Noname manuscript No.
(will be inserted by the editor)

Robust Estimation of the Number of Components for Mixtures of Linear Regression Models

Meng Li · Sijia Xiang · Weixin Yao

Received: date / Accepted: date

Abstract In this paper, we investigate a robust estimation of the number of components in the mixture of regression models using trimmed information criteria. Compared to the traditional information criteria, the trimmed criteria are robust and not sensitive to outliers. The superiority of the trimmed methods in comparison with the traditional information criterion methods is illustrated through a simulation study. Two real data applications are also used to illustrate the effectiveness of the trimmed model selection methods.

Keywords Mixture of linear regression models · Model selection · Robustness · Trimmed likelihood estimator

1 Introduction

Within the family of mixture models, the mixture of linear regression models has been studied extensively. The mixture of linear regression models was first introduced by Goldfeld and Quandt (1973) as a very general form of switching regression. The unknown parameters were estimated based on moment-generating functions, from a likelihood point of view. Jones and McLachlan

M. Li
Department of Statistics, Kansas State University, Manhattan, KS 66506

S. Xiang
Corresponding Author, School of Mathematics and Statistics, Zhejiang University of Finance and Economics, Hangzhou, Zhejiang 310018, PR China
Tel.: +86-571-87557158
Fax: +86-571-87557157
E-mail: fxbxsj@live.cn

W. Yao
Department of Statistics, University Of California, Riverside, CA 92521

(1992) applied the mixture of regressions in a data analysis and used EM algorithm to fit these models. For a general introduction to mixture models, see Lindsay (1995), Böhning (1999), and McLachlan and Peel (2000).

Choosing the number of components for mixture models has long been considered as an important but very difficult research problem. Many methods have been proposed. See, for example, the AIC and BIC methods (Leroux, 1992), distance measures based methods (Chen and Kalbfleisch, 1996; James et al., 2001; Charnigo and Sun, 2004; Woo and Sriram, 2006; Ray and Lindsay, 2008), and hypothesis testing based methods (Chen et al., 2001, 2004). Hawkins et al. (2001) proposed to choose the number of components in the mixture of linear regression models using the likelihood equations. Recently, Chen and Li (2009) and Li and Chen (2010) proposed an EM test approach for testing the order of finite mixtures.

However, most of the above model selection methods are not robust in the presence of outliers. Even a single outlier can totally change the result. In this article, we mainly focus on the information criteria based model selection methods for mixtures of regressions and consider a robust version of these methods based on the trimmed likelihood estimate (TLE, Neykov et al., 2007). A simulation study and two real data applications show that the new robust model selection methods work comparably to traditional information criteria based methods when the data are not contaminated but have superior performance when there are outliers.

The rest of the paper is organized as follows. In Section 2, we give an introduction of five traditionally used information criteria for model selection and introduce their corresponding robust versions based on TLE. A simulation study and two real data applications are used in Section 3 to demonstrate the effectiveness of the proposed robust model selection methods. A discussion section ends the paper.

2 Robust Model Selection Information Criteria for Mixtures of Regressions

2.1 Introduction of mixtures of regressions

Let Z be a latent class variable with $P(Z = j|\mathbf{x}) = \pi_j$, $j = 1, \cdots, m$, where $\mathbf{x}$ is a p-dimensional vector and m is the number of components. Given $Z = j$, the response y depends on the $p-$dimensional predictor $\mathbf{x}$ in a linear way:

$$y = \mathbf{x}^T\boldsymbol{\beta}_j + \epsilon_j,$$

where $\boldsymbol{\beta}_j = (\beta_{1j}, \cdots, \beta_{pj})^T$ and $\epsilon_j \sim N(0, \sigma_j^2)$. Here, we assume that $\mathbf{x}$ includes both the constant 1 and predictors. The conditional distribution of Y given $\mathbf{x}$ without observing Z can be written as:

$$f(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \phi(y; \mathbf{x}^T\boldsymbol{\beta}_j, \sigma_j^2), \tag{1}$$

where $\phi(y;\mu,\sigma^2)$ denotes the normal density with mean μ and variance σ^2, and $\boldsymbol{\theta} = (\pi_1, \sigma_1^2, \boldsymbol{\beta}_1, \cdots, \pi_m, \sigma_m^2, \boldsymbol{\beta}_m)^T$.

If the number of components m in the mixture of linear regression models was known, $\boldsymbol{\theta}$ could be estimated by maximizing the log-likelihood,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2) \right\}. \tag{2}$$

Note that the maximizer of (2) does not have an explicit solution and is usually estimated by the EM algorithm (Dempster et al., 1977). It is well known that the log-likelihood function (2) is unbounded and goes to infinity if one observation lies on one of the component lines and the corresponding component variance goes to zero. When the likelihood is unbounded, we define the MLE as the maximum interior/local mode (Hathaway, 1985, 1986; Chen, Tan, and Zhang, 2008; Yao, 2010).

2.2 Some information criteria for model selection

If the number of components m is unknown for mixture models, many methods have been proposed to determine the order m. Among them, information criteria have been popularly used to choose the number of components for mixture models due to their simplicity. If the log-likelihood (2) is treated as an objective function, one might tend to choose the model that maximizes the log-likelihood of the observed data. However, as pointed out by Celeux and Soromenho (1996), the log-likelihood is an increasing function of m. Therefore, the log-likelihood (2) can not be directly used to determine the number of components for mixture models. Many papers, such as Akaike (1974), Bozdogan (1993), and Rissanen (1986, 1987) have sought methods to remedy this problem by adding a penalty term to the log-likelihood.

Akaika's information criterion (AIC) is one of the most popular measures, and was proposed by Bozdogan and Sclove(1984) and Sclove (1987) in the mixture context. It takes the form:

$$AIC = -2\ell(\hat{\boldsymbol{\theta}}) + 2k,$$

where k is the number of parameters to be estimated and $\ell(\hat{\boldsymbol{\theta}})$ is the maximized value of the log-likelihood function for the estimated model.

Bayesian information criterion (BIC), proposed by Schwarz (1978), is another commonly used criterion, an approximation to twice the log Bayes factor (Fraley and Raftery, 1998). The approximation relies on regularity conditions that do not hold in the mixture models setting, but BIC has been shown to provide a consistent estimate of the number of components in mixture models (Keribin, 2000). It is defined by

$$BIC = -2\ell(\hat{\boldsymbol{\theta}}) + k \log n,$$

where n is the number of observations, or equivalently, the sample size.

As an alternative to AIC and BIC, the Hannan-Quinn information criterion (HQIC) is another criterion for model selection. It is given as

$$HQIC = -2\ell(\hat{\boldsymbol{\theta}}) + 2k\log(\log n),$$

where k is the number of parameters to be estimated and n is the number of observations. Although often cited, HQIC does not seem to have been popularly used in practice (Burnham and Anderson, 2002), and like BIC, is not asymptotically efficient (Claeskens and Hjort, 2008).

AIC and BIC are typically derived from approximations based on asymptotic arguments (Kass and Raftery, 1995). They penalize the log-likelihood by an additive factor and are relatively simple to implement. Although there are theoretical limitations on the applicability of these two methods, they have been proven to work quite well for model selection in mixture models.

BIC works well with the case that each mixture component corresponds to a separate cluster. However, if the number of clusters in the data set is different from the number of components, Biernacki et al. (2000) proposed the integrated complete likelihood (ICL) criterion as a modification. Let z_{ij} be the component label indicator,

$$z_{ij} = \begin{cases} 1, \text{ if } i^{th} \text{ observation is from } j^{th} \text{ component;} \\ 0, \text{ otherwise.} \end{cases}$$

and $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})$. Then, the complete-data is $\mathbf{x}_c = (\mathbf{x}_1, \mathbf{z}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)$ and its complete log-likelihood is given by:

$$\ell_c(\boldsymbol{\theta}; \mathbf{x}_c) = \sum_{i=1}^{n} \log \prod_{j=1}^{m} \left\{\pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2)\right\}^{z_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \log\left\{\pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2)\right\}.$$

ICL criterion penalizes the complexity of the mixture model, thus ensuring the partitioning of data with the greatest evidence.

$$ICL = -2\ell_c(\hat{\boldsymbol{\theta}}; \mathbf{x}_c) + k\log n,$$

where the missing data $\mathbf{z}$ have been replaced by their most probable values $\hat{\mathbf{z}}$, given the parameter estimate $\hat{\boldsymbol{\theta}}$, i.e., $\hat{z}_{ij} = 1$ if $\hat{p}_{ij} \geq \hat{p}_{ik}$ for all $k \neq j$ and 0 otherwise, where

$$\hat{p}_{ij} = \frac{\hat{\pi}_j \phi(y_i; \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)}{\sum_{j=1}^{m} \hat{\pi}_j \phi(y_i; \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)}.$$

In the BIC approach, only the observed likelihood is maximized, however, in the ICL approach, the complete log-likelihood is used. ICL appears to be more robust than BIC to the violation of some of mixture model assumptions and can select a number of clusters leading to a sensible partition of the data.

Bozdogan (1993) provided an analytic extension of AIC, without violating Akaika's principle of minimizing the Kulback-Leibler information quantity. The new selection criterion, called consistent AIC (CAIC), is defined as:

$$CAIC = -2\ell(\hat{\boldsymbol{\theta}}) + k(\log n + 1),$$

where k is the number of free parameters to be estimated, and n is the sample size.

The R package "mixtools" (Benaglia et al., 2009) uses AIC, BIC, ICL and CAIC to do model selection for mixture models. In the examples of Section 3, we use "mixtools" to implement the above four information criteria and calculate HQIC based on the information provided.

2.3 Trimmed information criteria

It is well known that the maximum likelihood estimate (MLE) via the expectation-maximization (EM) algorithm works well for the finite mixture of distributions. However, it is sensitive to outliers. Even a single outlier can cause at least one of the component parameters to become arbitrarily large. Therefore, the traditional information criteria introduced in Section 2.2 are sensitive to outliers in the data set. In this section, we consider a trimmed version of those information criteria to robustify the model selection procedures.

Assume that $(1-\alpha)\times 100\%$ of the observations in the data set are regular observations, and the remaining $\alpha \times 100\%$ are unpredictable outliers. The trimmed likelihood estimate (TLE) of mixture models, proposed by Neykov et al. (2007), only uses $(1-\alpha)\times 100\%$ of the data to fit the model, and removes the remaining $\alpha \times 100\%$ observations that are highly unlikely to occur if the fitted model were true. That is,

$$\max_{I_\alpha} \max_{\boldsymbol{\theta}} \sum_{i \in I_\alpha} f(y_i|\mathbf{x}_i, \boldsymbol{\theta}), \tag{3}$$

where $f(y \mid \mathbf{x}, \boldsymbol{\theta})$ is the density defined in (1), and I_α is the subset of $\{1, \ldots, n\}$ and only contains $\lfloor n(1-\alpha) \rfloor$ distinct elements of $\{1, \ldots, n\}$.

By combining the ideas of the trimmed likelihood estimate and the information criteria introduced in Section 2.2, we consider the trimmed versions of AIC, BIC, HQIC, ICL, and CAIC to robustly estimate the number of components for the mixture of regression models as follows:

$$TAIC = -2 \sum_{i \in \hat{I}_\alpha} f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}) + 2k, \tag{4}$$

$$TBIC = -2 \sum_{i \in \hat{I}_\alpha} f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}) + k \log n, \tag{5}$$

$$THQIC = -2 \sum_{i \in \hat{I}_\alpha} f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}) + 2k \log(\log n), \tag{6}$$

$$TICL = -2\sum_{i\in\hat{I}_\alpha} f_c(y_i|\mathbf{x}_i,\hat{\boldsymbol{\theta}},\hat{\mathbf{z}}) + k\log n, \tag{7}$$

$$TCAIC = -2\sum_{i\in\hat{I}_\alpha} f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\theta}}) + k(\log n + 1), \tag{8}$$

where $\hat{\boldsymbol{\theta}}$ is the trimmed likelihood estimator, $\hat{I}_\alpha$ is the corresponding index set, and $f_c(y_i|\mathbf{x}_c,\boldsymbol{\theta}) = \sum_{j=1}^{m} z_{ij}\log\left\{\pi_j\phi(y_i;\mathbf{x}_i^T\boldsymbol{\beta}_j,\sigma_j^2)\right\}$.

The combinatorial nature of the TLE, that is, all possible $\binom{n}{\lfloor n(1-\alpha)\rfloor}$ partitions of the data have to be fitted by the MLE, makes the TLE computationally expensive, and infeasible for large data sets. The FAST-TLE algorithm (Müller and Neykov, 2003; Neykov et al., 2007) was proposed to obtain an approximative TLE solution, which involves repeated iterations of a two-step procedure: a trial step followed by a refinement step. Next, we give the computation procedure to calculate (4)-(8) using FAST-TLE (Neykov et al., 2007).

Algorithm

Input: $\mathbf{x}, y, \alpha$, upk (upper limit for number of components), numini (number of initial values)
Output: selected models.
for $i = 1, ...,$ upk **do**
  for $j = 1, ...,$ numini **do**
    Find an initial value for $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}_0$.
    while change of (3) $\geq$ acc **do**
      For a given estimator $\hat{\boldsymbol{\theta}}$, sort $f(y|\mathbf{x},\hat{\boldsymbol{\theta}})$ as $f(y_{\nu(1)}|\mathbf{x}_{\nu(1)},\hat{\boldsymbol{\theta}}) \geq \cdots \geq f(y_{\nu(n)}|\mathbf{x}_{\nu(n)},\hat{\boldsymbol{\theta}})$, then $\{\nu(1),\cdots,\nu(\lfloor n(1-\alpha)\rfloor)\}$ forms the index set $\hat{I}_\alpha$.
      Given an index set $\hat{I}_\alpha$, update the estimator of $\boldsymbol{\theta}$, which maxi -mizes $\sum_{i\in\hat{I}_\alpha} f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$.
    end while
  end for
  Select $\hat{\boldsymbol{\theta}}$ and $\hat{I}_\alpha$ with the largest (3) value.
  Apply $\hat{\boldsymbol{\theta}}$ and $\hat{I}_\alpha$ to (4)-(8) to calculate the corresponding information criteria.
end for
return Selected model based on each information criterion.

The initial value $\boldsymbol{\theta}_0$ might be found by fitting the mixture of linear regressions to a random subsample of size d from the data, where d is a value larger than k. In our examples, we tried 40 initial values in total.

We want to point out that we are not the first ones to use the idea of trimmed information criteria. Neykov et al. (2007) have briefly discussed the trimmed BIC in their simulation study. Based on their limited empirical experience, the trimmed BIC works well and could robustly estimate the number of mixture components. In this paper, we extend such trimmed idea to some other commonly used information criteria and give more simulation study and two real data applications in the next section to demonstrate the effectiveness of the trimmed information criteria.

3 Simulation Study and Real Data Application

3.1 Simulation study

In this section, we investigate the effectiveness of the trimmed information criteria and compare them with the traditional information criteria for model selection for mixture models when outliers are present. To be more specific, the following five new methods are considered: trimmed AIC (TAIC), trimmed BIC (TBIC), trimmed HQIC (THQIC), trimmed CAIC (TCAIC), and trimmed ICL (TICL). The performance of the new methods are compared to AIC, BIC, HQIC, CAIC, and ICL, whose likelihoods are calculated based on the maximum likelihood estimate (MLE). The trimming proportion α is set to be 5% for all information criteria. Similar to TLE, the proportion α is an important tuning parameter. Usually a conservative α is desired. In our simulation study, the proportion of outliers is never greater than 0.05. In Section 3.2, we use a real data set to illustrate how to data adaptively choose the α using the graphical tool proposed in Neykov et al. (2007).

To compare the performance of different model selection methods, we report the percentage of times when the number of components is correctly estimated. In addition, we also report the lower quartile (LQ), the median (MD), and the upper quartile (UQ) of the estimated number of components for each method.

We consider the following two mixture of linear regression models:
Example 1:

$$Y = \begin{cases} 0 + X_1 + X_2 + \epsilon_1, \text{ if } Z = 1, \\ 0 - X_1 - X_2 + \epsilon_2, \text{ if } Z = 2. \end{cases}$$

Example 2:

$$Y = \begin{cases} 3 + 3X_1 + 4X_2 + \epsilon_1, & \text{if } Z = 1, \\ 1 + X_1 + X_2 + \epsilon_2, & \text{if } Z = 2, \\ -1 - X_1 - X_2 + \epsilon_3, & \text{if } Z = 3, \\ -3 - 3X_1 - 4X_2 + \epsilon_4, & \text{if } Z = 4. \end{cases}$$

The mixing proportions are 0.4 and 0.6 in *Example 1*, and are all equal to 0.25 in *Example 2*. In both examples, $X_{ik} \sim N(0, 1)$ for $k = 1, 2$. The sample sizes of $n = 100$ and $n = 200$ are conducted over 500 repetitions for *Example 1* and 200 repetitions for *Example 2*. The proportions of outliers are $\alpha_0 = 0.05, 0.03$ and 0.01, and the outliers are generated by shifting the generated Y up by a length randomly generated from $U(7, 10)$ for *Example 1* and $U(20, 30)$ for *Example 2*. Note that the trimming proportion is 0.05 for all trimmed information criteria. Using the above three proportions of outliers, we can check how the trimmed information criteria work for both the cases where the trimming proportions are correct and the cases where the trimming proportions are conservative.

For the error distributions, we consider two scenarios. In Scenario 1, the errors have the distribution as $\epsilon \sim N(0, 1)$. In addition, in Scenario 2, we also considered contaminated normals as the distribution of error. To be more specific, both $\epsilon \sim 0.95N(0, 1) + 0.05N(0, 3^2)$ and $\epsilon \sim 0.95N(0, 1) + 0.05N(0, 5^2)$

are used as the error distributions.

Scenario 1:

Table 1 reports the percentage of times when the number of components in the mixture of linear regression models is correctly estimated. From Table 1, we can see that the percentages of correct estimates of number of components by TBIC, TCAIC, and TICL are much higher than the five traditional methods, TAIC and THQIC. Both AIC and its trimmed version fail terribly to estimate the correct number of components. The trimmed versions of BIC, CAIC, and ICL, on the other hand, can greatly improve the performance of their original versions when the data is contaminated. In addition, the performance of THQIC is largely affected by the sample size. To our surprise, the traditional methods perform better when the sample size is small in our examples, but TBIC, TCAIC, and TICL show better results when the sample size is large.

To understand better how the methods perform, we also report the lower quartile, the median, and the upper quartile of the estimated numbers of components, in Table 2 and Table 3, for $n = 100$ and $n = 200$, respectively. Clearly, TBIC, TCAIC, and TICL give much better performance than the other methods. In addition, AIC, BIC, HQIC, CAIC, ICL, and TAIC tend to overestimate the number of components. The reason why TAIC and THQIC fail might be due to the small penalty term used as compared to other information criteria.

Scenario 2:

Since the contaminated normals already mimics the outlier case, we did not add any additional outliers. $\alpha = 0.05$ is still used as the trimming proportion, and the simulation results are reported in Table 4 - Table 7. It is clear that TBIC, TCAIC and TICL still perform the best compared to the traditional information criteria, TAIC and THQIC. The increment of sample size improves the performance of all the trimmed information criteria to a large extent, but no unanimous conclusion can be drawn to traditional information criteria.

In addition, thanks to the comment made by the reviewer, note that in Table 4, the correct percentage for untrimmed criteria are now much higher than Table 1. This happens, because for the mixture $0.95N(0,1)+0.05N(0,3^2)$ error model, all points have the same mean functions but potential outliers may have an inflated variance. Then the proportion of outliers may be much less than 5%.

3.2 Real data analysis

Example 1 (Crabs data). We use the crabs data from R-package "MASS" as an example to compare different information criteria and their trimmed versions. Five morphological measurements were taken from 200 crabs of the species Leptograpsus variegatus, collected at Fremantle, Australia. Following García-Escudero et al. (2010), we only focus on analyzing two variables, RW

	$n = 100$			$n = 200$		
	$\alpha_0 = 0.05$	$\alpha_0 = 0.03$	$\alpha_0 = 0.01$	$\alpha_0 = 0.05$	$\alpha_0 = 0.03$	$\alpha_0 = 0.01$
Example 1	$(m = 2)$					
AIC	0%	0.5%	2.5%	0%	0%	0%
BIC	4%	2.5%	18%	0%	0%	8.5%
HQIC	0%	0.5 %	2.5%	0%	0%	0%
CAIC	7.5%	5%	23%	0%	0%	10%
ICL	3.5%	1.5%	17.5%	0%	0%	7%
TAIC	0%	0%	0%	5%	0.5%	0%
TBIC	97.5%	95.5%	86%	100%	99%	99%
THQIC	27.5%	12 %	7.5%	87%	77%	60%
TCAIC	96%	92.5%	88%	100%	100%	98.5%
TICL	97%	93.5%	84%	99.5%	98.5%	98.5%
Example 2	$(m = 4)$					
AIC	5.5%	0%	2.5%	3%	0%	0%
BIC	23%	7.5%	9%	18.5%	5%	1%
HQIC	5.5%	0%	2.5%	2.5%	0%	0%
CAIC	27.5%	10.5%	14%	28%	6%	1%
ICL	23.5%	7.5%	9%	18%	4%	1%
TAIC	2%	2%	1%	25%	16.5%	7%
TBIC	86%	83%	75.5%	100%	98%	98%
THQIC	45%	30.5%	28.5%	86.5%	84.5%	71%
TCAIC	91%	85%	73.5%	99.5%	99.5%	99%
TICL	86%	82%	74.5%	100%	97.5%	97%

Table 1 Percentages of times when the number of components in the mixture of regression models is correctly estimated when $\epsilon \sim N(0, 1)$.

(rear width in mm) and CL (carapace length mm), with the objective of distinguishing between the two crabs sexes, without the other variables. The variable CL is considered as the response variable while RW is considered as the explanatory variable. The scatter plot of the data is shown in Figure 1, where squares and circles denote the two groups of crabs based on their sexes.

Let us suppose that the sexes of the crabs were unknown, and then estimate the number of components, using $\alpha = 0.05$ as the trimming proportion for all the trimmed information criteria. The results are reported in the first row of Table 8 ($\alpha_0 = 0$, without outliers). It can be seen that all trimmed criteria, except for TAIC, provide correct estimates of the number of components, when there are no outliers in the data.

To check the robustness of different model selection methods, similar to McLachlan and Peel (2000), we artificially add some random outliers to the original data set, 1%, 3% and 5% outliers, to be more specific. The outliers are generated uniformly in the rectangle decided by the maximums and minimums

	$\alpha_0 = 0.05$			$\alpha_0 = 0.03$			$\alpha_0 = 0.01$		
	LQ	MD	UQ	LQ	MD	UQ	LQ	MD	UQ
Example 1	($m = 2$)								
AIC	3	4	7	6	8	10	6	8	9
BIC	3	3	3	3	3	3	3	3	3
HQIC	3	4	7	6	8	9	4	7	9
CAIC	3	3	3	3	3	3	2	3	3
ICL	3	3	3	3	3	3	3	3	3
TAIC	9	10	10	9	10	10	9	10	10
TBIC	2	2	2	2	2	2	2	2	2
THQIC	2	8	10	8	9	10	8	9	10
TCAIC	2	2	2	2	2	2	2	2	2
TICL	2	2	2	2	2	2	2	2	2
Example 2	($m = 4$)								
AIC	5	5	7	5	6	8	5	6	7
BIC	4	5	5	5	5	6	5	5	5
HQIC	5	5	6	5	6	8	5	6	7
CAIC	4	5	5	5	5	5	5	5	5
ICL	4	5	5	5	5	6	5	5	5
TAIC	9	10	10	9	10	10	9	10	10
TBIC	4	4	4	4	4	4	4	4	4
THQIC	4	5	9	4	6	9	4	7	9
TCAIC	4	4	4	4	4	4	4	4	4
TICL	4	4	4	4	4	4	4	4	4

Table 2 The lower quartile (LQ), the median (MD), and the upper quartile (UQ) of the estimated numbers of components when $n = 100$ and $\epsilon \sim N(0, 1)$.

of RW and CL. That is, after generating outliers from $U(0,1) * U(0,1)$, we multiply the x-axis by the range of the RW, and add the mean of RW, and did the same thing for y-axis with CL. Figure 1 shows an example when 5% outliers, denoted by dots, were added into the data. The model selection results are also reported in Table 8, from which we can see that the proposed TBIC, THQIC, TCAIC, and TICL perform much better than the rest of the methods. In addition, it is interesting to note that AIC and TAIC tend to overestimate the number of components, which is consistent with the simulation results.

Example 2 (Australian Institute of Sport). Next, we consider the Australian Institute of Sport (AIC) data, available from R-package "alr3". The dataset describes physical and hematological measurements on 202 athletes (102 male and 100 female) at the Australian Institute of Sport. As suggested by Cook and Critchley (2000), here we consider an athlete's lean body mass as the response, and regress it on three predictors: height (in cm), weight (in kg), and red cell count.

	$\alpha_0 = 0.05$			$\alpha_0 = 0.03$			$\alpha_0 = 0.01$		
	LQ	MD	UQ	LQ	MD	UQ	LQ	MD	UQ
Example 1	$(m = 2)$								
AIC	3	3	4	3	3	5	3	5	9
BIC	3	3	3	3	3	3	3	3	3.25
HQIC	3	3	4	3	3	5	3	5	8
CAIC	3	3	3	3	3	3	3	3	3
ICL	3	3	3	3	3	3	3	3	4
TAIC	9	10	10	9	10	10	9	10	10
TBIC	2	2	2	2	2	2	2	2	2
THQIC	2	2	2	2	2	2	2	2	4.25
TCAIC	2	2	2	2	2	2	2	2	2
TICL	2	2	2	2	2	2	2	2	2
Example 2	$(m = 4)$								
AIC	5	5	7	5	5	6	5	5	6
BIC	5	5	6	5	5	5.25	5	5	5
HQIC	5	5	6	5	5	6	5	5	6
CAIC	5	5	5	5	5	5	5	5	5
ICL	5	5	6	5	5	6	5	5	5.25
TAIC	5	8	10	6	9	10	8	9	10
TBIC	4	4	4	4	4	4	4	4	4
THQIC	4	4	4	4	4	4	4	4	5
TCAIC	4	4	4	4	4	4	4	4	4
TICL	4	4	4	4	4	4	4	4	4

Table 3 The lower quartile (LQ), the median (MD), and the upper quartile (UQ) of the estimated numbers of components when $n = 200$ and $\epsilon \sim N(0, 1)$.

	Example 1		Example 2	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
AIC	13%	11.5%	34%	42.5%
BIC	69%	68.5%	65.5%	80.5%
HQIC	13%	11.5%	32.5%	42.5%
CAIC	72.5%	77.5%	60.5%	83.5%
ICL	68.5%	68%	66%	80%
TAIC	0%	0.5%	0%	6%
TBIC	87.5%	98.5%	66.5%	96%
THQIC	7.5%	68%	17.5%	70%
TCAIC	89.5%	100%	66.5%	94%
TICL	85.5%	98.5%	64%	95%

Table 4 Percentages of times when the number of components in the mixture of regression models is correctly estimated when $\epsilon \sim 0.95N(0, 1) + 0.05N(0, 3^2)$.

	$n = 100$			$n = 200$		
	LQ	MD	UQ	LQ	MD	UQ
Example 1						
AIC	4	7	9	3	5	8
BIC	2	2	2	2	2	3
HQIC	4	6	9	3	5	7
CAIC	2	2	2	2	2	2
ICL	2	2	2	2	2	3
TAIC	9	10	10	9	10	10
TBIC	2	2	2	2	2	2
THQIC	8	9	10	2	2	3
TCAIC	2	2	2	2	2	2
TICL	2	2	2	2	2	2
Example 2						
AIC	4	5	7	4	5	6
BIC	4	4	4	4	4	4
HQIC	4	5	7	4	5	6
CAIC	3	4	4	4	4	4
ICL	4	4	4	4	4	4
TAIC	9	10	10	8	9	10
TBIC	4	4	4	4	4	4
THQIC	6	8	10	4	4	5
TCAIC	3	4	4	4	4	4
TICL	4	4	4	4	4	4

Table 5 The lower quartile (LQ), the median (MD), and the upper quartile (UQ) of the estimated numbers of components when $\epsilon \sim 0.95N(0,1) + 0.05N(0,3^2)$.

	Example 1		Example 2	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
AIC	6%	3.5%	21.5%	12%
BIC	29.5%	20%	46%	41%
HQIC	6%	3.5%	18.5%	12%
CAIC	34.5%	25%	36.5%	44.5%
ICL	29.5%	20%	45%	41%
TAIC	0%	1%	2%	10.5%
TBIC	90%	99.5%	67%	97.5%
THQIC	8%	76.5%	22.5%	75%
TCAIC	91%	99.5%	69%	97.5%
TICL	90%	99%	65.5%	97%

Table 6 Percentages of times when the number of components in the mixture of regression models is correctly estimated when $\epsilon \sim 0.95N(0,1) + 0.05N(0,5^2)$.

	$n = 100$			$n = 200$		
	LQ	MD	UQ	LQ	MD	UQ
Example 1						
AIC	5	7	9	4	6	8
BIC	2	3	3	3	3	4
HQIC	4	7	9	4	6	8
CAIC	2	2	3	2	3	4
ICL	2	3	4	3	3	4
TAIC	9	10	10	9	10	10
TBIC	2	2	2	2	2	2
THQIC	9	9	10	2	2	2
TCAIC	2	2	2	2	2	2
TICL	2	2	2	2	2	2
Example 2						
AIC	5	6	8	5	6	7
BIC	4	4	5	4	5	6
HQIC	4	5	7	5	6	7
CAIC	2	4	4	4	4	5
ICL	4	4	5	4	5	6
TAIC	9	10	10	7	9	10
TBIC	4	4	4	4	4	4
THQIC	4	9	10	4	4	4.25
TCAIC	3	4	4	4	4	4
TICL	4	4	4	4	4	4

Table 7 The lower quartile (LQ), the median (MD), and the upper quartile (UQ) of the estimated numbers of components when $\epsilon \sim 0.95N(0,1) + 0.05N(0,5^2)$.

Table 8 The estimated number of components for crabs data when using traditional and trimmed information criteria.

	AIC	BIC	HQIC	CAIC	ICL	TAIC	TBIC	THQIC	TCAIC	TICL
α_0=0	2	2	2	2	2	4	2	2	2	2
α_0=0.01	7	2	7	2	2	7	2	2	2	2
α_0=0.03	8	2	5	4	4	9	2	2	2	2
α_0=0.05	8	4	8	4	4	4	2	2	2	2

According to Cook and Critchley (2000), the 12 males participating in field events can be considered as one component, and there should be two components in the remaining 190 athletes, based on their genders. The estimated number of components for the AIS data using traditional information criteria are reported in Table 9. After classifying the data based on the foregoing criterion, we did residual analysis within each component, and the Bonferroni

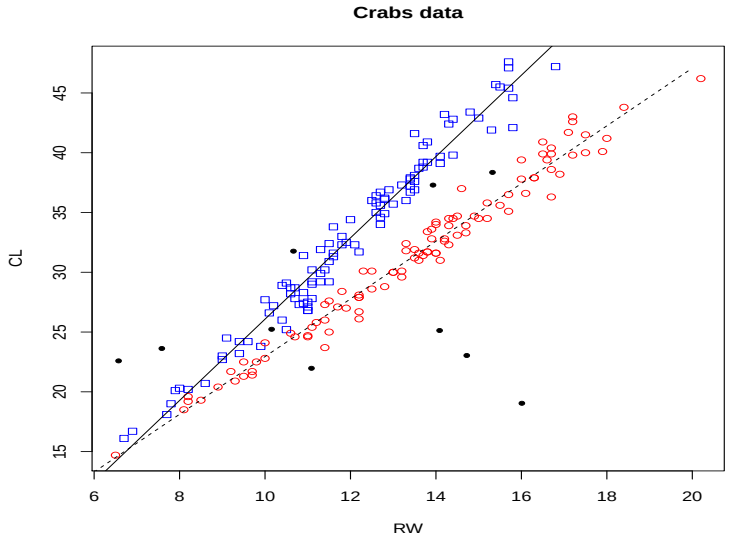

Fig. 1 Crabs data: squares and circles denote the two groups of crabs based on their sexes, and dots denote the randomly generated outliers.

outlier test shows that there are 4 outliers in total. That is, $\alpha_0 = 0.02$ in this case. Therefore, we tried $\alpha = 0.01, 0.03$ and 0.05 to the data, and the model selection results are also reported in Table 9. Clearly, the results by the traditional information criteria do not provide informative selection result, but the results by TBIC, TCAIC and TICL are satisfactory.

Table 9 The estimated number of components for Australian Institute of Sports data when using traditional and trimmed information criteria.

	AIC	BIC	HQIC	CAIC	ICL
	5	4	1	1	4
	TAIC	TBIC	THQIC	TCAIC	TICL
α=0.01	9	3	9	2	3
α=0.03	10	3	8	3	3
α=0.05	10	3	8	3	3

In all the above examples, we have fixed the trimming proportion α in advance. To explore how to data adaptively select the trimming proportions, we apply the graphical tool proposed in Neykov et al. (2007). That is, we fit the crabs data with a 2-component mixture of linear regression models using a grid points of α values, ranging from 0% to 50% in steps of 1%. Figure 2 shows the trimmed likelihoods versus α's with $m = 2$, when the actual percentages of

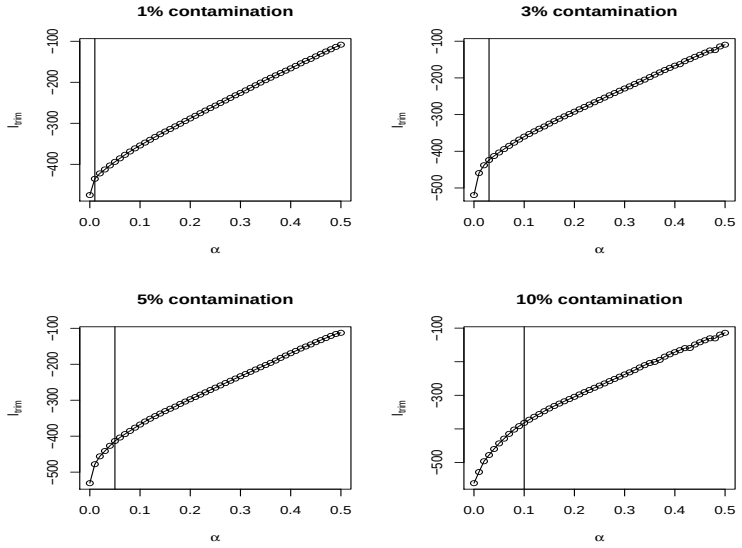

Fig. 2 Plot of trimmed likelihood versus the trimming proportion with $m = 2$. The vertical line corresponds to the true percentages of outliers.

outliers are 1%, 3%, 5% and 10%, respectively, denoted by the vertical lines. The percentage of outliers can be estimated by the largest α at which the slope of the curve changes. From Figure 2, we can see that the slope-changing locations are very close to the true percentages of outliers (corresponding to the vertical lines). To check how sensitive of the plot to the number of components, we also draw a similar plot in Figure 3 with $m = 3$. In this example, Figure 2 and Figure 3 give similar finding. In practice, however, we recommend to first choose m based on a conservative α. Then we can choose a better α based on the found m. In addition, we might also iterate the above two steps several times to improve the accuracy. Please refer to Section 4 for more discussion.

4 Discussion

In this paper, we investigated some trimmed versions of information criteria to robustly estimate the number of components for mixtures of linear regression models when outliers are present in the data set. We demonstrated the superiority of the trimmed methods in comparison with the traditional methods when the data are contaminated using a simulation study and two real data examples.

However, in this article, we mainly focus on the information criteria based model selection methods. It requires more research to see whether the trimmed idea can be used to robustify some other model selection methods, such as EM

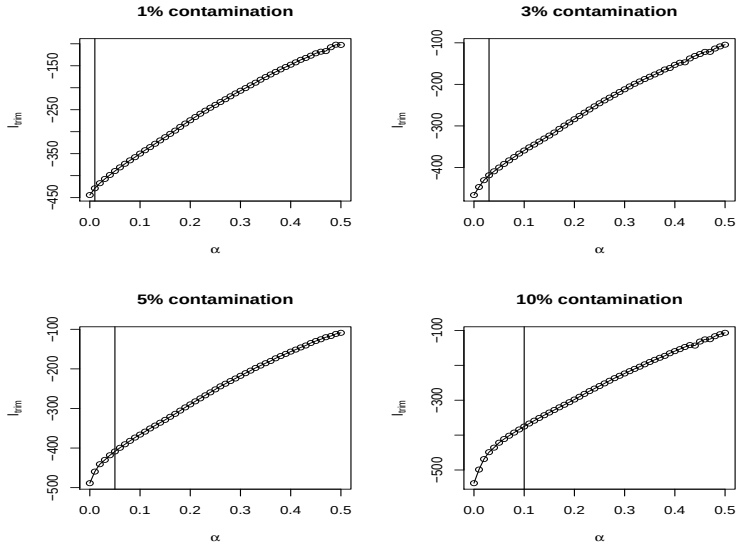

Fig. 3 Plot of trimmed likelihood versus the trimming proportion with $m = 3$. The vertical line corresponds to the true percentages of outliers.

test (Chen and Li, 2009; Li and Chen, 2010). Chen and Khalili (2009) used a penalized likelihood to select the number of components for mixture models.

It will be interesting to see whether robust order selection (determine the number of components) can be achieved if we apply similar penalty functions to some existing robust mixtures of regression models, where robust error distributions are used (e.g., t-distribution in Yao et al., 2004; Laplace distribution in Song et al., 2014), so that the mixture models are robust in case of outliers (Neykov et al. 2007; Bai et al. 2012; Bashir and Carter, 2012).

We have applied the graphical tool proposed in Neykov et al. (2007) to choose the trimming proportion α in the crabs data application. However, as pointed out by Neykov et al. (2007), such estimated trimming proportions tend to underestimate the true values in some cases. Therefore, it requires further study about how to adaptively choose an optimal or a conservative trimming proportion α analytically.

In this article, we mainly deal with the normal error data. However, we believe all the proposed trimming idea for model selection criteria can be also extended to non-normal error data, such as mixtures of poisson regression and mixtures of logistic regression. The only difference is the definition of likelihood function used in information criteria.

References

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

2. Bai, X., Yao, W. and Boyer, J. E. (2012). Robust fitting of mixture regression models. *Computational Statistics and Data Analysis*, 56, 2347-2359.
3. Bashir, S. and Carter, E. (2012). Robust mixture of linear regression models. *Communications in Statistics-Theory and Methods*, 41, 3371-3388.
4. Benaglia, B., Chauveau, D., Hunter, D.R. and Young, D. (2009). Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6), 1-29.
5. Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719-725.
6. Böhning, D. (1999), *Computer-Assisted Analysis of Mixtures and Applications*, Boca Raton, FL: Chapman and Hall/CRC.
7. Bozdogan, H. and Sclove, S. L. (1984). Multi-sample cluster analysis using Akaike's information criterion. *Annals of the Institute of Statistical Mathematics*, 36, 163-180.
8. Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In Opitz, O., Lausen, B. and Klar, R. (eds), *Information and Classification: Concepts, Methods and Applications*, 44-54. Springer-Verlag, Berlin.
9. Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag.
10. Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13, 195-212.
11. Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the L2-distance between competing models. *Journal of the American Statistical Association*, 99, 488-498.
12. Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of Royal Statistical Society, Ser. B*, 63, 19-29.
13. Chen, H., Chen, J., and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of Royal Statistical Society, Ser. B*, 66, 95-115.
14. Chen, J. and Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *The Canadian Journal of Statistics*, 24, 167-175.
15. Chen, J. and Khalili, A. (2009). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 104,187-196.
16. Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37, 2523-2542.
17. Chen, J., Tan, X., and Zhang, R. (2008). Inference for normal mixture in mean and variance. *Statistica Sincia*, 18, 443-465.
18. Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*, Cambridge.
19. Cook, R.D. and Critchley, F. (2000). Identifying regression outliers and mixtures graphically. *Journal of the American Statistical Association*, 95, 781-794.
20. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society, Ser. B*, 39, 1-38.
21. Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41, 578-588.
22. García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A. and San Martín, R. (2010). Robust clusterwise linear regression through trimming. *Computational Statistics and Data Analysis*, 54, 3057-3069.
23. Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1, 3-6.
24. Hathaway, R.J. (1985). A constrained formulation of maximum-likelihood estimation for nomral mixture distributions. *Annals of Statistics*, 13, 795-800.
25. Hathaway, R.J. (1986). A constrained EM algorithm for univariate mixutres. *Journal of Statistical Computation and Simulation*, 23, 211-230.
26. Hawkins, D. S., Allen, D. M. and Stomberg, A. J. (2001). Determining the number of components in mixtures of linear models. *Computational Statistical and Data Analysis*, 38, 15-48.
27. James, L. F., Priebe, C. E., and Marchette, D. J. (2001). Consistent estimation of mixture complexity. *The Annals of Statistics*, 29, 1281-1296.

28. Jones, P. N. and Mclachlan G. J. (1992). Fitting finite mixture models in a regression context. *Australian Journal of Statistics*, 34, 233-240.
29. Keribin, C. (2000). Consistent estimation of the order of mixture models. *The Indian Journal of Statistics*, 62, 49-66.
30. Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
31. Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20, 1350-1360.
32. Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105, 1084-1092.
33. Lindsay, B. G., (1995), *Mixture Models: Theory, Geometry, and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics v 5, Hayward, CA: Institure of Mathematical Statistics.
34. McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* John Wiley and Sons, New York.
35. Müller, C. and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference*, 116, 503-519.
36. Neykov, N., Filzmoser, P., Dimova, R. and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics and Data Analysis*, 52, 299-308.
37. Ray, S. and Lindsay, B. G. (2008). Model selection in high-dimensions: A quadratic-risk based approach. *Journal of Royal Statistical Society, Ser. B*, 70, 95-118.
38. Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statisics*, 14, 1080-1100.
39. Rissanen, J. (1987). Stochastic complexity. *Journal of Royal Statistical Society, Ser. B*, 49, 223-239.
40. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
41. Sclove, S. L. (1987). Application of some model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.
42. Song, W., Yao, W. and Xing, Y. (2014). Robust mixture regression model fitting by Laplace distribution. *Computational Statistics and Data Analysis*, 71, 128-137.
43. Woo, M. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101, 1475-1485.
44. Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*, 140, 2089-2098.
45. Yao, W., Wei, Y. and Yu, C. (2014). Robust mixture regression using t-distribution. *Computational Statistics and Data Analysis*, 71, 116-127.