**Computational Statistics manuscript No.**
(will be inserted by the editor)

# Robust order selection of mixtures of regression models with random effects

**Luísa Novais · Susana Faria**

Received: date / Accepted: date

**Abstract** Finite mixtures of regression models with random effects are a very flexible statistical tool to model data, as these models allow to model the heterogeneity of the population and to account for multiple correlated observations from the same individual at the same time. The selection of the number of components for these models has been a long-standing challenging problem in statistics. However, the majority of the existent methods for the estimation of the number of components are not robust and, therefore, are quite sensitive to outliers. In this article we study a robust estimation of the number of components for mixtures of regression models with random effects, investigating the performance of trimmed information and classification criteria comparatively to the performance of the traditional information and classification criteria. The simulation study and a real-world application showcase the superiority of the trimmed information and classification criteria in the presence of contaminated data.

**Keywords** Mixture models · Number of components · Trimmed Information Criteria · Trimmed Classification Criteria · Robustness

L. Novais
Department of Mathematics and Centre of Molecular and Environmental Biology, University of Minho, 4800-058 Guimarães, Portugal
E-mail: luisa_novais92@hotmail.com

S. Faria
Department of Mathematics and Centre of Molecular and Environmental Biology, University of Minho, 4800-058 Guimarães, Portugal
Tel.: +351 253510434
Fax: +351 253510401

## 1 Introduction

Mixture models have been extensively studied in the modelling and analysis of data from a heterogeneous population, in other words, a population divided into subpopulations present in unknown proportions. Within the context of mixtures of regression models, mixtures of regression models with random effects can be applied to the most variety of subjects, since they allow the explaination of correlations between observations of the same individual, through the incorporation of random effects, and, at the same time, to model the unobserved heterogeneity between the distinct individuals.

In the literature there are several examples of application of mixture models for the statistical modelling of different phenomena, assuming particular relevance in the areas of astronomy, biology, marketing, economics, medicine, among others (see Frühwirth-Schnatter 2006; Celeux et al. 2005 and Young and Hunter 2015). A comprehensive review of finite mixture models can be found in McLachlan and Peel (2000).

The estimation of the number of components is one of the most important problems in the context of mixture models because the statistical inference about the resulting model is highly sensitive to the value of the number of components (see Kasahara and Shimotsu 2015). For maximum likelihood estimation of finite mixture models, information and classification criteria present one of the simplest ways to estimate the number of components, which made their use quite popular. Depraetere and Vandebroek (2014) study different information and classification criteria, carrying out a large simulation study for mixtures of regression models and they verify that the performance of these criteria depends greatly on the model, concluding that there is not a single criterion that works well for all the simulated scenarios. Hui et al. (2015) study the behaviour of information criteria for order selection in finite mixture models, based on either the observed or the complete likelihood and propose a new order consistent criteria based on the observed likelihood, the $AIC_{mix}$. The authors show in their simulation study the poor finite-sample performance of the complete likelihood criteria, while showing the strong performance of $BIC$ and their $AIC_{mix}$ criterion. McLachlan and Rathnayake (2014) review various methods that have been proposed to select the number of components in a Gaussian mixture model, mainly focusing in information and classification criteria and in resampling approaches as the likelihood ratio test. Celeux et al. (2019) focus on the Bayesian solutions to the different interpretations of selecting the correct number of components for mixture models, reviewing well-known methods such as the reversible jump Markov Chain Monte Carlo (MCMC) to more recent ideas. Cappozzo et al. (2019) study a robust approach to model-based classification, introducing a robust modification to the Model-Based Classification framework, by employing impartial trimming and constraints. The authors propose a robust information criterion and underline the benefits of their method in real and simulated data. Li et al. (2016) study the use of trimmed information criteria to robustly estimate the number of components in mixtures of linear regressions, concluding that these crite-

ria are robust and not sensitive to outliers in comparison to the traditional criteria.

With this article, we intend to extend the work of Li et al. (2016) for mixtures of regression models with random effects. For that, we focus on order selection by using different information and classification criteria, both in their traditional version and in their robust version, to determine the number of components for finite mixtures of regression models with random effects. A simulation study and a real data application demonstrate that the two versions of the criteria perform similarly when there are no outliers present but the robust criteria perform much better than the traditional criteria in the presence of outliers.

The remainder of the paper is organized as follows. In Section 2 we present an overview of mixtures of regression models with random effects. In Section 3 we give an introduction of two versions of a series of information and classification criteria for order selection. In Section 4 we provide a simulation study to compare the performance of both versions of the different criteria in the selection of the number of components for mixtures of regression models with random effects. In Section 5 we use a real-world application to demonstrate the effectiveness of the robust criteria. A discussion section ends the paper.

## 2 Finite mixtures of regression models with random effects

We assume that $I$ is the number of individuals in the study, where we observe each individual $n_i$ times. It is also assumed that the population is heterogeneous and can be divided into $m$ somewhat homogeneous subpopulations. For each individual $i$, that is, for $i = 1, ..., I$, $Z_i$ is a latent variable varying from $1, ..., m$ with probabilities $\pi_1, ..., \pi_m$, respectively. In short, $Z_i$ is an unobserved variable representing the subpopulation to which the individual $i$ belongs, such that $P(Z_i = j) = \pi_j$ for $j = 1, ..., m$. However, in the estimation of the parameters using the EM algorithm (Dempster et al. 1977), it is convenient to use a $m$-dimensional vector $\boldsymbol{Z}_i$ instead of using the latent variable $Z_i$, where the $j$-th element of $\boldsymbol{Z}_i$, $Z_{ij}$, is defined as being equal to one or zero, whether the individual belongs, or not, to subpopulation $j$.

Given $Z_i = j$, the response variable $\boldsymbol{y}_i \in \mathbb{R}^{n_i}$ follows a linear mixed model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_j + \boldsymbol{U}_i\boldsymbol{b}_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{1}$$

where $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times p}$ and $\boldsymbol{U}_i \in \mathbb{R}^{n_i \times q}$ are, respectively, the fixed and random-effects covariate matrix, $\boldsymbol{\beta}_j \in \mathbb{R}^p$ and $\boldsymbol{b}_{ij} \in \mathbb{R}^q$ are, respectively, a fixed and random-effects vector, and $\boldsymbol{\varepsilon}_{ij} \in \mathbb{R}^{n_i}$ is the random error vector.

We also consider that $\boldsymbol{b}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ are independent, for $i = 1, ..., I$ and $j = 1, ..., m$, and that $\boldsymbol{b}_{ij} \sim N_q(\boldsymbol{0}, \boldsymbol{\Psi}_j)$ and $\boldsymbol{\varepsilon}_{ij} \sim N_{n_i}(\boldsymbol{0}, \boldsymbol{\Lambda}_{ij})$. In this study we assume that $\boldsymbol{\Lambda}_{ij} = \sigma_j^2 \boldsymbol{I}_{n_i}$.

Therefore, the conditional distribution of $\boldsymbol{y_i}$ given $\boldsymbol{X}_i$, $\boldsymbol{U}_i$ and $\boldsymbol{\theta}$ without observing $Z_i$ can be written as

$$f(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}_j, \boldsymbol{U}_i\boldsymbol{\Psi}_j\boldsymbol{U}_i^T + \sigma_j^2\boldsymbol{I}_{n_i}), \tag{2}$$

for $i = 1, ..., I$, where $\pi_j$ is the mixing proportion for $j = 1, ..., m$ with $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{m} \pi_j = 1$, and $\boldsymbol{\theta}$ is the parameter vector. In this case we have that $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\pi} = (\pi_1, ..., \pi_m)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m)$, $\boldsymbol{\Psi} = (\text{vech}(\boldsymbol{\Psi}_1), ..., \text{vech}(\boldsymbol{\Psi}_m))$, where $vech(.)$ is the half-vectorization function giving the lower triangular portion of a symmetric matrix in form of a vector, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_m^2)$.

Nonetheless, for mixture models maximizing the log-likelihood function can be very complex, since the log-likelihood function is unbounded, producing situations where the maximum likelihood estimator may not exist, at least in a global way. To solve this problem iterative methods are commonly used, in particular the EM algorithm of Dempster et al. (1977), which consists in an iterative calculation of the expectation and maximization of the complete log-likelihood function. For finite mixtures of linear mixed models, Grün (2008) describes the estimation procedure with the EM algorithm and outlines an alternative version of the EM algorithm where only the component membership is treated as missing data, as opposed to the traditional EM algorithm where both the component membership and the random effects are treated as missing data.

## 3 Information and classification criteria for order selection

Due to its simplicity, the most common parametric methods to select the number of components of a mixture model consist of using information or classification criteria. In order to accomplish that, mixture models with different numbers of components are fitted to the data and the number of components that corresponds to the smallest value of each information or classification criterion is selected. Information criteria are based on penalizing the logarithm of the likelihood function, also known as the observed log-likelihood function, which can be written as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{I} \ln\left\{\sum_{j=1}^{m} \pi_j N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}_j, \boldsymbol{U}_i\boldsymbol{\Psi}_j\boldsymbol{U}_i^T + \sigma_j^2\boldsymbol{I}_{n_i})\right\}, \tag{3}$$

while classification criteria are based on penalizing the complete log-likelihood function, also called classification log-likelihood function, which can be written as

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{j=1}^{m} \ln\left\{\left(\pi_j N_{n_i}(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_j + \boldsymbol{U}_i\boldsymbol{b}_{ij}, \sigma_j^2\boldsymbol{I}_{n_i}) N_q(\boldsymbol{b}_{ij}; \boldsymbol{0}, \boldsymbol{\Psi}_j)\right)^{Z_{ij}}\right\}. \tag{4}$$

In this study, we use the following information criteria: Akaike Information Criterion (*AIC*) (Akaike 1974), Bayesian Information Criterion (*BIC*) (Schwarz 1978), which are the two most widely used information criteria, the consistent version of the Akaike Information Criterion (*CAIC*) (Bozdogan 1987), Hannan-Quinn Information Criterion (*HQIC*) (Hannan and Quinn 1979), which is also consistent, Kullback Information Criterion (*KIC*) (Cavanaugh 1999), Akaike Information Criterion 4 (*AIC4*) (Bhansali and Downham 1977), which consists of increasing the penalty of the *AIC* from $2k$ to $4k$, Adjusted Bayesian Information Criterion (*aBIC*) (Sclove 1987), Corrected Akaike Information Criterion ($AIC_c$) (Hurvich and Tsai 1989), Corrected Kullback Information Criterion ($KIC_c$) (Cavanaugh 2004), which are, respectively, alternatives to *BIC*, *AIC* and to *KIC* for small samples, Minimum Description Length Criterion 2 (*MDL2*) (Liang et al. 1992) and Minimum Description Length Criterion 5 (*MDL5*) (Liang et al. 1992), which consist of multiplying the penalty term of *BIC* by 2 and 5, respectively. The following classification criteria are also used: Classification Likelihood Criterion (*CLC*) (Biernacki and Govaert 1997), Approximate Weight of Evidence Criterion (*AWE*) (Banfield and Raftery 1993), which can be similar to *BIC* when the components are well separated, Normalized Entropy Criterion (*NEC*) (Celeux and Soromenho 1996), which uses the entropy directly, and the large sample *BIC* approximation to *ICL* (*ICL-BIC*) (Biernacki et al. 2000), which the authors found that it presents very similar results to their *ICL* criterion.

More details on these criteria can be found in Novais and Faria (2021).

### 3.1 Trimmed information and classification criteria

Since the maximum likelihood estimation is sensitive to outliers, the information and classification criteria stated in the previous section may be influenced by outliers. As a result, the presence of a single outlier may cause changes in the estimated number of components (Li et al. 2016). Thus, in this section we propose to use a robust version of those information and classification criteria, based on trimmed maximum likelihood estimates.

Assuming that $\alpha \times 100\%$ of the observations consist in outliers, the trimmed maximum likelihood estimate (TLE) for mixture models, proposed by Neykov et al. (2007) and Müller and Neykov (2003), uses only $(1-\alpha) \times 100\%$ of the observations to fit the model, removing the remaining ones, that is, $\max_{\boldsymbol{I}_\alpha} \max_{\boldsymbol{\theta}} \sum_{i \in \boldsymbol{I}_\alpha} \ln(f(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{\theta}))$ where $\boldsymbol{I}_\alpha$ is the set of all the $\lfloor I(1-\alpha) \rfloor$-subsets of $\{1, ..., I\}$ and $f(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{\theta})$ is defined in (2) (see Li et al. 2016 and Yu et al. 2020 for details on the TLE).

Hence, the robust version of the information and classification criteria stated in Section 3 can be found on Table 1, where $k$ is the number of parameters to be estimated, $n$ is the number of observations, $\hat{\boldsymbol{\theta}}_t = (\hat{\boldsymbol{\pi}}_t, \hat{\boldsymbol{\beta}}_t, \hat{\boldsymbol{\Psi}}_t, \hat{\boldsymbol{\sigma}}_t^2)$ is the trimmed maximum likelihood estimate and $l(\hat{\boldsymbol{\theta}}_t)$ and $l_c(\hat{\boldsymbol{\theta}}_t)$ are, respectively, the maximum value of the observed and complete log-likelihood function

for the estimated mixture model. For *TCLC* and *TNEC*, $\hat{z}_{ij}$ can be written as

$$\hat{z}_{ij} = E(Z_{ij} \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \boldsymbol{y}) = \frac{\hat{\pi}_j N_{n_i}(\boldsymbol{y}_i; \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_j, \boldsymbol{U}_i\hat{\boldsymbol{\Psi}}_j\boldsymbol{U}_i^T + \hat{\sigma}_j^2\boldsymbol{I}_{n_i})}{\sum_{j=1}^{m} \hat{\pi}_j N_{n_i}(\boldsymbol{y}_i; \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_j, \boldsymbol{U}_i\hat{\boldsymbol{\Psi}}_j\boldsymbol{U}_i^T + \hat{\sigma}_j^2\boldsymbol{I}_{n_i})}. \quad (5)$$

For *TNEC*, $l_{(1)}$ is the maximized log-likelihood function for the 1-component mixture model. Since this criterion is not set for $m = 1$, as Biernacki et al. (1999) do, we consider that *TNEC* takes the value 1 for these cases.

**Table 1** Robust information and classification criteria

| Trimmed information criteria | |
|---|---|
| Criteria | Formula |
| *TAIC* | $-2l(\hat{\boldsymbol{\theta}}_t) + 2k$ |
| *TBIC* | $-2l(\hat{\boldsymbol{\theta}}_t) + k\ln(n)$ |
| *TCAIC* | $-2l(\hat{\boldsymbol{\theta}}_t) + k(\ln(n) + 1)$ |
| *THQIC* | $-2l(\hat{\boldsymbol{\theta}}_t) + 2k\ln(\ln(n))$ |
| *TKIC* | $-2l(\hat{\boldsymbol{\theta}}_t) + 3k$ |
| *TAIC4* | $-2l(\hat{\boldsymbol{\theta}}_t) + 4k$ |
| *TaBIC* | $-2l(\hat{\boldsymbol{\theta}}_t) + k\ln\left(\frac{n+2}{24}\right)$ |
| $TAIC_c$ | $TAIC + \frac{2k(k+1)}{n-k-1}$ |
| $TKIC_c$ | $-2l(\hat{\boldsymbol{\theta}}_t) + n\ln\left(\frac{n}{n-k+1}\right) + \frac{n((n-k+1)(2k+1)-2)}{(n-k-1)(n-k+1)}$ |
| *TMDL2* | $-2l(\hat{\boldsymbol{\theta}}_t) + 2k\ln(n)$ |
| *TMDL5* | $-2l(\hat{\boldsymbol{\theta}}_t) + 5k\ln(n)$ |
| Trimmed classification criteria | |
| Criteria | Formula |
| *TCLC* | $-2l(\hat{\boldsymbol{\theta}}_t) - 2\sum_{j=1}^{m}\sum_{i=1}^{I}(\hat{z}_{ij})_t \ln((\hat{z}_{ij})_t)$ |
| *TAWE* | $TCLC + 2k\left(\frac{3}{2} + \ln(n)\right)$ |
| *TNEC* | $\frac{-\sum_{j=1}^{m}\sum_{i=1}^{I}(\hat{z}_{ij})_t \ln((\hat{z}_{ij})_t)}{l(\hat{\boldsymbol{\theta}}_t) - l_{(1)}(\hat{\boldsymbol{\theta}}_t)}$ |
| *TICL-BIC* | $TCLC + k\ln(n)$ |

The fact that all the possible $\binom{I}{\lfloor I(1-\alpha)\rfloor}$ partitions of the data have to be fitted by the maximum likelihood estimate (MLE) makes the procedure computationally very expensive. The *FAST-TLE* algorithm (Neykov and Müller 2003 and Neykov et al. 2007) was proposed in order to avoid adjusting all partitions. The *FAST-TLE* algorithm involves repeated iterations of a two-step procedure - a trial step followed by a refinement step, allowing an approximate solution of the *TLE*, and being computationally much less demanding, especially for large samples. In the trial step a subsample is randomly selected from the data sample and then the model is fitted to that subsample in order to get a trial maximum likelihood estimate. In the refinement step, the cases with the smallest log-likelihoods based on the current estimate are found, starting with the trial maximum likelihood estimator as the initial estimator, and then followed by fitting the model to these cases in order to obtain an improved fit, which has larger trimmed likelihood than the original model fit (see Neykov

et al. 2007 and Yu et al. 2020 for more details). Although there is no guarantee that the estimate will be a global maximum, it will always be a good approximation.

The choice of the trimming proportion $\alpha$ plays an important role for the TLE and should be predetermined. Thus, if $\alpha$ is too large, the TLE will lose much efficiency and, on the other hand, if $\alpha$ is too small and the percentage of outliers is bigger than $\alpha$ the TLE will fail (Yu et al. 2020).

### 3.2 *FAST-TLE* algorithm

Li et al. (2016) enunciate the *FAST-TLE* algorithm adapted to the calculation of robust information criteria. Considering that the authors only apply the algorithm to mixtures of regression models, some changes were made to the algorithm in order to allow the computation of the robust classification criteria and to improve its computational performance in the application to mixtures of regression models with random effects.

Therefore, for a given data set, for a given maximum number of components ($\max(m_a)$), for a given number of initial values ($v$), and for a given trimming proportion ($\alpha$), the new version of the *FAST-TLE* algorithm adapted to the calculation of the information and classification criteria in mixtures of regression models with random effects consists of the following steps:

For $i = 1, ...,$ maximum number of components ($\max(m_a)$):
  For $j = 1, ...,$ number of initial values (v):
    Find an initial value for $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_0$;
    While the change of $l(\hat{\boldsymbol{\theta}})$, in absolute value, is greater than a certain value:
      For a given estimate $\hat{\boldsymbol{\theta}}$, calculate each of the terms of the sum of $l(\hat{\boldsymbol{\theta}})$, corresponding to each individual, and sort the terms in descending order;
      Select the subsample corresponding to the individuals of the first $\lfloor n(1-\alpha) \rfloor$ sorted terms of the sum of $l(\hat{\boldsymbol{\theta}})$;
      Update the estimate of $\boldsymbol{\theta}$ using the subsample;
    Save in a vector the value obtained for $l(\hat{\boldsymbol{\theta}})$ and the respective value of $EN(\hat{z}) = -\sum_{j=1}^{m}\sum_{i=1}^{I} \hat{z}_{ij} \ln(\hat{z}_{ij})$;
  Select the largest of the $l(\hat{\boldsymbol{\theta}})$, denoted by $l(\hat{\boldsymbol{\theta}}_t)$, and the correspondent value of $EN(\hat{z})$, denoted by $EN(\hat{z}_t)$;
  Calculate the robust information and classification criteria.

The initial value of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_0$, can be found by fitting a mixture model to any random subsample of dimension $d$, where $d$ is greater than the number of parameters to be estimated (Li et al. 2016).

## 4 Simulation study

In this section, we use a simulation study to assess the performance of the proposed robust version for the information and classification criteria, stated in Table 1, in the presence of outliers. In order to do so, we study the efficiency of different information and classification criteria, both in their traditional form and in their robust version, in the determination of the number of components of mixtures of regression models with random effects.

To develop the simulation study, and also for the real-world application, we used the statistical software R (*R Development Core Team* 2018).

### 4.1 Design of the simulation study

The design of the simulation study is the following:

- Number of replicates ($n_i$): 4 and 8;
- Number of fixed-effects covariates ($p$): 1 and 4. The rows of the covariates $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times p}$ are independently generated from $N_p(\boldsymbol{0}, \boldsymbol{I})$;
- Fixed-effects vector ($\boldsymbol{\beta}_j$):
  $\boldsymbol{\beta}_1 = (3)^T$, $\boldsymbol{\beta}_2 = (-3)^T$ and $\boldsymbol{\beta}_3 = (0)^T$ for $p = 1$;
  $\boldsymbol{\beta}_1 = (3, 3, 0, 0)^T$, $\boldsymbol{\beta}_2 = (0, 0, 1, 1)^T$ and $\boldsymbol{\beta}_3 = (1, 1, -1, -1)^T$ for $p = 4$;
- Real number of components ($m$): 2 and 3;
- Fitted number of components ($m_a$): 1, 2, 3 and 4;
- Mixing proportions ($\boldsymbol{\pi}$):

$\boldsymbol{\pi} = (0.6, 0.4)$ for $m = 2$;
$\boldsymbol{\pi} = (0.4, 0.4, 0.2)$ for $m = 3$;
- Trimming proportion ($\alpha$): 10 %;
- Sample size ($I$): 100;
- Initial subsample size ($d$): 80;
- Number of random-effects covariates ($q$): 2. The rows of $\boldsymbol{U}_i \in \mathbb{R}^{n_i \times 2}$ are independently generated from $N_2(\boldsymbol{0}, \boldsymbol{I})$;
- The random-effects and error distributions:
  **Scenario I:** $\boldsymbol{b}_{ij} \sim N_2(\boldsymbol{0}, \boldsymbol{\Psi}_j)$, where $\boldsymbol{\Psi}_j = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, and $\boldsymbol{\varepsilon}_{ij} \sim N_{n_i}(\boldsymbol{0}, 4\boldsymbol{I})$;
  **Scenario II:** $\boldsymbol{b}_{ij} \sim 0.95N_2(\boldsymbol{0}, \boldsymbol{\Psi}_j) + 0.05N_2(\boldsymbol{0}, 25\boldsymbol{I})$, where $\boldsymbol{\Psi}_j = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, and $\boldsymbol{\varepsilon}_{ij} \sim 0.95N_{n_i}(\boldsymbol{0}, \boldsymbol{I}) + 0.05N_{n_i}(\boldsymbol{0}, 25\boldsymbol{I})$;
- Number of initial values ($v$): 15;
- Stopping criterion of the while loop in the *FAST-TLE* algorithm: The loop ends when the change of $l(\hat{\boldsymbol{\theta}})$, in absolute value, is smaller or equal to $10^{-2}$;
- Number of simulations ($S$): 200.

It should be noted that for a real number of components equal to $m$, the first $m$ $\boldsymbol{\beta}_j$ were used for each case, that is, $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m$.

For the trimming proportion we chose $\alpha = 10\%$ since for Scenario II we considered a contaminated Normal distribution with a level of contamination of 5% for both the distribution of the random effects and the distribution of the random errors. Therefore, if none of the contaminations coincide in the same individual, the percentage of the sample that is contaminated is, at most, 10%.

In order to be able to fit each of the mixture models we used the EM algorithm with 50 random initializations, thus avoiding convergence to a local maximum, and for each case we selected the mixture with the highest value of the log-likelihood function. As a stopping criterion, the algorithm was stopped when, in a given iteration, the difference between the log-likelihood of a given iteration and the previous one was smaller than $10^{-6}$.

The simulation process for the traditional information and classification criteria is as follows:

1. Generate a data set of $I$ individuals with $n_i$ replicates from a mixture of $m$ components, obtaining a data set with size $I \times n_i$;
2. Fit 1 to 4-component mixtures to the generated data;
3. Calculate the information and classification criteria for each of the mixtures obtained;
4. Select the mixture that provides the smallest value for each information and classification criterion;
5. Repeat the previous steps $S$ times;
6. Calculate the proportions for the fitted models with $m_a$ components.

It should also be mentioned that a similar simulation process is found in Novais and Faria (2021), where the authors studied the problem of determining

the number of components in mixtures of linear mixed models using some of these criteria but without the presence of outliers.

The simulation process for the robust information criteria is as follows:

1. Generate a data set of $I$ individuals with $n_i$ replicates from a mixture of $m$ components, obtaining a data set with size $I \times n_i$;
2. Use the *FAST-TLE* algorithm as described in Section 4.1;
3. Select the mixture that provides the lowest value for each robust information and classification criterion;
4. Repeat the previous steps $S$ times;
5. Calculate the proportions for the fitted models with $m_a$ components.

## 4.2 Simulation results

**Table 2** Proportions for the fitted models with $m_a$ components from a 2-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 4$, $\boldsymbol{b}_{ij} \sim N_2(\mathbf{0}, \boldsymbol{\Psi}_j)$ and $\boldsymbol{\varepsilon}_{ij} \sim N_4(\mathbf{0}, 4\boldsymbol{I})$)

| I | 100 | | | | | | I | 100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | 4 | | | p | 1 | | | 4 | | |
| $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 | $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 |
| *AIC* | - | **0.93** | 0.07 | - | **0.66** | 0.34 | *TAIC* | - | **0.86** | 0.14 | - | **0.53** | 0.47 |
| *BIC* | - | **1** | - | - | **1** | - | *TBIC* | - | **1** | - | - | **1** | - |
| *CAIC* | - | **1** | - | - | **1** | - | *TCAIC* | - | **1** | - | - | **1** | - |
| *HQIC* | - | **1** | - | - | **0.99** | 0.01 | *THQIC* | - | **0.99** | 0.01 | - | **0.98** | 0.02 |
| *KIC* | - | **1** | - | - | **0.96** | 0.04 | *TKIC* | - | **0.99** | 0.01 | - | **0.93** | 0.07 |
| $AIC_4$ | - | **1** | - | - | **1** | - | $TAIC_4$ | - | **0.99** | 0.01 | - | **1** | - |
| *aBIC* | - | **1** | - | - | **0.94** | 0.06 | *TaBIC* | - | **0.98** | 0.02 | - | **0.89** | 0.11 |
| $AIC_c$ | - | **0.97** | 0.03 | - | **0.84** | 0.16 | $TAIC_c$ | - | **0.89** | 0.11 | - | **0.64** | 0.36 |
| $KIC_c$ | - | **1** | - | - | **0.98** | 0.02 | $TKIC_c$ | - | **0.99** | 0.01 | - | **0.96** | 0.04 |
| *MDL2* | - | **1** | - | 0.02 | **0.98** | - | *TMDL2* | - | **1** | - | 0.20 | **0.80** | - |
| *MDL5* | - | **1** | - | **1** | - | - | *TMDL5* | 0.11 | **0.89** | - | **1** | - | - |
| *CLC* | - | **0.98** | 0.02 | 0.08 | **0.78** | 0.14 | *TCLC* | - | **0.85** | 0.15 | 0.04 | **0.65** | 0.31 |
| *AWE* | - | **1** | - | **0.98** | 0.02 | - | *TAWE* | 0.03 | **0.97** | - | **0.96** | 0.04 | - |
| *NEC* | - | **0.99** | 0.01 | 0.04 | **0.86** | 0.10 | *TNEC* | - | **0.92** | 0.08 | - | **0.76** | 0.24 |
| *ICL-BIC* | - | **1** | - | 0.41 | **0.59** | - | *TICL-BIC* | - | **1** | - | 0.38 | **0.62** | - |

Tables 2 to 9 give the proportions for the fitted models with $m_a$ components that each information and classification criterion, both in its traditional and robust version, estimates a number of components of 200 samples simulated under different data sets configurations.

As expected, it can be seen that different simulated scenarios influence the performance of the information and classification criteria, in the selection of the number of components. Thus, for both versions of the criteria, it can be seen that when the number of components and the number of fixed-effects covariates increase, the performance of all the information and classification criteria decreases, while the increase in the number of replicates improves the performance of these criteria.

Starting by analysing the 2-component mixture models, Tables 2 to 5, the first conclusion to be drawn is that for Scenario I (no contamination) it

**Table 3** Proportions for the fitted models with $m_a$ components from a 2-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 4$, $\boldsymbol{b}_{ij} \sim 0.95N_2(\mathbf{0}, \boldsymbol{\Psi}_j) + 0.05N_2(\mathbf{0}, 25\boldsymbol{I})$ and $\boldsymbol{\varepsilon}_{ij} \sim 0.95N_4(\mathbf{0}, \boldsymbol{I}) + 0.05N_4(\mathbf{0}, 25\boldsymbol{I})$)

| I | 100 | | | | | | I | 100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | 4 | | | p | 1 | | | 4 | | |
| $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 | $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 |
| *AIC* | - | 0.01 | **0.99** | - | 0.01 | **0.99** | *TAIC* | - | **0.61** | 0.39 | - | 0.29 | **0.71** |
| *BIC* | - | 0.04 | **0.96** | - | 0.07 | **0.93** | *TBIC* | - | **0.99** | 0.01 | - | **0.97** | 0.03 |
| *CAIC* | - | 0.05 | **0.95** | - | 0.13 | **0.87** | *TCAIC* | - | **0.99** | 0.01 | - | **0.97** | 0.03 |
| *HQIC* | - | 0.03 | **0.97** | - | 0.02 | **0.98** | *THQIC* | - | **0.95** | 0.05 | - | **0.92** | 0.08 |
| *KIC* | - | 0.02 | **0.98** | - | 0.01 | **0.99** | *TKIC* | - | **0.88** | 0.12 | - | **0.84** | 0.16 |
| *AIC4* | - | 0.03 | **0.97** | - | 0.02 | **0.98** | *TAIC4* | - | **0.97** | 0.03 | - | **0.93** | 0.07 |
| *aBIC* | - | 0.01 | **0.99** | - | 0.01 | **0.99** | *TaBIC* | - | **0.86** | 0.14 | - | **0.84** | 0.16 |
| $AIC_c$ | - | 0.01 | **0.99** | - | 0.01 | **0.99** | $TAIC_c$ | - | **0.67** | 0.33 | - | 0.34 | **0.66** |
| $KIC_c$ | - | 0.02 | **0.98** | - | 0.01 | **0.99** | $TKIC_c$ | - | **0.90** | 0.10 | - | **0.86** | 0.14 |
| *MDL2* | - | 0.24 | **0.76** | - | 0.45 | **0.55** | *TMDL2* | - | **1** | - | - | **1** | - |
| *MDL5* | - | **0.86** | 0.14 | **0.53** | 0.47 | - | *TMDL5* | - | **1** | - | 0.05 | **0.95** | - |
| *CLC* | - | 0.07 | **0.93** | - | 0.03 | **0.97** | *TCLC* | - | **0.81** | 0.19 | - | **0.53** | 0.47 |
| *AWE* | - | 0.41 | **0.59** | - | **0.63** | 0.37 | *TAWE* | - | **1** | - | - | **1** | - |
| *NEC* | - | **0.58** | 0.42 | - | 0.28 | **0.72** | *TNEC* | - | **0.96** | 0.04 | - | **0.81** | 0.19 |
| *ICL-BIC* | - | 0.17 | **0.83** | - | 0.20 | **0.80** | *TICL-BIC* | - | **0.99** | 0.01 | - | **1** | - |

**Table 4** Proportions for the fitted models with $m_a$ components from a 2-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 8$, $\boldsymbol{b}_{ij} \sim N_2(\mathbf{0}, \boldsymbol{\Psi}_j)$ and $\boldsymbol{\varepsilon}_{ij} \sim N_8(\mathbf{0}, 4\boldsymbol{I})$)

| I | 100 | | | | | | I | 100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | 4 | | | p | 1 | | | 4 | | |
| $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 | $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 |
| *AIC* | - | **0.90** | 0.10 | - | **0.83** | 0.17 | *TAIC* | - | **0.87** | 0.13 | - | **0.76** | 0.24 |
| *BIC* | - | **1** | - | - | **1** | - | *TBIC* | - | **1** | - | - | **1** | - |
| *CAIC* | - | **1** | - | - | **1** | - | *TCAIC* | - | **1** | - | - | **1** | - |
| *HQIC* | - | **1** | - | - | **1** | - | *THQIC* | - | **0.99** | 0.01 | - | **1** | - |
| *KIC* | - | **0.99** | 0.01 | - | **0.97** | 0.03 | *TKIC* | - | **0.98** | 0.02 | - | **0.99** | 0.01 |
| *AIC4* | - | **1** | - | - | **1** | - | *TAIC4* | - | **1** | - | - | **1** | - |
| *aBIC* | - | **1** | - | - | **0.98** | 0.02 | *TaBIC* | - | **0.99** | 0.01 | - | **1** | - |
| $AIC_c$ | - | **0.91** | 0.09 | - | **0.87** | 0.13 | $TAIC_c$ | - | **0.89** | 0.11 | - | **0.83** | 0.17 |
| $KIC_c$ | - | **0.99** | 0.01 | - | **0.98** | 0.02 | $TKIC_c$ | - | **0.99** | 0.01 | - | **1** | - |
| *MDL2* | - | **1** | - | - | **1** | - | *TMDL2* | - | **1** | - | - | **1** | - |
| *MDL5* | - | **1** | - | - | **1** | - | *TMDL5* | - | **1** | - | 0.14 | **0.86** | - |
| *CLC* | - | **0.97** | 0.03 | - | **0.89** | 0.11 | *TCLC* | - | **0.89** | 0.11 | - | **0.80** | 0.20 |
| *AWE* | - | **1** | - | - | **1** | - | *TAWE* | - | **1** | - | - | **1** | - |
| *NEC* | - | **0.99** | 0.01 | - | **0.93** | 0.07 | *TNEC* | - | **0.98** | 0.02 | - | **0.94** | 0.06 |
| *ICL-BIC* | - | **1** | - | - | **1** | - | *TICL-BIC* | - | **1** | - | - | **1** | - |

can be found that both versions of the information and classification criteria yield similar results (Tables 2 and 4). *AIC*, $AIC_c$, *CLC*, *MDL2* and *MDL5* decrease their performance in the robust version, especially with the increase in the number of fixed-effects covariates. The remaining criteria present small differences from one version to the other.

On the other hand, as expected, the two versions of the information and classification criteria present quite different results in the presence of contaminated samples (Tables 3 and 5). For these cases, all the criteria overestimate the number of components, estimating it to be 3, with high proportions of overestimation, with the exception of *MDL5*, *AWE* and *NEC* for some of the simulated scenarios. Thus, in the presence of contamination, the proportion of

**Table 5** Proportions for the fitted models with $m_a$ components from a 2-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 8$, $\boldsymbol{b}_{ij} \sim 0.95N_2(\boldsymbol{0}, \boldsymbol{\Psi}_j) + 0.05N_2(\boldsymbol{0}, 25\boldsymbol{I})$ and $\boldsymbol{\varepsilon}_{ij} \sim 0.95N_8(\boldsymbol{0}, \boldsymbol{I}) + 0.05N_8(\boldsymbol{0}, 25\boldsymbol{I})$)

| I | 100 | | | | | | I | 100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | 4 | | | p | 1 | | | 4 | | |
| $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 | $m_a$ | 1 | 2 | 3 | 1 | 2 | 3 |
| *AIC* | - | 0.03 | **0.97** | - | 0.02 | **0.98** | *TAIC* | - | **0.61** | 0.39 | - | 0.39 | **0.61** |
| *BIC* | - | 0.05 | **0.95** | - | 0.03 | **0.97** | *TBIC* | - | **0.95** | 0.05 | - | **0.99** | 0.01 |
| *CAIC* | - | 0.05 | **0.95** | - | 0.04 | **0.96** | *TCAIC* | - | **0.97** | 0.03 | - | **1** | - |
| *HQIC* | - | 0.03 | **0.97** | - | 0.03 | **0.97** | *THQIC* | - | **0.87** | 0.13 | - | **0.92** | 0.08 |
| *KIC* | - | 0.03 | **0.97** | - | 0.02 | **0.98** | *TKIC* | - | **0.83** | 0.17 | - | **0.87** | 0.13 |
| *AIC4* | - | 0.03 | **0.97** | - | 0.03 | **0.97** | *TAIC4* | - | **0.89** | 0.11 | - | **0.94** | 0.06 |
| *aBIC* | - | 0.03 | **0.97** | - | 0.03 | **0.97** | *TaBIC* | - | **0.88** | 0.12 | - | **0.89** | 0.11 |
| $AIC_c$ | - | 0.03 | **0.97** | - | 0.02 | **0.98** | $TAIC_c$ | - | **0.66** | 0.34 | - | 0.46 | **0.54** |
| $KIC_c$ | - | 0.03 | **0.97** | - | 0.02 | **0.98** | $TKIC_c$ | - | **0.84** | 0.16 | - | **0.89** | 0.11 |
| *MDL2* | - | 0.07 | **0.93** | - | 0.08 | **0.92** | *TMDL2* | - | **0.99** | 0.01 | - | **1** | - |
| *MDL5* | - | 0.26 | **0.74** | - | 0.49 | **0.51** | *TMDL5* | - | **1** | - | - | **1** | - |
| *CLC* | - | 0.04 | **0.96** | - | 0.03 | **0.97** | *TCLC* | - | **0.84** | 0.16 | - | **0.69** | 0.31 |
| *AWE* | - | 0.10 | **0.90** | - | 0.11 | **0.89** | *TAWE* | - | **0.99** | 0.01 | - | **1** | - |
| *NEC* | - | **0.80** | 0.20 | - | 0.40 | **0.60** | *TNEC* | - | **1** | - | - | **0.91** | 0.09 |
| *ICL-BIC* | - | 0.05 | **0.95** | - | 0.04 | **0.96** | *TICL-BIC* | - | **0.98** | 0.02 | - | **1** | - |

**Table 6** Proportions for the fitted models with $m_a$ components from a 3-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 4$, $\boldsymbol{b}_{ij} \sim N_2(\boldsymbol{0}, \boldsymbol{\Psi}_j)$ and $\boldsymbol{\varepsilon}_{ij} \sim N_4(\boldsymbol{0}, 4\boldsymbol{I})$)

| I | 100 | | | | | | | | I | 100 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | | 4 | | | | p | 1 | | | | 4 | | | |
| $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| *AIC* | - | 0.11 | **0.83** | 0.06 | - | 0.07 | **0.66** | 0.27 | *TAIC* | - | 0.22 | **0.66** | 0.12 | - | 0.04 | 0.29 | **0.67** |
| *BIC* | - | **0.87** | 0.13 | - | - | **0.97** | 0.03 | - | *TBIC* | - | **0.97** | 0.03 | - | - | **0.98** | 0.02 | - |
| *CAIC* | - | **0.95** | 0.05 | - | - | **0.99** | 0.01 | - | *TCAIC* | - | **0.99** | 0.01 | - | 0.02 | **0.98** | - | - |
| *HQIC* | - | 0.43 | **0.57** | - | - | **0.52** | 0.45 | 0.03 | *THQIC* | - | **0.67** | 0.33 | - | - | **0.52** | 0.44 | 0.04 |
| *KIC* | - | 0.30 | **0.70** | - | - | 0.34 | **0.62** | 0.04 | *TKIC* | - | 0.46 | **0.53** | 0.01 | - | 0.32 | **0.52** | 0.16 |
| *AIC4* | - | **0.53** | 0.47 | - | - | **0.65** | 0.34 | 0.01 | *TAIC4* | - | **0.77** | 0.23 | - | - | **0.74** | 0.24 | 0.02 |
| *aBIC* | - | 0.27 | **0.73** | - | - | 0.30 | **0.63** | 0.07 | *TaBIC* | - | 0.42 | **0.58** | 0.01 | - | 0.28 | **0.56** | 0.16 |
| $AIC_c$ | - | 0.14 | **0.85** | 0.01 | - | 0.13 | **0.70** | 0.17 | $TAIC_c$ | - | 0.28 | **0.65** | 0.07 | - | 0.14 | **0.54** | 0.32 |
| $KIC_c$ | - | 0.34 | **0.66** | - | - | 0.44 | **0.53** | 0.03 | $TKIC_c$ | - | **0.58** | 0.42 | - | - | 0.40 | **0.54** | 0.06 |
| *MDL2* | - | **1** | - | - | 0.49 | **0.51** | - | - | *TMDL2* | - | **1** | - | - | **0.74** | 0.26 | - | - |
| *MDL5* | 0.38 | **0.62** | - | - | **1** | - | - | - | *TMDL5* | **0.74** | 0.26 | - | - | **1** | - | - | - |
| *CLC* | 0.03 | **0.96** | 0.01 | - | **0.76** | 0.21 | 0.02 | 0.01 | *TCLC* | 0.03 | **0.92** | 0.03 | 0.02 | **0.54** | 0.16 | 0.12 | 0.18 |
| *AWE* | **0.51** | 0.49 | - | - | **1** | - | - | - | *TAWE* | **0.50** | **0.50** | - | - | **1** | - | - | - |
| *NEC* | 0.03 | **0.97** | - | - | **0.73** | 0.24 | 0.02 | 0.01 | *TNEC* | 0.03 | **0.95** | 0.01 | 0.01 | **0.44** | 0.24 | 0.12 | 0.20 |
| *ICL-BIC* | 0.13 | **0.87** | - | - | **0.98** | 0.02 | - | - | *TICL-BIC* | 0.16 | **0.84** | - | - | **0.94** | 0.06 | - | - |

times that the traditional version of each information and classification criterion correctly estimates the number of components is very low. As expected, the highest rates of success relate to the criteria most likely to underestimate the number of components while, on the contrary, the criteria with the greatest tendency to overestimate are those that present the worst results (*AIC* and $AIC_c$). However, all robust information and classification criteria correctly estimate the number of components for all scenarios, with the exception of *AIC* and $AIC_c$ when the number of fixed effects covariates is equal to 4. Although they have significantly improved their performance compared to their usual version, these two criteria continue to overestimate the number of components. All the other robust criteria correctly estimate the number of components, doing so with rates of success above 80 % for the great majority of scenarios.

**Table 7** Proportions for the fitted models with $m_a$ components from a 3-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 4$, $\boldsymbol{b}_{ij} \sim 0.95N_2(\mathbf{0}, \boldsymbol{\Psi}_j) + 0.05N_2(\mathbf{0}, 25\boldsymbol{I})$ and $\boldsymbol{\varepsilon}_{ij} \sim 0.95N_4(\mathbf{0}, \boldsymbol{I}) + 0.05N_4(\mathbf{0}, 25\boldsymbol{I})$)

| I | 100 | | | | | | | | I | 100 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | | 4 | | | | p | 1 | | | | 4 | | | |
| $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| *AIC* | - | - | 0.08 | **0.92** | - | - | 0.04 | **0.96** | *TAIC* | - | - | **0.79** | 0.21 | - | - | 0.37 | **0.63** |
| *BIC* | - | - | 0.17 | **0.83** | - | 0.01 | 0.32 | **0.67** | *TBIC* | - | 0.02 | **0.98** | - | - | 0.10 | **0.90** | - |
| *CAIC* | - | - | 0.21 | **0.79** | - | 0.08 | 0.40 | **0.52** | *TCAIC* | - | 0.03 | **0.97** | - | - | 0.20 | **0.80** | - |
| *HQIC* | - | - | 0.13 | **0.87** | - | - | 0.15 | **0.85** | *THQIC* | - | - | **0.98** | 0.02 | - | - | **0.98** | 0.02 |
| *KIC* | - | - | 0.11 | **0.89** | - | - | 0.12 | **0.88** | *TKIC* | - | - | **0.96** | 0.04 | - | - | **0.84** | 0.16 |
| *AIC*$_4$ | - | - | 0.14 | **0.86** | - | - | 0.16 | **0.84** | *TAIC*$_4$ | - | - | **0.99** | 0.01 | - | - | **0.98** | 0.02 |
| *aBIC* | - | - | 0.11 | **0.89** | - | - | 0.11 | **0.89** | *TaBIC* | - | - | **0.96** | 0.04 | - | - | **0.78** | 0.22 |
| *AIC*$_c$ | - | - | 0.09 | **0.91** | - | - | 0.07 | **0.93** | *TAIC*$_c$ | - | - | **0.83** | 0.17 | - | - | **0.50** | **0.50** |
| *KIC*$_c$ | - | - | 0.11 | **0.89** | - | - | 0.15 | **0.85** | *TKIC*$_c$ | - | - | **0.97** | 0.03 | - | - | **0.96** | 0.04 |
| *MDL2* | - | 0.03 | **0.57** | 0.40 | 0.01 | **0.64** | 0.32 | 0.03 | *TMDL2* | - | 0.28 | **0.72** | - | - | **0.80** | 0.20 | - |
| *MDL5* | 0.01 | **0.94** | 0.05 | - | **0.97** | 0.03 | - | - | *TMDL5* | - | **1** | - | - | **0.92** | 0.08 | - | - |
| *CLC* | - | - | 0.12 | **0.88** | - | 0.02 | 0.13 | **0.85** | *TCLC* | - | 0.06 | **0.74** | 0.20 | - | - | 0.32 | **0.58** |
| *AWE* | - | **0.67** | 0.20 | 0.13 | **0.69** | 0.29 | 0.02 | - | *TAWE* | - | **0.82** | 0.18 | - | 0.28 | **0.70** | 0.02 | - |
| *NEC* | - | **0.49** | 0.11 | 0.40 | - | 0.24 | 0.11 | **0.65** | *TNEC* | - | **0.84** | 0.12 | 0.04 | - | **0.36** | 0.32 | 0.32 |
| *ICL-BIC* | - | 0.06 | 0.17 | **0.77** | 0.05 | 0.23 | 0.18 | **0.54** | *TICL-BIC* | - | 0.30 | **0.70** | - | - | 0.30 | **0.68** | 0.02 |

**Table 8** Proportions for the fitted models with $m_a$ components from a 3-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 8$, $\boldsymbol{b}_{ij} \sim N_2(\mathbf{0}, \boldsymbol{\Psi}_j)$ and $\boldsymbol{\varepsilon}_{ij} \sim N_8(\mathbf{0}, 4\boldsymbol{I})$)

| I | 100 | | | | | | | | I | 100 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | | 4 | | | | p | 1 | | | | 4 | | | |
| $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| *AIC* | - | - | **0.93** | 0.07 | - | - | **0.87** | 0.13 | *TAIC* | - | 0.01 | **0.92** | 0.07 | - | 0.01 | **0.78** | 0.21 |
| *BIC* | - | 0.04 | **0.96** | - | - | 0.08 | **0.92** | - | *TBIC* | - | 0.30 | **0.70** | - | - | 0.43 | **0.57** | - |
| *CAIC* | - | 0.05 | **0.95** | - | - | 0.13 | **0.87** | - | *TCAIC* | - | 0.41 | **0.59** | - | - | **0.58** | 0.42 | - |
| *HQIC* | - | - | **1** | - | - | - | **1** | - | *THQIC* | - | 0.05 | **0.95** | - | - | 0.09 | **0.91** | - |
| *KIC* | - | - | **0.99** | 0.01 | - | - | **0.98** | 0.02 | *TKIC* | - | 0.05 | **0.94** | 0.01 | - | 0.03 | **0.97** | - |
| *AIC*$_4$ | - | - | **1** | - | - | - | **1** | - | *TAIC*$_4$ | - | 0.05 | **0.95** | - | - | 0.11 | **0.89** | - |
| *aBIC* | - | - | **0.99** | 0.01 | - | - | **1** | - | *TaBIC* | - | 0.05 | **0.95** | - | - | 0.04 | **0.96** | - |
| *AIC*$_c$ | - | - | **0.97** | 0.03 | - | - | **0.92** | 0.08 | *TAIC*$_c$ | - | 0.01 | **0.93** | 0.06 | - | 0.01 | **0.86** | 0.13 |
| *KIC*$_c$ | - | - | **0.99** | 0.01 | - | - | **0.99** | 0.01 | *TKIC*$_c$ | - | 0.05 | **0.95** | - | - | 0.04 | **0.96** | - |
| *MDL2* | - | 0.47 | **0.53** | - | - | **0.83** | 0.17 | - | *TMDL2* | - | **0.72** | 0.28 | - | - | **0.97** | 0.03 | - |
| *MDL5* | - | **1** | - | - | **0.65** | 0.35 | - | - | *TMDL5* | - | **1** | - | - | **0.78** | 0.22 | - | - |
| *CLC* | - | **0.65** | 0.34 | 0.01 | - | 0.23 | **0.70** | 0.07 | *TCLC* | - | **0.83** | 0.15 | 0.02 | - | 0.35 | **0.53** | 0.12 |
| *AWE* | - | **1** | - | - | 0.26 | **0.72** | 0.02 | - | *TAWE* | - | **1** | - | - | 0.21 | **0.79** | - | - |
| *NEC* | - | **1** | - | - | - | **0.65** | 0.32 | 0.03 | *TNEC* | - | **1** | - | - | - | **0.80** | 0.16 | 0.04 |
| *ICL-BIC* | - | **0.88** | 0.12 | - | 0.03 | **0.58** | 0.39 | - | *TICL-BIC* | - | **1** | - | - | 0.01 | **0.78** | 0.21 | - |

Regarding the analysis of 3-component mixture models, Tables 6 to 9, it appears that, as in the simulation study of Novais and Faria (2021), the performance of both the information and classification criteria is worse when compared to their performance for the 2-component mixture models, that is, the proportion of times that these criteria correctly estimate the number of components is lower than the same proportion for 2-component mixture models, regardless of the version used.

Starting by analysing the cases in which there is no contamination, Tables 6 and 8, it can be verified once again that both versions of the generality of the criteria produce similar results. However, the performance of the robust version of some criteria decreases, especially with the increase in the number of fixed-effects covariates and for a lower number of replicates, as is the case of the criteria *AIC*, *AIC*$_c$, *MDL2* and *MDL5* and, on some occasions, for most of the classification criteria, *BIC* and *CAIC*. In particular, as with the 2-component mixture models, the robust version of *AIC* and *AIC*$_c$ denotes an even more marked trend towards overestimating the number of compo-

**Table 9** Proportions for the fitted models with $m_a$ components from a 3-component mixture of linear mixed models for both versions of the information and classification criteria ($n_i = 8$, $\boldsymbol{b}_{ij} \sim 0.95N_2(\mathbf{0}, \boldsymbol{\Psi}_j) + 0.05N_2(\mathbf{0}, 25\boldsymbol{I})$ and $\boldsymbol{\varepsilon}_{ij} \sim 0.95N_8(\mathbf{0}, \boldsymbol{I}) + 0.05N_8(\mathbf{0}, 25\boldsymbol{I})$)

| I | 100 | | | | | | | | I | 100 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 1 | | | | 4 | | | | p | 1 | | | | 4 | | | |
| $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | $m_a$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| *AIC* | - | - | 0.02 | **0.98** | - | - | - | **1** | *TAIC* | - | - | **0.66** | 0.34 | - | - | **0.50** | **0.50** |
| *BIC* | - | - | 0.04 | **0.96** | - | - | 0.06 | **0.94** | *TBIC* | - | - | **0.98** | 0.02 | - | - | **1** | - |
| *CAIC* | - | - | 0.05 | **0.95** | - | - | 0.09 | **0.91** | *TCAIC* | - | - | **0.98** | 0.02 | - | - | **1** | - |
| *HQIC* | - | - | 0.02 | **0.98** | - | - | 0.03 | **0.97** | *THQIC* | - | - | **0.86** | 0.14 | - | - | **0.93** | 0.07 |
| *KIC* | - | - | 0.02 | **0.98** | - | - | 0.01 | **0.99** | *TKIC* | - | - | **0.81** | 0.19 | - | - | **0.79** | 0.21 |
| $AIC_4$ | - | - | 0.02 | **0.98** | - | - | 0.03 | **0.97** | $TAIC_4$ | - | - | **0.88** | 0.12 | - | - | **0.93** | 0.07 |
| *aBIC* | - | - | 0.02 | **0.98** | - | - | 0.02 | **0.98** | *TaBIC* | - | - | **0.86** | 0.14 | - | - | **0.90** | 0.10 |
| $AIC_c$ | - | - | 0.02 | **0.98** | - | - | - | **1** | $TAIC_c$ | - | - | **0.66** | 0.34 | - | - | **0.57** | 0.43 |
| $KIC_c$ | - | - | 0.02 | **0.98** | - | - | 0.02 | **0.98** | $TKIC_c$ | - | - | **0.82** | 0.18 | - | - | **0.84** | 0.16 |
| *MDL2* | - | - | 0.09 | **0.91** | - | - | 0.16 | **0.84** | *TMDL2* | - | - | **1** | - | - | - | **1** | - |
| *MDL5* | - | 0.03 | **0.59** | 0.38 | - | **0.56** | 0.42 | 0.02 | *TMDL5* | - | 0.12 | **0.88** | - | - | **0.52** | 0.48 | - |
| *CLC* | - | - | 0.04 | **0.96** | - | - | 0.04 | **0.96** | *TCLC* | - | - | **0.82** | 0.18 | - | - | **0.68** | 0.32 |
| *AWE* | - | 0.01 | 0.09 | **0.90** | - | - | 0.20 | **0.80** | *TAWE* | - | 0.01 | **0.99** | - | - | 0.01 | **0.99** | - |
| *NEC* | - | 0.36 | **0.64** | - | - | 0.13 | 0.08 | **0.79** | *TNEC* | - | 0.46 | **0.52** | 0.02 | - | 0.29 | **0.57** | 0.14 |
| *ICL-BIC* | - | 0.01 | 0.06 | **0.93** | - | - | 0.08 | **0.92** | *TICL-BIC* | - | - | **1** | - | - | - | **0.98** | 0.02 |

nents, something particularly notorious with the increase in the number of fixed-effects covariates. The remaining criteria, on the other hand, show low proportion fluctuations from one version to another.

In the presence of contamination, Tables 7 and 9, the two versions of the information and classification criteria show very different results. For contaminated samples, all criteria in their usual version overestimate the number of components as being 4, with very high overestimation proportions, with the exception of *MDL2*, *MDL5*, *AWE* and *NEC* for some of the scenarios. For these scenarios, these criteria underestimate the number of components even in the presence of contamination, something that is not surprising given that they are criteria with a notorious trend to underestimate the number of components, as demonstrated in Novais and Faria (2021). On the other hand, in their robust version, all criteria correctly estimate the number of components, with the main exceptions being *MDL5* and, to a lesser extent, *MDL2*, *AWE* and *NEC*, which still continue to underestimate the number of components in some of the scenarios and, on the opposite direction, *AIC* and $AIC_c$, continue to overestimate the number of components in their robust version for a large part of the scenarios, particularly for models with a greater number of fixed-effects covariates. As such, in the presence of contamination, the proportion of times that each criterion in its usual version effectively determines the number of components is extremely reduced while, on the other hand, in their robust version the information and classification criteria are capable of detecting the number of components correctly for the great majority of the studied scenarios, with rates of success above 80 % for most of the criteria.

In conclusion, the robust information and classification criteria produce similar results to the traditional information criteria when there is no contamination but present significantly better results in the presence of contamination. However, one of the main drawbacks of the robust information criteria is their computational time since the *FAST-TLE* algorithm is very demanding. Therefore, if there is no contamination it does not compensate to use the

robust information and classification criteria given the computational effort required and the similarities in the results obtained for the two versions of the criteria. Contrastingly, in the presence of contamination, the computational effort is clearly compensated in the results, much better than the results of the traditional information and classification criteria.

## 5 Real-world application

In this section we compare the performance of the two versions of different information and classification criteria in a real-world application, the "PenicillinC" data set. The data set can be loaded into the software R from the package "robustlmm" of Koller (2016) with the command source(system.file("doc/Penicillin.R", package = "robustlmm")). The original data set can be found in the package "lme4" of Bates et al. (2007) as the "Penicillin" data set. However, to emphasize the effect of his robust method, Koller (2016) slightly modified the data set in order to contain contaminated data and called it the "PenicillinC" data set.

As first reported by Davies and Goldsmith (1972) for the original data set, the goal is to assess the variability between samples of penicillin by the *B. subtilis* method. In this test method a bulk-inoculated nutrient agar medium is poured into a Petri dish of approximately 90 mm. diameter, known as a plate. When the medium has set, six small hollow cylinders or pots (about 4 mm. in diameter) are cemented onto the surface at equally spaced intervals. A few drops of the penicillin solutions to be compared are placed in the respective cylinders, and the whole plate is placed in an incubator for a given time. Penicillin diuses from the pots into the agar, and this produces a clear circular zone of inhibition of growth of the organisms, which can be readily measured. The diameter of the zone is related in a known way to the concentration of penicillin in the solution.

Thus, this data set contains 144 observations, where we consider the diameter as the response variable, a variable varying from 15.20mm to 27mm with a mean of 22.77mm, and we have two types of crossed random effects: the sample with 6 levels and the plate with 24 levels. In the data set of Koller (2016) there is also a fourth variable, called contaminated, indicating whether or not an observation has been modified. Out of the 144 observations, 7 of them were changed, which means that for this data set we have almost 5% of outliers.

Since the number of components is unknown, it has to be determined. In order to determine it, we fit mixtures of regression models with random effects with a number of components varying from 1 to 4, that is, for $m = 1, ..., 4$, and we use the information and classification criteria of Section 3 to identify the most suitable mixture.

Table 10 shows the information and classification criteria for each of the mixtures and in bold is the mixture selected by each criterion. We see that the majority of the information and classification criteria select the 2-component

**Table 10** Information and classification criteria for the 4 mixtures of regression models with random effects

| | Number of components | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| *AIC* | 645.84 | **632.95** | 640.95 | 648.95 |
| *BIC* | 654.75 | **653.74** | 673.62 | 693.50 |
| *CAIC* | **657.75** | 660.74 | 684.62 | 708.50 |
| *HQIC* | 649.46 | **641.40** | 654.22 | 667.05 |
| *KIC* | 648.84 | **639.95** | 651.95 | 663.95 |
| *AIC4* | 651.84 | **646.95** | 662.95 | 678.95 |
| *aBIC* | 645.25 | **631.59** | 638.81 | 646.03 |
| $AIC_c$ | 646.01 | **633.77** | 642.95 | 652.70 |
| $KIC_c$ | 649.04 | **640.94** | 654.39 | 668.54 |
| *MDL2* | **669.66** | 688.53 | 728.29 | 768.04 |
| *MDL5* | **714.39** | 792.89 | 892.29 | 991.69 |
| *CLC* | 639.84 | **618.95** | 797.52 | 918.99 |
| *AWE* | **678.66** | 709.53 | 939.85 | 1113.08 |
| *NEC* | 1.00 | $\mathbf{1.01 \times 10^{-8}}$ | 8.55 | 14.36 |
| *ICL-BIC* | 654.75 | **653.74** | 852.18 | 993.53 |

mixture, while only 4 of the criteria select the 1-component mixture, that is, a linear mixed model. In Novais and Faria (2021) these 4 criteria tended to underestimate the number of components, so it is not surprising that these criteria simply select the linear mixed model.

It is important to note that the 2-component mixture chosen by the majority of the criteria consists in one component containing 138 out of the 144 observations, in which 137 of them are the original observations and the remaining one is an outlier, and the other component contains 6 observations and the 6 of them are outliers. This fact is not surprising since, as already stated, the information and classification criteria are not robust to outliers, so the presence of outliers may cause the number of components to change by generating, at least, an additional component for the outliers.

Thus, to prove the effectiveness of the robust criteria, we calculate the robust information and classification criteria, as showed in Table 11. In order to accomplish it, and since we know that we have almost 5% of outliers, we use a trimming proportion of 5%, that is, $\alpha = 5\%$.

It can be seen that, as expected, all the robust criteria select the 1-component mixture, the linear mixed model, thus corroborating the simulation study in the sense that, when the sample is contaminated, the robust criteria, calculated using the FAST-TLE algorithm, are capable of determining the correct number of components, unlike the traditional criteria.

Since we know that almost 5% of the sample consists of outliers we used a trimming proportion of 5%, but sometimes we do not know the percentage of outliers beforehand, so further analysis of the data is needed in order to select the most suitable trimming proportion.

To illustrate the important role of the trimming proportion, in Table 12 we present the same robust information and classification criteria, but using 3% instead of 5% as the trimming proportion.

**Table 11** Robust information and classification criteria for the 4 mixtures of regression models with random effects, for $\alpha = 5\%$

| | Number of components | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| *TAIC* | **586.36** | 594.36 | 602.36 | 610.36 |
| *TBIC* | **595.27** | 615.15 | 635.03 | 654.91 |
| *TCAIC* | **598.27** | 622.15 | 646.03 | 669.91 |
| *THQIC* | **589.98** | 602.81 | 615.63 | 628.46 |
| *TKIC* | **589.36** | 601.36 | 613.36 | 625.36 |
| *TAIC4* | **592.36** | 608.36 | 624.36 | 640.36 |
| *TaBIC* | **585.77** | 593.00 | 600.22 | 607.44 |
| $TAIC_c$ | **586.53** | 595.18 | 604.36 | 614.11 |
| $TKIC_c$ | **589.56** | 602.35 | 615.80 | 629.94 |
| *TMDL2* | **610.18** | 649.94 | 689.69 | 729.45 |
| *TMDL5* | **654.91** | 754.30 | 853.70 | 953.09 |
| *TCLC* | **580.36** | 769.91 | 876.51 | 933.61 |
| *TAWE* | **619.18** | 860.49 | 1018.84 | 1127.70 |
| *TNEC* | **1.00** | $9.94 \times 10^6$ | $1.63 \times 10^8$ | $3.33 \times 10^7$ |
| *TICL-BIC* | **595.27** | 804.70 | 931.18 | 1008.16 |

**Table 12** Robust information and classification criteria for the 4 mixtures of regression models with random effects, for $\alpha = 3\%$

| | Number of components | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| *TAIC* | 619.40 | **614.17** | 622.17 | 630.18 |
| *TBIC* | **628.31** | 634.96 | 654.84 | 674.72 |
| *TCAIC* | **631.31** | 641.96 | 665.84 | 689.72 |
| *THQIC* | 623.02 | **622.62** | 635.45 | 648.28 |
| *TKIC* | 622.40 | **621.17** | 633.17 | 645.18 |
| *TAIC4* | **625.40** | 628.17 | 644.17 | 660.18 |
| *TaBIC* | 618.82 | **612.81** | 620.04 | 627.26 |
| $TAIC_c$ | 619.58 | **615.00** | 624.17 | 633.93 |
| $TKIC_c$ | 622.60 | **622.17** | 635.61 | 649.76 |
| *TMDL2* | **643.22** | 669.75 | 709.51 | 749.27 |
| *TMDL5* | **687.95** | 774.12 | 873.51 | 972.91 |
| *TCLC* | 613.40 | **600.18** | 789.05 | 883.17 |
| *TAWE* | **652.22** | 690.76 | 931.38 | 1077.26 |
| *TNEC* | 1.00 | $\mathbf{2.76 \times 10^{-4}}$ | 14.28 | 21.39 |
| *TICL-BIC* | **628.31** | 634.97 | 843.71 | 957.72 |

As it can be seen on Table 12, for a trimming proportion of 3%, 8 out of the 15 criteria chose the 2-component mixture, while just 7 of these criteria still chose the linear mixed model even for a mixing proportion smaller than the proportion of outliers. However, the majority of these criteria are known to underestimate the number of components. Thus, comparing Table 11 to Table 12, the magnitude of the role played by the trimming proportion becomes clear, so the value of $\alpha$ must be chosen with care, that is, after a careful analysis of the data.

## 6 Conclusions

In this article, we investigated the performance of two different versions of a variety of information and classification criteria, the traditional version and a robust version, in the selection of the number of components of mixtures of regression models with random effects through a simulation study and a real-world application.

In the simulation study it was evident that both versions of the information and classification criteria perform similarly when there are no outliers. However, in the presence of outliers the robust information and classification criteria show a better performance since these methods correctly estimate the number of components for the majority of the scenarios. In other words, the presence of outliers does not seem to affect the performance of these criteria. On the other hand, the traditional information and classification criteria overestimate the number of components for almost every scenario containing outliers. In the same direction, the real-world application corroborated the conclusions drawn from the simulation study and demonstrated the importance of an adequate choice for the trimming proportion.

Regarding the choice for the trimming proportion, it is important to note that without a proper choice of the trimming proportion, a correct identification of the outliers may not happen. For instance, if the trimming proportion to use is too large, part of the observations will be incorrectly identified as outliers and, thus, will also be trimmed along with the real outliers. As such, in order to distinguish these observations from the outliers, an inspection of the FAST-TLE posterior weights may be needed, which will lead to an even more challenging computational effort. However, the technological advances of the recent years, namely in terms of processor power and memory, mean that nowadays one is able to afford it.

Novais and Faria (2021) demonstrated that $aBIC$, $KIC$ and $KIC_c$ showed to be the most reliable criteria in the estimation of the number of components of mixtures of linear mixed models. In this work, it can be seen that these criteria also produce good results in their robust version in the presence of contaminated data, so their use is recommended in any scenario.

Therefore, determining the correct number of components in a mixture model is not an easy question, even more so in the presence of outliers, and different scenarios clearly influence the performance of the information and classification criteria. Despite the challenging computational performance of the *FAST-TLE* algorithm, which can be very demanding and, as such, may constitute a drawback to its use without an adequate computer, the superiority of the robust information and classification criteria was clear when the data is contaminated, so their use is always recommended in the presence of outliers.

**Acknowledgements** The research of L. Novais was financed by FCT - Fundação para a Ciência e a Tecnologia, through the PhD scholarship with reference SFRH/BD/139121/2018.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics pp 803–821

Bates D, Sarkar D, Bates MD, Matrix L (2007) The lme4 package. R package version 2(1):74

Bhansali RJ, Downham DY (1977) Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion. Biometrika 64(3):547–551

Biernacki C, Govaert G (1997) Using the classification likelihood to choose the number of clusters. Computing Science and Statistics 29(2):451–457

Biernacki C, Celeux G, Govaert G (1999) An improvement of the NEC criterion for assessing the number of clusters in a mixture model. Pattern Recognition Letters 20(3):267–272

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the Integrated Completed Likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7):719–725

Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52(3):345–370

Cappozzo A, Greselin F, Murphy TB (2019) A robust approach to model-based classification based on trimming and constraints. Advances in Data Analysis and Classification pp 1–28

Cavanaugh JE (1999) A large-sample model selection criterion based on Kullback's symmetric divergence. Statistics & Probability Letters 42(4):333–343

Cavanaugh JE (2004) Criteria for linear model selection based on Kullback's symmetric divergence. Australian & New Zealand Journal of Statistics 46(2):257–274

Celeux G, Soromenho G (1996) An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification 13(2):195–212

Celeux G, Martin O, Lavergne C (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. Statistical Modelling 5(3):243–267

Celeux G, Frühwirth-Schnatter S, Robert CP (2019) Model selection for mixture models–perspectives and strategies. In: Handbook of mixture analysis, Chapman and Hall/CRC, Boca Raton, pp 117–154

Davies O, Goldsmith P (1972) Statistical Methods in Research and Production, 4th edn. Hafner Publishing Company, New York

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (methodological) 39(1):1–38

Depraetere N, Vandebroek M (2014) Order selection in finite mixtures of linear regressions. Statistical Papers 55(3):871–911

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer Science & Business Media, New York

Grün B (2008) Fitting finite mixtures of linear mixed models with the EM algorithm. In P Brito (Eds), Compstat 2008 - International Conference on Computational Statistics (pp 165-173), Springer, Heidelberg

Hannan EJ, Quinn BG (1979) The determination of the order of an autoregression. Journal of the Royal Statistical Society Series B (Methodological) 41(2):190–195

Hui FK, Warton DI, Foster SD (2015) Order selection in finite mixture models: complete or observed likelihood information criteria? Biometrika 102(3):724–730

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76(2):297–307

Kasahara H, Shimotsu K (2015) Testing the number of components in Normal mixture regression models. Journal of the American Statistical Association 110(512):1632–1645

Koller M (2016) robustlmm: an R package for robust estimation of linear mixed-effects models. Journal of Statistical Software 75(6):1–24

Li M, Xiang S, Yao W (2016) Robust estimation of the number of components for mixtures of linear regression models. Computational Statistics 31(4):1539–1555

Liang Z, Jaszczak RJ, Coleman RE (1992) Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing. IEEE Transactions on Nuclear Science 39(4):1126–1133

McLachlan G, Peel D (2000) Finite Mixture Models. John Wiley & Sons, New York

McLachlan GJ, Rathnayake S (2014) On the number of components in a gaussian mixture model. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4(5):341–355

Müller CH, Neykov N (2003) Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. Journal of Statistical Planning and Inference 116(2):503–519

Neykov N, Müller CH (2003) Breakdown point and computation of trimmed likelihood estimators in generalized linear models. Developments in Robust Statistics 142(1):277–286

Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. Computational Statistics & Data Analysis 52(1):299–308

Novais L, Faria S (2021) Selection of the number of components for finite mixtures of linear mixed models. Journal of Interdisciplinary Mathematics

pp 1–32
Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464
Sclove SL (1987) Application of model-selection criteria to some problems in multivariate analysis. Psychometrika 52(3):333–343
*R Development Core Team* (2018) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing
Young DS, Hunter DR (2015) Random effects regression mixtures for analyzing infant habituation. Journal of Applied Statistics 42(7):1421–1441
Yu C, Yao W, Yang G (2020) A selective overview and comparison of robust mixture regression estimators. International Statistical Review 88(1):176–202