Flexible Robust Mixture Regression Modeling

Marcus G. Lavagnole Nascimento†, Carlos A. Abanto-Valle† and Victor H. Lachos⋆

† Department of Statistics, Federal University of Rio de Janeiro, Caixa Postal 68530, CEP: 21945-970, Rio de Janeiro, Brazil

⋆ Department of Statistics, University of Connecticut, U-4120, Storrs, CT 06269, USA

Abstract

This paper describes a flexible methodology for the class of finite mixture of regressions with scale mixture of skew-normal errors (SMSN-MRM) introduced by Zeller et al. (2016). Bayesian inference based on the data augmentation principle is derived and a Markov chain Monte Carlo (MCMC) algorithm is developed. These procedures are proposed with the aim of understanding the possible effects caused by the restrictions commonly imposed in the context of robust mixture regression modeling. In order to make the comparisons between the results possible, the Tone Perception data is analysed.

Keywords: Finite mixture of regressions, scale mixture of skew-normal distributions, Markov chain Monte Carlo.

1 Introduction

Finite mixture of regression models (FMRM) enable investigating the association between variables coming from several unknown latent homogeneous groups. First introduced under the titles "switching regression" or "clusterwise linear regression" (Quandt, 1972; Spath, 1979), FMRM of Gaussian distributions are frequently applied in areas like marketing (DeSarbo and Cron, 1988; DeSarbo et al., 1992) and economics (Cosslett and Lee, 1985; Hamilton, 1989). In order to model properly data sets arising from a class or several classes with heavy tails observations, Song et al. (2014) and Yao et al. (2014) proposed a robust estimation procedure for mixture linear regression models assuming that the error terms follow, respectively, a Laplace and a Student-t distribution. As an attempt to accommodate asymmetric observations, Liu and Lin (2014) proposed a version of the FMRM based on skew-normal (Azzalini, 1985) errors.

More recently, as an attractive way to deal with both skewness as well as heavy tails, Zeller et al. (2016) proposed a mixture regression model based on scale mixtures of skew-normal distributions (Branco and Dey, 2001, SMSN) as follow:

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

where $g(\cdot|\mathbf{x}_i, \boldsymbol{\theta}_j)$ denotes the SMSN$(\mathbf{x}_i\boldsymbol{\beta}_j+\mu_j, \sigma_j^2, \lambda_j, \nu_j)$ probability density function, $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2, \lambda_j, \nu_j)$, the specific parametric vector for the component j, $\eta_j \geq 0$, $j = 1, \ldots, G$, $\sum_{j=1}^{G} \eta_j = 1$, $\boldsymbol{\vartheta}$ and $\boldsymbol{\eta}$ denote the unknown parameters with $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_G)$. Nevertheless, Zeller et al. (2016) impose the constraints $\gamma_1^2 = \ldots = \gamma_G^2$ and $\nu_1 = \ldots = \nu_G$ about the parameters during the estimation process in which $\gamma_j^2 = \sigma_j^2 - \sigma_j^2\delta_j^2$ and $\delta_j = \lambda_j/(\sqrt{1+\lambda_j^2})$.

The aim of this paper, therefore, is to propose a flexible version for the mixture of regressions based on scale mixtures of skew-normal distributions introduced by Zeller et al. (2016) and make an empirical analysis about the possible effects caused by the restrictions imposed by the previous authors. Towards this end, Bayesian inference is developed using the ideas of the data augmentation principle, the stochastic representation in terms of a random-effects model (Azzalini, 1986; Henze, 1986) and the standard hierarchical representation of a finite mixture model introduced by Diebolt and Robert (1994).

The remainder of the paper is organized as follows. Section 2 is related to the development of a flexible methodology for the mixture regression models based on scale mixture of skew-normal (SMSN-MRM) distributions from a Bayesian perspective. Section 3 is devoted to a real data set application and comparison among the results obtained by the present work and Zeller et al. (2016). Finally, some concluding remarks and suggestions for future developments are given in Section 4.

2 Mixture regression model based on scale mixtures of skew-normal distributions

2.1 The model

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be a random sample from a G-component mixture model ($G > 1$), $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$, 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-MRM) as defined by the equation 1. Introducing the allocation vector $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_n)$, 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 $\mathbf{S}_i = (S_{i1}, \ldots, S_{iG})^\top$, with

$$S_{ij} = \begin{cases} 1, & \text{if } Y_i \text{ belongs to component } j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

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

$$p(\mathbf{S}_i|\boldsymbol{\eta}) = \eta_1^{S_{i1}} \eta_2^{S_{i2}} \ldots (1 - \eta_1 - \cdots - \eta_{G-1})^{S_{iG}}. \tag{3}$$

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

$$f(\mathbf{y}, \mathbf{s}|\mathbf{x}, \boldsymbol{\vartheta}) = \prod_{j=1}^{G} \prod_{i=1}^{n} [\eta_j g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j)]^{S_{ij}}. \tag{4}$$

From the stochastic representation, a random variable Y_i 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

$$\begin{aligned} Y_i|S_{ij}=1,\mathbf{x}_i,w_i,u_i,\boldsymbol{\theta}_j &\sim N(\mathbf{x}_i\boldsymbol{\beta}_j+\mu_j+\sigma_j\delta_j w_i, k(u_i)\sigma_j\sqrt{1-\delta_j^2}), \\ W_i|S_{ij}=1,u_i &\sim TN_{[0,+\infty)}(0,k(u_i)), \\ U_i|S_{ij}=1,\nu_j &\sim h(\cdot;\nu_j), \end{aligned} \tag{5}$$

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

$$f(\mathbf{y},\mathbf{s},\mathbf{w},\mathbf{u}|\mathbf{x},\boldsymbol{\vartheta},\boldsymbol{\eta})=\prod_{j=1}^{G}\left[\prod_{i=1}^{n}[\eta_j f(y_i|\boldsymbol{\theta}_j,\mathbf{x}_i,w_i,u_i)f(w_i|u_i)f(u_i|\nu_j)]^{S_{ij}}\right]p(\mathbf{s}|\boldsymbol{\eta}). \tag{6}$$

Along the following sections, the restriction to the case in that $k(U)=U^{-1}$ is made, since it leads to good mathematical properties. Without loss of generality, the distributions skew normal (Azzalini, 1985, SN), skew-t (Azzalini and Capitanio, 2003, ST) and skew-slash (Wang and Genton, 2006, SSL) are studied, 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.

Last but not least, following Fruhwirth-Schnatter and Pyne (2010), a parameterization in terms of $\boldsymbol{\theta}_j^*=(\boldsymbol{\beta}_j,\psi_j,\tau_j^2,\nu_j)$, where $\psi_j=\sigma_j\delta_j$ and $\tau_j^2=\sigma_j^2(1-\delta_j^2)$, is applied for the scale mixtures of skew-normal distributions. The original parametric vector $\boldsymbol{\theta}_j=(\boldsymbol{\beta}_j,\sigma_j^2,\lambda_j,\nu_j)$ is recovered through

$$\lambda_j=\frac{\psi_j}{\tau_j},\quad \sigma_j^2=\tau_j^2+\psi_j^2. \tag{7}$$

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 (Fruhwirth-Schnatter, 2006). In addition, as noticed by Jennison (1997), 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. For these reasons, as in Fruhwirth-Schnatter and Pyne (2010), it was adopted the hierarchical

priors introduced by Richardson and Green (1997) 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: $\boldsymbol{\eta} \sim D(e_0, \ldots, e_0)$, $(\boldsymbol{\beta}_j, \psi_j)|\tau_j^2 \sim N_{p+1}(\mathbf{b}_0, \tau_j^2\mathbf{B}_0)$, $\tau_j^2|C_0 \sim IG(c_0, C_0)$ and $C_0 \sim G(g_0, G_0)$, where e_0, $\mathbf{b}_0 \in \Re^2$, $\mathbf{B}_0 \in \Re^{2\times 2}$, c_0, g_0 and G_0 are known hyper parameters, $N_q(\cdot,\cdot)$, $D(\cdot,\ldots,\cdot)$ and $IG(\cdot,\cdot)$ indicate the q-variate normal, the dirichlet and inverse gamma distributions. Considering the parameter ν priors, $p(\nu_j) \propto \nu_j/(\nu_j + d)^3 \mathbb{1}_{(2,40)}(\nu_j)$ (Juárez and Steel, 2010) and $\nu_j \sim G_{(2,40)}(\alpha, \gamma)$, 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-MRM and SSL-MRM respectively.

The joint posterior density of parameters and latent unobservable variables can be written as

$$p(\boldsymbol{\vartheta}, \boldsymbol{\eta}, \mathbf{w}, \mathbf{u}, \mathbf{s}|\mathbf{y}, \mathbf{x}) \propto \left\{ \prod_{j=1}^{G} \left[\prod_{i=1}^{n} \left[\eta_j f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i) f(w_i|u_i) f(u_i|\nu_j)\right]^{S_{ij}} \right] p(\boldsymbol{\theta}_j^*) \right\} p(\mathbf{s} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}), \quad (8)$$

where $p(\boldsymbol{\theta}_j^*) = p(\boldsymbol{\beta}_j, \psi_j|\tau_j^2)p(\tau_j^2|C_0)p(C_0)p(\nu_j)$. As expressed in Tanner and Wong (1987), 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 unobservable variables for the mixture regression models based on the SMSN distributions are written as follows:

$$\begin{aligned}
p(\boldsymbol{\eta}|\mathbf{s}) &\propto p(\mathbf{s}|\boldsymbol{\eta})p(\boldsymbol{\eta}) && (9)\\
p(w_i|S_{ij}=1,\cdots) &\propto \left[f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i) f(w_i|u_i)\right]^{S_{ij}}, && (10)\\
p(u_i|S_{ij}=1,\cdots) &\propto \left[f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i) f(w_i|u_i) f(u_i|\nu_j)\right]^{S_{ij}}, && (11)\\
p(\boldsymbol{\beta}_j, \psi_j|\cdots) &\propto \prod_{\{i:S_{ij}=1\}} f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i) p(\boldsymbol{\beta}_j, \psi_j|\tau_j^2), && (12)\\
p(\tau_j^2|\cdots) &\propto \prod_{\{i:S_{ij}=1\}} f(y_i|\boldsymbol{\theta}_j^*, \mathbf{x}_i, w_i, u_i) p(\tau_j^2|C_0), && (13)\\
p(C_0|\cdots) &\propto \prod_{j=1}^{G} p(\tau_j^2|C_0)p(C_0), && (14)\\
p(\nu_j|\cdots) &\propto \prod_{\{i:S_{ij}=1\}} f(u_i|\nu_j)p(\nu_j). && (15)
\end{aligned}$$

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

In furtherance of making Bayesian analysis feasible for parameter estimation in the SMSN-MRM 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 unobservable variables.

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

1 Set $t = 1$ *and get starting values for* $\mathbf{S}^{(0)}$, $(\boldsymbol{\theta}_1^{*(0)}, \ldots, \boldsymbol{\theta}_G^{*(0)})$, $\boldsymbol{\eta}^{(0)}$, $\mathbf{w}^{(0)}$ *and* $\mathbf{u}^{(0)}$*;*

2 Parameter simulation conditional on the classification $\mathbf{S}^{(t-1)}$*:*

2.1 Sample $\boldsymbol{\eta}^{(t)}$ *from* $p(\boldsymbol{\eta}|\mathbf{s}^{(t-1)})$*;*

2.2 Sample the component latent variables $w_i^{(t)}$ *and* $u_i^{(t)}$, $i = 1, \ldots, n$, *from the full conditionals* (11)-(12) *and the component parameters* $\boldsymbol{\beta}_j^{*(t)}, \psi_j^{*(t)}, \tau_j^{2^{*(t)}}, \nu_j^{*(t)}$, $j = 1, \ldots, G$, *from the full conditionals* (13)-(16)*.*

3 Sample $S_i^{(t)}$ *independently for each* $i = 1, \ldots, n$ *from*

$$Pr(S_i = l|y_i, \mathbf{x}_i, \boldsymbol{\vartheta}) = \frac{g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j^*)Pr(S_i = l|\boldsymbol{\vartheta})}{\sum_{j=1}^{G} g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_j^*)Pr(S_i = j|\boldsymbol{\vartheta})}. \tag{16}$$

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

3 Application

In order to explore the interval memory hypothesis and the partial matching hypothesis, Cohen (1984) 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 data set has been analyzed in many articles which explored the mixture of linear regression framework (DeVeaux, 1989; Viele and Tong, 2002; Hunter and Young, 2012). More recently, Yao

et al. (2014) fitted a robust mixture regression model using the $t-$distribution and Zeller et al. (2016), a robust mixture regression based on the SMSN class of distributions. Conducive to make comparisons with the results in Zeller et al. (2016) possible, the methods proposed in this paper are applied to the tone perception data. Additionally, in order to compare the fit of the different models considered, we compute two classical comparison criteria, the Akaike Information Criterion (Akaike, 1974, AIC) and the Bayesian Information Criterion (Schwarz, 1978, BIC), and two versions proposed by Gelman et al. (2014) of the Bayesian criteria known as Watanabe-Akaike Information Criterion (Watanabe, 2010, WAIC).

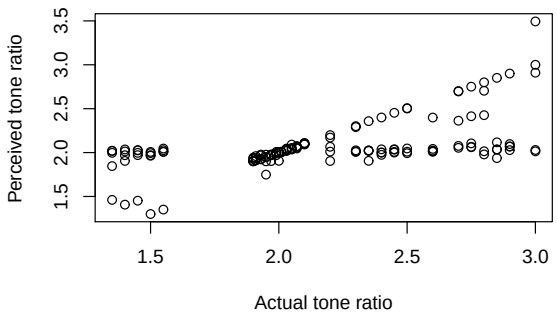

Figure 1: Tone perception data scatter plot.

Considering the estimation process for the SN-MRM, ST-MRM and SSL-MRM, the priors hyperparameters set was specified as: $e_0 = 4$, $\mathbf{b}_0 = (0, 0, 0)$, $\mathbf{B}_0 = Diag(100, 100, 100)$, $c_0 = 0.01$, $g_0 = 0.01$, $G_0 = 0.01$. For the ST-MRM, $d = 4/(1+\sqrt{4})$ was chosen and, for the SSL-MRM, $\alpha = 6$ and $\gamma = 0.8$ were specified. A MCMC simulation for 50000 iterations was drawn, the first 10000 draws were discarded as a burn-in period, and then the next 40000 were recorded. In order to reduce the autocorrelation between successive values of the simulated chain, only every 40th values of the chain were stored. With the resulting 1000 we calculated the posterior estimates.

Table 1 contains the maximum a posteriori estimation of the parameters of the models under analysis: SN-MRM, ST-MRM and SSL-MRM, besides their corresponding 95% high posterior density credibility interval. It is important to mention that, because of the two well defined com-

ponents, the label switching (Redner and Walker, 1984) problem was not identified. Furthermore, we computed the BIC, AIC, WAIC_1 and WAIC_2 as models comparison criteria. The criteria values indicate that the T-MRM has the best fitting result followed by the ST-MRM model.

Table 1: Estimation results for fitting the SMSN-MRM under analysis to the tone data.

Parameters	N-MRM		T-MRM		SN-MRM		ST-MRM		SSL-MRM	
	MODE	95%	MODE	95%	MODE	95%	MODE	95%	MODE	95%
$\beta_{0,1}$	1.9088	(1.8604,1.9585)	1.9330	(1.8829,1.9826)	1.9036	(1.8564,1.9604)	1.9313	(1.8793,1.9907)	1.9118	(1.8653,1.9713)
$\beta_{1,1}$	0.0459	(0.0234,0.0680)	0.0363	(0.0153,0.0602)	0.0450	(0.0226,0.0670)	0.0375	(0.0177,0.0643)	0.0457	(0.0203,0.0637)
$\beta_{0,2}$	-0.0126	(-0.2468,0.2053)	0.0191	(-0.0221,0.0516)	-0.0055	(-0.2553,0.2119)	0.0167	(-0.0276,0.0804)	0.0150	(-0.1491,0.1450)
$\beta_{1,2}$	0.9876	(0.9018,1.0929)	0.9903	(0.9750,1.0084)	0.9829	(0.8950,1.0981)	0.9879	(0.9625,1.0096)	0.9757	(0.9129,1.0440)
σ_1^2	0.0025	(0.0019,0.0035)	0.0020	(0.0012,0.0031)	0.0028	(0.0020,0.0043)	0.0023	(0.0014,0.0037)	0.0024	(0.0017,0.0038)
σ_2^2	0.0183	(0.0121,0.0311)	0.0004	(0.0002,0.0008)	0.0239	(0.0143,0.0546)	0.0008	(0.0003,0.0021)	0.0085	(0.0025,0.0271)
λ_1	-	-	-	-	0.0840	(-0.8727,0.7990)	-0.0269	(-0.6846,0.5915)	0.0480	(-0.7889,0.7278)
λ_2	-	-	-	-	0.5222	(-1.7761,1.9262)	-0.3730	(-1.2133,0.3347)	-1.8254	(-3.4831,0.7764)
η_1	0.7108	(0.6045,0.7835)	0.5704	(0.4579,0.6467)	0.7026	(0.6208,0.7955)	0.5675	(0.4507,0.6549)	0.6426	(0.5379,0.7520)
η_2	0.2891	(0.2165,0.3955)	0.4296	(0.3532,0.5421)	0.2974	(0.2045,0.3792)	0.4325	(0.3451,0.5493)	0.3574	(0.2480,0.4621)
ν_1	-	-	2.8313	(2.0018,17.0252)	-	-	5.4843	(2.0016,29.6809)	7.9826	(3.2876,14.2028)
ν_2	-	-	2.1167	(2.0000,2.7508)	-	-	2.1196	(2,0000,2.7728)	3.0652	(2.0005,7.4982)
BIC	-246.2410		**-326.4319**		-232.4607		-302.4382		-240.5561	
AIC	-267.3155		**-353.5277**		-259.5565		-335.5551		-273.6730	
$WAIC_1$	-265.3233		**-354.0518**		-254.7645		-331.6707		-269.6907	
$WAIC_2$	-289.9573		**-378.4551**		-289.7004		-368.9228		-313.3396	

In comparison with Zeller et al. (2016), the ST-MRM model presented the best fitting performance. In general, the coefficients $\boldsymbol{\beta}$ estimates are in line with the obtained by Zeller et al. (2016), however, the results for the scale, skewness and degrees of freedom parameters are quite different, mainly, if we consider the skewness parameter. These divergences are caused by the restrictions imposed by Zeller et al. (2016). Figure 2 illustrates that the introduction of the skewness parameter is unnecessary considering the data set under analysis, fact that is not observed for the models under constraints proposed by Zeller et al. (2016). Hence, it is possible to affirm that the more flexible estimation process introduced in this work contradicts the results observed by Zeller et al. (2016) for the tone perception data set.

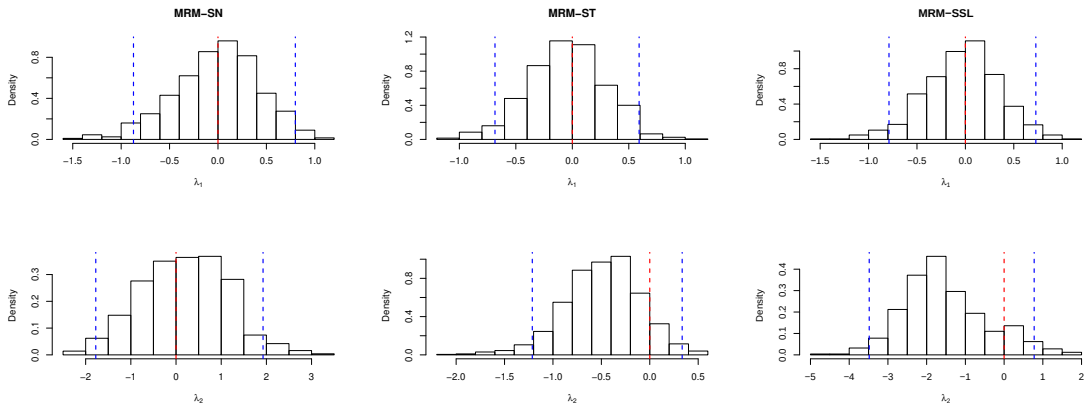

Figure 2: Skewness parameters posterior samples.

4 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 Zeller et al. (2016) 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 is analysed in order to verify the advantages that the additional flexibility introduced by the methodology developed in this article have. In fact, this paper provided divergent results in comparison with Zeller et al. (2016) and brought an empirical illustration about the possible effects of imposing constraint for this class of

models.

An interesting extension which will be pursued in a future research is to develop fully Bayesian inference, i. e., to consider the number of components as an unknown quantity of interest. Also the proposed methods can be extended to multivariate settings, such as the recent proposals of Galimberti and Soffritti (2014) for mixtures of multivariate Student-t distributions and to models capable to deal with longitudinal data as discussed in Verbeke and Lesaffre (1996).

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

Considering the FM-SN model and assuming $\mathbf{F}_{n\times(p+1)} = (\mathbf{x}\ \mathbf{w})$, for each $k = 1, \ldots, K$, construct a matrix $\mathbf{F}_k \in \Re^{N_k\times(p+1)}$, $N_k = \sum_{i=1}^{n} S_{ik}$. Similarly, construct an observation matrix $\mathbf{y}_k \in \Re^{N_k\times 1}$. Hence, by the Bayes theorem, the full conditionals are

- $\boldsymbol{\eta}|\mathbf{s} \sim D(e_0 + N_1, \ldots, e_0 + N_K)$;
- $(\boldsymbol{\beta}_k, \psi_k)|\mathbf{s}, \mathbf{y}, \mathbf{w}, \tau_k^2 \sim N_{p+1}(\mathbf{b}_k, \mathbf{B}_k)$;

  $\mathbf{B}_k = \left(\frac{1}{\tau_k^2}\mathbf{B}_0^{-1} + \frac{1}{\tau_k^2}(\mathbf{F}_k'\mathbf{F}_k)\right)^{-1}$

  $\mathbf{b}_k = \mathbf{B}\left(\frac{1}{\tau_k^2}\mathbf{B}_0^{-1}\mathbf{b}_0 + \frac{1}{\tau_k^2}(\mathbf{F}_k'(\mathbf{y}_k - \mu_k))\right)$
- $\tau_k^2|\mathbf{s}, \mathbf{y}, \mathbf{w}, C_0, \boldsymbol{\beta}_k, \psi_k \sim IG(c_k, C_k)$;

  $c_k = c_0 + \frac{N_k}{2} + \frac{1}{2}$

  $C_k = C_0 + \frac{(\mathbf{y}_k - \mathbf{F}_k\boldsymbol{\beta}_k^* - \mu_k)'(\mathbf{y}_k - \mathbf{F}_k\boldsymbol{\beta}_k^* - \mu_k) + (\boldsymbol{\beta}_k^* - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\boldsymbol{\beta}_k^* - \mathbf{b}_0)}{2}$
- $C_0|\tau_1^2, \ldots, \tau_K^2 \sim G(g, G)$.

  $g = g_0 + Kc_0$

  $G = G_0 + \sum_{k=1}^{K}\frac{1}{\tau_k^2}$

where $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_k\ \psi_k)'$. Considering now the latent variable $\mathbf{W}$

- $W_i|S_{ik}=1, y_i, \boldsymbol{\beta}_k, \psi_k, \tau_k^2 \sim TN_{[0,+\infty)}(a, A)$;

  $a = \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta}_k - \mu_k)\psi_k}{\tau_k^2+\psi_k^2}$

  $A = \frac{\tau_k^2}{\tau_k^2+\psi_k^2}$

For the FM-ST and the FM-SSL models the full conditionals are almost the same, the difference is that $\mathbf{F}$ is replaced by $\mathbf{F}^w_{n\times 2} = (\sqrt{\mathbf{u}}\mathbf{x}\ \sqrt{\mathbf{u}}\mathbf{w})$ 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 $\mathbf{W}$

- $W_i|S_{ik}=1, y_i, u_i, \boldsymbol{\beta}_k, \psi_k, \tau_k^2 \sim TN_{[0,+\infty)}(a, A/u_i)$.

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

- Skew-T

  $U_i|S_{ik}=1, y_i, w_i, \nu_k, \boldsymbol{\beta}_k, \psi_k, \tau_k^2 \sim G\left(\frac{\nu_k}{2}+1, \frac{\nu_k}{2} + \frac{(y_i-\mu_k-\mathbf{x}_i\boldsymbol{\beta}_k-\psi_k w_i)^2}{2\tau^2} + \frac{w_i^2}{2}\right)$;

- Skew-Slash

  $U_i|S_{ik}=1, y_i, w_i, \nu_k, \boldsymbol{\beta}_k, \psi_k, \tau_k^2 \sim G_{(0,1)}\left(\nu_k+1, \frac{(y_i-\mu_k-\mathbf{x}_i\boldsymbol{\beta}_k-\psi_k w_i)^2}{2\tau^2} + \frac{w_i^2}{2}\right)$;

  $\nu_k|\mathbf{s}, \mathbf{u} \sim G_{(2,40)}(\alpha + N_k, \gamma - \sum_{i:S_{ik}=1} u_i)$

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_k, $k = 1, \ldots, K$ a normal log random walk proposal was used

$$log(\nu_k^{new} - 2) \sim N(log(\nu_k - 2), c_{\nu_k}) \tag{17}$$

with adaptive width parameter c_{ν_k} (Shaby and Wells, 2010). The proposal was shifted away from 0, as it is advisable to avoid values for ν_k that are close to 0, see Fernández and Steel (1999).

References

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

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

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

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

Bouguila, N., Ziou, D., and Vaillancourt, J. (2004), "Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application," *IEEE Transactions on Image Processing*, 13, 1533–1543.

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

Cohen, E. A. (1984), "Some Effects of Inharmonic Partials on Interval Perception," *Music Perception*, 1, 323–349.

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

da Paz, R. F., Bazán, J. L., and Milan, L. A. (2017), "Bayesian estimation for a mixture of simplex distributions with an unknown number of components: HDI analysis in Brazil," *Journal of Applied Statistics*, 44, 1630–1643.

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

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

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

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

Fernández, C., and Steel, M. F. J. (1999), "Multivariate student-t regression models: Pitfalls and inference," *Biometrika*, 86, 153–167.

Fruhwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, 1 edn, New York: Springer.

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

Fu, R., Dey, D. K., and Holsinger, K. E. (2011), "A Beta-Mixture Model for Assessing Genetic Population Structure," *Biometrics*, 67, 1073–1082.

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.

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

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

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

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

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

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, 52–66.

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

Quandt, R. (1972), "A new approach to estimating switching regressions," *Journal of the American Statistical Association*, 67, 306–310.

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

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, 731–792.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

Shaby, B. A., and Wells, M. T. (2010), Exploring an Adaptive Metropolis Algorithm,, Technical report, Duke University, Department of Statistical Science.

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

Spath, H. (1979), "Algorithm 39 clusterwise linear regression," *Computing*, 67, 367–373.

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

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, 217–221.

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

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

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

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

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.

Zhang, H., Wu, Q. M. J., and Nguyen, T. M. (2013), "Incorporating Mean Template Into Finite Mixture Model for Image Segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 24, 328–335.