Advances in Data Analysis and Classification (2023) 17:181–210
https://doi.org/10.1007/s11634-022-00495-6

REGULAR ARTICLE

Robust mixture regression modeling based on two-piece scale mixtures of normal distributions

Atefeh Zarei[1] · Zahra Khodadadi[1] · Mohsen Maleki[2] · Karim Zare[1]

Received: 28 February 2021 / Revised: 29 December 2021 / Accepted: 10 February 2022 /
Published online: 23 March 2022

Abstract

The inference of mixture regression models (MRM) is traditionally based on the normal (symmetry) assumption of component errors and thus is sensitive to outliers or symmetric/asymmetric lightly/heavy-tailed errors. To deal with these problems, some new mixture regression models have been proposed recently. In this paper, a general class of robust mixture regression models is presented based on the two-piece scale mixtures of normal (TP-SMN) distributions. The proposed model is so flexible that can simultaneously accommodate asymmetry and heavy tails. The stochastic representation of the proposed model enables us to easily implement an EM-type algorithm to estimate the unknown parameters of the model based on a penalized likelihood. In addition, the performance of the considered estimators is illustrated using a simulation study and a real data example.

Keywords ECME algorithm · Mixture regression models · Penalized likelihood · Two-piece scale mixtures of normal distributions

Mathematics Subject Classification 62H30 · 62J20 · 62E17 · 62F10 · 62J05

1 Introduction

Mixture regression models (*MRM*) have broad applications in many fields including Engineering, Biology, Biometrics, Genetics, Medicine, Econometrics, Psychology and Marketing. These models are used to investigate the relationship between variables which come from several unknown latent homogeneous groups. The *MRM* was first introduced by Quandt (1972) and Quandt and Ramsey (1978) as switching regression

✉ Mohsen Maleki
m.maleki.stat@gmail.com; m.maleki@mcs.ui.ac.ir

[1] Department of Statistics, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran

[2] Department of Statistics, Faculty of Mathematics and Statistics, University of Isfahan, 81746-73441 Isfahan, Iran

models and Späth (1979) as cluster wise linear regression models. For a comprehensive survey see McLachlan and Peel (2000).

The maximum likelihood (*ML*) estimation of the parameter of the *MRM* is usually based on the normality assumption. In this regard, many extensive literatures are available. Applications include marketing (DeSarbo and Cron 1988; DeSarbo et al. 1992; Naik et al. 2007), finance (Engel and Hamilton 1990), economics (Cosslett and Lee 1985; Hamilton 1989), agriculture (Turner 2000), nutrition (Arellano-Valle et al. 2008), psychometrics (Liu et al. 2011), health (Maleki et al. 2019a), sports (Maleki et al. 2019b; Maleki and Wraith 2019), telecommunication (Hajrajabi and Maleki 2019; Maleki et al. 2020a; Mahmoudi et al. 2020). The estimators of the parameters of the normal *MRM* work well when the error distribution is indeed normal, but these estimators are very sensitive to the departures from normality. These departures often appear when the datasets contain outliers, or the error distribution displays an asymmetric shape or heavy tail. To deal with the departures from normality, many extensions of this classic model have been proposed. For example, Markatou (2000) proposed a weight function to robustly estimate the mixture regression parameters. Bai et al. (2012) used a robust estimation procedure based on *M*-regression estimation to robustly estimate the mixture regression parameters. Yao et al. (2014) studied the *MRM* assuming that the error terms follow a *t* distribution which is a generalization of the mixture of t distribution proposed by Peel and McLachlan (2000). Also, Song et al. (2014) introduced a robust model and method to estimate the parameters of *MRM* when the error distribution is a mixture of Laplace distribution. Another robust MRM based on the skew normal distribution has been studied by Liu and Lin (2014). Recently, Zeller et al. (2016) proposed a unified robust *MRM* when the error term follows scale mixtures of skew-normal distributions and examined the performance of the estimation procedure. In this regard, Doğru and Arslan (2017) investigated a *MRM* based on the skew-*t* distribution as a special case of the model proposed by Zeller et al. (2016).

In this paper, a general class of robust mixture regression models based on two-piece scale mixtures of normal (*TP-SMN*) distributions proposed by Maleki and Mahmoudi (2017) is presented. The class of *TP-SMN* distributions is a rich class of distributions that includes the well-known family of scale mixtures of normal (*SMN*; Andrews and Mallows 1974) distributions which covers symmetrical/asymmetrical and lightly/heavy-tailed distributions (see also e.g., Arellano-Valle et al. (2005), Maleki and Mahmoudi (2017), Moravveji et al. (2019), Bazrafkan et al. (2021), Hoseinzadeh et al. (2021), Maleki et al. (2021, 2022) and Maleki (2022)). Here, the family of two-piece scale mixtures of normal distributions is considered and this class of distribution is extended to the mixture regression setting.

In addition, the class of *TP-SMN* distributions is an attractive family for modeling the skewed and heavy-tailed data sets in a much wider range (see e.g., Maleki et al. (2019c, 2020b), Ghasami et al. (2020) and Maleki (2022)). So, our mixture regression model based on the two-piece scale mixtures of normal (*TP-SMN-MRM*) is very flexible and robust, and can efficiently deal with skewness and heavy-tailed-ness in the *MRM* setting. In this work, a penalized likelihood function is also considered to set the best number of component and after using the stochastic representation of the suggested model, two extensions of the *EM*-algorithm (Dempster et al. 1977) are developed, including the *ECM* algorithm (Meng and Rubin 1993) and the *ECME* algorithm (Liu and Rubin 1994).

The rest of this paper is organized as follows. In Sect. 2, the researchers review some properties of the *TP-SMN* distributions. In Sect. 3, the *TP-SMN-MRM* is introduced and maximum penalized estimates (*MPL*) of the proposed model based on an *EM*-type algorithm are obtained. In Sect. 4, numerical studies involving simulations with some applications of the proposed models and estimates on real datasets are presented. In addition, comparison is made with well-known normal competitor and then symmetrical/asymmetrical and lightly/heavy-tailed scale mixtures of skew-normal (*SMSN*; Branco and Dey, 2001) family in Zeller et al. (2016) which had been studied previously. Some conclusive remarks are presented in Sect. 4.

2 TP-SMN distributions

2.1 Preliminaries

The two-piece scale mixtures of normal (*TP-SMN*) family of distributions were constructed by the celebrated well-known scale mixtures of normal (*SMN*; Andrews and Mallows 1974) family, based on the methodology of constructing the general two-piece distributions. The *SMN* random variable X has the following probability density function (pdf) and denoted by $X \sim SMN(\mu, \sigma, \boldsymbol{\nu})$:

$$f_{SMN}(x;\mu,\sigma,\boldsymbol{\nu}) = \int_0^\infty \phi\Big(x;\mu,k(u)\sigma^2\Big)dH(u;\boldsymbol{\nu}),\ x \in R, \tag{1}$$

where $\phi\big(\cdot;\mu,\sigma^2\big)$ represents the pdf of $N\big(\mu,\sigma^2\big)$ distribution, $H(\cdot;\boldsymbol{\nu})$ is the cumulative distribution function (cdf) of the scale mixing random variable U which was indexed by parameter $\boldsymbol{\nu}$. By letting $k(u) = 1/u$, some suitable mathematical properties (such as appropriate hierarchical forms in the classical inferences and closed form posteriors in the Bayesian inferences, (see e.g., Zeller et al. (2016) and Barkhordar et al. (2020)) are obtained. Also $X \sim SMN(\mu, \sigma, \boldsymbol{\nu})$ has the stochastic representation given by

$$X = \mu + \sigma k^{1/2}(U)W, \tag{2}$$

where W is a standard normal random variable that is assumed independent of U.

The *TP-SMN* is a rich family of distributions that covers the symmetric/asymmetric lightly/heavy-tailed distributions and its main members are two-piece normal (*TP-N* or Epsilon-Skew-Normal: Mudholkar and Hutson 2000; Maleki and Nematollahi 2017), two-piece *t* (*TP-T*), and two-piece slash (*TP-SL*) distributions.

Definition 2.1 Following general two-piece distributions from Arellano-Valle et al. (2005), the pdf of random variable $Y \sim TP - SMN(\mu, \sigma, \gamma, \boldsymbol{\nu})$, for $y \in R$ is represented as.

$$f(y;\mu,\sigma,\gamma,\boldsymbol{\nu}) = \begin{cases} 2(1-\gamma)f_{SMN}(y;\mu,\sigma(1-\gamma),\boldsymbol{\nu}),\ y \le \mu, & y \le \mu, \\ 2\gamma f_{SMN}(y;\mu,\sigma\gamma,\boldsymbol{\nu}), & y > \mu, \end{cases} \tag{3}$$

where $\gamma \in (0, 1)$ is the slant parameter, $f_{SMN}(\cdot;\mu,\sigma,\boldsymbol{\nu})$ is given by (1).

Proposition 2.1 Let $Y \sim TP - SMN(\mu, \sigma, \gamma, \boldsymbol{\nu})$, then Y has a stochastic representation given by.

$$Y = \mu - \sigma(1-\gamma)S_1 k^{1/2}(U)|W| + \sigma\gamma S_2 k^{1/2}(U)|W|, \tag{4}$$

where W is a standard normal random variable that is assumed independent of scale mixing random variable $U \sim H(u; \boldsymbol{\nu})$, and under reparameterization $\sigma_1 = \sigma(1-\gamma)$ and $\sigma_2 = \sigma\gamma$, $\boldsymbol{S} = (S_1, S_2)^\top \sim \text{Multinomial}(1, \frac{\sigma_1}{\sigma_1+\sigma_2}, \frac{\sigma_2}{\sigma_1+\sigma_2})$, with the following probability mass function (pmf):

$$P(\boldsymbol{S} = \boldsymbol{s}) = \frac{\sigma_1^s \sigma_2^{1-s}}{\sigma_1 + \sigma_2}; s_1 = 1 - s_2 = s = 0, 1.$$

Proof The pdf of the $Y \sim TP - SMN(\mu, \sigma, \gamma, \boldsymbol{\nu})$ in (3) is a piecewise function, which according to the Eq. (2), on its top piece, the $2f_{SMN}(y; \mu, \sigma(1-\gamma), \boldsymbol{\nu})fory \leq \mu$, pdf, has the following stochastic representation.

$$\left[\mu - \sigma(1-\gamma)k^{1/2}(U)|W|\right] \sim SMN(\mu, \sigma(1-\gamma), \boldsymbol{\nu})I(-\infty, \mu],$$

and also on its bottom piece, $2f_{SMN}(y; \mu, \sigma\gamma, \boldsymbol{\nu})fory > \mu$, has the following stochastic representation

$$\left[\mu + \sigma\gamma k^{1/2}(U)|W|\right] \sim SMN(\mu, \sigma\gamma, \boldsymbol{\nu})I(\mu, +\infty).$$

So the random variable Y can obey from the top piece with probability $(1-\gamma)\left(= \frac{\sigma_1}{\sigma_1+\sigma_2}\right)$ when $S_1 = 1$, and can obey from the bottom piece with probability $(\gamma)\left(= \frac{\sigma_2}{\sigma_1+\sigma_2}\right)$ when $S_2 = 1$. So combining these stochastic representations and latent variable $\boldsymbol{S} = (S_1, S_2)^\top$ conclude the (4) $\ominus$

Proposition 2.2 Let $Y \sim TP - SMN(\mu, \sigma, \gamma, \boldsymbol{\nu})$,

$E(Y) = \mu - b\Delta;$
$\text{Var}(Y) = \sigma^2\left[c_2 k_2(\boldsymbol{\nu}) - b^2 c_1^2\right],$

where $\Delta = \sigma(1-2\gamma)$, $b = \sqrt{2/\pi}k_1(\boldsymbol{\nu})$, $c_r = \gamma^{r+1} + (-1)^r(1-\gamma)^{r+1}$ and $k_r(\boldsymbol{\nu}) = E\left(U^{-r/2}\right)$, for which U is the scale mixing variable in (2).

Proof Considering the Proposition 2.4. from Maleki and Mahmoudi (2017), these results have been obtained $\ominus$

More statistical properties along with the details of the *TP-SMN* family were introduced by Arellano-Valle et al. (2005) and Maleki and Mahmoudi (2017).

Proposition 2.3 The *TP-SMN* distributions with the pdf given in (3) can be represented as the two-component mixture of left and right half *SMN* distributions with special component probabilities as follows:

$$f\left(y;\mu,\sigma_1,\sigma_2,\boldsymbol{\nu}\right)=2\frac{\sigma_1}{\sigma_1+\sigma_2}f_{SMN}\left(y;\mu,\sigma_1,\boldsymbol{\nu}\right)I_{(-\infty,\mu]}\left(y\right)\\
+2\frac{\sigma_2}{\sigma_1+\sigma_2}f_{SMN}\left(y;\mu,\sigma_2,\boldsymbol{\nu}\right)I_{(\mu,+\infty)}\left(y\right),$$

where as in (4), $\sigma_1=\sigma(1-\gamma)$, $\sigma_2=\sigma\gamma$, and the scale parameter σ and skewness parameter γ in (3) are recovered in the form of $\sigma=\sigma_1+\sigma_2$ and $\gamma=\sigma_2/(\sigma_1+\sigma_2)$.

Proof Considering the pdf (3) and reparameterization $\sigma_1=\sigma(1-\gamma)$ and $\sigma_2=\sigma\gamma$, the results have been obtained.

Note that in the symmetric positions ($\gamma=0.5$), the *TP-SMN* distributions are the well-known *SMN* distributions attributed to Andrews and Mallows (1971).

2.2 Examples of the TP-SMN distributions

In this section, some particular cases of *TP-SMN* distributions are considered. Let $Y\sim TP-SMN(\mu,\sigma,\gamma,\boldsymbol{\nu})$, different members of the *TP-SMN* family accordance of different distributions for the scale mixing variable U in (4) are as follows:

- Two-piece normal (*TP-N*):

  In this case U=1, with the following pdf,

$$f(y;\mu,\sigma,\gamma)=\begin{cases}2(1-\gamma)\phi\big(y;\mu,\sigma^2(1-\gamma)^2\big),\ y\le\mu;\\[2ex] 2\gamma\phi\big(y;\mu,\sigma^2\gamma^2\big),\ y>\mu.\end{cases}$$

- Two-piece t (*TP-T*) with ν degrees of freedom:

  In this case $U\sim\text{Gamma}\big(\frac{\nu}{2},\frac{\nu}{2}\big)$, for which $k_r(\boldsymbol{\nu})=\big(\frac{\nu}{2}\big)^{r/2}\frac{\Gamma(\frac{\nu-r}{2})}{\Gamma(\frac{\nu}{2})}$, $\nu>r$, with the following pdf,

$$f(y;\mu,\sigma,\gamma,\boldsymbol{\nu})=\begin{cases}2\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{y-\mu}{\sigma(1-\gamma)}\right)^2\right)^{-\frac{\nu+1}{2}},\ y\le\mu;\\[3ex] 2\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{y-\mu}{\sigma\gamma}\right)^2\right)^{-\frac{\nu+1}{2}},\ y>\mu.\end{cases}$$

- Two-piece slash (*TP-SL*):

  In this case $U\sim\text{Beta}(\nu,1)$, for which $k_r(\boldsymbol{\nu})=\frac{2\nu}{2\nu-r}$, $\nu>\frac{r}{2}$, with the following pdf,

$$f(y;\mu,\sigma,\gamma,\boldsymbol{\nu})=\begin{cases}2\nu(1-\gamma)\int_0^1u^{\nu-1}\phi\big(y;\mu,u^{-1}\sigma^2(1-\gamma)^2\big)du, & y\le\mu;\\ 2\nu\gamma\int_0^1u^{\nu-1}\phi\big(y;\mu,u^{-1}\sigma^2\gamma^2\big)du, & y>\mu.\end{cases}$$

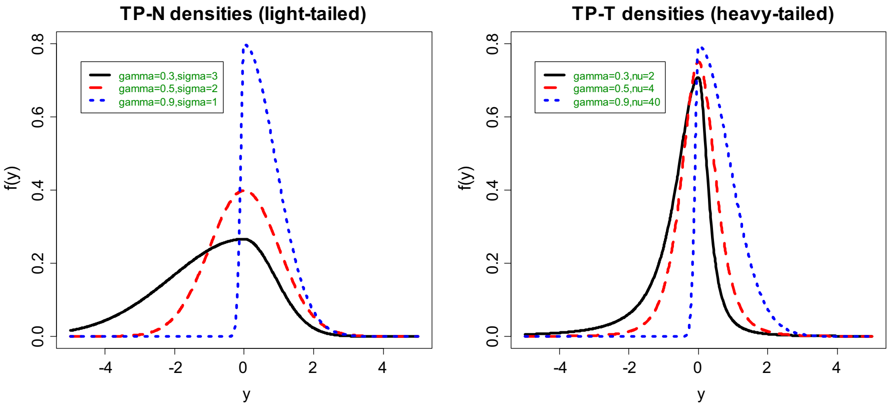

Fig. 1 Some typical graphs of the light-tailed *TP-N* densities (left) and heavy-tailed *TP-T* densities (right) with various shape, scale and degrees of freedom parameters

Note that the *TP-N* is a light-tailed density while the *TP-T* and *TP-SL* are heavy-tailed densities. Some asymmetry (with various values of shape parameter $\gamma = 0.3and0.9,$) and symmetry ($\gamma = 0$) graphs of the light-tailed *TP-N* and heavy-tailed *TP-T* densities with various scale ($\sigma = 1, 2, 3$) and degrees of freedom ($\nu = 2, 4, 40$) parameters are provided in Fig. 1.

Proposition 2.4 Let $Y \sim TP - SMN(\mu, \sigma, \gamma, \boldsymbol{\nu})$. Considering the stochastic representation (4) and $k(U) = 1/U$, conditional expectation $\tau = E[SU|y]$ for the *TP-SMN* distribution members are given by:

- *TP-N*: $\tau = I_{(-\infty,\mu]}(y)$,
- *TP-T*: $\tau = \frac{\nu+1}{\nu+d}$,
- *TP-SL*: $\tau = \frac{2\nu+1}{d} \frac{P_1(\nu+3/2, d/2)}{P_1(\nu+1/2, d_j/2)}$,

where $d = \left(\frac{y-\mu}{m_1\sigma_1+m_2\sigma_2}\right)^2$, for which $m_1 = I_{(-\infty,\mu]}(y)$ and $m_2 = 1 - m_1$, and $P_x(a, b)$ denote the distribution function of the Gamma (a, b) distribution evaluated at x. Note the conditional expectations in Proposition 2.4 are used in the E-step of the EM-algorithm to obtain the *MPL* estimates.

3 Mixture Regression model using the TP-SMN distributions

3.1 The TP-SMN-MRM

In this section, the mixture regression model where the random errors follow the two-piece scale mixtures of normal distributions (*TP-SMN-MRM*) is examined. It is defined as

$$Y|\big(Z_g = 1\big) = \boldsymbol{x}^\top \boldsymbol{\beta}_g + \varepsilon_g, g = 1, \ldots, G, \tag{5}$$

where G is the number of components (groups) in mixture regression model, $Z_{\mathrm{g}} = 1, \mathrm{g} = 1, \ldots, G$, set for the gth component, such that $P\big(Z_{\mathrm{g}} = 1\big) = \pi_{\mathrm{g}}, \mathrm{g} = 1, \ldots, G$, $\boldsymbol{\beta}_{\mathrm{g}} = (\beta_{1\mathrm{g}}, \ldots, \beta_{p\mathrm{g}})^{\top}$ is a vector of regression coefficient (fixed explanatory variables) parameters, Y is a response variable, and $\boldsymbol{x} = (x_1, \ldots, x_p)^{\top}$ is a vector of fixed explanatory variables which is independent of the random errors ε_{g}. In the presented methodology $\varepsilon_{\mathrm{g}} \sim TP - SMN\big(\mu_{\mathrm{g}}, \sigma_{\mathrm{g}}, \gamma_{\mathrm{g}}, \boldsymbol{\nu}_{\mathrm{g}}\big), \mathrm{g} = 1, \ldots, G$, where $\mu_{\mathrm{g}} = b_{\mathrm{g}}\big(1 - 2\gamma_{\mathrm{g}}\big)\sigma_{\mathrm{g}}(\mathrm{or}\mu_{\mathrm{g}} = b_{\mathrm{g}}\Delta_{\mathrm{g}}, \Delta_{\mathrm{g}} = \big(1 - 2\gamma_{\mathrm{g}}\big)\sigma_{\mathrm{g}})$ for which $b_{\mathrm{g}} = \sqrt{2/\pi}k_1\big(\boldsymbol{\nu}_{\mathrm{g}}\big)$, and $k_1(\cdot)$ was defined in proposition 2.2. Also, note that due to the Proposition 2.2., the errors have zero mean $\big(E\big(\varepsilon_{\mathrm{g}}\big) = 0\big)$. For computational convenience, the parameter of mixing distribution $H\big(\cdot; \boldsymbol{\nu}_{\mathrm{g}}\big), \mathrm{g} = 1, \ldots, G$ are assumed equal as $\boldsymbol{\nu}_1 = \cdots = \boldsymbol{\nu}_{\mathrm{G}} = \boldsymbol{\nu}$. The identifiability of finite mixtures has been studied by Teicher (1963) to ensure that our *MRM* is identifiable. In addition, in this study, the maximum likelihood inferential paradigm is used and so label switching has no practical implications and arises only as a theoretical identifiability issue that can usually be resolved by specifying some ordering on the mixing proportions in the form of $\pi_1 > \ldots > \pi_G$. Note that in cases where mixing proportions are equal, a total ordering on other model parameters can be considered.

Using an auxiliary random variable $\boldsymbol{Z} = (Z_1, \ldots, Z_G)^{\top}$ (independent of $\boldsymbol{x}$), for which $Z_{\mathrm{g}} = 1, \mathrm{g} = 1, \ldots, G$, set the regression model in (5) for the gth component, such that $P\big(Z_{\mathrm{g}} = 1\big) = \pi_{\mathrm{g}}, \mathrm{g} = 1, \ldots, G$, then the density of response variable Y is given by

$$f_{MR}(y; \boldsymbol{x}, \boldsymbol{\Theta}) = \sum_{\mathrm{g}=1}^{G} \pi_{\mathrm{g}} f\big(y; \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{g}}\big), \tag{6}$$

where $f\big(\cdot; \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{g}}\big)$ is the pdf of $TP - SMN\big(\boldsymbol{x}^{\top}\boldsymbol{\beta}_{\mathrm{g}} + \mu_{\mathrm{g}}, \sigma_{\mathrm{g}}, \gamma_{\mathrm{g}}, \boldsymbol{\nu}\big)$ and $\boldsymbol{\theta}_{\mathrm{g}} = \Big(\boldsymbol{\beta}_{\mathrm{g}}^{\top}, \sigma_{\mathrm{g}}, \gamma_{\mathrm{g}}, \boldsymbol{\nu}^{\top}\Big)$, $\mathrm{g} = 1, \ldots, G$ or according to the representation of Proposition 2.3, $\boldsymbol{\theta}_{\mathrm{g}} = \Big(\boldsymbol{\beta}_{\mathrm{g}}^{\top}, \sigma_{1\mathrm{g}}, \sigma_{2\mathrm{g}}, \boldsymbol{\nu}^{\top}\Big), \mathrm{g} = 1, \ldots, G$ and $\boldsymbol{\Theta} = \big(\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1^{\top}, \ldots, \boldsymbol{\theta}_G^{\top}\big)^{\top}$. In the viewpoint of classical inferences, using the observations $(Y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, the parameter $\boldsymbol{\Theta}$ is traditionally estimated by maximization of the log-likelihood of an IID sample $(\boldsymbol{Y}, \boldsymbol{x})^{\top}$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\top}$ and $\boldsymbol{x} = \big(\boldsymbol{x}_1^{\top}, \ldots, \boldsymbol{x}_n^{\top}\big)^{\top}$ as

$$\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log f_{MR}(y_i; \boldsymbol{x}_i, \boldsymbol{\Theta}).$$

In applications, existence of too many components imply that the mixture models may overfit the data and yield poor interpretations, while existence of too few components, imply that the mixture models may not be flexible enough to approximate the true underlying data structure. So, estimating the true number of components in the mixture models is very important. In order to solve this issue, we have used a penalized log-likelihood function to avoid overestimating or underestimating them, given by

$$\ell_P(\boldsymbol{\Theta}) = \ell(\boldsymbol{\Theta}) - n\lambda D_{f.MR}\sum_{\mathrm{g}=1}^{G}\left[\log\left(\epsilon + \pi_{\mathrm{g}}\right) - \log(\epsilon)\right], \tag{7}$$

where $\ell(\boldsymbol{\Theta})$ is the log-likelihood function, λ is a tuning parameter, ϵ is a very small positive number, say 10^{-6}, and $D_{f.\mathrm{MR}}$ is the number of free parameters for each component. For the *TP-N-MRM*, *TP-T-MRM* and *TP-SL-MRM*, each component has $D_{f.MR} = p + 4$, and for *TP-CN-MRM* each component has $D_{f.MR} = p + 5$ number of free parameters. Huang et al. (2017) had used this penalty term in the structure of likelihood function of the mixture of Gaussian model.

To obtain the proposed maximizer given by penalized log-likelihood (7), there is not an explicit solution, so an *EM*-type algorithm (Dempster et al. 1977; McLachlan and Peel, 2000) is considered.

3.2 The observed information matrix

In this section, the observed information matrix of the *TP-SMN-MRM*, defined as $\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{y}) = -\frac{\partial^2 \ell_P(\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top}$, where $\ell_P(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\ell_{Pi}(\boldsymbol{\Theta})$, for which

$$\ell_{Pi}(\boldsymbol{\Theta}) = \log\sum_{\mathrm{g}=1}^{G}\pi_{\mathrm{g}} f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right) - \lambda D_{f.MR}\sum_{\mathrm{g}=1}^{G}\left[\log\left(\epsilon + \pi_{\mathrm{g}}\right) - \log(\epsilon)\right].$$

It is well known that, under some regularity conditions, the covariance matrix of the *MPL* estimates $\widehat{\boldsymbol{\Theta}}$ can be approximated by the inverse of $\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{y})$. So, the square roots of its diagonal elements have been considered as the standard deviations of the *MPL* estimates in the real applications. Thus, following Basford et al. (1997) and Lin et al. (2007),

$$\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{y}) = \sum_{i=1}^{n}\widehat{\boldsymbol{j}}_i^\top\widehat{\boldsymbol{j}}_i,$$

where $\widehat{\boldsymbol{j}}_i = \left.\frac{\partial\ell_{Pi}(\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}$, and now consider the vector $\widehat{\boldsymbol{j}}_i$ which is partitioned into components corresponding to all the parameters in $\boldsymbol{\Theta}$ as

$$\widehat{\boldsymbol{j}}_i = \left(\widehat{j}_{i,\pi_1}, \ldots, \widehat{j}_{i,\pi_{G-1}}, \widehat{\boldsymbol{j}}_{i,\boldsymbol{\beta}_1}^\top, \ldots, \widehat{\boldsymbol{j}}_{i,\boldsymbol{\beta}_G}^\top, \widehat{j}_{i,\sigma_1}, \ldots, \widehat{j}_{i,\sigma_G}, \widehat{j}_{i,\gamma_1}, \ldots, \widehat{j}_{i,\gamma_G}, \widehat{\boldsymbol{j}}_{i,\boldsymbol{\nu}}^\top\right)^\top,$$

where its coordinate elements for $\mathrm{g} = 1, \ldots, G$ are given by

$$\widehat{j}_{i,\pi_{\mathrm{g}}} = \frac{f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right) - f(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_G, \boldsymbol{\nu})}{f_{MR}(y_i|\boldsymbol{x}_i, \boldsymbol{\Theta})} - \lambda D_{f.MR}\left[\frac{1}{\epsilon + \pi_{\mathrm{g}}} - \frac{1}{\epsilon + \pi_{\mathrm{G}}}\right],$$

$$\widehat{\boldsymbol{j}}_{i,\boldsymbol{\beta}_{\mathrm{g}}} = \frac{\pi_{\mathrm{g}} D_{\boldsymbol{\beta}_{\mathrm{g}}}\left(f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right)}{f_{MR}\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Theta}\right)}, \widehat{j}_{i,\sigma_{\mathrm{g}}}$$
$$= \frac{\pi_{\mathrm{g}} D_{\sigma_{\mathrm{g}}}\left(f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right)}{f_{MR}\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Theta}\right)}, \widehat{j}_{i,\gamma_{\mathrm{g}}} = \frac{\pi_{\mathrm{g}} D_{\gamma_{\mathrm{g}}}\left(f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right)}{f_{MR}\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Theta}\right)},$$

and

$$\widehat{\boldsymbol{j}}_{i,\boldsymbol{\nu}} = \frac{\sum_{\mathrm{g}=1}^{G} \pi_{\mathrm{g}} D_{\boldsymbol{\nu}}\left(f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right)}{f_{MR}(y_i; \boldsymbol{x}_i, \boldsymbol{\Theta})},$$

for which $D_{\boldsymbol{\alpha}}\left[f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right] = \partial f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)/\partial \boldsymbol{\alpha}$, for $\boldsymbol{\alpha} = \boldsymbol{\beta}_{\mathrm{g}}, \sigma_{\mathrm{g}}, \gamma_{\mathrm{g}}, \boldsymbol{\nu}$. To determine the coordinate elements of the $\widehat{\boldsymbol{j}}_i$, let us define $\zeta_{i\mathrm{g}}(\omega) = E_H\left[u^{\omega}\exp\left(-\frac{1}{2}um_{i\mathrm{g}}\right)\right]$, where $m_{i\mathrm{g}} = \frac{d_{i\mathrm{g}}^2}{\sigma_{\mathrm{g}}^2\rho_{\mathrm{g}}^2}$ is the Mahalanobis distances for which $d_{i\mathrm{g}} = y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta}_{\mathrm{g}} - \mu_{\mathrm{g}}$, and hereafter $\rho_{\mathrm{g}} = 1 - \gamma_{\mathrm{g}}$ if $d_{i\mathrm{g}} \le 0$ and $\rho_{\mathrm{g}} = \gamma_{\mathrm{g}}$ if $d_{i\mathrm{g}} > 0$. So, we have

$$D_{\boldsymbol{\beta}_{\mathrm{g}}}\left[f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right] = \frac{2}{\sqrt{2\pi}}\left[\frac{1}{\sigma_{\mathrm{g}}^3}\zeta_{i\mathrm{g}}\left(\frac{3}{2}\right)d_{i\mathrm{g}}\boldsymbol{x}_i\right]$$

$$D_{\sigma_{\mathrm{g}}}\left[f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right] = \frac{2}{\sqrt{2\pi}}\left[\frac{1}{\sigma_{\mathrm{g}}^4\rho_{\mathrm{g}}^2}(d_{i\mathrm{g}} + \mu_{\mathrm{g}}/2)\zeta_{i\mathrm{g}}\left(\frac{3}{2}\right) - \frac{1}{\sigma_{\mathrm{g}}^2}\zeta_{i\mathrm{g}}\left(\frac{1}{2}\right)\right],$$

$$D_{\gamma_{\mathrm{g}}}\left[f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right] = \frac{2}{\sqrt{2\pi}}\left(\mathrm{sign}\left(d_{i\mathrm{g}}\right)\frac{d_{i\mathrm{g}}}{\sigma_{\mathrm{g}}^3\rho_{\mathrm{g}}^3} - \frac{b}{\sigma_{\mathrm{g}}^2\rho_{\mathrm{g}}^2}\right)\zeta_{i\mathrm{g}}\left(\frac{3}{2}\right),$$

where $\zeta_{i\mathrm{g}}(\cdot)$ in the above relations, and also $D_{\boldsymbol{\nu}}\left[f\left(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right]$ for the *TP-SMN-MRM* members, are given by:

(i) *TP-N-MRM*:

$$\zeta_{i\mathrm{g}}(\omega) = \exp\left(-\frac{1}{2}m_{i\mathrm{g}}\right),$$

$$D_{\boldsymbol{\nu}}\left[f\left(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\mathrm{g}}, \boldsymbol{\nu}\right)\right] = 0;$$

(ii) *TP-T-MRM*:

$$\zeta_{i\mathrm{g}}(\omega) = \frac{2^{\omega}\nu^{\nu/2}\Gamma(\nu/2 + \omega)}{\Gamma(\nu/2)\left(\nu + m_{i\mathrm{g}}\right)^{\nu/2+\omega}},$$

$$D_{\boldsymbol{\nu}}\left[f\left(y_i;\boldsymbol{x}_i,\boldsymbol{\theta}_{\mathrm{g}},\boldsymbol{\nu}\right)\right]=\frac{1}{2}f\left(y_i;\boldsymbol{x}_i,\boldsymbol{\theta}_{\mathrm{g}},\boldsymbol{\nu}\right)\Bigg[\psi\left(\frac{\nu+1}{2}\right)-\psi\left(\frac{\nu}{2}\right)-\frac{1}{\nu}$$
$$-\log\left(1+\frac{m_{i\mathrm{g}}}{\nu}\right)+(\nu$$
$$+1)\,\frac{m_{i\mathrm{g}}+\mu_{\mathrm{g}}\left[1+\nu\psi\left(\frac{\nu-1}{2}\right)-\nu\psi\left(\frac{\nu}{2}\right)\right]\sqrt{m_{i\mathrm{g}}}}{\nu^2+\nu m_{i\mathrm{g}}}\Bigg];$$

(iii) *TP-SL-MRM*:

$$\zeta_{li\mathrm{g}}(\omega)=\frac{\nu\Gamma(\nu+\omega)}{\left(m_{i\mathrm{g}}/2\right)^{\nu+\omega}}P_1\left(\nu+\omega,m_{i\mathrm{g}}/2\right),$$

$$D_{\boldsymbol{\nu}}\left[f\left(y_i;\boldsymbol{x}_i,\boldsymbol{\theta}_{\mathrm{g}},\boldsymbol{\nu}\right)\right]=\nu^{-1}f\left(y_i;\boldsymbol{x}_i,\boldsymbol{\theta}_{\mathrm{g}},\nu\right)+\nu f\left(y_i;\boldsymbol{x}_i,\boldsymbol{\theta}_{\mathrm{g}},\nu-1\right);$$

where $P_x(a,b)$ denotes the distribution function of the Gamma (a,b) distribution evaluated at x.

3.3 Maximum penalized estimation of the model parameters

In this section, an efficient *EM*-type algorithm for *MPL* estimation of the parameters of *TP-SMN-MRM* is developed using an incomplete-data framework. To do this procedure, beside all the observations $(Y_i,\boldsymbol{x}_i), i=1,\ldots,n$ defines the latent random vector as $\boldsymbol{Z}_i=(Z_{i1},\ldots,Z_{iG})^{\top}, i=1,\ldots,n$, where

$$Z_{ig}=\begin{cases}1, & \text{if the } i\text{th observation belongs the } g\text{th component;}\\ 0, & \text{otherwise.}\end{cases}$$

Therefore, under the above approach the latent random vector $\boldsymbol{Z}_i, i=1,\ldots,n$ has the following multinomial pmf:

$$P(\boldsymbol{Z}_i=z_i)=\prod_{\mathrm{g}=1}^{G}\pi_{\mathrm{g}}^{z_{i\mathrm{g}}};i=1,\ldots,n,$$

such that $\sum_{\mathrm{g}=1}^{G}\pi_{\mathrm{g}}=1,\pi_{\mathrm{g}}>0,\mathrm{g}=1,\ldots,G$ and

$$Y_i\big|z_{i\mathrm{g}}=1\sim TP-SMN\Big(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_{\mathrm{g}}+\mu_{\mathrm{g}},\sigma_{\mathrm{g}},\gamma_{\mathrm{g}},\boldsymbol{\nu}\Big),\mathrm{g}=1,\ldots,G.$$

So, using the stochastic representation of the *TP-SMN* family given by (4), the following hierarchical representation is considered

$$Y_i|U_i,S_{ij}=1,Z_{i\mathrm{g}}=1\underset{\sim}{ind.}N\Big(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_{\mathrm{g}}+\mu_{\mathrm{g}},u_i^{-1}\sigma_{\mathrm{g}j}^2\Big)I_{A_i}(y_i)^{2-j}I_{A_i^c}(y_i)^{j-1},$$

$$U_i \big| Z_{i\mathrm{g}} = 1 \overset{ind.}{\sim} H(u_i; \boldsymbol{\nu}),$$

$$\boldsymbol{S}_i | Z_{i\mathrm{g}} = 1 \overset{ind.}{\sim} \text{Multinomial}\left(1, \frac{\sigma_{\mathrm{g}1}}{\sigma_{\mathrm{g}1} + \sigma_{\mathrm{g}2}}, \frac{\sigma_{\mathrm{g}2}}{\sigma_{\mathrm{g}1} + \sigma_{\mathrm{g}2}}\right),$$

$$\boldsymbol{Z}_i \overset{i.i.d.}{\sim} \text{Multinomial}(1, \pi_1, \ldots, \pi_G), \tag{8}$$

for $i = 1, \ldots, n$, $\mathrm{g} = 1, \ldots, G$ and $j = 1, 2$, where $A_i = \left(-\infty, \boldsymbol{x}_i^\top \boldsymbol{\beta}_\mathrm{g} + \mu_\mathrm{g}\right]$ and $N(\cdot)I_A(\cdot)$ denotes the univariate normal distribution truncated on the interval A.

The hierarchical representation (8) of the *TP-SMN-MRM* is used to obtain the *MPL* estimates via an *EM*-algorithm called *ECME* algorithm. It is a generalization of the *ECM* algorithm introduced by Meng and Rubin (1993). It can be obtained by replacing some *CM*-steps which maximize the constrained expected complete-data penalized log-likelihood function with steps that maximize the correspondingly constrained actual likelihood function.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$,$\boldsymbol{s} = (\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_n^\top)^\top$ and $\boldsymbol{z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$ for which $\boldsymbol{s}_i = (s_{i1}, s_{i2})^\top$, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})^\top$ for $i = 1, \ldots, n$, so considering the complete data $\boldsymbol{y}_c = (\boldsymbol{y}^\top, \boldsymbol{u}^\top, \boldsymbol{s}^\top, \boldsymbol{z}^\top)^\top$ and using the hierarchical representation in (8) of the *TP-SMN-MRM*, the complete log-likelihood function is given by

$$\begin{aligned}\ell_{cp}\left(\boldsymbol{\Theta}|\boldsymbol{y}_c\right) = c + \sum_{i=1}^{n}\sum_{\mathrm{g}=1}^{G} z_{i\mathrm{g}} \log\pi_\mathrm{g} - \sum_{i=1}^{n}\sum_{\mathrm{g}=1}^{G} z_{i\mathrm{g}} \log\left(\sigma_{\mathrm{g}1} + \sigma_{\mathrm{g}2}\right)\\ - \frac{1}{2}\sum_{i=1}^{n}\sum_{\mathrm{g}=1}^{G}\sum_{j=1}^{2} \frac{z_{i\mathrm{g}} s_{ij} u_i}{\sigma_{\mathrm{g}j}^2}\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\mathrm{g} - \mu_\mathrm{g}\right)^2\\ - n\lambda D_{f.MR}\sum_{\mathrm{g}=1}^{G}\left[\log\left(\epsilon + \pi_\mathrm{g}\right) - \log\left(\epsilon\right)\right],\end{aligned}$$

where c is a constant and independent of $\boldsymbol{\Theta}$.

Letting $\widehat{\boldsymbol{\Theta}}^{(k)}$ the estimates of $\boldsymbol{\Theta}$ at the kth iteration, the conditional expectation of complete log-likelihood function ignoring constant is given by

$$\begin{aligned}Q\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)}\right) = \sum_{i=1}^{n}\sum_{\mathrm{g}=1}^{G} \hat{z}_{i\mathrm{g}}^{(k)} \log\pi_\mathrm{g} - \sum_{i=1}^{n}\sum_{\mathrm{g}=1}^{G} \hat{z}_{i\mathrm{g}}^{(k)} \log\left(\sigma_{\mathrm{g}1} + \sigma_{\mathrm{g}2}\right)\\ - \frac{1}{2}\sum_{i=1}^{n}\sum_{\mathrm{g}=1}^{G}\sum_{j=1}^{2} \frac{\widehat{zsu}_{i\mathrm{g}j}^{(k)}}{\sigma_{\mathrm{g}j}^2}\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\mathrm{g} - \mu_\mathrm{g}\right)^2\end{aligned}$$

$$- n\lambda D_{f.MR} \sum_{\mathrm{g}=1}^{G} \left[\log\left(\epsilon + \pi_{\mathrm{g}}\right) - \log(\epsilon)\right],$$

where$\widehat{z}_{i\mathrm{g}}^{(k)} = E\left[Z_{i\mathrm{g}} \middle| y_i, \widehat{\mathbf{\Theta}}^{(k)} \right]$is determined by using known properties of conditional expectation, as

$$\widehat{z}_{i\mathrm{g}}^{(k)} = \frac{\widehat{\pi}_{\mathrm{g}}^{(k)} f\left(y_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_{\mathrm{g}}^{(k)}\right)}{\sum_{\mathrm{g}=1}^{G} \widehat{\pi}_{\mathrm{g}}^{(k)} f\left(y_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_{\mathrm{g}}^{(k)}\right)}; i = 1, \ldots, n, \mathrm{g} = 1, \ldots G,$$

for which $f\left(\cdot; \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{g}}\right)$ was defined in the (6), and $\widehat{zsu}_{i\mathrm{g}j}^{(k)} = E\left[Z_{i\mathrm{g}} S_{ij} U_i \middle| y_i, \widehat{\mathbf{\Theta}}^{(k)}\right] = \widehat{z}_{i\mathrm{g}}^{(k)} \widehat{\tau}_{i\mathrm{g}j}^{(k)}$, for which $\widehat{\tau}_{i\mathrm{g}j}^{(k)}$ values can be easily derived from the Proposition 2.4.

Now, this *EM*-type algorithm (*ECME*) is described to obtain the *MPL* estimates of the parameters of *TP-SMN-MRM.*

E-step Given $\mathbf{\Theta} = \widehat{\mathbf{\Theta}}^{(k)}$ and using the above calculations, we compute $\widehat{z}_{i\mathrm{g}}^{(k)}$ and $\widehat{zsu}_{i\mathrm{g}j}^{(k)}$ for$j = 1, 2$, $\mathrm{g} = 1, \ldots, G$ and$i = 1, ..., n$.

CM-step Update $\widehat{\mathbf{\Theta}}^{(k+1)}$ by maximizing $Q\left(\mathbf{\Theta}|\widehat{\mathbf{\Theta}}^{(k)}\right)$ over $\mathbf{\Theta}$ with the following updates:

Update $\widehat{\pi}_{\mathrm{g}}$; $\mathrm{g} = 1, \ldots, G$, with given ϵ is very close to zero, by using straightforward calculations, we obtain

$$\widehat{\pi}_{\mathrm{g}}^{(k+1)} = Max\left\{0, \frac{1}{1 - \lambda G D_{f.MR}} \left[\frac{\sum_{i=1}^{n} \widehat{z}_{i\mathrm{g}}^{(k)}}{n} - \lambda D_{f.MR}\right]\right\}.$$

The penalized log-likelihood and the number of effective clusters (with non-zero proportions) evolved during the iterations of the ECME algorithm works as follows: it starts with a pre-specified large number of components (for example G = 10 in the last section), and whenever a mixing probability is shrunk to zero by CM-step (for example $\widehat{\pi}_{\mathrm{g}}^{(k)} < 0.01 forg = 1, 2, \ldots Ginthelastsection$), the corresponding component is deleted, thus fewer components are retained for the remaining ECME iterations. Here we abuse the notation G for the number of components at beginning of each ECME iteration, and through the updating process, G becomes smaller and smaller. For a given ECME iteration step, it is possible that none, one, or more than one components are deleted (see e.g., Huang et al. (2017)). Note that our proposed penalized likelihood method is significantly different from various Bayesian methods in the objective function and theoretical properties. When a component is eliminated, i.e., the mixing weight of that component is shrunk to zero, the objective function of our proposed method changes continuously. So above estimation of π_{g} is different of any maximum a posteriori (MAP) estimation of them.

Update $\widehat{\boldsymbol{\beta}}_{\mathrm{g}}; \mathrm{g} = 1, \dots, G$, by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{g}}^{(k+1)} = \left(\sum_{i=1}^{n} \widehat{\varrho}_{i\mathrm{g}}^{(k)} \boldsymbol{x}_i \boldsymbol{x}_i^{\top}\right)^{-1} \sum_{i=1}^{n} \widehat{\varrho}_{i\mathrm{g}}^{(k)} \left(Y_i - \widehat{\mu}_{\mathrm{g}}^{(k)}\right) \boldsymbol{x}_i,$$

where $\widehat{\varrho}_{i\mathrm{g}}^{(k)} = \widehat{zsu}_{i\mathrm{g}1}^{(k)}/\sigma_{\mathrm{g}1}^2 + \widehat{zsu}_{i\mathrm{g}2}^{(k)}/\sigma_{\mathrm{g}2}^2$.

Update $\widehat{\sigma}_{\mathrm{g}j}; \mathrm{g} = 1, \dots, G$, $j = 1, 2$, by solving the following equations

$$\sum_{i=1}^{n} \widehat{z}_{i\mathrm{g}}^{(k)} \left(\sigma_{\mathrm{g}1} + \sigma_{\mathrm{g}2}\right)^{-1} = \sum_{i=1}^{n} \left[\sigma_{\mathrm{g}j}^{-3} \widehat{zsu}_{i\mathrm{g}j}^{(k)} \widehat{e}_{i\mathrm{g}}^{2(k+1)} + (-1)^{j+1} b \widehat{\varrho}_{i\mathrm{g}}^{(k)} \widehat{e}_{i\mathrm{g}}^{(k+1)}\right],$$

where $e_{i\mathrm{g}} = Y_i - \boldsymbol{x}_i^{\top} \boldsymbol{\beta}_{\mathrm{g}} - b\left(\sigma_{\mathrm{g}1} - \sigma_{\mathrm{g}2}\right)$. Note that the above equation is a cubic equation for each $\sigma_{\mathrm{g}j}$ in the form of $\sigma_{\mathrm{g}j}^3 + c_1 \sigma_{\mathrm{g}j} + c_2 = 0$ such that $c_1, c_2 < 0$, so this cubic equation has unique root in the $(0, +\infty)$ interval.

CML-step In the last step, update $\widehat{\boldsymbol{\nu}}$ by maximizing the actual marginal log-likelihood function, as

$$\boldsymbol{\nu}^{(k+1)} = \mathrm{argmax}_{\boldsymbol{\nu}} \sum_{i=1}^{n} \log \sum_{\mathrm{g}=1}^{G} \widehat{\pi}_{\mathrm{g}}^{(k)} f\left(y_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}_{\mathrm{g}}^{(k+1)}, \widehat{\sigma}_{\mathrm{g}1}^{(k+1)}, \widehat{\sigma}_{\mathrm{g}2}^{(k+1)}, \boldsymbol{\nu}\right),$$

where $f\left(\cdot; \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{g}}\right)$ is defined in (6).

The proposed *ECME* algorithm works as follows: it starts with a pre-specified large number of components, and due to updating $\widehat{\pi}_{\mathrm{g}}^{(k+1)}$, $\mathrm{g} = 1, \dots, G$, whenever a mixing probability is shrunk to zero, the corresponding component is deleted, and as result fewer components are retained for the remaining *EM* iterations. The iterations are repeated until a suitable convergence rule is satisfied, e.g., $\left|\ell\left(\widehat{\boldsymbol{\Theta}}^{(k+1)}\right)/\ell\left(\widehat{\boldsymbol{\Theta}}^{(k)}\right) - 1\right| \leq 10^{-4}$ where $\ell(\cdot)$ is the actual log-likelihood, was defined in the Sect. 3.1.

3.4 Selection of tuning parameter and model selection

To obtain the final estimate of the mixture model by maximizing (7), one needs to select the tuning parameter λ. For standard *LASSO* (Tibshirani 1996) and *SCAD* (Fan and Li 2001) penalized regressions, there are many methods to select λ, and in this work we have used *BIC* function in Wang et al. (2007). Here we define a $BIC(\lambda)$ value as

$$BIC(\lambda) = \sum_{i=1}^{n} \log \sum_{\mathrm{g}=1}^{\widehat{G}} \widehat{\pi}_{\mathrm{g}} f\left(y_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}_{\mathrm{g}}, \widehat{\sigma}_{\mathrm{g}1}, \widehat{\sigma}_{\mathrm{g}2}, \widehat{\boldsymbol{\nu}}\right) - \frac{1}{2} \widehat{G} D_{f.MR} \log n,$$

and estimate λ by

$$\widehat{\lambda} = \text{argmax}_{\lambda} BIC(\lambda),$$

where $\widehat{G}$ is the estimate of the number of *TP-SMN-MRM* components.

The $BIC(\lambda)$ value is useful for selecting an appropriate model with the best number of components, for the given data with adequate sample size, but in this study, four criteria are also considered in simulations in order to select the best fitted *MRM*. They are maximized log-likelihood values, the Akaike information criterion (*AIC*; Akaike 1974), the Bayesian information criterion (*BIC*; Schwarz 1978) and the efficient determination criterion (*EDC*; Resende and Dorea 2016). The above criteria have the general following form

$$kr_n - 2\ell\left(\widehat{\boldsymbol{\Theta}}\middle|\boldsymbol{y}\right),$$

where $\ell\left(\widehat{\boldsymbol{\Theta}}\middle|\boldsymbol{y}\right)$ is the actual log-likelihood, k is the number of free parameters that has to be estimated in the model and the penalty term r_n is a convenient sequence of positive numbers. Additionally, the values $r_n = 2$, $r_n = \log n$ and $r_n = 0.2\sqrt{n}$, for the *AIC*, *BIC* and *EDC* are used respectively. Fewer values of the *AIC*, *BIC* and *EDC* criteria indicate choosing the best model.

4 Numerical study

In this section, some simulations and a real dataset to show the satisfactory performances of the proposed model are considered.

4.1 Simulations

In this section, three parts of simulations are presented. In the first part, we have some simulations for *TP-SMN-MRM* parameters recovery by simulating from them and estimating the proposed *MPL* estimates to show the satisfaction of the proposed estimations. In the second part, by choosing some various sample sizes, the consistency properties of the proposed model and estimation methods are shown. Finally, in the third part of simulations, using an asymmetry and heavy-tailed distribution that belong to the class of scale mixtures of skew-normal (*SMSN*) distributions, a similar *MRM* to ours is generated to show the performances (robustness, misspecification and right classification) of our models to model the data with unknown structure. Note that in the all parts of numerical studies, the search range of tuning parameter is interval of (0, 10), and the maximum initial (pre-specified) number of components is set to be 10.

4.1.1 Part1: recovery of parameters

The following *TP-SMN-MRM* with two components was considered in three scenarios. In the first one, both components had skewed behavior between week up to moderate, in the second one, both components had skewed behavior between moderate up to strong, and in the third one, a component had skewed behavior between week up to moderate and another component had skewed behavior between moderate up to strong. The simulated model is given by

$$\begin{cases} Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \varepsilon_1, \text{ with Probability } \pi \\ Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_2 + \varepsilon_2, \text{ with Probability } 1-\pi, \end{cases}$$

where $\boldsymbol{x}_i^\top = (1, x_{i1}, x_{i2})$ for $i = 1, \ldots, n$, such that $x_{i1} \sim U(0, 1)$ and independent of $x_{i2} \sim N(0, 1)$, and, ε_1 and ε_2 follow the *TP-SMN* distributions, as the assumption given in (5).

700 samples were generated from the above model with $n = 400$ from the *TP-N*, *TP-T* and *TP-SL* models the following parameter values:

$$\begin{aligned} \boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \beta_{21})^\top = (1, 3, 5)^\top, \boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12}, \beta_{22})^\top \\ = (5, -2, -6)^\top, \pi = 0.4, \sigma_1 = \sigma_2 = 2, \end{aligned}$$

and, $\gamma_1 = 0.45, \gamma_2 = 0.55$ (for the first scenario), $\gamma_1 = 0.05, \gamma_2 = 0.95$ (for the second scenario) and $\gamma_1 = 0.1, \gamma_2 = 0.6$ (for the third scenario), for which $\nu = 4$ has used in the *TP-T-MRM* and *TP-SL-MRM.*

The maximum likelihood estimation via the proposed *ECME* algorithm for each sample was calculated, and the average values of *MPL* estimates and the corresponding standard deviations (SD) of the *MPL* estimates across all samples were computed and recorded in Tables 1, 2 and 3. The results indicated us that all the point estimates are

Table 1 Mean and standard deviations (SD) of *MPL* estimates based on 700 samples from the *TP-SMN-MRM* with true values of parameters in the parentheses (Scenario 1)

Model	*TP-N-MRM*		*TP-T-MRM*		*TP-SL-MRM*	
Parameter	Mean	SD	Mean	SD	Mean	SD
$\beta_{01}(1)$	1.0004	0.1307	0.9971	0.1403	0.9988	0.1692
$\beta_{11}(3)$	3.0041	0.2012	2.9901	0.2133	3.0234	0.2923
$\beta_{21}(5)$	4.9922	0.1972	5.0024	0.3018	5.0408	0.3239
$\beta_{02}(5)$	5.0014	0.1653	5.0032	0.2893	4.9918	0.3432
$\beta_{12}(-2)$	− 2.0033	0.1319	− 1.9961	0.1492	− 2.0103	0.1360
$\beta_{22}(-6)$	− 5.9170	0.0867	− 6.0073	0.1069	− 6.0005	0.1233
$\sigma_1(2)$	2.1301	0.0823	2.1083	0.0715	1.9702	0.0911
$\sigma_2(2)$	1.9650	0.0904	2.1045	0.0643	2.1217	0.0908
$\gamma_1(0.45)$	0.4483	0.0213	0.4502	0.0176	0.4511	0.0209
$\gamma_2(0.55)$	0.5514	0.0225	0.5507	0.0200	0.5511	0.0223
$\nu(4)$	–	–	4.0122	0.3904	4.3420	1.0218
$\pi(0.4)$	0.4001	0.0176	0.4006	0.0200	0.3997	0.0197

Table 2 Mean and standard deviations (SD) of *MPL* estimates based on 700 samples from the *TP-SMN-MRM* with true values of parameters in the parentheses (Scenario 2)

Model	*TP-N-MRM*		*TP-T-MRM*		*TP-SL-MRM*	
Parameter	Mean	SD	Mean	SD	Mean	SD
$\beta_{01}(1)$	0.9984	0.1283	0.9968	0.1301	1.0012	0.1593
$\beta_{11}(3)$	3.0037	0.1973	3.0045	0.2065	2.9905	0.3001
$\beta_{21}(5)$	5.0032	0.1345	4.9972	0.1837	5.0082	0.2329
$\beta_{02}(5)$	4.9963	0.2043	5.0021	0.2532	5.0015	0.2574
$\beta_{12}(-2)$	− 2.0030	0.1432	− 2.0052	0.1504	− 1.9971	0.1378
$\beta_{22}(-6)$	− 6.0009	0.1045	− 6.0064	0.1122	− 5.9182	0.1276
$\sigma_1(2)$	1.9732	0.0838	2.0983	0.0563	1.9843	0.0813
$\sigma_2(2)$	2.0109	0.0546	2.0837	0.0838	2.1193	0.0781
$\gamma_1(0.05)$	0.0467	0.0013	0.0511	0.0053	0.0513	0.0076
$\gamma_2(0.95)$	0.9484	0.0025	0.9510	0.0073	0.9580	0.0054
$\nu(4)$	–	–	3.8720	0.4038	4.2530	0.9873
$\pi(0.4)$	0.3989	0.0227	0.4021	0.0301	0.4083	0.0234

Table 3 Mean and standard deviations (SD) of ML estimates based on 700 samples from the *TP-SMN-MRM* with true values of parameters in the parentheses (Scenario 3)

Model	*TP-N-MRM*		*TP-T-MRM*		*TP-SL-MRM*	
Parameter	Mean	SD	Mean	SD	Mean	SD
$\beta_{01}(1)$	1.0021	0.1098	1.0042	0.1411	1.0044	0.1446
$\beta_{11}(3)$	2.9863	0.2018	2.9898	0.2042	3.0526	0.2878
$\beta_{21}(5)$	4.9968	0.1564	4.9963	0.2019	5.0103	0.2409
$\beta_{02}(5)$	5.0064	0.2101	5.0037	0.2555	4.9979	0.2365
$\beta_{12}(-2)$	− 1.9980	0.1290	− 2.0066	0.1432	− 1.9984	0.1201
$\beta_{22}(-6)$	− 5.9907	0.1211	− 6.0073	0.1232	− 6.0100	0.1341
$\sigma_1(2)$	2.0657	0.0837	2.0757	0.0802	2.0937	0.0901
$\sigma_2(2)$	1.9873	0.0838	2.1002	0.1092	2.1219	0.0979
$\gamma_1(0.1)$	0.1083	0.0020	0.0973	0.0031	0.1108	0.0108
$\gamma_2(0.6)$	0.6108	0.0053	0.5936	0.0044	0.6110	0.0098
$\nu(4)$	–	–	4.1093	0.3109	4.2018	0.6094
$\pi(0.4)$	0.4087	0.0198	0.4074	0.0422	0.4093	0.0277

quite accurate in all the three considered scenarios. Thus, the results suggest that the proposed *EM*-type algorithm produced satisfactory estimates of the proposed models on the all proposed scenarios.

4.1.2 Part2: consistency of estimations and convergence of BIC

In the further simulation study with various sample sizes, generating the following model given by

$$\begin{cases} Y_i = 1 - 2x_{i1} + \varepsilon_1, withProbability \pi = 1/2 \\ Y_i = 2 + 3x_{i1} + \varepsilon_2, withProbability 1 - \pi = 1/2, \end{cases}$$

for $i = 1, \ldots, n$, such that $x_{i1} \sim U(0, 1)$, and, ε_1 and ε_2 follow the *TP-SMN* distributions with the following parameters and as the assumption given in (5),

$$\sigma_1 = 1, \sigma_2 = 2, \gamma_1 = 0.25, \gamma_2 = 0.75, \nu = 3.$$

1000 samples from the above model for sample sizes $n = 50, 100, 250$ and $n = 450$ were generated respectively. Table 4 reports the mean squared errors (*MSE*) and the absolute bias (*Bias*) of the *MPL* estimates in each sample $j (= 1, \ldots, 1000)$ in a way that for each parameter $\theta \in \mathbf{\Theta}$, is defined respectively by.

$$Bias(\theta) = \frac{1}{1000} \sum_{j=1}^{1000} \left| \widehat{\theta}_j - \theta_j \right| \text{and} MSE(\theta) = \frac{1}{1000} \sum_{j=1}^{1000} \left(\theta_j - \widehat{\theta}_j \right)^2.$$

As it can be noticed from the Table 4, by increasing the sample size, the absolute biases and *MSE* of the *MPL* estimates tend to approach zero. These results indicate that the proposed *MPL* estimates of the *TP-SMN-MRM* based on the *ECME* algorithm do possess good consistency properties.

We consider further simulations with 100 samples with lengths of $n = 300$ from the above *TP-SMN-MRM*, where ε_1 and ε_2 follow the proposed *TP-T* distribution. We plotted the $BIC(\lambda)$ for each sample during the ECME algorithm in Fig. 2 (left) and also Barplot of mean of estimated numbers of components from 100 samples in Fig. 2 (right). Diagrams of $BIC(\lambda)$ show their monotonic behavior and converging during the ECME algorithm. Also Barplot of mean of estimated numbers of components show the true number of components (which is two-components) has the most frequency, which are convergence of the number of components during the ECME algorithm. These results together show the performances of the proposed estimates of the work with reasonability of choosing the best number of components.

4.1.3 Part3: robustness, misspecification and classification

In this part, the performance of the *TP-SMN-MRM* to cluster observations with unknown structure in the weakly and strongly separated datasets (homogeneous and heterogeneous, respectively) was investigated. In addition, a comparison was made to

Table 4 Absolute bias and MSE (in parentheses) of point estimates

n	β_{01}	β_{11}	β_{12}	β_{12}	σ_1	σ_2	γ_1	γ_2	ν
TP-N-MRM									
50	0.0064 (0.3409)	0.0086 (0.4034)	0.0101 (0.4054)	0.0098 (0.3965)	0.0072 (0.0654)	0.0074 (0.0875)	0.0008 (0.0070)	0.0012 (0.0121)	–
100	0.0056 (0.2865)	0.0079 (0.3207)	0.0088 (0.3709)	0.0086 (0.3054)	0.0068 (0.0457)	0.0066 (0.0511)	0.0006 (0.0054)	0.0008 (0.0065)	–
250	0.0032 (0.1293)	0.0035 (0.1982)	0.0043 (0.2018)	0.0037 (0.1982)	0.0031 (0.0198)	0.0041 (0.0231)	0.0005 (0.0048)	0.0006 (0.0051)	–
450	0.0029 (0.1218)	0.0031 (0.1603)	0.0037 (0.1791)	0.0034 (0.1203)	0.0027 (0.0140)	0.0029 (0.0156)	0.0005 (0.0039)	0.0004 (0.0040)	–
TP-T-MRM									
50	0.0073 (0.2386)	0.0080 (0.2981)	0.0079 (0.2566)	0.0082 (0.2773)	0.0081 (0.0598)	0.0073 (0.0602)	0.0010 (0.0086)	0.0014 (0.0098)	0.0076 (0.3647)
100	0.0061 (0.2004)	0.0068 (0.1954)	0.0071 (0.2011)	0.0069 (0.1850)	0.0070 (0.0455)	0.0061 (0.0499)	0.0009 (0.0080)	0.0012 (0.0091)	0.0061 (0.3018)
250	0.0028 (0.1341)	0.0033 (0.1117)	0.0026 (0.1098)	0.0030 (0.1221)	0.0033 (0.0211)	0.0049 (0.0198)	0.0005 (0.0043)	0.0005 (0.0057)	0.0034 (0.2091)
450	0.0025 (0.1245)	0.0030 (0.1098)	0.0025 (0.1041)	0.0030 (0.1198)	0.0029 (0.0166)	0.0040 (0.0187)	0.0003 (0.0040)	0.0004 (0.0049)	0.0030 (0.1675)
TP-SL-MRM									
50	0.0081 (0.3092)	0.0092 (0.3402)	0.0083 (0.2978)	0.0095 (0.3604)	0.0088 (0.0702)	0.0079 (0.0723)	0.0014 (0.0102)	0.0015 (0.0111)	0.0123 (0.3994)
100	0.0064 (0.2864)	0.0067 (0.3002)	0.0070 (0.2745)	0.0073 (0.3026)	0.0059 (0.0511)	0.0063 (0.0547)	0.0011 (0.0093)	0.0013 (0.0102)	0.0111 (0.3029)
250	0.0031 (0.1276)	0.0029 (0.1365)	0.0033 (0.1576)	0.0037 (0.1101)	0.0028 (0.0237)	0.0031 (0.0244)	0.0006 (0.0039)	0.0006 (0.0048)	0.0058 (0.1873)
450	0.0028 (0.1206)	0.0025 (0.1003)	0.0026 (0.1397)	0.0031 (0.1086)	0.0025 (0.0198)	0.0031 (0.0203)	0.0005 (0.0037)	0.0003 (0.0042)	0.0054 (0.1732)

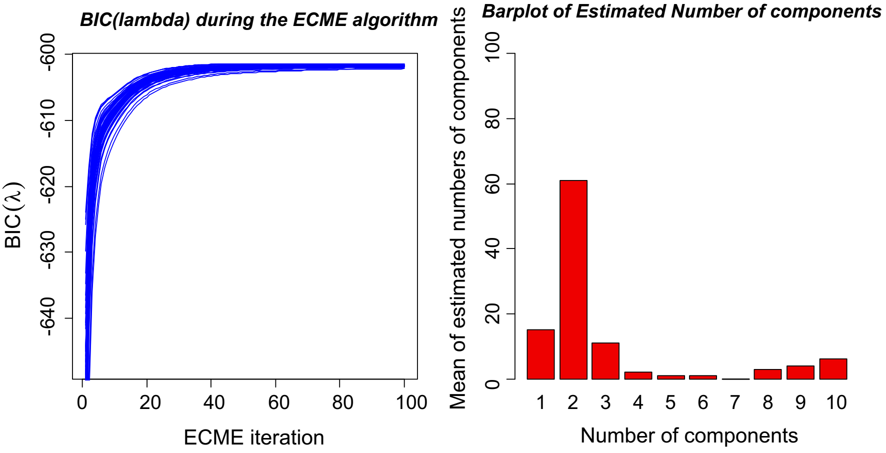

Fig. 2 BIC(λ) of 100 samples during the ECME algorithm (left) and Barplot of mean of estimated numbers of components from 100 samples (right)

find the applicability of some classic procedures to choose between the underlying *TP-SMN-MRM* for simulated data from another similar model which is an *MRM* based on the skew-*t* distributions (Branco and Day, 2001). To do the proposed simulations, the number of components ($G = 2$), sample size ($n = 700$) and the following parameter values were fixed in the two schemes strongly and weakly separated models. Then, without loss of generality, 700 samples from the proposed skew-*t-MRM* were artificially generated and, for each sample, the Normal-*MRM*, *TP-N-MRM*, *TP-T-MRM* and the *TP-SL-MRM* were fitted. The proposed skew-*t-MRM* had the asymmetric and heavy tails behavior and it was expected that the *TP-T-MRM* and (possibly the *TP-SL-MRM*) has the best fitting on them to have a robust inference.

Also, the quality of the classification of each mixture model is important. In this study, the methodology proposed by Liu and Lin (2014) is followed. The *correct classification rate* (CCR) index is based on the estimate of the posterior probability ($\widehat{z}_{ig}$) assigned to each subject, i.e., the maximum value of the $\widehat{z}_{ig}$, g $= 1, \ldots, G$ determines that an observation y_i belongs to its corresponding component of the mixture. So for tth ($t = 1, \ldots, 700$) sample of the 700 samples, the number of correct allocations (which are known in simulations) divided by the sample size $n = 700$, has been embedded as CCR_t and *mean of correct classification rate* ($MCCR$) was computed using the mathematical average of correct classification rate in the form of $MCCR = \frac{1}{700}\sum_{t=1}^{700} CCR_t$. Also *mean of the number of the correct allocation* (MCA) which is the average number of correct allocations on 700 samples has been considered.

Two schemes of the strongly and weakly separated models are given by:

- Strongly separated model:

$$\begin{cases} Y_i = 3 + 2x_{i1} + \varepsilon_1, with\ Probability\ \pi = 0.3 \\ Y_i = -1 - 2x_{i1} + \varepsilon_2, with\ Probability\ 1 - \pi = 0.7, \end{cases}$$

for $i = 1, \ldots, 700$, such that $x_{i1} \sim U(0, 1)$, and, ε_1 and ε_2 follow the skew-*t* distributions with zero mean, scale parameters $\sigma_1 = 1, \sigma_2 = 1$, shape parameters $\lambda_1 = -3, \lambda_2 = +3$, and degrees of freedom $\nu = 4$. Figure 3 shows a scatter plot and a histogram for one of these simulated samples.

- Weakly separated model:

$$\begin{cases} Y_i = 3 + 2x_{i1} + \varepsilon_1, with Probability\ \pi = 0.3 \\ Y_i = 1 - 1x_{i1} + \varepsilon_2, with Probability\ 1 - \pi = 0.7, \end{cases}$$

for $i = 1, \ldots, 700$, such that $x_{i1} \sim U(0, 1)$, and, ε_1 and ε_2 follow the skew-*t* distributions with zero mean, scale parameters $\sigma_1 = 2, \sigma_2 = 1$, shape parameters $\lambda_1 = -5, \lambda_2 = +5$, and degrees of freedom $\nu = 2$. Figure 3 shows scatter plots and histograms for one of these simulated samples on each scheme.

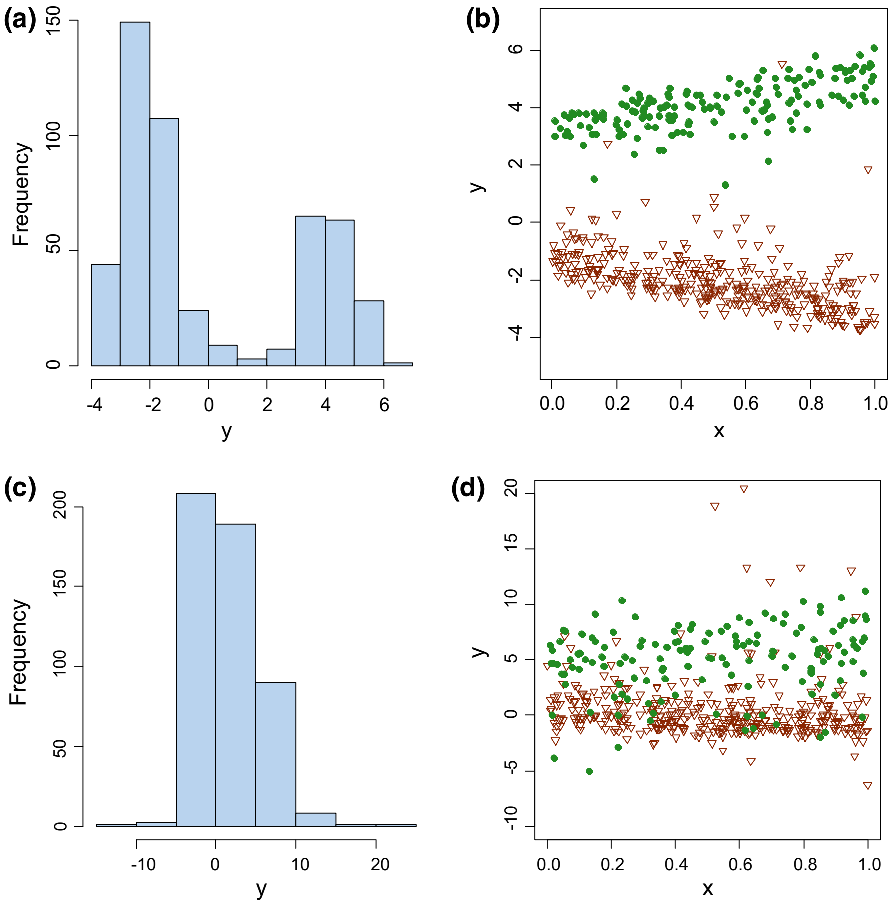

Fig. 3 **a** Histogram and **b** scatterplot of the strongly separated simulated skew-*t* MRM. **c** Histogram and **d** scatterplot of the weakly separated simulated *skew-t* MRM

Table 5 Correctness of classification analysis of the *TP-SMN-MRM* for 700 samples artificially generated from the skew-*t*-*MRM*

Schemes	Strongly separated model			Weakly separated model		
Fitted model	MCA	SD of MCA	MCCR	MCA	SD of MCA	MCCR
Normal-MRM	586.5461	98.0560	0.8379	449.8742	120.0874	0.6427
TP-N-MRM	611.9475	61.0933	0.8742	508.2844	87.0947	0.7261
TP-T-MRM	661.4231	42.7584	0.9449	597.7452	77.0931	0.8539
TP-SL-MRM	649.6031	57.9094	0.9280	586.3846	82.7463	0.8377

Table 6 Percentages that the best fitted *TP-SMN-MRM* are chosen using some model selection criteria

Schemes	Strongly separated model			Weakly separated model		
Condition examined	AIC (%)	EDC (%)	Log-likelihood (%)	AIC (%)	EDC (%)	Log-likelihood (%)
TP-T vs Normal	100	100	100	100	100	100
TP-T vs TP-N	98.57	98.57	98.57	98.85	98.85	98.85
TP-T vs TP-SL	97.86	97.86	97.86	98.49	99.49	99.49

Fitting the several models that belong to the *TP-SMN-MRM* on the generated datasets from the skew-*t*-*MRM* in the both strongly and weakly separated schemes, the *MCA* and standard deviation (SD) of correct allocations on 700 samples, as well as the *MCCR* are presented in Table 5. Note larger values indicate better classification results.

For each fitted model, the *AIC*, *EDC* and the log-likelihood criterion were computed. The percentage rates at which the best model was chosen for each criterion are recorded in Table 6. Note that as it was expected, all the criteria have satisfactory behavior, in that, they favor the best model, that is, the *TP-T-MRM*. Figure 4 shows the *AIC* values for each sample and the best (expected and robust) *TP-T-MRM* and *TP-N-MRM*.

4.2 Application

In this section, the proposed models and methods on datasets which the first represent the perception of musical tones by musicians are illustrated as they are described in Cohen (1984), and the second represent the US census population and poverty percentage estimates by county.

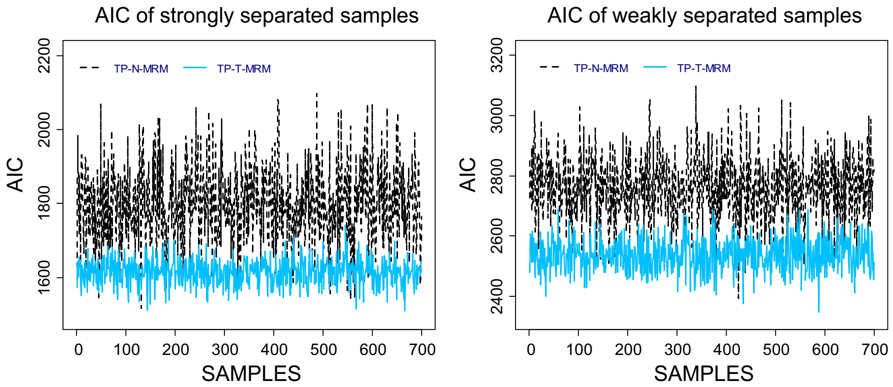

Fig. 4 *AIC* values of 700 samples with blue line for *TP-T-MRM* and black dashed line for *TP-N-MRM*

4.2.1 Tone perception data

In the well-known data, a pure fundamental tone with electronically generated overtones added was played to a trained musician. In this experiment, the subjects were asked to tune an adjustable tone to one octave above the fundamental tone and their perceived tone was recorded versus the actual tone. A number of 150 trials from the same musician were recorded in this experiment. The overtones were determined by a stretching ratio which is the ratio between the adjusted and the fundamental tone. The experiment was designed to find out how the tuning ratio affects the perception of the tone and decide which of the two musical perception theories was reasonable. So we consider the actual tune ratio as the explanatory variable x and perceived tone ratio as the response variable Y.

The scatter plot and the histogram of the perceived tone ratio are plotted in Fig. 5.

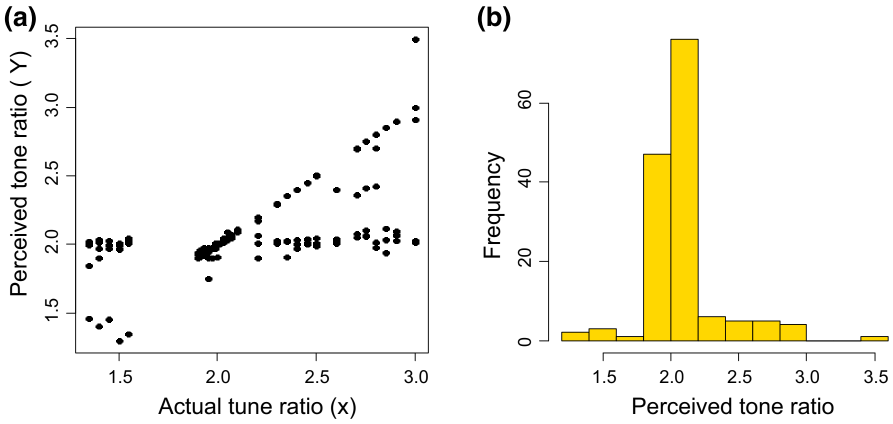

Fig. 5 **a** Scatterplot and **b** histogram of the tone perception data

These plots demonstrate that there are two groups with separate trends in the dataset and they have a non-normal distribution. Based on the realizations of the data, Cohen (1984) discussed two hypotheses which were called the interval memory hypothesis and the partial matching hypothesis. Many have considered and modeled this data using a mixture of linear regressions framework, see DeVeaux (1989), Viele and Tong (2002), Hunter and Young (2012), Yao et al. (2014), Zeller et al. (2016) and Doğru and Arslan (2017). Zeller et al. (2016) and Doğru and Arslan (2017) propose the robust mixture regression using the *SMSN* distributions which are similar counterparts of the *TP-SMN* distributions.

The proposed *TP-SMN-MRM* was expanded to model the data. Using the *ECME* algorithm, the *MPL* estimates together with their corresponding standard errors (based on the square root of invers of the observed information matrix form Sect. 3.2) of the parameters from the Normal-*MRM*, *TP-N-MRM*, *TP-T-MRM*, *TP-SL-MRM* and the skew-*t-MRM* (as asymmetry heavy-tailed competitor) are presented in Table 7. According to the recorded model selection criteria, numbers and elapsed time (s) of algorithm iterations (N.I. and E.T., respectively) in Table 8, the best fitted *TP-SMN-MRM* of the tone perception data is the *TP-T-MRM*. Observing the estimated parameters of the best fitted model, it is concluded that the model which is based on the asymmetric distribution with heavier tails provides a better fit compared to the ordinary, normal and the *TP-N* distribution.

Figure 6 shows the scatter plot of the data set with the lightly and heavy tailed fitted *TP-N-MRM* and *TP-T-MRM*, respectively, and clustering of the dataset. Clustering of the data based on the fitted skew-*t-MRM* is also in the Fig. 7. In Fig. 8, we plot the profile log-likelihood of the parameter ν for the *TP-T-MRM* and skew-*t-MRM* in all of *ECME* algorithm iterations.

4.2.2 US population and poverty percentage counties data

In this subsection we consider a dataset which is provided in "*usmap*" package from *R* software called "*countypop*" and "*countypov*" which are the 2015 population estimate (in number of people) for the corresponding county (see also, https://www.census.gov/programs-surveys/popest.html), and the 2014 poverty percentage estimate (in percent of county population) for the corresponding county (see also, https://www.census.gov/topics/income-poverty/poverty.html), respectively. We consider the logarithm of population estimate as the explanatory variable and poverty estimate as the response variable. *MPL* estimates and their corresponding standard errors of the parameters from the *TP-T-MRM* (the best fitted *TP-SMN-MRM*) and the skew-*t-MRM* on the dataset are presented in Table 9. The estimations of the shape parameters (γ_g, $g = 1, 2$) and degrees of freedom (ν) of the fitted *TP-T-MRM*, show that both of the fitted components have skew behavior and are heavy-tailed. Also, the estimation of regression coefficients of components and Fig. 9, which are the clustering the US counties dataset based on the fitted *TP-T-MRM* the skew-*t-MRM*, demonstrates us that, in the first component the levels of poverty percentage are more than the second. Also in the first component by increasing the population estimates, poverty percentage estimates are decreasing, while in the second component it seems the population estimates are not effective on

Table 7 *MPL* estimation results with their standard errors for fitting several *TP-SMN-MRM* and skew-*t*-*MRM* on the tone perception data

Model	Normal-MRM		TP-N-MRM		TP-T-MRM		TP-SL-MRM		Skew-*t*-*MRM*	
Parameter	Est	S.E	Est	S.E	Est	S.E	Est	S.E	Est	S.E
β_{01}	− 0.0439	0.0649	− 0.0334	0.0702	0.0045	0.0086	− 0.0130	0.0201	0.0043	0.0198
β_{11}	1.0113	0.0293	0.9971	0.0238	0.9976	0.0193	0.9981	0.0199	0.9986	0.0241
β_{02}	1.8902	0.0411	1.9204	0.0382	1.9371	0.0541	1.9311	0.0764	1.9419	0.0812
β_{12}	0.0575	0.0179	0.0400	0.0160	0.0335	0.0102	0.0371	0.0213	0.0335	0.0242
σ_1	0.0839	0.0009	0.2529	0.0007	0.0078	0.0008	0.0165	0.0013	0.0044	0.0023
σ_2	0.0839	0.0008	0.0874	0.0007	0.0777	0.0009	0.0612	0.0015	0.0385	0.0019
γ_1	–	–	0.4548	0.0052	0.3520	0.0079	0.0449	0.0098	− 0.0440	0.9048
γ_2	–	–	0.3313	0.0039	0.4641	0.0065	0.5145	0.0113	− 0.0085	0.0511
ν	–	–	–	–	2.1001	0.0383	1.0637	0.0398	2.1001	0.1028
π	0.3238	0.0478	0.3164	0.0468	0.3828	0.0321	0.4108	0.0534	0.3847	0.0702
λ	0.1921	–	0.1666	–	0.1668	–	0.1803	–	0.1645	–

Table 8 Some model selection criteria, numbers and elapsed time of algorithm iterations (N.I. and E.T., respectively) for the fitted *TP-SMN-MRM* and skew-*t-MRM* of the tone perception data

Model Selection criteria	Normal-MRM	TP-N-MRM	TP-T-MRM	TP-SL-MRM	Skew-*t-MRM*
Log-likelihood	105.2510	144.2888	217.5689	195.7090	215.9857
BIC(λ)	80.1978	119.2357	187.5090	165.6452	186.0273
AIC	− 194.5020	− 270.5778	− 417.1378	− 371.4180	− 413.9714
BIC	− 170.4169	− 243.4821	− 390.0420	− 341.3116	− 386.8757
EDC	− 190.9061	− 266.5324	− 413.0924	− 366.9231	− 409.9260
N.I	21	20	8	18	12
E.T	29.04	28.73	20.34	36.92	27.24

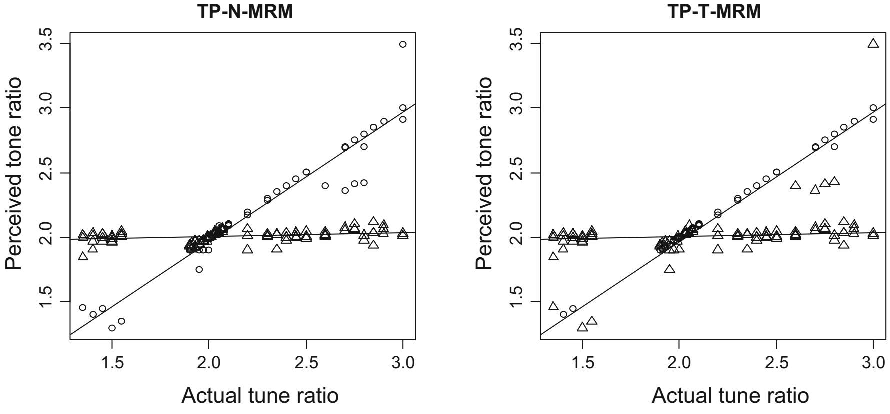

Fig. 6 The scatterplots and clustering of the tone perception data based on the lightly and heavily tailed fitted *TP-N-MRM* and *TP-T-MRM*

the poverty percentage estimates. Clustering the US counties based on the proposed fitted *TP-T-MRM* is provided in the US map in Fig. 10.

5 Conclusion

Finite mixture of regression models is a research area with several applications. In the current study, a model called the *TP-SMN* distributions was proposed based on a flexible class of symmetric/asymmetric and lightly/heavy tailed distribution. In fact, the proposed model is a generalization of the work carried out by Yao et al. (2014) and Liu and Lin (2014) that can efficiently and simultaneously deal with skewness and heavy-tailed-ness in the mixture regression model setting. Also we have used the penalized likelihood to have the best number of components, and the robust proposed model allows the researchers on different areas to analyze data in an extremely flexible

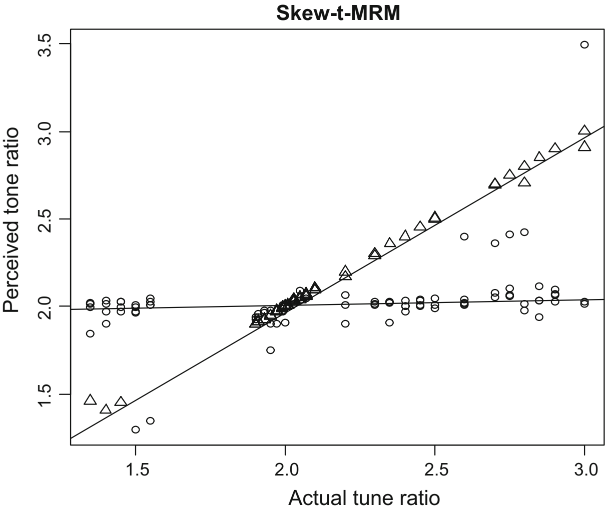

Fig. 7 The scatter plot and clustering of the tone perception data based on the skew-*t*-*MRM*

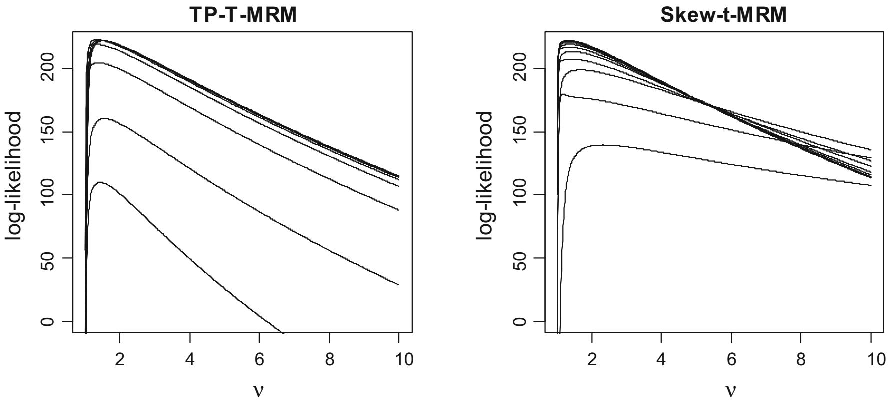

Fig. 8 Plots of the profile log-likelihood on the *EM*-algorithm iterations of the parameter ν for fitting the perception data with a two component *TP*-*T*-*MRM* (left) and skew-*t*-*MRM* (right)

methodology. An *EM*-type algorithm was employed and some simulation studies were presented to show that this algorithm gives reasonable estimates. After obtaining the *MPL* estimates via the *ECME* algorithm, they were easily implemented and coded with existing statistical software such as the *R* package, and the R code is available from us upon request. Results of the work indicated that using the *TP*-*SMN*-*MRM* leads to a better fit, solves the outliers' issues and gives a more precise picture of robust inferences. It is intended to pursue a fully Bayesian inference via the Markov chain Monte Carlo method on the proposed model in future research.

Table 9 MPL estimation results with their standard errors for fitting *TP-T-MRM* on the US population and poverty percentage counties data

Model	TP-T-MRM				Skew-t-MRM			
Component	g = 1		g = 2		g = 1		g = 2	
Parameter	Est	S.E	Est	S.E	Est	S.E	Est	S.E
β_{0g}	30.7825	0.0903	14.4010	0.0714	11.6434	0.0965	27.0811	0.0932
β_{1g}	– 0.7610	0.0110	– 0.0417	0.0094	0.0068	0.0135	– 0.6070	0.0114
σ_g	9.1269	0.0814	7.5820	0.0633	3.1360	0.0813	5.2881	0.0620
γ_g	0.6858	0.0064	0.6713	0.0055	0.1740	0.0401	0.0940	0.0076
π_g	0.4350	0.0463	0.5650	0.0463	0.4558	0.0470	0.5442	0.0417
ν	6.4796	0.0741	6.4796	0.0741	7.2925	0.0752	7.2925	0.052
λ	0.1699				0.2501			
Log-likelihood	– 10,336.63				– 10,425.65			
BIC(λ)	– 10,384.48				– 10,476.14			
AIC	20,691.26				20,869.30			
BIC	20,745.74				20,923.78			
EDC	20,774.16				20,952.20			
N.I	12				13			
E.T	82.34				452.56			

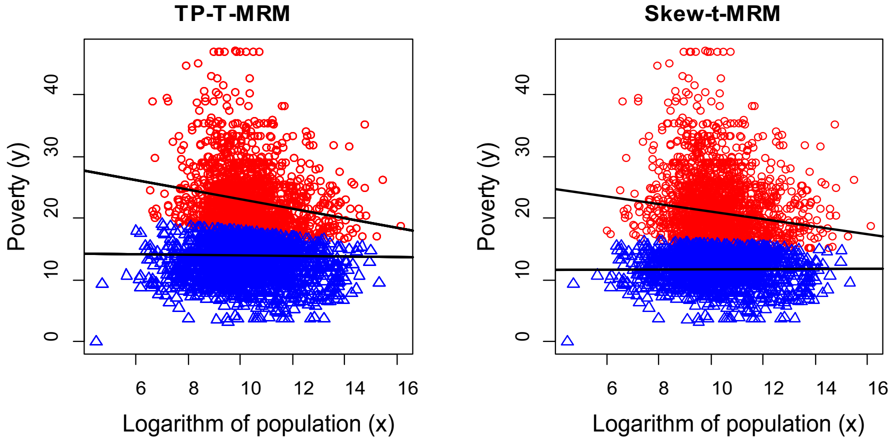

Fig. 9 Clustering of the US counties based on the estimated *TP-T-MRM* to population and poverty percentage estimate data (light color is due to the first cluster and dark color is due to the second cluster

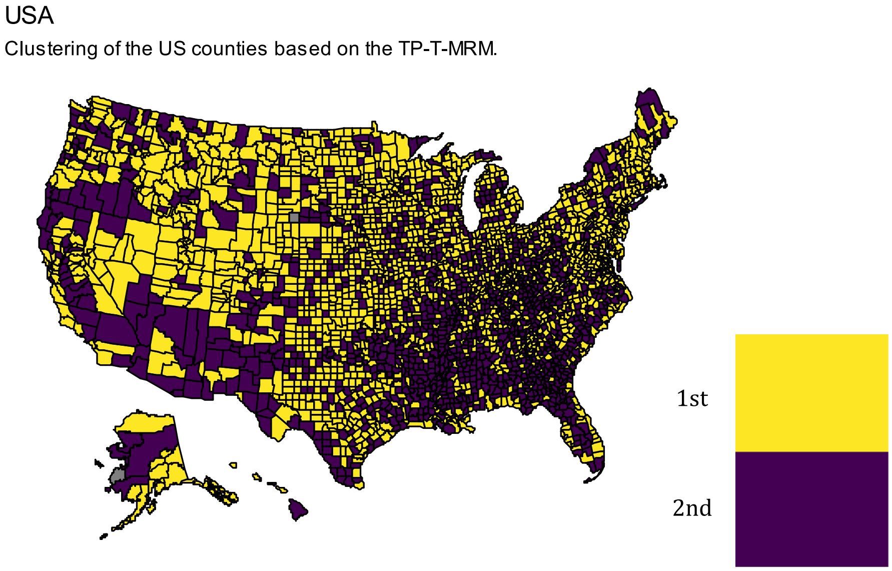

Fig. 10 The scatterplot and clustering of the US population and poverty percentage estimate data based on the *TP-T-MRM* (left) and skew-*t-MRM*. Top members (red color) are due to the first cluster and bottom members (blue color) are due to the second cluster

Acknowledgements We would like to express our very great appreciation to editor, associate editor and two anonymous reviewers for their valuable and constructive suggestions during the planning and development of this research work.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Andrews DR, Mallows CL (1974) Scale mixture of normal distribution. J R Stat Soc B 36:99–102

Arellano-Valle RB, Gómez H, Quintana FA (2005) Statistical inference for a general class of asymmetric distributions. J Stat Plan Inference 128:427–443

Arellano-Valle RB, Castro LM, Genton MG, Gómez HW (2008) Bayesian inference for shape mixtures of skewed distributions, with application to regression analysis. Bayesian Anal 3(3):513–539

Bai X, Yao W, Boyer JE (2012) Robust fitting of mixture regression models. Comput Stat Data Anal 56:2347–2359

Barkhordar Z, Maleki M, Khodadadi Z, Wraith D, Negahdari F (2020) A Bayesian approach on the two-piece scale mixtures of normal homoscedastic nonlinear regression models. J Appl Stat. https://doi.org/10.1080/02664763.2020.1854203

Basford KE, Greenway DR, Mclachlan GJ, Peel D (1997) Standard errors of fitted component means of normal mixtures. Comput Stat 12:1–17

Bazrafkan M, Zare K, Maleki M, Khodadadi Z (2021) Partially linear models based on heavy-tailed and asymmetrical distributions. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02101-1

Branco MD, Dey DK (2001) A general class of multivariate skew-elliptical distributions. J Multivar Anal 79:99–113

Cohen E (1984) Some effects of inharmonic partials on interval perception. Music Percept 1:323–349

Cosslett SR, Lee L-F (1985) Serial correlation in latent discrete variable models. J Econom 27(1):79–97

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Methodol 39:1–22

DeSarbo WS, Cron WL (1988) A maximum likelihood methodology for clusterwise linear regression. J Classif 5:248–282

DeSarbo WS, Wedel M, Vriens M, Ramaswamy V (1992) Latent class metric conjoint analysis. Mark Lett 3(3):273–288

DeVeaux RD (1989) Mixtures of linear regressions. Comput Stat Data Anal 8(3):227–245

Doğru FZ, Arslan O (2017) Robust mixture regression based on the skew t distribution. Revista Colombiana De Estadística 40(1):45–64. https://doi.org/10.15446/rce.v40n1.53580

Engel C, Hamilton JD (1990) Long swings in the Dollar: are they in the data and do markets know it? Am Econ Rev 80(4):689–713

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Ghasami S, Maleki M, Khodadadi Z (2020) Leptokurtic and platykurtic class of robust symmetrical and asymmetrical time series models. J Comput Appl Math. https://doi.org/10.1016/j.cam.2020.112806

Hajrajabi A, Maleki M (2019) Nonlinear semiparametric autoregressive model with finite mixtures of scale mixtures of skew normal innovations. J Appl Stat 46(11):2010–2029

Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384

Hoseinzadeh A, Maleki M, Khodadadi Z (2021) Heteroscedastic nonlinear regression models using asymmetric and heavy tailed two-piece distributions. AStA Adv Stat Anal 105:451–467

Huang T, Peng H, Zhang K (2017) Model selection for Gaussian mixture models. Stat Sin 27(1):147–169

Hunter DR, Young DS (2012) Semiparametric mixtures of regressions. J Nonparametric Stat 24(1):19–38

Lin TI, Lee JC, Hsieh WJ (2007) Robust mixture modelling using the skew t distribution. Stat Comput 17:81–92

Liu M, Lin T-I (2014) A skew-normal mixture regression model. Educ Psychol Meas 74:139–162

Liu C, Rubin DB (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika 81:633–648

Liu M, Hancock GR, Harring JR (2011) Using finite mixture modeling to deal with systematic measurement error: a case study. J Mod Appl Stat Methods 10(1):249–261

Mahmoudi MR, Maleki M, Baleanu D, Nguyen VT, Pho KH (2020) A Bayesian approach to heavy-tailed finite mixture autoregressive models. Symmetry 12(6):929

Maleki M (2022) Time series modelling and prediction of the coronavirus outbreaks (COVID-19) in the World. In: Azar AT, Hassanien AE (eds) Modeling, control and drug development for COVID-19 outbreak prevention: studies in systems, decision and control, vol 366. Springer, Cham. https://doi.org/10.1007/978-3-030-72834-2_2

Maleki M, Mahmoudi MR (2017) Two-piece location-scale distributions based on scale mixtures of normal family. Commun Stat Theory Methods 46(24):12356–12369

Maleki M, Nematollahi AR (2017) Bayesian approach to epsilon-skew-normal family. Commun Stat Theory Methods 46(15):7546–7561

Maleki M, Wraith D (2019) Mixtures of multivariate restricted skew-normal factor analyzer models in a Bayesian framework. Comput Stat 34:1039–1053

Maleki M, Barkhordar Z, Khodadado Z, Wraith D (2019a) A robust class of homoscedastic nonlinear regression models. J Stat Comput Simul 89(14):2765–2781

Maleki M, Contreras-Reyes JE, Mahmoudi MR (2019b) Robust mixture modeling based on two-piece scale mixtures of normal family. Axioms 8(2):38. https://doi.org/10.3390/axioms8020038

Maleki M, Wraith D, Arellano-Valle RB (2019c) Robust finite mixture modeling of multivariate unrestricted skew-normal generalized hyperbolic distributions. Stat Comput 29(3):415–428

Maleki M, Hajrajabi A, Arellano-Valle RB (2020a) Symmetrical and asymmetrical mixture autoregressive processes. Braz J Probab Stat 34(2):273–290

Maleki M, Mahmoudi MR, Wraith D, Pho KH (2020b) Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel Med Infect Dis. https://doi.org/10.1016/j.tmaid.2020.101742

Maleki M, McLachlan G, Lee S (2021) Robust clustering based on finite mixture of multivariate fragmental distributions. Stat Model. https://doi.org/10.1177/1471082X211048660

Maleki M, Bidram H, Wraith D (2022) Robust clustering of COVID-19 cases across U.S. counties using mixtures of asymmetric time series models with time varying and freely indexed covariates. J Appl Stat. https://doi.org/10.1080/02664763.2021.2019688

Markatou M (2000) Mixture models, robustness, and the weighted likelihood methodology. Biometrics 56:483–486

McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York

Meng X, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80:267–278

Moravveji B, Khodadadi Z, Maleki M (2019) A Bayesian analysis of two-piece distributions based on the scale mixtures of normal family. Iran J Sci Technol Trans Science 43(3):991–1001

Mudholkar GS, Hutson AD (2000) The epsilon-skew-normal distribution for analyzing near-normal data. J Stat Plan Inference 83(2):291–309

Naik PA, Shi P, Tsai C-L (2007) Extending the Akaike information criterion to mixture regression models. J Am Stat Assoc 102(477):244–254

Quandt RE (1972) A new approach to estimating switching regressions. J Am Stat Assoc 67:306–310

Quandt RE, Ramsey JB (1978) Estimating mixtures of normal distributions and switching regressions. J Am Stat Assoc 73(364):730–738

Resende PAA, Dorea CCY (2016) Model identification using the efficient determination criterion. J Multivar Anal 150:229–244

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Song W, Yao W, Xing Y (2014) Robust mixture regression model fitting by Laplace distribution. Comput Stat Data Anal 71:128–137

Späth H (1979) Algorithm 39 clusterwise linear regression. Computing 22(4):367–373

Teicher H (1963) Identifiability of finite mixtures. Ann Math Stat 34(4):1265–1269

Tibshirani RJ (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B 58:267–288

Turner TR (2000) Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. J R Stat Soc Ser C (appl Stat) 49(3):371–384

Viele K, Tong B (2002) Modeling with mixtures of linear regressions. Stat Comput 12(4):315–330

Wang H, Li R, Tsai C-L (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94:553–568

Yao W, Wei Y, Yu C (2014) Robust mixture regression using the t-distribution. Comput Stat Data Anal 71:116–127

Zeller CB, Cabral CRB, Lachos VH (2016) Robust mixture regression modeling based on scale mixtures of skew-normal distributions. TEST 25:375–396

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.