REVSTAT – Statistical Journal
Volume 20, Number 1, January 2022, 101–115

FLEXIBLE ROBUST MIXTURE REGRESSION MODELING

Authors: Marcus G. Lavagnole Nascimento
– Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
marcus@dme.ufrj.br

Carlos A. Abanto-Valle
– Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
cabantovalle@im.ufrj.br

Received: April 2019 Revised: December 2019 Accepted: December 2019

Abstract:

- This paper provides a flexible methodology for the class of finite mixture of regressions with scale mixture of skew-normal errors (SMSN-FMRM) introduced by [42], relaxing the constraints imposed by the authors during the estimation process. Based on the data augmentation principle and Markov chain Monte Carlo (MCMC) algorithms, a Bayesian inference procedure is developed. A simulation study is implemented in order to understand the possible effects caused by the restrictions and an example with a well known dataset illustrates the performance of the proposed methods.

Keywords:

- *finite mixture of regressions; scale mixture of skew-normal distributions; Markov chain Monte Carlo.*

AMS Subject Classification:

- 62F15, 62J05, 62H30.

1. INTRODUCTION

Finite mixture regression models (FMRM) provide a flexible tool for modeling data that arise from a heterogeneous population, where a single regression model is not enough for capturing the complexities of the conditional distribution of the observed sample given the features. FMRM of Gaussian distributions, using maximum likelihood methods for parameter estimation, have been extensively used in the literature in different fields like marketing [11, 12], economics [10, 21], agriculture [36], psychometrics [28], among others.

From a Bayesian perspective, there is a wide range of nonparametric methods, in particular, methods in which the error follows a mixture of Dirichlet process [27] or a mixture of Polya trees [22]. However, in comparison with these methodologies, the finite mixture of regressions presents the advantage of classifying the observations over the components of the mixture in a natural way. This classification, in a range of applications, is the main topic of interest and provides for practitioners a clear interpretation of the results, besides facilitating the implementation.

Extensions of FMRM of Gaussian distributions have been proposed to broaden the applicability of the model to more general structures like skewed or heavy tailed errors. In this regard, [4] modified the EM algorithm for normal mixtures, replacing the least squares criterion in the M step with a robust one. [33] and [41], in turn, implemented an estimation procedure for finite mixture of linear regression models assuming that the error terms follow a Laplace and a Student-t distribution, respectively. As an attempt to accommodate asymmetric observations, [29] introduced a FMRM based on skew-normal distributions [1].

More recently, as an attractive way to deal with skewness and heavy tails simultaneously, [42] introduced a finite mixture regression model based on scale mixtures of skew-normal distributions [6, SMSN] as follow:

$$f(y_i|\mathbf{x}_i, \boldsymbol{\vartheta}, \boldsymbol{\eta}) = \sum_{j=1}^{G} \eta_j g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j), \tag{1.1}$$

where the probability density function $g(\cdot|\mathbf{x}_i, \boldsymbol{\theta}_j)$ comes from the same member of the SMSN$(\mathbf{x}_i\boldsymbol{\beta}_j + \mu_j, \sigma_j^2, \lambda_j, \nu_j)$ family, $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2, \lambda_j, \nu_j)$ is the specific parametric vector for the component j, $\eta_j > 0$, $j = 1, ..., G$, $\sum_{j=1}^{G} \eta_j = 1$, $\boldsymbol{\vartheta}$ and $\boldsymbol{\eta}$ denote the unknown parameters with $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_G)$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_G)$. However, [42] impose the constraints $\tau_1^2 = \cdots = \tau_G^2$ and $\nu_1 = \cdots = \nu_G$ about the parameters during the estimation procedure in which $\tau_j^2 = \sigma_j^2(1 - \delta_j^2)$ and $\delta_j = \lambda_j/(\sqrt{1 + \lambda_j^2})$.

The aim of this paper, therefore, is to provide a flexible version for the mixture of regressions based on scale mixtures of skew-normal distributions introduced by [42], relaxing the restrictions described above and verifying empirically how our ideas improve the estimation process. Bayesian inference is developed applying ideas like the data augmentation principle, stochastic representation in terms of a random-effects model [2, 23], standard hierarchical representation of a finite mixture model [14] and MCMC methods.

The remainder of the paper is organized as follows. Section 2 is related to the development of a flexible methodology for the mixture regression model based on scale mixture of skew-normal (SMSN-FMRM) distributions from a Bayesian perspective. In order to make comparisons between the methodology proposed in the present work and the one proposed by [42] feasible, Sections 3 and 4 present the analysis of a simulation study and a real dataset respectively. Finally, some concluding remarks and suggestions for future developments are given in Section 5.

2. MIXTURE REGRESSION MODEL BASED ON SCALE MIXTURE OF SKEW-NORMAL DISTRIBUTIONS

2.1. The model

Let $\mathbf{y} = (y_1, ..., y_n)^T$ given $\mathbf{x} = (\mathbf{x}_1^T, ..., \mathbf{x}_n^T)^T$ be a random sample from a G-component mixture model, $\mathbf{x}_i$ is a p-dimensional vector of explanatory variables, and consider a mixture regression model in which the random errors follow a scale mixtures of skew-normal distributions (SMSN-FMRM) as defined by the equation 1.1. Let $\mathbf{S} = (\mathbf{S}_1, ..., \mathbf{S}_n)$ be the allocation vector, i. e., the vector containing the information about in which group the observation y_i of the random variable Y_i is. The indicator variable $\mathbf{S}_i = (S_{i1}, ..., S_{iG})^T$, with

$$S_{ij} = \begin{cases} 1, & \text{if } Y_i \text{ belongs to component } j \\ 0, & \text{otherwise} \end{cases}$$

and $\sum_{j=1}^{G} S_{ij} = 1$. Given the weights vector $\boldsymbol{\eta}$, the latent variables $\mathbf{S}_1, ..., \mathbf{S}_n$ are independent with multinomial distribution

$$p(\mathbf{S}_i|\boldsymbol{\eta}) = \eta_1^{S_{i1}} \eta_2^{S_{i2}} \cdots (1 - \eta_1 - \cdots - \eta_{G-1})^{S_{iG}}.$$

The joint density of $\mathbf{Y} = (Y_1, ..., Y_n)$ and $\mathbf{S} = (\mathbf{S}_1, ..., \mathbf{S}_n)$ is given by

$$f(\mathbf{y}, \mathbf{s}|\mathbf{x}, \boldsymbol{\vartheta}, \boldsymbol{\eta}) = \prod_{j=1}^{G} \prod_{i=1}^{n} \left[\eta_j g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j)\right]^{S_{ij}}.$$

From the stochastic representation in terms of a random-effects model introduced by [2] and [23], a random variable drawn from the scale mixture of skew-normal distributions has a hierarchical representation. Hence, the individual Y_i belonging to the j-th component can be written as

$$\begin{aligned} Y_i|S_{ij} = 1, \mathbf{x}_i, w_i, u_i, \boldsymbol{\theta}_j &\sim N(\mathbf{x}_i\boldsymbol{\beta}_j + \mu_j + \sigma_j\delta_j w_i, k(u_i)\sigma_j\sqrt{1-\delta_j^2}), \\ W_i|S_{ij} = 1, u_i &\sim TN_{[0,+\infty)}(0, k(u_i)), \\ U_i|S_{ij} = 1, \nu_j &\sim h(\cdot; \nu_j), \end{aligned}$$

where $\mu_j = -\sqrt{\frac{2}{\pi}} m_{1,j}\sigma_j\delta_j$, $m_1 = E[U^{-1/2}]$, which corresponds to the regression model where the error distribution has zero mean and hence the regression parameters are all comparable.

Thus, the joint density of $\mathbf{Y}$ and the latent variables $\mathbf{S}$, $\mathbf{W}$ and $\mathbf{U}$ is

$$f(\mathbf{y}, \mathbf{s}, \mathbf{w}, \mathbf{u}|\mathbf{x}, \boldsymbol{\vartheta}, \boldsymbol{\eta}) = \prod_{j=1}^{G} \left[\prod_{i=1}^{n} [\eta_j f(y_i|\boldsymbol{\theta}_j, \mathbf{x}_i, w_i, u_i) f(w_i|u_i) f(u_i|\nu_j)]^{S_{ij}} \right].$$

In this article, $k(U) = U^{-1}$ is used since it leads to good mathematical properties. Without loss of generality, the distributions skew normal [1, SN], skew-t [3, ST] and skew-slash [39, SSL] are considered here, it means that mixing variables are chosen as: $U = 1$, $U \sim G(\frac{\nu}{2}, \frac{\nu}{2})$ and $U \sim Be(\nu, 1)$, where $G(\cdot,\cdot)$ and $Be(\cdot,\cdot)$ indicate the gamma and beta distributions respectively.

As in [17], we introduce a new parameterization in terms of the component-specific parameters $\boldsymbol{\theta}_j^* = (\boldsymbol{\beta}_j, \psi_j, \tau_j^2, \nu_j)$, where $\psi_j = \sigma_j \delta_j$ and $\tau_j^2 = \sigma_j^2(1 - \delta_j^2)$. The original parametric vector $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2, \lambda_j, \nu_j)$, on its turn, is recovered through

$$\lambda_j = \frac{\psi_j}{\tau_j}, \quad \sigma_j^2 = \tau_j^2 + \psi_j^2,$$

since $\psi_j/\tau_j = \sigma_j \delta_j / (\sigma_j \sqrt{1 - \delta_j^2}) = \lambda_j$ and $\tau_j^2 + \psi_j^2 = \sigma_j^2(1 - \delta_j^2) + \sigma_j^2 \delta_j^2 = \sigma_j^2$.

2.2. Bayesian inference

Performing a Bayesian analysis, an important step is the priors distributions selection. In the context of finite mixture models, in particular, mixture regression models, a special attention on these choices is quite relevant since it is not possible to choose an improper prior because it implies in an improper posterior density [16]. In addition, as pointed by [25], it is recommended to avoid be as "noninformative as possible" by choosing large prior variances because the number of components is highly influenced by the prior choices. Consequently, in order to avoid identifiability problems, it was adopted the hierarchical priors introduced by [31] for mixtures of normal distributions to reduce sensitivity with respect to choosing the prior variances.

Hence, considering the parametric vector $\boldsymbol{\theta}_j^* = (\boldsymbol{\beta}_j, \psi_j, \tau_j^2, \nu_j)$ for an arbitrary mixture component j, the prior set was specified as: $\boldsymbol{\eta} \sim D(e_0, ..., e_0)$, $(\boldsymbol{\beta}_j, \psi_j)|\tau_j^2 \sim N_{p+1}(\mathbf{b}_0, \tau_j^2 \mathbf{B}_0)$, $\tau_j^2|C_0 \sim IG(c_0, C_0)$ and $C_0 \sim G(h_0, H_0)$, where e_0, $\mathbf{b}_0 \in \mathbb{R}^{(p+1)}$, $\mathbf{B}_0 \in \mathbb{R}^{(p+1)\times(p+1)}$, c_0, h_0 and H_0 are known hyper parameters, $N_q(\cdot,\cdot)$, $D(\cdot, ..., \cdot)$ and $IG(\cdot,\cdot)$ indicate the q-variate normal, the Dirichlet and inverse gamma distributions. Considering the parameter ν priors, $p(\nu_j) \propto \nu_j/(\nu_j + d)^3 \mathbb{1}_{(2,40)}(\nu_j)$ [26] and $\nu_j \sim G_{(1,40)}(\alpha, \gamma)$, where α and γ are known hyper parameters and $G_A(\cdot,\cdot)$ denotes the truncated gamma on set A, are specified for the ST-FMRM and SSL-FMRM respectively.

The Bayesian approach for estimating the parameters uses the data augmentation principle [35], which considers $\mathbf{W}, \mathbf{U}$ and $\mathbf{S}$ as latent unobserved variables. The joint posterior density of parameters and latent variables can be written as

$$p(\boldsymbol{\vartheta}^*, \boldsymbol{\eta}, \mathbf{w}, \mathbf{u}, \mathbf{s}|\mathbf{y}, \mathbf{x}) \propto \left\{ \prod_{j=1}^{G} \left[\prod_{i=1}^{n} [\eta_j f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i) f(w_i|u_i) f(u_i|\nu_j)]^{S_{ij}} \right] p(\boldsymbol{\theta}_j^*) \right\} p(\boldsymbol{\eta}),$$

where $p(\boldsymbol{\theta}_j^*) = p(\boldsymbol{\beta}_j, \psi_j|\tau_j^2)p(\tau_j^2|C_0)p(C_0)p(\nu_j)$ and $\boldsymbol{\vartheta}^* = (\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_G^*)$. In light of the data augmentation technique, conditional on the allocation vector $\mathbf{S}$, the parameters estimation may be executed independently for each parametric component $\boldsymbol{\theta}_j^*$ and for the weights distribution $\boldsymbol{\eta}$. As a consequence, the full conditionals of the parameters and the latent unobserved variables for the mixture regression models based on the SMSN distributions are written as follows:

$$p(\boldsymbol{\eta}|\mathbf{s}) \propto p(\mathbf{s}|\boldsymbol{\eta})p(\boldsymbol{\eta}) \tag{2.1}$$

$$p(w_i|S_{ij} = 1, \cdots) \propto \left[f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i)f(w_i|u_i)\right]^{S_{ij}}, \tag{2.2}$$

$$p(u_i|S_{ij} = 1, \cdots) \propto \left[f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i)f(w_i|u_i)f(u_i|\nu_j)\right]^{S_{ij}}, \tag{2.3}$$

$$p(\boldsymbol{\beta}_j, \psi_j|\cdots) \propto \prod_{\{i:S_{ij}=1\}} f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i)p(\boldsymbol{\beta}_j, \psi_j|\tau_j^2), \tag{2.4}$$

$$p(\tau_j^2|\cdots) \propto \prod_{\{i:S_{ij}=1\}} f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i)p(\tau_j^2|C_0), \tag{2.5}$$

$$p(C_0|\cdots) \propto \prod_{j=1}^{G} p(\tau_j^2|C_0)p(C_0), \tag{2.6}$$

$$p(\nu_j|\cdots) \propto \prod_{\{i:S_{ij}=1\}} f(u_i|\nu_j)p(\nu_j). \tag{2.7}$$

Additional details about the derivations of the full conditionals are available in Appendix A.1.

In furtherance of making Bayesian analysis feasible for parameter estimation in the SMSN-FMRM class of models, random samples from the posterior distributions of $(\boldsymbol{\vartheta}, \boldsymbol{\eta}, \mathbf{w}, \mathbf{u}, \mathbf{s})$ given $(\mathbf{y}, \mathbf{x})$ are drawn through Monte Chain Monte Carlo simulation methods. Algorithm 1 describes the sampling scheme from the full conditionals distributions of the parameters and the latent unobserved variables.

Algorithm 1. MCMC for finite mixture of scale mixtures of skew-normal.

```
1. Set $k = 1$ and get starting values for $\mathbf{S}^{(0)}$, $(\boldsymbol{\theta}_1^{*(0)}, ..., \boldsymbol{\theta}_G^{*(0)})$, $\boldsymbol{\eta}^{(0)}$, $\mathbf{w}^{(0)}$ and $\mathbf{u}^{(0)}$;
2. Parameter simulation conditional on the classification $\mathbf{S}^{(k-1)}$:
    2.1. Sample $\boldsymbol{\eta}^{(k)}$ from $p(\boldsymbol{\eta}|\mathbf{s}^{(k-1)})$;
    2.2. Sample the component latent variables $w_i^{(k)}$ and $u_i^{(k)}$, $i = 1, ..., n$, from the full
         conditionals (2.2)–(2.3) and the component parameters $\boldsymbol{\beta}_j^{(k)}, \psi_j^{(k)}, \tau_j^{2(k)}, \nu_j^{(k)}$,
         $j = 1, ..., G$, from the full conditionals (2.4)–(2.7).
3. Sample $S_i^{(k)}$ independently for each $i = 1, ..., n$ from

    $\Pr(S_{il} = 1|y_i, \mathbf{x}_i, \boldsymbol{\vartheta}^*) = \frac{g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_l^*) \Pr(S_{il} = 1|\boldsymbol{\vartheta}^*)}{\sum_{j=1}^{G} g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j^*) \Pr(S_{ij} = 1|\boldsymbol{\vartheta}^*)}.$

4. Set $k = k + 1$ and repeat the steps $2, 3$ and $4$ until convergence is achieved.
```

Introduced by [30] into the mixture models background, the term *label switching* refers to the invariance of the mixture likelihood function under relabeling the components. Considering the maximum likelihood estimation, where we are looking for the corresponding modes of the likelihood function, label switching is not an object of interest. From the Bayesian point of view, however, it is a topic of concern because the labeling of the unobserved categories changes during the sample process of the mixture posterior distribution. Post-processed the MCMC, in order to deal with the label switching problem, the Kullback–Leibler algorithm [34] is applied over this paper.

3. SIMULATION STUDY

In this section, a simulated scenario is considered for three purposes:

- **(i)** verifying if the true parameter values are recovered accurately by using the methodology described on Section 2;
- **(ii)** comparing the estimation performance of the unconstrained and constrained models;
- **(iii)** formulating a sensitivity analysis study to the hyperparameters specification.

To that end, datasets are artificially generated as follow:

$$\begin{cases} Y_i = \mathbf{x}_i\boldsymbol{\beta}_1 + \varepsilon_1, \;\; S_{i1} = 1, \\ Y_i = \mathbf{x}_i\boldsymbol{\beta}_2 + \varepsilon_2, \;\; S_{i2} = 1, \end{cases}$$

where S_{ij} is a component indicator of Y_i with $\Pr(S_{ij} = 1) = \eta_j$, $j = 1, 2$, $\mathbf{x}_i = (1, x_{i1})$, $i = 1, ..., n$. Finally, ε_1 and ε_2 follow a distribution in the SMSN family. According to this procedure, 100 random samples of size $n = 500$ are generated from the SN-FMRM, ST-FMRM and SSL-FMRM models with the following parameter values: $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11})^T = (20, 0)^T$, $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12})^T = (-4, 3)^T$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\lambda_1 = 0$, $\lambda_2 = 5$, $\eta_1 = 0.4$, $\eta_2 = 0.6$. In addition, for the ST-FMRM and SSL-FMRM models, $\boldsymbol{\nu} = (\nu_1, \nu_2) = (8, 3)$ and $\boldsymbol{\nu} = (6, 2)$, respectively.

During the estimation process for the SMSN-FMRM models, the unconstrained version proposed in this paper and the constrained version of [42] were considered and it was adopted the four different hyperparameters specifications described in Table 1 for both. For each sample, 20000 iterations from Algorithm 1 were conducted. The first 10000 were discarded as a burn-in period. In order to reduce the autocorrelation within the successive values of the simulated chain, it was required a thin equals to 10. Finally, based on 1000 records, the posterior mean were obtained.

Table 1: Prior sets hyperparameters specifications.

Specification	e_0	$\mathbf{b}_0$	$\mathbf{B}_0$	c_0	h_0	H_0	d	α	γ
P_1	4	(0,0,0)	Diag(100,100,100)	0.01	0.01	0.01	$4/(1+\sqrt{2})$	6	1
P_2	4	(0,0,0)	Diag(10,10,10)	0.01	0.01	0.01	$4/(1+\sqrt{2})$	6	1
P_3	4	(0,0,0)	Diag(100,100,100)	2.5	0.75	$\frac{0.75}{0.5s_y^2}$	$4/(1+\sqrt{2})$	6	1
P_4	4	(0,0,0)	Diag(100,100,100)	0.01	0.01	0.01	$9/(1+\sqrt{2})$	4	1

Table 2: MSE and coverage percentage in parenthesis for the MCMC estimates based on the 100 samples from the SMSN-FMRM.

Parameters		SN-FMRM		ST-FMRM		SSL-FMRM	
		$\tau_1^2 \neq \tau_2^2$	$\tau_1^2 = \tau_2^2$	$\tau_1^2 \neq \tau_2^2, \nu_1 \neq \nu_2$	$\tau_1^2 = \tau_2^2, \nu_1 = \nu_2$	$\tau_1^2 \neq \tau_2^2, \nu_1 \neq \nu_2$	$\tau_1^2 = \tau_2^2, \nu_1 = \nu_2$
$\beta_{0,1}$	P_1	0.0143(1.00)	0.0148(0.99)	0.0221(0.99)	0.0271(0.96)	0.0234(1.00)	0.0458(0.97)
	P_2	0.0222(0.98)	0.0225(0.97)	0.0311(0.98)	0.0293(1.00)	0.0426(0.98)	0.0497(0.97)
	P_3	0.0253(1.00)	0.0272(0.98)	0.0286(0.99)	0.0364(0.94)	0.0312(0.99)	0.0434(0.97)
	P_4	—	—	0.0228(0.99)	0.0284(0.98)	0.0378(0.99)	0.0499(0.97)
$\beta_{1,1}$	P_1	0.0000(0.97)	0.0001(0.98)	0.0001(0.96)	0.0001(0.96)	0.0001(0.93)	0.0001(0.94)
	P_2	0.0001(0.95)	0.0001(0.93)	0.0002(0.95)	0.0002(0.92)	0.0002(0.95)	0.0002(0.92)
	P_3	0.0001(0.94)	0.0001(0.89)	0.0002(0.95)	0.0002(0.95)	0.0001(0.97)	0.0001(0.94)
	P_4	—	—	0.0001(0.99)	0.0001(0.94)	0.0002(0.91)	0.0002(0.92)
$\beta_{0,2}$	P_1	0.0142(0.94)	0.0170(0.97)	0.0454(0.84)	0.0545(0.84)	0.1869(0.91)	0.1932(0.84)
	P_2	0.0156(0.94)	0.0204(0.99)	0.0351(0.94)	0.0697(0.83)	0.1461(0.91)	0.2866(0.65)
	P_3	0.0157(0.93)	0.0163(0.96)	0.0369(0.91)	0.0477(0.90)	0.1502(0.91)	0.1502(0.90)
	P_4	—	—	0.0316(0.90)	0.0429(0.88)	0.1708(0.96)	0.1170(0.95)
$\beta_{1,2}$	P_1	0.0000(0.94)	0.0001(0.96)	0.0001(0.97)	0.0001(0.99)	0.0001(0.90)	0.0001(0.92)
	P_2	0.0000(0.96)	0.0001(0.95)	0.0001(0.97)	0.0001(0.99)	0.0001(0.99)	0.0001(0.99)
	P_3	0.0000(0.95)	0.0000(0.98)	0.0001(0.97)	0.0001(0.96)	0.0001(0.98)	0.0001(0.99)
	P_4	—	—	0.0001(0.90)	0.0001(0.98)	0.0001(0.95)	0.0001(0.96)
σ_1^2	P_1	0.0956(0.95)	0.9566(0.34)	0.0523(0.99)	0.0756(0.98)	0.0943(0.99)	0.4890(0.77)
	P_2	0.1233(0.37)	0.0823(0.89)	0.2337(0.17)	0.0335(0.98)	0.1374(0.42)	0.0612(0.96)
	P_3	0.1385(0.90)	0.8905(0.29)	0.0600(0.99)	0.1026(0.95)	0.1015(0.98)	0.4174(0.84)
	P_4	—	—	0.0593(0.98)	0.1311(0.95)	0.0348(1.00)	0.2948(0.89)
σ_2^2	P_1	0.1760(0.91)	0.5010(0.62)	1.3980(0.84)	0.7495(0.86)	2.5937(0.84)	1.5439(0.85)
	P_2	0.2358(0.82)	1.9505(0.09)	0.7075(0.90)	0.5111(0.94)	2.3527(0.72)	1.8668(0.74)
	P_3	0.2076(0.89)	0.5872(0.55)	0.9655(0.87)	0.7671(0.85)	2.3347(0.80)	1.3687(0.88)
	P_4	—	—	0.9523(0.91)	0.7049(0.88)	1.4102(0.92)	0.9150(0.88)
λ_1	P_1	0.0648(1.00)	2.4698(0.54)	0.0622(1.00)	0.6500(0.85)	0.1128(1.00)	1.7950(0.62)
	P_2	0.0122(1.00)	0.0820(1.00)	0.0128(1.00)	0.0589(1.00)	0.0142(1.00)	0.0965(1.00)
	P_3	0.1465(1.00)	2.3544(0.48)	0.0781(0.99)	0.6287(0.89)	0.1241(1.00)	1.2843(0.78)
	P_4	—	—	0.0620(1.00)	0.7134(0.83)	0.0547(1.00)	1.3541(0.72)
λ_2	P_1	1.6120(0.96)	6.1614(0.00)	3.1617(0.98)	6.3709(0.00)	2.3855(0.94)	4.1220(0.04)
	P_2	1.8628(0.52)	14.1231(0.00)	0.7802(0.92)	10.7803(0.00)	0.5909(0.94)	8.9064(0.00)
	P_3	1.0375(0.86)	6.6829(0.00)	0.9961(0.96)	6.6883(0.01)	0.7847(0.97)	4.6046(0.02)
	P_4	—	—	3.2051(0.96)	6.4518(0.00)	1.8351(1.00)	4.7205(0.01)
η_1	P_1	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
	P_2	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
	P_3	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
	P_4	—	—	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
η_2	P_1	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
	P_2	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
	P_3	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
	P_4	—	—	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)	0.0000(1.00)
ν_1	P_1	—	—	3.6146(0.97)	19.8156(0.14)	1.3108(1.00)	7.2708(0.73)
	P_2			11.0161(1.00)	14.5225(0.25)	1.9544(1.00)	3.2971(0.89)
	P_3			5.1347(0.96)	20.0174(0.12)	1.7550(1.00)	8.0294(0.66)
	P_4			5.9129(0.97)	19.0113(0.16)	4.9511(0.97)	10.5920(0.46)
ν_2	P_1	—	—	1.0621(1.00)	0.9324(0.92)	3.3490(0.86)	4.7921(0.65)
	P_2			1.7930(0.95)	3.4168(0.63)	4.4666(0.69)	14.4817(0.18)
	P_3			1.0371(0.99)	1.3937(0.92)	3.0896(0.79)	3.7640(0.67)
	P_4			1.9016(0.95)	2.2195(0.91)	1.0183(0.96)	1.7100(0.84)

Table 2 shows the mean squared error (MSE) and coverage percentage for the MCMC estimates based on the 100 samples, in which the coverage percentage is the proportion of the time that the credibility interval contains the true value of interest. The first important fact that is possible to observe from the table is that with high probability the true parameter values are recovered, particularly if the unconstrained methodology is considered. Comparing the unconstrained methodology proposed in this work with the restricted version, there is a significant improvement on the MSE and coverage percentage, specially for the scale, symmetry and kurtosis parameters. Taking λ_2, for example, the coverage percentage is zero or almost zero in all cases and the MSE is more than ten times greater in specific cases.

Taking the hyperparameters specification P_1 as a baseline, a sensitivity analysis study is built. The specification P_2 consists in reducing the values of $\mathbf{B}_0$, and almost no impact on the results of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ is observed, however, looking to the unconstrained model, a significant decrease in the coverage percentage for the scale and symmetry parameters is noticed. The specification P_3 follows [31], the results are similar compared with the P_1 ones, but there is a gain on the MSE for λ_2 in the heavy tailed distributions and unconstrained model. Lastly, a degradation on the MSE for $\boldsymbol{\nu}$ is noted when the changes made in P_4 for d, α and γ are assumed.

4. EMPIRICAL ANALYSIS

In order to explore the interval memory hypothesis and the partial matching hypothesis, [9] designed an experiment in which a pure fundamental tone with electronically generated overtones added was played to a trained musician. The overtones were determined by a stretching ratio, corresponding to the harmonic pattern usually heard in traditional definite pitched instruments. The musician was asked to tune an adjustable tone to the octave above the fundamental tone and 150 trials were recorded as the ratio of the adjusted tone to the fundamental.

This dataset has been analysed in many articles which explored the mixture of linear regression framework [13, 38, 24]. More recently, [41] fitted a robust mixture regression model using the t-distribution and [42], a robust mixture regression based on the SMSN class of distributions. Conducive to make comparisons with the results in [42] possible, the methods proposed in this paper are applied to the tone perception data.

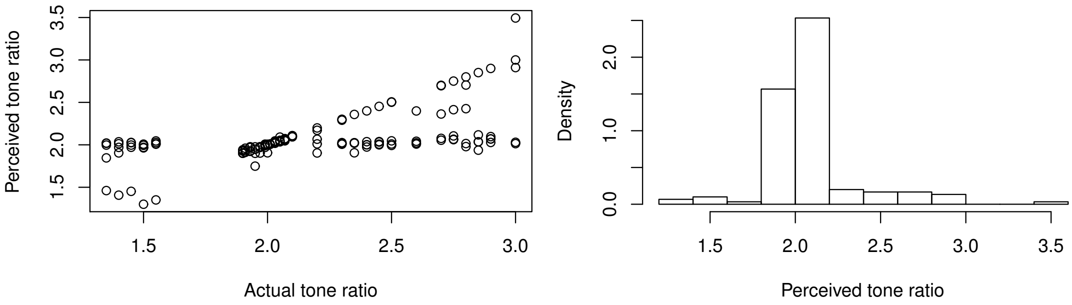

Figure 1: Tone perception data scatterplot and histogram.

Considering the estimation process for the SN-FMRM, ST-FMRM and SSL-FMRM, the hyperparameters specification P_3 presented in Table 1 was chosen. From the MCMC scheme described in Section 2.2, 20000 iterations were drawn. The first 10000 draws were discarded as a burn-in period. In order to reduce the autocorrelation between successive values of the simulated chain, only every 10th values of the chain were stored and from the resulting 1000 we calculated the posterior estimates. It is worth mentioning that, because of the two well defined components, the label switching problem was not identified.

Table 3: Estimation results for fitting the SMSN-FMRM under analysis to the tone data. First row: maximum a posteriori. Second row: 95% high posterior density credibility interval. Third row: convergence test Z-scores.

Parameters	N-FMRM	T-FMRM	SL-FMRM	SN-FMRM	ST-FMRM	SSL-FMRM
$\beta_{0,1}$	1.9107 (1.8586,1.9569) −1.2777	1.9325 (1.8832,1.9771) 0.0250	1.9167 (1.8703,1.9689) −0.0878	1.9044 (1.8532,1.9664) −0.7281	1.9291 (1.8757,1.9846) 0.2527	1.9147 (1.8679,1.9703) −0.2157
$\beta_{1,1}$	0.0457 (0.0243,0.0688) 1.0459	0.0387 (0.0175,0.0595) −0.2666	0.0425 (0.0196,0.0649) −0.2561	0.0447 (0.0205,0.0672) 0.4088	0.0365 (0.0151,0.0618) −0.2404	0.0431 (0.0202,0.0641) 1.1750
$\beta_{0,2}$	−0.0188 (−0.2054,0.2059) −0.7409	0.0153 (−0.0186,0.0704) 1.2719	0.0477 (−0.0317,0.1359) −0.6623	0.0208 (−0.2457,0.2495) −0.1860	0.0136 (−0.0358,0.0849) −0.9211	0.0194 (−0.1075,0.1276) −1.1195
$\beta_{1,2}$	0.9893 (0.9070,1.0802) 0.3946	0.9928 (0.9669,1.0079) −1.5883	0.9745 (0.9304,1.0061) 1.0861	0.9796 (0.8899,1.0971) 0.3949	0.9869 (0.9615,1.0141) 0.0831	0.9729 (0.9228,1.0212) 1.7043
σ_1^2	0.0027 (0.0019,0.0036) −0.4449	0.0020 (0.0012,0.0029) 1.7121	0.0019 (0.0014,0.0029) −1.5685	0.0028 (0.0019,0.0042) 0.6334	0.0021 (0.0013,0.0035) 0.4865	0.0022 (0.0015,0.0034) 1.8521
σ_2^2	0.0173 (0.0105,0.02676) 0.1553	0.0005 (0.0002,0.0010) 1.5999	0.0011 (0.0004,0.0026) −0.9927	0.0269 (0.0127,0.0621) 1.1119	0.0009 (0.0003,0.0024) 0.2782	0.0032 (0.0008,0.0141) −0.2783
λ_1	—	—	—	0.0800 (−0.7634,0.7341) −0.3516	−0.0972 (−0.8113,0.5411) −1.6838	0.0186 (−0.7843,0.5725) 0.1532
λ_2	—	—	—	1.0045 (−1.7427,2.7095) −0.7809	−0.3676 (−1.3333,0.0821) −1.3453	−1.2264 (−2.6623,0.3076) 0.4094
η_1	0.6908 (0.6030,0.7733) −0.2578	0.5606 (0.4700,0.6516) 1.6709	0.5805 (0.4820,0.6876) 1.7261	0.7045 (0.6103,0.7901) 0.3072	0.5691 (0.4538,0.6564) 0.4209	0.6296 (0.5223,0.7383) −0.6418
η_2	0.3091 (0.2266,0.3969) 0.2578	0.4393 (0.3483,0.5299) −1.6709	0.4194 (0.3123,0.5179) −1.7261	0.2954 (0.2098,0.3896) −0.3072	0.4308 (0.3435,0.5461) −0.4209	0.3703 (0.2616,0.4776) 0.6418
ν_1	—	3.0280 (2.0015,24.7743) 0.7383	5.8212 (2.1481,11.7897) −1.0693	—	5.5252 (2.0678,21.7135) 1.5870	6.2337 (3.1571,11.5048) 1.4383
ν_2	—	2.1162 (2.0001,2.6451) 0.8492	1.4630 (1.4000,1.7509) 0.9953	—	2.1281 (2.0000,2.6977) −1.8332	1.5494 (1.4000,3.0780) −0.3276
WAIC$_1$ WAIC$_2$	−263.9868 −288.2918	**−349.6941** **−372.0548**	−301.1313 −329.3142	−253.9442 −290.7716	−329.4679 −361.6124	−283.6500 −330.5183

Table 3 contains the maximum a posteriori estimation of the parameters of the models under analysis: SN-FMRM, ST-FMRM and SSL-FMRM in addition to their corresponding 95% high posterior density credibility interval and the Z-scores for the convergence test introduced by [20]. Additionally, in order to compare the fit of the different models, two versions proposed by [19] of the Watanabe–Akaike Information Criterion [40, WAIC] were computed, indicating that the T-FMRM has the best fitting, conclusion that goes in opposition to the ST-FMRM model observed by [42]. More details about these criteria are available in Appendix A.2. Figure 2 illustrates the scatterplots of the dataset with the six fitted models and the equivalent 95% high posterior density credibility intervals.

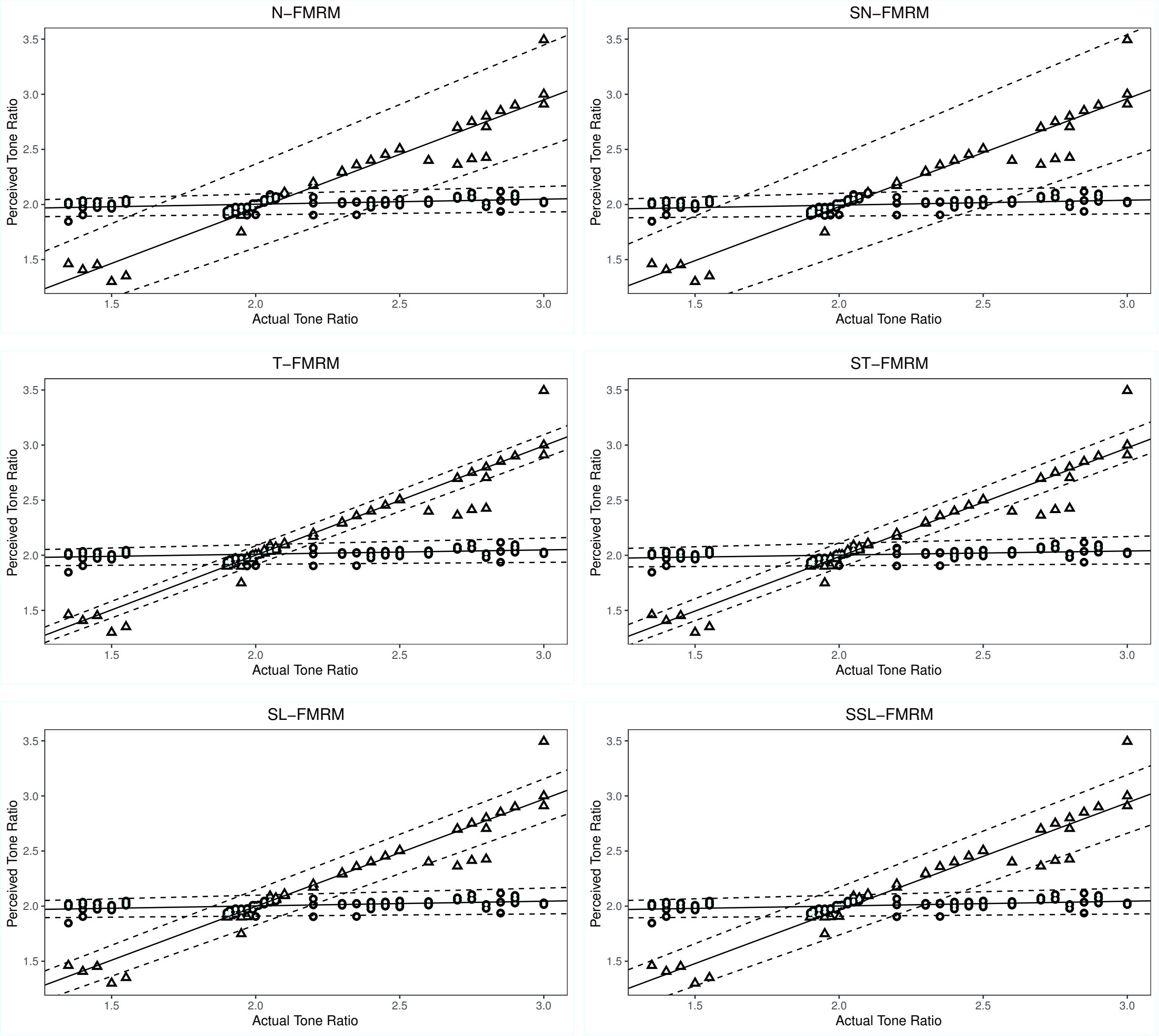

Figure 2: Tone perception data scatterplot and the fitted SMSN-FMRM models.

In comparison with [42], the coefficients $\boldsymbol{\beta}$ estimates are quite similar. However, for the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$, in line with the results observed on the previous section, the estimates diverge. [42] outcomes point to the presence of asymmetry for at least of one the components when the SN-FMRM, ST-FMRM and SSL-FMRM are considered. Nevertheless, as Figure 3 illustrates, when the flexible version proposed in this paper is applied, it is possible to verify that the introduction of a skewness parameter is not effective considering the dataset under analysis.

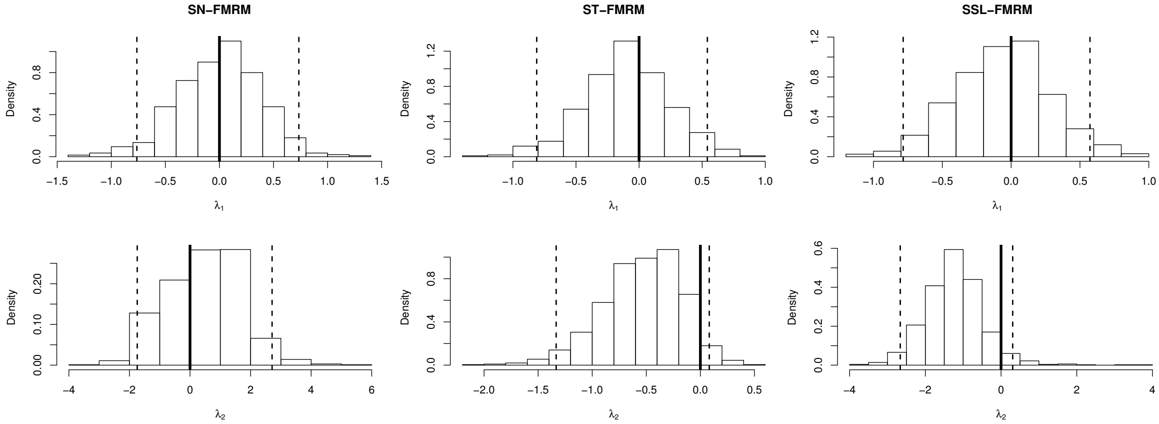

Figure 3: Skewness parameters posterior samples.

5. CONCLUSION

In this work a flexible Bayesian methodology is developed for the mixture regression models based on scale mixtures of skew-normal distributions proposed by [42] with the aim of understanding the possible effects caused by the restrictions commonly imposed in the context of robust mixture regression modeling. The tone perception data and an artificial dataset are analysed in order to verify the advantages that the additional flexibility introduced by the methodology developed in this article has. In fact, this paper presents divergent results in comparison with [42] and the empirical analysis illustrates the possible effects of imposing constraints for this class of models.

Extensions of the contributions made in this article are possible. First, the number of components might be consider as an unknown quantity of interest, estimating it in a full Bayesian framework. Also the proposed methods may be extended to multivariate settings, such as the recent proposals of [18] for mixtures of multivariate Student-t distributions and to models capable to deal with longitudinal data as discussed in [37]. Contemplating extensions able to deal with nonlinear effects of the covariates [7, 8, 5] is also a stimulating topic for further research.

A. APPENDIX

A.1. Mixture regression based on scale mixtures of skew-normal full conditional distributions

Considering the SN-FMRM model and assuming $\mathbf{F}_{n\times(p+1)} = (\mathbf{x}\ \mathbf{w})$, for each $j = 1, ..., G$, construct a matrix $\mathbf{F}_j \in \mathbb{R}^{N_j\times(p+1)}$, $N_j = \sum_{i=1}^n S_{ij}$. Similarly, construct an observation matrix $\mathbf{y}_j \in \mathbb{R}^{N_j\times 1}$. Hence, by the Bayes theorem, the full conditionals are:

- $\boldsymbol{\eta}|\mathbf{s} \sim D(e_0 + N_1, ..., e_0 + N_G)$;
- $(\boldsymbol{\beta}_j, \psi_j)|\mathbf{s}, \mathbf{y}, \mathbf{w}, \tau_k^2 \sim N_{p+1}(\mathbf{b}_j, \mathbf{B}_j)$,

  $\mathbf{B}_j = \left(\frac{1}{\tau_j^2}\mathbf{B}_0^{-1} + \frac{1}{\tau_j^2}(\mathbf{F}_j{}^T\mathbf{F}_j)\right)^{-1}$,

  $\mathbf{b}_j = \mathbf{B}_j\left(\frac{1}{\tau_j^2}\mathbf{B}_0^{-1}\mathbf{b}_0 + \frac{1}{\tau_j^2}(\mathbf{F}_j{}^T(\mathbf{y}_k - \mu_k))\right)$;
- $\tau_j^2|\mathbf{s}, \mathbf{y}, \mathbf{w}, C_0, \boldsymbol{\beta}_j, \psi_j \sim IG(c_j, C_j)$,

  $c_j = c_0 + \frac{N_j}{2} + \frac{1}{2}$,

  $C_j = C_0 + \frac{(\mathbf{y}_j - \mathbf{F}_j\boldsymbol{\beta}_j^* - \mu_j)^T(\mathbf{y}_j - \mathbf{F}_j\boldsymbol{\beta}_j^* - \mu_j) + (\boldsymbol{\beta}_j^* - \mathbf{b}_0)^T\mathbf{B}_0^{-1}(\boldsymbol{\beta}_j^* - \mathbf{b}_0)}{2}$;
- $C_0|\tau_1^2, ..., \tau_G^2 \sim G(h, H)$,

  $h = h_0 + Gc_0$,

  $H = H_0 + \sum_{j=1}^G \frac{1}{\tau_j^2}$;

where $\boldsymbol{\beta}_j^* = (\boldsymbol{\beta}_j\ \psi_j)^T$. Considering now the latent variable $\mathbf{W}$:

- $W_i|S_{ij} = 1, y_i, \boldsymbol{\beta}_j, \psi_j, \tau_j^2 \sim TN_{[0,+\infty)}(a, A)$,

  $a = \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta}_j - \mu_j)\psi_j}{\tau_j^2 + \psi_j^2}$,

  $A = \frac{\tau_j^2}{\tau_j^2 + \psi_j^2}$.

For the ST-FMRM and the SSL-FMRM models the full conditionals are almost the same, the difference is that $\mathbf{F}$ is replaced by $\mathbf{F}^w_{n\times(p+1)} = (\sqrt{\mathbf{u}}\mathbf{x}\ \sqrt{\mathbf{u}}\mathbf{w})$ and $\mathbf{y}$, by $\mathbf{y}^w = \sqrt{\mathbf{u}}\mathbf{y}$, where $\sqrt{\mathbf{u}}$ is the square root element by element. Considering now the latent variable $\mathbf{W}$:

- $W_i|S_{ij} = 1, y_i, u_i, \boldsymbol{\beta}_j, \psi_j, \tau_j^2 \sim TN_{[0,+\infty)}(a, A/u_i)$.

Lastly, for the latent variable $\mathbf{U}$ and the parameters ν:

- Skew-t:

  $U_i|S_{ij} = 1, y_i, w_i, \nu_j, \boldsymbol{\beta}_j, \psi_j, \tau_j^2 \sim G\left(\frac{\nu_j}{2} + 1, \frac{\nu_j}{2} + \frac{(y_i - \mu_j - \mathbf{x}_i\boldsymbol{\beta}_j - \psi_j w_i)^2}{2\tau_j^2} + \frac{w_i^2}{2}\right)$;
- Skew-slash:

  $U_i|S_{ij} = 1, y_i, w_i, \nu_j, \boldsymbol{\beta}_j, \psi_j, \tau_j^2 \sim G_{(0,1)}\left(\nu_j + 1, \frac{(y_i - \mu_j - \mathbf{x}_i\boldsymbol{\beta}_j - \psi_j w_i)^2}{2\tau_j^2} + \frac{w_i^2}{2}\right)$,

  $\nu_j|\mathbf{s}, \mathbf{u} \sim G_{(2,40)}(\alpha + N_j, \gamma - \sum_{i:S_{ij}=1} u_i)$.

For the degrees of freedom in skew-t is not possible to find a closed form to the full conditionals, so a Metropolis–Hastings step is required. To sample ν_j, $j = 1, ..., G$ a normal log random walk proposal was used

$$\log(\nu_j^{new} - 2) \sim N(\log(\nu_j - 2), c_{\nu_j}) \tag{A.1}$$

with adaptive width parameter c_{ν_j} [32]. The proposal was shifted away from 0, as it is advisable to avoid values for ν_j that are close to 0, see [15].

A.2. Watanabe–Akaike information criterion

Define the predictive accuracy of the fitted model to data as

$$p(\mathbf{y}) = \sum_{i=1}^{n} \log \int f(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

To compute this predictive density, it is possible to evaluate the expectation using draws from the usual posterior simulations:

$$\overline{p(\mathbf{y})} = \sum_{i=1}^{n} \log\left(\frac{1}{T}\sum_{t=1}^{T} f(y_i|\boldsymbol{\theta}^{(t)})\right).$$

Introduced by [40], the Watanabe–Akaike information criterion (WAIC) consists on the posterior predictive density in addition to a correction for effective number of parameters to adjust for overfitting. [19] describes two adjustments. The first one is a difference:

$$WAIC_1^* = 2\sum_{i=1}^{n}\left(\log\left(E_{(\boldsymbol{\theta}|\mathbf{y})} f(y_i|\boldsymbol{\theta})\right) - E_{(\boldsymbol{\theta}|\mathbf{y})}\left(\log(f(y_i|\boldsymbol{\theta}))\right)\right),$$

which can be computed from simulations by replacing the expectations by averages over the posterior draws, it means,

$$\overline{WAIC_1^*} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{T}\sum_{t=1}^{T} f(y_i|\boldsymbol{\theta}^{(t)})\right) - \frac{1}{T}\sum_{t=1}^{T}\log f(y_i|\boldsymbol{\theta}^{(t)})\right).$$

The second is based on the variance of individual terms in the log predictive density summed over the n data observations:

$$WAIC_2^* = \sum_{i=1}^{n} \mathrm{var}_{(\boldsymbol{\theta}|\mathbf{y})}\left(\log f(y_i|\boldsymbol{\theta})\right).$$

In practice, the posterior variance of the log predictive density for each data point y_i, that is, $V_{t=1}^{T} \log f(y_i|\theta^{(t)})$, where $V_{t=1}^{T}$ is the sample variance, $V_{t=1}^{T} a_{(t)} = \frac{1}{T-1}\sum_{t=1}^{T}(a_{(t)} - \bar{a})^2$. Summing over all the data observations, the effective number of parameters is:

$$\overline{WAIC_2^*} = \sum_{i=1}^{n} V_{t=1}^{T}\left(\log f(y_i|\theta^{(t)})\right).$$

Finally, either $WAIC_1^*$ or $WAIC_2^*$ are applied as a bias correction:

$$WAIC_q = -2(p(\mathbf{y}) - WAIC_q^*). \tag{A.2}$$

REFERENCES

[1] Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**(2), 171–178.

[2] Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones, *Statistica*, **46**(2), 199–208.

[3] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution, *Journal of the Royal Statistical Society, Series B*, **65**(2), 367–389.

[4] Bai, X.; Yao, W. and Boyer, J.E. (2012). Robust fitting of mixture regression models, *Computational Statistics and Data Analysis*, **56**(7), 2347–2359.

[5] Baldacchino, T.; Worden, K. and Rowson, J. (2017). Robust nonlinear system identification: Bayesian mixture of experts using the t-distribution, *Mechanical Systems and Signal Processing*, **85**, 977–992.

[6] Branco, M.D. and Dey, D.K. (2001). A general class of multivariate skew-elliptical distributions, *Journal of Multivariate Analysis*, **79**(1), 99–113.

[7] Carvalho, A.X. and Tanner, M.A. (2006). Modeling nonlinearities with mixtures-of-experts of time series models, *International Journal of Mathematics and Mathematical Sciences*, **2006**, 1–22.

[8] Chamroukhi, F. (2016). Robust mixture of experts modeling using the t-distribution, *Neural Networks*, **79**, 20–36.

[9] Cohen, E.A. (1984). Some effects of inharmonic partials on interval perception, *Music Perception*, **1**(3), 323–349.

[10] Cosslett, S.R. and Lee, L.F. (1985). Serial correlation in latent discrete variable models, *Journal of Econometrics*, **27**(1), 79–97.

[11] DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression, *Journal of Classification*, **5**, 249–282.

[12] DeSarbo, W.S.; Wedel, M.; Vriens, M. and Ramaswamy, V. (1992). Latent class metric conjoint analysis, *Marketing Letters*, **3**(3), 273–288.

[13] DeVeaux, R.D. (1989). Mixtures of linear regressions, *Computational Statistics and Data Analysis*, **8**(3), 227–245.

[14] Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society, Series B*, **56**(2), 363–375.

[15] Fernandez, C. and Steel, M.F.J. (1999). Multivariate Student-t regression models: pitfalls and inference, *Biometrika*, **86**(1), 153–167.

[16] Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, New York.

[17] Fruhwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions, *Biostatistics*, **11**(2), 317–336.

[18] Galimberti, G. and Soffritti, G. (2014). A multivariate linear regression analysis using finite mixtures of t distributions, *Computational Statistics and Data Analysis*, **71**, 138–150.

[19] Gelman, A.; Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models, *Statistics and Computing*, **24**, 997–1016.

[20] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *Bayesian Statistics*, **4**, 169–193.

[21] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, **57**(2), 357–384.

[22] Hanson, T. and Johnson, W.O. (2002). Modeling regression error with a mixture of Polya trees, *Journal of the American Statistical Association*, **97**(460), 1020–1033.

[23] Henze, N. (1986). A probabilistic representation of the skew-normal distribution, *Scandinavian Journal of Statistics*, **13**(4), 271–275.

[24] Hunter, D.R. and Young, D.S. (2012). Semiparametric mixtures of regressions, *Journal of Nonparametric Statistics*, **24**(1), 19–38.

[25] Jennison, C. (1997). Discussion of the paper by Richardson and Green, *Journal of the Royal Statistical Society, Series B*, **59**(4), 778–779.

[26] Juárez, M.A. and Steel, M.F.J. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions, *Journal of Business & Economic Statistics*, **28**(1), 52–66.

[27] Kuo, L. and Mallick, B.K. (1997). Bayesian semiparametric inference for the accelerated failure time model, *Canadian Journal of Statistics*, **25**(4), 457–472.

[28] Liu, M.; Hancock, G.R. and Harring, J.R. (2011). Using finite mixture modeling to deal with systematic measurement error: a case study, *Journal of Modern Applied Statistical Methods*, **10**(1), 249–261.

[29] Liu, M. and Lin, T.I. (2014). A skew-normal mixture regression model, *Educational and Psychological Measurement*, **74**(1), 139–162.

[30] Redner, R.A. and Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, **26**(2), 195–239.

[31] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society, Series B*, **59**(4), 731–792.

[32] Shaby, B.A. and Wells, M.T. (2010). *Exploring an adaptive Metropolis algorithm*, Technical Report, Duke University, Department of Statistical Science.

[33] Song, W.; Yao, W. and Xing, Y. (2014). Robust mixture regression model fitting by Laplace distribution, *Computational Statistics and Data Analysis*, **71**, 128–137.

[34] Stephens, M. (2000). Dealing with label switching in mixture models, *Journal of the Royal Statistical Society, Series B*, **62**(4), 795–809.

[35] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**(398), 528–540.

[36] Turner, T.R. (2000). Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions, *Applied Statistics*, **49**(3), 371–384.

[37] Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population, *Journal of the American Statistical Association*, **91**(433), 217–221.

[38] Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions, *Statistics and Computing*, **12**, 315–330.

[39] Wang, J. and Genton, M.G. (2006). The multivariate skew-slash distribution, *Journal of Statistical Planning and Inference*, **136**(1), 209–220.

[40] 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.

[41] Yao, W.; Wei, Y. and Yu, C. (2014). Robust mixture regression using the t-distribution, *Computational Statistics and Data Analysis*, **71**, 116–127.

[42] Zeller, C.B.; Cabral, C.R.B. and Lachos, V.H. (2016). Robust mixture regression modeling based on scale mixtures of skew-normal distributions, *TEST*, **25**, 375–396.