the mth iteration after the burn-in. Then, for each $m = 1, 2, \ldots$, firstly sample $(u_{n+1}^{(m)} | \mathbf{Z}^{(m)}, v_t^{(m)}, v_s^{(m)})$ from (8) and finally sample

$$Y_{n+1}^{(m)} \sim \mathcal{N}(\mathbf{X}_{n+1}\boldsymbol{\beta}^{(m)}, \sigma^{2(m)}(\mathbf{Z}'^{(m)}\boldsymbol{\gamma}^{(m)})(u_{n+1}^{(m)})^{-1}), \tag{17}$$

where $\mathbf{Z}'^{(m)}$ is a row vector and $\boldsymbol{\gamma}^{(m)}$ is a column vector.

One can also consider the posterior predictive distribution of Y_{n+1} under one particular model, for example, the one with the highest posterior probability. In that case, it is enough to consider the sub-sample of the sample above corresponding to the chosen model.

4 Simulated examples

In this section, we introduce synthetic data examples to better understand the properties of the proposed methodology. Our goal is twofold: to provide strong empirical evidence that, (1) as long as information is available, the true model is selected using the proposed methodology and (2) this selects the correct or more adequate model more often than some traditional criteria.

We firstly present a study to see how well the proposed methodology correctly identifies the true model. A second study shows the performance of the criteria when the model has correlated covariates, which may cause problems in the estimate of the fixed effects and variance parameter. Finally, a third synthetic data set is generated from a residual mixture model to investigate if the model that better approximates the true mixture distribution is chosen.

4.1 Study I

In this study, data is generated from one of the proposed distributions: Normal, Student-t and Slash. We consider an intercept and two covariates, i.e. $\mathbf{X}_{i\cdot} = (1, X_{i1}, X_{i2})$, where X_{i1} is a standard Normal random variable, X_{i2} is a Bernoulli random variable with parameter 0.5 and $i = 1, \ldots, n$. The regression coefficients are $\boldsymbol{\beta}^\top = (1, 2, -2)$. Finally, for all models, the variance σ^2 is set to 1. The synthetic data were generated from each of the following distributions:

1. Normal;
2. Student-t with degrees of freedom $v_t = 15$ and $v_t = 3$;
3. Slash with degrees of freedom $v_s = 3.36$ and $v_s = 1.25$.

Different sample sizes n are also considered—100, 500, 1000 and 5000, giving a total of 20 scenarios. The degrees of freedom for the Slash were chosen to minimise the Kullback–Leibler divergence between the Student-t with $v_t = 15$ and $v_t = 3$, respectively.

For each simulated scenario a Markov chain runs for 110k iterations, with a burn-in of 10k giving a total posterior chain of 100k iterations. Convergence is checked using Geweke's criterion (Geweke, 1992) since we only ran one chain. The same chain size, burn-in period and convergence verification were performed for all the examples in the paper. Notice that the parametrisation adopted allows some parameters to be estimated using the whole chain, independently of the model that is visited in each iteration. This favors the chain convergence and the Monte Carlo variance of the estimates of $(\boldsymbol{\beta}, \sigma^2)$.

The summary posterior results of one run are presented in Table 1. They show that as the sample size increases the proposed methodology selects the correct model. Moreover, in the case where data is generated from the Normal distribution, not only the correct model is correctly chosen in all but one case, but also the estimated degrees of freedom of the Student-t and of the Slash distributions are high—making these distributions similar to Gaussian. Another important feature presented in the Table 1 is that the degrees of freedom parameter of the generating model is well estimated. For the non-generating model, the degrees of freedom parameter is reasonably estimated, in the sense of making the respective model as close as possible to the true one. For example, when the data is generated from the Student-t with $v_t = 15$ the v_s is estimated close to 3.36, which is the value that minimises the Kullback–Leibler divergence between the two distributions. Table 1 also emphasises that, for small sample sizes $n = 100$ or $n = 500$, there is not enough information about the tail behavior to clearly distinguish among the models.

To check the capability of the proposed methodology in selecting the correct model, we performed a Monte Carlo study with 50 replicates of each of the 20 generation schemes. Table 2 presents the Mean Square Error (MSE) of the posterior estimates of the model pa-

Table 1 *Results for Study I. Estimates* (*posterior mean*) *refer to one of the* 50 *replications*

Model	Sample size	$\beta^{\top} = (1, 2, -2)$	$\sigma^2 = 1$	(v_t, v_s)	$\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$
Normal	100	(1.153, 1.991, −2.305)	1.121	(10.62, 2.09)	(0.103, 0.537, 0.360)
	500	(0.997, 2.047, −2.065)	0.979	(29.47, 4.43)	(0.882, 0.073, 0.045)
	1000	(1.004, 1.981, −1.986)	0.999	(31.20, 4.45)	(0.644, 0.280, 0.076)
	5000	(0.990, 1.979, −1.965)	0.980	(44.25, 5.32)	(0.749, 0.097, 0.154)
Student-t	100	(1.236, 1.829, −2.148)	1.267	(9.86, 1.86)	(0.044, 0.439, 0.517)
($v_t = 15$)	500	(1.074, 2.029, −2.042)	1.038	(28.64, 4.14)	(0.777, 0.151, 0.072)
	1000	(1.012, 2.006, −1.991)	0.982	(21.24, 3.72)	(0.123, 0.609, 0.268)
	5000	(1.014, 2.000, −1.999)	0.993	(16.19, 3.19)	(0.000, 0.807, 0.193)
Student-t	100	(1.116, 1.865, −2.045)	1.389	(3.22, 1.22)	(0.000, 0.371, 0.629)
($v_t = 3$)	500	(0.978, 2.031, −1.923)	1.244	(3.36, 1.20)	(0.000, 0.679, 0.321)
	1000	(1.001, 2.005, −1.959)	0.861	(3.30, 1.25)	(0.000, 0.990, 0.010)
	5000	(1.024, 2.007, −2.035)	1.029	(2.95, −)	(0.000, 1.000, 0.000)
Slash	100	(0.968, 2.097, −1.902)	1.049	(17.37, 2.76)	(0.369, 0.357, 0.274)
($v_s = 3.36$)	500	(0.976, 2.003, −2.039)	0.963	(19.90, 3.30)	(0.167, 0.450, 0.383)
	1000	(1.004, 1.997, −2.010)	1.015	(17.72, 3.22)	(0.020, 0.626, 0.354)
	5000	(1.029, 2.000, −2.044)	0.963	(22.61, 3.65)	(0.000, 0.230, 0.770)
Slash	100	(1.012, 1.988, −1.957)	0.454	(18.04, 2.75)	(0.344, 0.367, 0.289)
($v_s = 1.25$)	500	(1.033, 2.026, −2.015)	0.904	(3.91, 1.29)	(0.000, 0.280, 0.720)
	1000	(1.012, 2.012, −2.040)	0.839	(3.93, 1.35)	(0.000, 0.561, 0.439)
	5000	(1.017, 1.988, −2.011)	0.863	(−, 1.30)	(0.000, 0.000, 1.000)

Table 2 *Mean Square error estimates from the 50 replications for each parameter. PCS is the percentage of correct selection in the 50 replications. The ($\times 10^3$) means that the reported value is MSE $\times 10^3$*

Model	Sample size	$\beta (\times 10^3)$	$\sigma^2 (\times 10^3)$	ν_t or ν_s	ρ	PCS
Normal	100	(16.29, 10.15, 34.44)	32.12	–	0.303	80%
	500	(3.24, 1.84, 7.23)	3.96	–	0.207	86%
	1000	(2.24, 0.75, 5.03)	2.24	–	0.192	84%
	5000	(0.40, 0.24, 0.66)	0.46	–	0.123	88%
Student-t	100	(26.04, 11.27, 44.80)	72.04	15.852	0.456	30%
($\nu_t = 15$)	500	(4.39, 2.42, 9.83)	5.91	45.564	0.288	64%
	1000	(2.25, 0.82, 5.40)	2.84	36.736	0.284	68%
	5000	(0.35, 0.18, 0.67)	0.46	11.580	0.131	80%
Student-t	100	(10.82, 6.26, 23.04)	97.85	71.458	0.365	32%
($\nu_t = 3$)	500	(2.81, 0.87, 4.82)	83.92	0.580	0.226	62%
	1000	(0.95, 0.55, 1.98)	53.03	0.151	0.176	76%
	5000	(0.16, 0.09, 0.42)	5.79	0.022	0.000	100%
Slash	100	(21.90, 10.93, 33.79)	45.65	1.625	0.557	10%
($\nu_s = 3.36$)	500	(2.71, 1.82, 5.89)	5.27	0.417	0.447	32%
	1000	(1.49, 0.84, 3.35)	3.38	0.185	0.412	40%
	5000	(0.42, 0.19, 0.74)	0.70	0.176	0.268	54%
Slash	100	(7.17, 5.37, 16.46)	86.38	0.056	0.259	62%
($\nu_s = 1.25$)	500	(2.46, 1.16, 3.94)	40.40	0.021	0.193	72%
	1000	(0.95, 0.67, 1.58)	49.71	0.006	0.200	64%
	5000	(0.24, 0.09, 0.56)	23.65	0.001	0.206	78%

rameters. The MSE of the ν parameters was calculated considering only the replications in which the true model was selected. The MSE of ρ is the mean square error between the posterior estimate of the true model's probability and 1. From Table 2, we can see that, even for small sample sizes, the MSE values of β and σ^2 indicate a very good recovery of the true values. The MSE values of ν_t include some large values for small samples sizes when there is not enough information to estimate precisely the degrees of freedom. The difference in the magnitudes of ν_t and ν_s are explained by the difference in the scale of those parameters. Finally, the last column of the table shows the percentage of times that the correct model was selected (i.e., had the highest posterior probability).

4.2 Study II

This study investigates how the model selection procedure and the parameter estimation is affected in the presence of correlated covariates. We generate data from a model with $e_i \sim \mathcal{T}(0, 1, 3)$, $\beta = (1, 2, -2, 1)$, $\mathbf{X}_i = (1, X_{i1}, X_{i2}, X_{i3})$ and $X_{i3} = 2X_{i2} + \mathcal{N}(0, 0.5)$ which induces an average correlation of 0.9 between the two covariates.

We reproduce 50 replicates of this scenario with different sample sizes $n = 500, 1000, 2000$ and 5000. Table 3 shows the percentage of times that the proposed methodology selects each model and compares it with the other model selection criteria. All the criteria—WAIC, CPO, DIC, EAIC and EBIC, selected the same model in all the replications. It is clear that the traditional criteria have problems to distinguish between models with heavy tails even when the sample size increases, whilst the proposed methodology performs a robust selection specially for large sample sizes where tail information is more abundant.

Figure 1 shows some results regarding the estimation of the regression coefficients. They are quite similar between the proposed methodology and the other model selection criteria. The same, however, does not happen when we look at the estimates for the variance σ^2 (see

Table 3   *Percentage of the times each model was selected for different sample sizes in Study II*

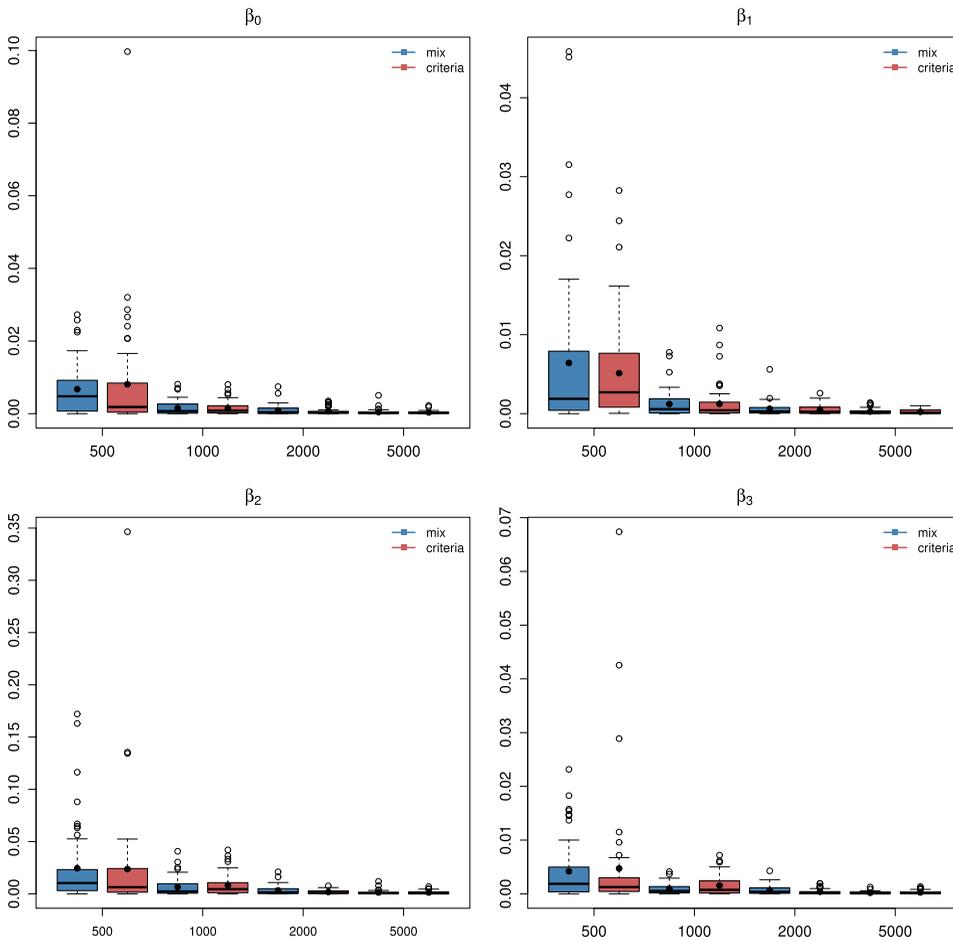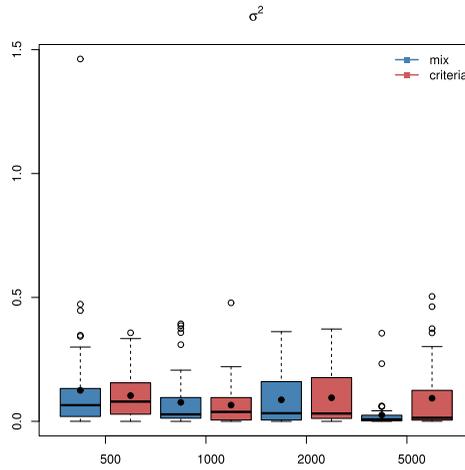| Sample size	Proposed methodology			WAIC/CPO/DIC/EAIC/EBIC		
	Normal	Student-t	Slash	Normal	Student-t	Slash
500	0%	62%	38%	6%	42%	52%
1000	0%	62%	38%	0%	46%	54%
2000	0%	94%	6%	0%	58%	42%
5000	0%	100%	0%	0%	54%	46%

Figure 1   *Study II—boxplots of the mean square error (mse) of the β estimates (posterior mean under the selected model) for the 50 replicates for different sample sizes. Colour blue refers to the proposed methodology and red to the other criteria. The black solid dots represent the mean.*

Figure 2). The poor performance of the other criteria in selecting the correct model is clearly reflected in the estimation of the variance, which is significantly overcome by the respective estimates obtained with the proposed methodology.

4.3 Study III

In this study, the generating distribution for the error term is not a specific distribution as in Sections 4.1 and 4.2, but a mixture of the Normal, Student-t and Slash distributions. More specifically, we consider $e_i \sim 0.1\mathcal{N}(0, 1) + 0.6\mathcal{T}(0, 1, 4.00) + 0.3\mathcal{S}(0, 1, 1.15)$, with

Figure 2 *Study II—boxplots of the mean square error (mse) of the σ^2 estimates (posterior mean under the selected model) for the 50 replicates for different sample sizes. Colour blue refers to the proposed methodology and red to the other criteria. The black solid dots represent the mean.*

Table 4 *Posterior results (mean) for Study III*

Sample size	$\beta^\top = (1, 2, -2)$	$\sigma^2 = 1$	(v_t, v_s)	$\rho = (\rho_1, \rho_2, \rho_3)$
500	(1.018, 2.014, −1.993)	1.265	(3.56, 1.24)	(0.000, 0.786, 0.214)
1000	(1.038, 1.972, −1.981)	0.839	(4.43, 1.51)	(0.000, 0.914, 0.086)
2000	(1.027, 2.012, −2.046)	0.946	(4.61, 1.53)	(0.000, 0.922, 0.078)
5000	(0.985, 1.979, −2.073)	0.918	(4.03, −)	(0.000, 1.000, 0.000)

the same $\mathbf{X}$, β's and σ^2 from Study I. The sample sizes n considered are the same as in Study II.

Again, 50 replications are generated. It is important to notice that our modeling framework to perform robust model selection cannot retrieve the generating model, since we assume that all the residuals must be from the same distribution. Nevertheless, a good fit may still be provided by one of the individual models. Table 4 shows the result of one of the 50 replications. It is clear that the posterior distribution identifies the Student-t distribution as the best candidate model, specially as the sample size increases.

For sample size 2000, Figure 3 shows the fit of the selected model (Student-t) and the other two models, Normal and Slash, fitted individually. It also shows the true generating distribution for the error term. It is clear that, although the posterior distribution is different from the true generating distribution, by definition, it approximates fairly very well the original one.

Table 5 shows that the proposed model consistently chooses the dominating model (Student-t) as the sample size increases. The same does not happen for the other model selection criteria.

5 Application

5.1 AIS

In this section, we introduce a biomedical study from the Australian Institute of Sports (AIS) in 202 athletes (Cook and Weisberg, 1994). To exemplify our modeling, we consider the body
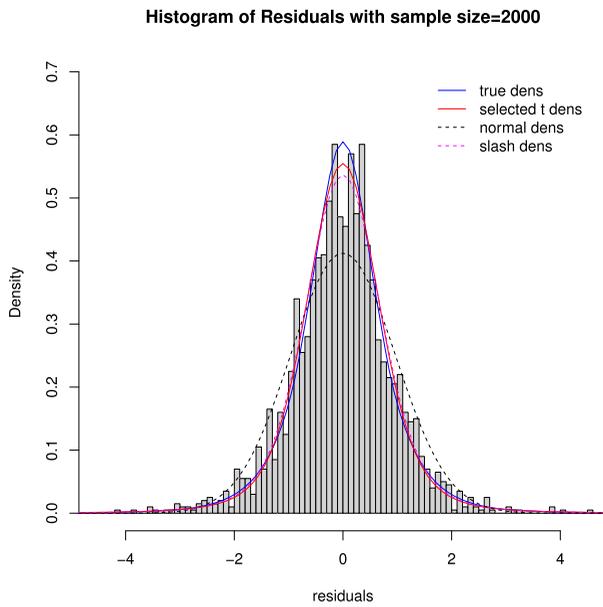
Figure 3 *Study III—residual histogram with true generating model (blue), selected Student-t model (red), Normal model (dashed black), and Slash model (dashed magenta) for sample size* 2000.

Table 5 *Percentage of the times each model was selected for different sample sizes (Study III)*

	Proposed methodology			*WAIC		
Sample size	Normal	Student-t	Slash	Normal	Student-t	Slash
500	0%	70%	30%	6%	50%	44%
1000	0%	98%	2%	4%	46%	50%
2000	0%	100%	0%	0%	62%	38%
5000	0%	100%	0%	0%	52%	48%

*WAIC = all the other criteria (CPO, DIC, EAIC and EBIC) select the same model as WAIC.

Table 6 *Model selection criterion for the fitting of the Normal, Student-t and Slash regression models. LMPL is the the log-marginal pseudo-likelihood*

Models	−LMPL	DIC	EAIC	EBIC	WAIC
Normal	498.497	2976.407	994.142	1000.758	996.971
Student-t	491.623	2935.009	982.059	991.984	983.210
Slash	491.033	2931.636	980.633	990.558	982.049

mass index (BMI) as our response and the percentage of body fat (Bfat) as our covariate. This way, we have a regression model with $\mathbf{X}_{i\cdot} = (1, \text{Bfat}_i)$ for $i = 1, \dots, 202$.

We fit each individual model separately and the proposed mixture model. Results are presented in Tables 6 and 7. Note that, although the Slash model is chosen by all the criteria, and the estimates of the regression coefficients are similar between the individual fit and our model, significant, though not large, differences can be found for the estimates of the variance σ^2. This highlights the model averaging feature of our approach, which is particularly appealing when one of the models is not chosen with very high probability—in this example, $\rho = (0.001, 0.304, 0.695)$.

Table 7 *Posterior results for the BMI analysis with Bfat as covariate for the robust mixture model. The posterior mean, median and standard deviation (Sd) are presented as well as the 95% high posterior density (HPD) interval*

Model	Parameters	Mean	Median	Sd	95% HPD interval
Slash model	β_0	21.810	21.810	0.419	(20.980, 22.620)
	β_1	0.070	0.070	0.028	(0.015, 0.126)
	σ^2	10.093	8.989	3.587	(5.702, 17.940)
	ν_s	1.705	1.612	0.442	(1.110, 2.569)
Slash selected	β_0	21.794	21.799	0.418	(21.022, 22.667)
model	β_1	0.071	0.071	0.028	(0.016, 0.128)
	σ^2	9.200	8.462	2.954	(5.543, 14.765)
	ν_s	1.716	1.628	0.434	(1.111, 2.549)

5.2 WAGE

The wage rate data set presented in Mroz (1987) is used to extend our modeling framework for censored data. The data consist of the wage of 753 married white women, with ages between 30 and 60 years old in 1975. Out of the 753 women considered in this study, 428 worked at some point during that year. When the wives did not work in 1975, the wage rates were set equal to zero. However, it is considered that they may had a cost in that year and, therefore, these observations are considered left-censored at zero. The considered response is Y_i—the wage rate, and the explanatory variables are the wife's age (X_{1i}), years of schooling (X_{2i}), number of children younger than six years old in the household (X_{3i}) and number of children between six and nineteen years old (X_{4i}). Thus, $\mathbf{X}_{i\cdot} = (1, X_{1i}, X_{2i}, X_{3i}, X_{4i})$, $i = 1, \ldots, 753$.

Since the Wage data is censored, we have the following characteristic for our response variables:

$$Y_{\text{obs}_i} = \begin{cases} \kappa_i, & \text{if } Y_i \leq \kappa_i, \\ Y_i & \text{if } Y_i > \kappa_i, \end{cases}$$

with $\kappa_i = 0$.

Suppose that, out of the n responses, C of them are censored as κ_i. From a Bayesian perspective, these observations, $\mathbf{Y}_C = (y_1, \ldots, y_C)$, can be viewed as latent and sampled at each step of the MCMC. Because of the model structure presented in (6)–(9), it is simple to notice that

$$\begin{aligned} (Y_c | Z_j = 1, u_c, \boldsymbol{\beta}, \sigma^2, \nu_j) &\sim \mathcal{TN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \gamma_j u_c^{-1}), \lfloor -\infty, \kappa_c \rfloor, \\ c &= 1, \ldots, C, \end{aligned} \tag{18}$$

where $\mathcal{TN}$ is a truncated Normal distribution with limits $\lfloor -\infty, \kappa_c \rfloor$. Therefore, we simply add a new sampling step in the blocking scheme as

$$(\mathbf{p}, \mathbf{Z}, \mathbf{U}), \qquad \mathbf{Y}_C, \qquad \boldsymbol{\beta}, \qquad \sigma^2, \qquad (\nu_t, \nu_s).$$

This simple extension allows our modeling framework to deal with any kind of censored data, where, for each type of censoring scheme, the new limits of (18) must be calculated.

To obtain our final chain with 100k observations, a Markov Chain of 110k iterations is run and the first 10k observations are discarded for burn-in. The posterior estimate for $\boldsymbol{\rho}$ is (0.000, 0.025, 0.975), which indicates the Slash distribution as the preferred one.

Table 8 summarises the posterior results. Garay et al. (2015) studied this data set from a Bayesian perspective fitting a variety of independent models in the NI family. In their study, the Slash distribution was selected as the preferred one as in our case. Moreover, the posterior

Table 8 *Posterior results for the Wage data analysis. The posterior mean, median and standard deviation (Sd) are presented as well as the 95% high posterior density (HPD) interval*

Parameters	Mean	Median	Sd	95% HPD interval
β_0	-1.174	-1.152	1.408	$(-3.952, 1.523)$
β_1	-0.109	-0.108	0.022	$(-0.155, -0.066)$
β_2	0.646	0.645	0.070	$(0.508, 0.783)$
β_3	-3.114	-3.103	0.387	$(-3.887, -2.381)$
β_4	-0.293	-0.294	0.129	$(-0.539, -0.039)$
σ^2	26.542	24.740	7.843	$(14.784, 42.624)$
ν_s	1.410	1.374	0.207	$(1.110, 1.788)$

mean estimates of the fixed effects parameters in Table 8 are very similar to the ones presented in Garay et al. (2015) as well as the statistical significance of each covariate. The wife's age, the number of children younger than six years old in the household and the number of children between six and nineteen years old tend to decrease the wage rate, while years of schooling tend to increase the salary. The posterior estimates encountered by Garay et al. (2015) for ν_s and ν_t, fitting separate models, were 1.438 and 5.279, respectively, which agree with our results. Our mixture model approach was able to correctly capture the Slash distribution without separately fitting the three models. Moreover, it provides a high computational gain given the high posterior probability of the Slash model.

6 Conclusions and some extensions

Our proposed methodology has shown considerable flexibility to perform model selection for heavy-tailed data explained by covariates under a regression framework. From theoretical arguments, simulation studies and application to real data sets, it is clear that the methodology provides a robust alternative to select the best model instead of relying on model selection criteria which can be unstable (Gelman, Hwang and Vehtari, 2014).

In Section 5.2, we extend the methodology to censored heavy-tailed regression, showing that the extension is straightforward and achieved by adding one simple step to the Gibbs sampler. Also, the extension of the algorithm described in Section 3.1 to include more distributions in the finite mixture is almost direct. Finally, it is clear from our results that this finite mixture idea can be used in a variety of problems where a common parametrisation exists for a family of distributions.

Besides the computational advantage of fitting one general model instead of K separated models, we also emphasise that our robust model selection framework automatically performs multiple comparison between the K models, which gives an advantage if one, instead, prefers to use the Bayes factor performing 2 by 2 comparisons in each individual model. Moreover, our approach also allows the use of model averaging.

Although the proposed methodology enriches the class of traditional censored regression models, we conjecture that it may not provide satisfactory result when the response exhibit asymmetry besides the non-normal behavior. To overcome this limitation, extending the work to account for skewness behavior is also a possibility, for example, by using the scale mixtures of skew-normal (SMSN) distributions proposed in Lachos, Ghosh and Arellano-Valle (2010). Nevertheless, a deeper investigation of those modifications in the parametrisation and implementations is beyond the scope of this paper, but provides stimulating topics for further research. Another possibility of future research is to generalise these modeling framework to linear mixed model, for example, clustered, temporal or spatial dependence. These extensions are being studied in a different manuscript.

Finally, a supplementary material with all the codes and the wage data is available for users who want to apply our methodology (Gonçalves, Prates and Lachos, 2020).

Appendix A: Proof of Lemma 1

The posterior density of $\mathbf{p}$ is given by

$$f(\mathbf{p}|\mathbf{Y} = \mathbf{y}) = \sum_{k=1}^{K} f(\mathbf{p}|\mathbf{Y} = \mathbf{y}, Z_k = 1) P(Z_k = 1|\mathbf{Y} = \mathbf{y}).$$

If we multiply both sides by p_j, integrate with respect to $\mathbf{p}$ and use the fact that $\mathbf{p}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, we get

$$\mathbb{E}[p_j|\mathbf{y}] = \sum_{k=1}^{K} \mathbb{E}[p_j|Z_k = 1] P(Z_k = 1|\mathbf{Y} = \mathbf{y}),$$

which is a weighted average of $\{\mathbb{E}[p_j|Z_k = 1]\}_{k=1}^{K}$ and, therefore, implies that

$$\mathbb{E}[p_j|\mathbf{y}] \in \left(\min_{k}\{\mathbb{E}[p_j|Z_k = 1]\}, \max_{k}\{\mathbb{E}[p_j|Z_k = 1]\}\right).$$

Now note that $(\mathbf{p}|Z_k = 1) \sim \text{Dir}(\alpha_1 + \mathbb{1}\{k = 1\}, \ldots, \alpha_K + \mathbb{1}\{k = K\})$ and $\mathbb{E}[p_j|Z_k = 1]$ is $\frac{\alpha_j}{\alpha_0+1}$ if $j \neq k$ and is $\frac{\alpha_j+1}{\alpha_0+1}$ if $j = k$, where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. This concludes the proof.

Appendix B: Model comparison criteria

The DIC (Spiegelhalter et al., 2002) is a generalisation of the Akaike information criterion (AIC) and is based on the posterior mean of the deviance, which is also a measure of goodness-of-fit. The DIC is defined by

$$\text{DIC} = \overline{\text{D}}(\boldsymbol{\theta}) + \rho_{\text{D}} = 2\overline{\text{D}}(\boldsymbol{\theta}) - \text{D}(\tilde{\boldsymbol{\theta}}),$$

where $\tilde{\boldsymbol{\theta}} = \text{E}[\boldsymbol{\theta}|\mathbf{y}]$, $\overline{\text{D}}(\boldsymbol{\theta})$ is the posterior expectation of the deviance and ρ_{D} is a measure of the effective number of parameters in the model. The effective number of parameters, ρ_{D}, is defined as $\rho_{\text{D}} = \overline{\text{D}}(\boldsymbol{\theta}) - \text{D}(\tilde{\boldsymbol{\theta}})$, with $\overline{\text{D}}(\boldsymbol{\theta}) = -2\text{E}[\log f(\mathbf{y}|\boldsymbol{\theta})|\mathbf{y}]$.

The computation of the integral $\overline{\text{D}}(\boldsymbol{\theta})$ is complex, a good solution can be obtained using the MCMC sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ from the posterior distribution. Thus, we can obtain an approximation of the DIC by first computing the sample posterior mean of the deviations $\overline{\text{D}} = -2\frac{1}{M}\sum_{m=1}^{M} \log f(\mathbf{y}|\boldsymbol{\theta}_m)$ and then $\widehat{\text{DIC}} = 2\overline{\text{D}} - \text{D}(\tilde{\boldsymbol{\theta}})$.

The expected Akaike information criterion (EAIC), and the expected Bayesian information criterion (EBIC) (see discussion at Spiegelhalter et al., 2002) are given by

$$\widehat{\text{EAIC}} = \overline{\text{D}} + 2\vartheta \quad \text{and} \quad \widehat{\text{EBIC}} = \overline{\text{D}} + \vartheta \log(n),$$

respectively, where ϑ is the number of model parameters and can be used for model comparison.

Recently, Watanabe (2010) introduced the Widely Applicable Information Criterion (WAIC). The WAIC is a fully Bayesian approach for estimating the out-of-sample expectation. The idea is to compute the log pointwise posterior predictive density (*lppd*) given by $lppd = \sum_{i=1}^{n} \log(\frac{1}{M}\sum_{m=1}^{M} f(y_i|\boldsymbol{\theta}_m))$, and then, to adjust for overfitting, add a term to correct for effective number of parameters $\rho_{\text{WAIC}} = \sum_{i=1}^{n} V_{m=1}^{M}(\log f(y_i|\boldsymbol{\theta}_m))$, where

$V_{m=1}^M(a) = \frac{1}{M-1} \sum_{m=1}^M (a_m - \bar{a})^2$. Finally, as proposed by Gelman, Hwang and Vehtari (2014), the WAIC is given by

$$\text{WAIC} = -2(lppd - \rho_{\text{WAIC}}).$$

So far, for the DIC, EAIC, EBIC and WAIC, the model that best fits a data set is the model with the smallest value of the criterion.

Another common alternative is the conditional predictive ordinate (*CPO*) approach (Geisser and Eddy, 1979). This statistic is based on the cross validation criterion to compare the models. Let $\mathbf{y} = \{y_1, \ldots, y_n\}$ be an observed sample from $f(\cdot|\boldsymbol{\theta})$. For the ith observation, the CPO_i can be written as:

$$CPO_i = p(y_i|\mathbf{y}_{(-i)}) = \int_{\boldsymbol{\theta} \in \Theta} f(y_i|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)}) d\boldsymbol{\theta} = \left\{ \int_{\boldsymbol{\theta} \in \Theta} \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{f(y_i|\boldsymbol{\theta})} d\boldsymbol{\theta} \right\}^{-1},$$

where $\mathbf{y}_{(-i)}$ is the $\mathbf{y}$ without the ith observation and $\pi(\boldsymbol{\theta}|\mathbf{y})$ denotes the posterior distribution of $\boldsymbol{\theta}$. Thus, the CPO_i has the idea of the leave one out cross validation, where each value is an indicator of the likelihood value given all the other observations. For this reason, low values of CPO_i must correspond to poorly fitted observations. For many models, the analytic calculation of the CPO is not available. However, Dey, Chen and Chang (1997) showed that an harmonic mean approach can be used to do a Monte Carlo approximation of the CPO_i by using a MCMC sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$. Therefore, the CPO_i approximation is given by

$$\widehat{CPO_i} = \left\{ \frac{1}{M} \sum_{m=1}^M \frac{1}{f(y_i|\boldsymbol{\theta}_m)} \right\}^{-1}.$$

Since the CPO_i is defined for each observation, the log-marginal pseudo likelihood (LMPL) given as

$$\text{LMPL} = \sum_{i=1}^n \log(\widehat{CPO_i}),$$

is used to summarise the CPO_i information and the larger the value of LMPL is, the better the fit of the model under consideration.

Supplementary Material

Supplement to "Robust Bayesian model selection for heavy-tailed linear regression using finite mixtures" (DOI: 10.1214/18-BJPS417SUPP; .zip). Supplementary information.

References

Andrews, D. F. and Mallows, S. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B* **36**, 99–102. MR0359122

Basso, R. M., Lachos, V. H., Cabral, C. R. B. and Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis* **54**, 2926–2941. MR2727724 https://doi.org/10.1016/j.csda.2009.09.031

Cabral, C. R. B., Lachos, V. H. and Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis* **56**, 126–142. MR2833042 https://doi.org/10.1016/j.csda.2011.06.026

Carlin, B. P. (2006). Comments to discussion of deviance information criteria for missing data. *Bayesian Analysis* **1**, 675–676. MR2282198 https://doi.org/10.1214/06-BA122A

Chen, M.-H. (2006). Comments to discussion of deviance information criteria for missing data. *Bayesian Analysis* **1**, 677–680.

Choy, S. T. B. and Chan, J. S. K. (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zeland Journal of Statistics* **50**, 135–146. MR2516871 https://doi.org/10.1111/j.1467-842X.2008.00504.x

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley. MR1285353 https://doi.org/10.1002/9780470316863

Dey, D. K., Chen, M. H. and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics* **53**, 1239–1252.

Fernandez, C. and Steel, M. F. J. (1999). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika* **86**, 153–167. MR1688079 https://doi.org/10.1093/biomet/86.1.153

Fonseca, T. C. O., Ferreira, M. A. R. and Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika* **95**, 325–333. MR2521587 https://doi.org/10.1093/biomet/asn001

Galea, M., Paula, G. A. and Cysneiros, F. J. A. (2005). On diagnostics in symmetrical nonlinear models. *Statistics & Probability Letters* **73**, 459–467. MR2187861 https://doi.org/10.1016/j.spl.2005.04.033

Galea, M., Paula, G. A. and Uribe-Opazo, M. (2003). On influence diagnostic in univariate elliptical linear regression models. *Statistical Papers* **44**, 23–45. MR1963364 https://doi.org/10.1007/s00362-002-0132-9

Garay, A. M., Bolfarine, H., Lachos, V. H. and Cabral, C. R. B. (2015). Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. *Journal of Applied Statistics* **42**, 2694–2714. MR3428840 https://doi.org/10.1080/02664763.2015.1048671

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection (Corr: V75 p765). *Journal of the American Statistical Association* **74**, 153–160. MR0529531

Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24**, 997–1016. MR3253850 https://doi.org/10.1007/s11222-013-9416-2

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **85**, 398–409.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (Disc: P189-193). In *Bayesian Statistics, Vol. 4. Proceedings of the Fourth Valencia International Meeting* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 169–188. New York: Clarendon Press [Oxford University Press]. MR1380276

Gonçalves, F. B., Gamerman, D. and Soares, T. M. (2013). Simultaneous multifactor DIF analysis and detection in item response theory. *Computational Statistics & Data Analysis* **59**, 144–160. MR3000048 https://doi.org/10.1016/j.csda.2012.10.011

Gonçalves, F. B., Prates, M. O. and Lachos, V. H. (2020). Supplement to "Robust Bayesian model selection for heavy-tailed linear regression using finite mixtures." https://doi.org/10.1214/18-BJPS417SUPP.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86. MR0039968 https://doi.org/10.1214/aoms/1177729694

Lachos, V. H., Angolini, T. and Abanto-Valle, C. (2011). On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. *Statistical Papers* **52**, 567–590. MR2821058 https://doi.org/10.1007/s00362-009-0270-4

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

Lange, K. L., Little, R. and Taylor, J. (1989). Robust statistical modeling using t distribution. *Journal of the American Statistical Association* **84**, 881–896. MR1134486

Lange, K. L. and Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* **2**, 175–198. MR1272391 https://doi.org/10.2307/1390698

Lin, J.-G. and Cao, C.-Z. (2013). On estimation of measurement error models with replication under heavy-tailed distributions. *Computational Statistics* **28**, 809–829. MR3064480 https://doi.org/10.1007/s00180-012-0330-4

Martins, T. G. and Rue, H. (2013). Prior for flexibility parameters: The Student's t case. Preprint 08/2013, Norwegian University of Science and Technology.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* **55**, 765–799.

Osorio, F., Paula, G. A. and Galea, M. (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics & Data Analysis* **51**, 4354–4368. MR2364450 https://doi.org/10.1016/j.csda.2006.06.004

Pinheiro, J. C., Liu, C. and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* **10**, 249–276. MR1939700 https://doi.org/10.1198/10618600152628059

Raftery, A., Madigan, D. and Volinsky, C. T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics, Vol. 5*, 323–349. New York: Oxford University Press. MR1425413

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability* **7**, 110–120. MR1428751 https://doi.org/10.1214/aoap/1034625254

Rosa, G., Padovani, C. and Gianola, D. (2003). Robust linear mixed models with normal/independent distributions and Bayesian mcmc implementation. *Biometrical Journal* **45**, 573–590. MR1998137 https://doi.org/10.1002/bimj.200390034

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science* **32**, 1–28. MR3634300 https://doi.org/10.1214/16-STS576

Spiegelhalter, D., Best, N., Carlin, B. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

Steel, M. F. J. and Fernandez, C. (1999). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika* **86**, 153–168. MR1688079 https://doi.org/10.1093/biomet/86.1.153

Tierney, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *The Annals of Applied Probability* **8**, 1–9. MR1620401 https://doi.org/10.1214/aoap/1027961031

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571–3594. MR2756194

F. B. Gonçalves
M. O. Prates
Departamento de Estatística
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - DEST/ICEx/UFMG
Belo Horizonte, Minas Gerais, 31270-901
Brazil
E-mail: fbgoncalves@est.ufmg.br

V. H. Lachos
Department of Statistics
University of Connecticut
316 CLAS Building
215 Glenbrook Rd. U-4120
Storrs, Connecticut 06269
USA
E-mail: hlachos@uconn.edu