International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

https://gssrr.org/index.php/JournalOfBasicAndApplied/index

On Model Selection Criterion for Finite Gaussian Mixture Models

Aya Ben Shatwan[a], Abdelbaset Abdalla[b*], Hatim Mohammed[c],Eisay Bin Ismaeil[d] ,Ahmed M. Mami[e]

[a]*Libyan National Oil Corporation*

[b,c,e]*Department of Statistics, Faculty of Science, University of Benghazi, Benghazi, Libya*

[d]*Department of Statistics, Faculty of Arts and Science, University of Ajdabiya, Ajdabiya, Libya*

[a]*Email: aya@uob.edu.ly,* [b]*Email: abdelbaset.abdalla@uob.edu.lyn,* [c]*Email: hatim.mohammed@uob.edu.ly*

[d]*Email: Issawow30@gmail.com,* [e]*Email*: *ahmed.mami@uob.edu.ly*

Abstract

This paper delves into the realm of model selection criteria for Finite Mixture Models (FMM), focusing on key evaluation methods such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Integrated Completed Likelihood (ICL), and the Bootstrap Likelihood Ratio Test (BLRT). These criteria aid in balancing model fit and complexity, guiding researchers in choosing the most appropriate FMM for analyzing simulated and real datasets. Through extensive simulation studies, the paper meticulously evaluates and contrasts the performance of these criteria under various parameter settings and sample sizes, offering valuable insights for advancements in statistical modeling. The study underscores the importance of selecting the right criterion tailored to the dataset characteristics and research objectives. It highlights the impact of sample size on model selection, noting AIC's tendency to favor complexity and potential overfitting, while BIC and ICL excel in handling sample size variations by penalizing complexity effectively. The utilization of BLRT for comparing models with different complexities aids in identifying the optimal model configuration. Statistical analyses, including p-value assessments and visual aids like scatter plots and density functions, enhance the understanding of model performance and complexity.

Received: *3/28/2024*
Accepted: *5/28/2024*
Published: *6/8/2024*

---

* Corresponding author.

Overall, the paper emphasizes the significance of informed model selection decisions, ensuring a robust and accurate representation of underlying models in regression analysis.

***Keywords*:** Finite Mixture Models (FMM); Akaike Information Criterion (AIC); Bayesian Information Criterion (BIC); Integrated Completed Likelihood (ICL);Bootstrap Likelihood Ratio Test (BLRT).

1. Introduction

Finite mixture models were introduced in the late 19th century by Newcomb in 1886 and further developed by [27] using a mixture of two univariate Gaussian distributions. Gaussian components have since become the most commonly used type of mixture model. In the 1990s, finite mixture models were expanded to include standard linear regression models and generalized linear models. [16] discussed non-parametric and semi-parametric maximum likelihood estimation in mixture models, while [18] addressed key issues related to mixture models. The use of finite mixture models has significantly increased over the past decade, attracting attention for both practical and theoretical applications. For a comprehensive historical overview and insights into applications of finite mixture models, valuable resources can be found in the works of [23,12]

Finite mixture models are essential for modeling data from heterogeneous populations and are used for model-based clustering to classify data into distinct groups. Each mixture component represents a different group of observations in the dataset. These models are widely utilized in various fields such as medicine and biology. The derivations and applications of finite mixture models are extensively discussed in the works of [21,12]-, and recent reviews by [26,25,19], which explore recent advancements and challenges in finite mixture models and model-based clustering. References include [20,11,31].

Finite mixture models, including the Finite Mixture of Regression (FMR) models, have been extensively researched and applied across various domains. FMR models involve fitting different regression models to segments of data showing similar behavior. The concept of mixtures of linear regression models was introduced by [29] and further developed by [9] using an EM approach. Jones and [14] applied combinations of regressions in data analysis using the EM algorithm. The flexibility of finite mixture models has led to successful applications in fields like astronomy, biology, medicine, economics, and marketing. The Expectation-Maximization (EM) algorithm is commonly used for fitting Gaussian Mixture Models (GMMs). It iterates between the E step, where it estimates the probability of each data point belonging to each component, and the M step, where it updates model parameters to maximize data likelihood. While the EM algorithm is known for global convergence, it can be sensitive to initial parameter values, emphasizing the importance of careful initialization for reliable results. Sensible outcomes are typically achieved starting from reasonable initial values. References include [33,3034].

In the field of statistical modeling, Finite Mixture Models (FMMs) are like versatile tools that help us unravel complex datasets by assuming that the data comes from a mix of different underlying distributions. When researchers want to figure out which FMM works best for a particular dataset, they rely on specific ways to evaluate them. These evaluation methods can be grouped into three main categories: information criteria, approximate likelihood ratio tests, and resampling techniques.

Information criteria include measures like the Akaike Information Criterion (AIC) developed by [2], the Bayesian Information Criterion (BIC) introduced by [32], and the Integrated Completed Likelihood (ICL) criterion by [4]. AIC and BIC help us balance how well the model fits the data with its complexity, with lower values indicating a better fit with fewer parameters. However, BIC might not work well with small sample sizes, while AIC might overcomplicate things. Even though it's more challenging to use, [26] have shown that the ICL method performs exceptionally well in many situations.

In the second category, approximate likelihood ratio tests include methods like the Bootstrap Likelihood Ratio Test (BLRT) created by [17]. This test compares how likely the data is under the model being tested compared to a more complex model, with a significant p-value suggesting that the more complex model is a better fit.

The third category involves resampling techniques, such as the k-fold cross-validation method introduced by [13]. This technique divides the data into subsets, trains the model on most of the subsets, and then tests it on the remaining subset to see how well it generalizes.

This thesis sets out to explore how these different evaluation methods can help us choose the best FMM for analyzing both simulated and real datasets. By examining the performance of various criteria, we aim to gain valuable insights and contribute to advancements in the field of statistical modeling. Our focus in this work lies on the first and second categories of this project. To evaluate and contrast the various proposed model selection criteria in the realm of FMM, we employ numerous simulation studies. We thoroughly examine the pros and cons of both model selection criteria based on parameter settings and sample sizes and draw compelling conclusions with key takeaways. In forthcoming work, we shall present our application on actual data sets.

2. Methodology

This section will lay the essential groundwork for finite mixture models and model selection criteria. Following this, the proposed methodology will be outlined.

2.1 Finite Mixture of Gaussian Regression Model

Suppose a random sample $\{(x_i, y_i), i = 1, \dots, n\}$ of independent identically distributed

i.i.d observations are drawn from a finite mixture of normal regression models. In this case, explanatory variables x_i are collected for each observation y_i. Then, the probability distribution function is given by

$$g(y_i; x_i; \psi) = \sum_{k=1}^{k} \alpha_k \, \emptyset(\, y_i;\, x_i\beta_k, \sigma_k^2) \qquad (2.1)$$

where K is the total number of mixture regression components, $\emptyset(\, y_i;\, x_i\beta_k, \sigma_k^2)$

is a Gaussian density function of the kth component with mean $x_i\beta_k$, and variance σ_k^2

The mixing proportions, $\alpha_k, k = 1, \dots, K$ have the following restrictions: $0 < \alpha_k \leq 1$ and $\sum_{k=1}^{K} \alpha_k = 1$ .Therefore, the parameter vector $\psi = \{\alpha_1, \dots, \alpha_{k-1}, \beta_1, \dots, \beta_k, \sigma_1^2, \dots, \sigma_k^2\}$ Where $\beta_1, \dots, \beta_k, \sigma_1^2, \dots, \sigma_k^2$ are the component-specific regressions coefficients and variances, respectively. The common goal of statistical inference in this setting is to estimate the model's parameters. Below we describe two estimation procedures. The first one is the traditional maximum likelihood approach which we will refer to as the 'unweighted MLE' and the second one is a pseudo-maximum likelihood approach which we call the weighted MLE'. We assume that K is unknown, and regard it as a parameter when performing model fitting. The matter of how best to select an appropriate K is considered part of our model fit and model selection

2.2 Maximum Likelihood Approach Via EM- algorithm

In this case, estimation of the parameters is typically performed through the maximum

likelihood approach. The log-likelihood function is given by

$$\ell(\psi) = \sum_{i=1}^{n} log\{\sum_{k=1}^{k} \alpha_k \, \emptyset(x_i;\, x_i\beta_k, \sigma_k^2)\} \qquad (2.2)$$

Due to the inconvenient form of (ψ) in the equation (2.2), the expectation-maximization algorithm [10]. which is based on a complete-data log-likelihood function, is employed. The complete-data setup is given i.i.d samples from $g(y_i;\, x_i, \psi)$; we define the latent variable Z_{ik} such that

$$Z_{ik} = \begin{cases} 1 \text{ if the } ith\ observation \in kth\ component \\ 0 \quad otherwise \end{cases}$$

Then, we can write the complete-data log-likelihood function as

$$l_c(\psi) = \sum_{i=1}^{n} \sum_{k=1}^{k} I(Z_{ik=1})\{log\ \alpha_k + log\ \emptyset(y_i; x_i\beta_k, \sigma_k^2)\} \qquad (2.3)$$

The EM algorithm is an iterative procedure of two steps, the Expectation (E) step, and

the Maximization (M) step. At the E-step, we calculate the conditional expectation of

the complete-data log-likelihood function given the observed data, E $(l_c(\psi) \mid$ y,x), which simplifies to

$$\mathrm{E}\big(I(Z_{ik=1}) \mid y_i, x_i, \psi^{(t-1)}\big) = P_r\big(Z_{ik=1} \mid y_i, x_i, \psi^{(t-1)}\big) \qquad (2.4)$$

This posterior probability will be denoted as τ_{ik} . The expression of τ_{ik} at the (t)th iteration of the E-step is given by

$$\tau_{ik}^{(t)} = \frac{\alpha_k^{(t-1)}\, \emptyset(y_i; x_i\beta_k^{(t-1)}, \sigma_k^{2(t-1)})}{\sum_{k`}^{k} \alpha_{k`}^{(t-1)} \emptyset(y_i; x_i\beta_{k`}^{(t-1)}, \sigma_{k`}^{2(t-1)})} \qquad (2.5)$$

At the M-step of the (t)th iteration, we maximize the conditional expectation of the complete data log-likelihood

function commonly known as the Q-function given by

$$Q(\psi;\psi^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{k} \tau_{ik}\left\{log\ \alpha_k^{(t-1)} + log\ \emptyset(x_i; x_i\beta_k, \sigma_k^2)\right\} \qquad (2.6)$$

The two steps are iterated until a predetermined convergence criterion is met. For a simple linear regression model, $y_i = \beta_{ko} + \beta_{k1}x_i + \epsilon_{ik}$, where $y_i$is the response variable value, x_i denotes a single explanatory variable and $\epsilon_{ik} \sim$ N (0, σ_k^2)

$$Q(\psi;\psi^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{k} \tau_{ik}\left\{log\alpha_k - \frac{n}{2}\log(2\pi\sigma_k^2) - \frac{(\text{yi} - \beta_{ko} - \beta_{k1}x_i)^2}{2\sigma_k^2}\right\} \qquad (2.7)$$

and the closed-form solutions for parameters at (t)th iteration of the M-step are given by

$$\alpha_k^{(t)} = \frac{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}}{\Sigma_{k=1}^{k}\Sigma_{i=1}^{n}\tau_{ik}^{(t)}} \qquad (2.8)$$

$$\beta_{k1}^{(t)} = \frac{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}\Sigma_{i=1}^{n}\tau_{ik}^{(t)}x_i\,y_i - \Sigma_{i=1}^{n}\tau_{ik}^{(t)}x_i\ \Sigma_{i=1}^{n}\tau_{ik}^{(t)}y_i}{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}\ \Sigma_{i=1}^{n}\tau_{ik}^{(t)}x_i^2 - (\Sigma_{i=1}^{n}\tau_{ik}^{(t)}x_i)^2} \qquad (2.9)$$

$$\beta_{k0}^{(t)} = \frac{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}y_i}{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}} - \beta_{k1}^{(t)}\ \frac{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}x_i}{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}} \qquad (2.10), \text{ and}$$

$$\sigma_k^{2(t)} = \frac{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}\ (\text{yi} - \beta_{k0}^{(t)} - \beta_{k1}^{(t)}\ x_i)^2}{\Sigma_{i=1}^{n}\tau_{ik}^{(t)}} \qquad (2.11)$$

Note that Equations 2.9, 2.10, and 2.11 are similar to least squares simple linear regression estimates except that they are weighted by the posterior probability from E-step. The E and M-steps are iterated until the convergence criterion is fulfilled. The criterion used in this paper is the relative difference between consecutive log-likelihood values which is given by

$$\frac{l(\psi^{(t)};x) - l(\psi^{(t-1)};x)}{|l(\psi^{(t-1)};x)|} < 10^{-8} \qquad (2.12)$$

where $\ell\ (\psi)$ is the log-likelihood value evaluated at ψ. We will use the unweighted approach in this work, following most of the notations from [1].

2.3 Model Selection Criteria

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are criteria for selecting models that are used to compare different models and choose the one that fits the best. The main goal of both criteria is to

find a balance between the goodness of fit of the model and its complexity in order to avoid overfitting or underfitting. The main difference between AIC and BIC lies in how they deal with the trade-off between goodness of fit and complexity. AIC is based on the principle of maximum likelihood and penalizes models that have too many parameters compared to the size of the data.

2.3.1 Akaike's Information Criteria (AIC)

One of the most commonly used information criteria is AIC. The idea of AIC [2] is to select the model that minimizes the negative likelihood penalized by the number of parameters as specified in the equation.

$$AIC = -2log(L) + 2k \qquad (2.13)$$

Where L refers to the likelihood under the fitted model k is the number of parameters in the model and n is the observations number. The goal is to find the model with the lowest AIC value since this indicates that the model has a good balance of goodness-of-fit and complexity. The AIC, developed by [2], is a methodology for model selection when multiple models have been fitted to data. It aims to find the best approximating model for the unknown true data-generating process. Its applications draw from the works of [2], [5], and [35]. This model selection is crucial to the objectives of the research, and it was developed by Akaike to screen candidate models in such situations. (Henry de-Graft Acquah, Department of Agricultural Economics and Extension, University of Cape Coast, Cape Coast)

2.3.2 Bayesian information criteria (BIC)

The Bayesian Information Criterion (BIC) is another commonly used information criterion. Unlike the Akaike Information Criterion, the BIC is derived within a Bayesian framework as an estimate of the Bayes factor for two competing models ([32[[15]. BIC is defined as:

$$BIC = -2log(L) + klog(n) \qquad (2.14)$$

where k is the number of parameters in the model, n is the number of data points, and L is the maximum likelihood of the model.

The goal is to find the model with the lowest BIC value since this indicates that the model has the best balance of goodness-of-fit and complexity. In general, BIC tends to penalize models with a large number of parameters more severely than AIC, so it is often used when the goal is to find a parsimonious model. However, both AIC and BIC can be used to compare different models and select the best one for a given dataset [2].

Superficially, BIC differs from AIC only in the second term which now depends on sample size n. Models that minimize the Bayesian Information Criteria are selected. From a Bayesian perspective, BIC is designed to find the most probable model given the data. The performance of the model selection criteria in selecting good models for the observed data is examined using simulation studies. Such a comparison is not straight forward and even its relevance could be questioned, given that the two criteria are based on different theoretical motivations and

objectives. The objective is to identify the model with the lowest BIC value, as it signifies the best trade-off between goodness-of-fit and complexity. BIC is inclined to penalize models with a higher number of parameters more harshly than AIC, making it valuable in identifying a simpler model. Both AIC and BIC can be utilized to compare different models and choose the most suitable one for a specific dataset.

BIC varies from AIC only in the second term, which now relies on the sample size, n. Models that minimize the Bayesian Information Criteria are chosen. From a Bayesian standpoint, BIC is formulated to uncover the most probable model given the data. The effectiveness of the model selection criteria in identifying good models for the observed data is assessed through simulation studies. Comparing the two criteria is not a straightforward task, as they are grounded on different theoretical motivations and goals. Nonetheless, for comparative purposes, both the Akaike Information Criteria and the Bayesian Information Criteria aim to pinpoint good models, even though they diverge in their precise definition of a "good model".

Hence, comparing them is warranted to evaluate how each criterion performs in terms of identifying the correct model or how they behave when both should favor the same model and application For comparative purposes, the Akaike Information Criteria and the Bayesian Information Criteria both aim to identify good models, even though they differ in their exact definition of a "good model." Therefore, comparing them is justified, at least to examine how each criterion performs in terms of identifying the correct model or how they behave when both should prefer the same model [6].

2.3.3 Integrated Completed Likelihood criterion(ICL)

In the realm of mixture model selection,[3] describe two possibly different optimal solutions: the Bayesian Information Criterion (BIC) and the Integrated Completed Likelihood (ICL) serve as critical tools, each with a unique focus. BIC is designed to estimate the number of components in a mixture model by assessing the likelihood of observed data and imposing a penalty for increased model complexity, thus promoting simpler models. ICL, on the other hand, builds upon BIC by adding a term for estimated mean entropy, which discourages the selection of models with overlapping clusters, thereby favoring distinct, well-separated groups. This makes ICL particularly effective for identifying clusters, as it considers the complete-data likelihood, which includes both observed and latent data structures. Consequently, ICL is invaluable for Gaussian mixture models applied through the Expectation Maximization algorithm, aiding in the discernment of discrete data groups and the determination of the appropriate number of clusters.

The formula for the Integrated Completed Likelihood criterion (ICL) can be expressed as:

$$\text{ICL} = log(L) - k * log(n) + 2*\text{entropy} \qquad (2.15)$$

Where L is the likelihood of the model K is the number of parameters in the model n is the number of observations in the dataset. Roughly speaking, the criterion ICL is the criterion BIC penalized by the estimated mean entropy.

In the context of Gaussian mixtures of regression models, entropy is a measure of uncertainty or randomness associated with the model's predictions. Specifically, it quantifies the amount of information needed to describe

the random variability within the clusters formed by the mixture component, for a Gaussian mixture model, the entropy H of a cluster can be defined as: $H = -\sum_{\{i=1\}}^{K} p(x_i) \log p(x_i)$, here, $p(x_i)$ represents the probability of a data point x_i belonging to a particular cluster, and K is the number of clusters or mixture components. he entropy is minimized when the clusters are well-separated and each data point is assigned to one cluster, indicating a model with high certainty in its predictions.

Likelihood Ratio Bootstrap

In addition to the information criteria just mentioned, the choice of the order of a mixture model for a specific component-covariances parameterization can be carried out by likelihood ratio testing (LRT). Suppose we want to test the null hypothesis

H_0 : $K = K_0$ against the alternative H_1: $K = K_1$ for some $K_1 > K_0$; usually, $K_1 = K_0$; + *1* as it is a common procedure to keep adding components sequentially. Let Ψb Gj be the MLE of Ψ calculated under H_j: $K = K_j$ *(for j = 0, 1).* The likelihood ratio test statistic (LRTS) can be written as

$$LRTS = -2\log\{\frac{L(\Psi_b G_0)}{L(\Psi_b G_1)}\} = 2\{l(\Psi_b G_1) - l(\Psi_b G_0)\} \qquad (2.17)$$

where large values of LRTS provide evidence against the null hypothesis. However, standard regularity conditions do not hold for the null distribution of the LRTS to have its usual chi-squared distribution.

([23], Chap. 6). Consequently, LRT significance is often estimated by a resampling approach to produce a p-value. [113] proposed using the bootstrap to obtain the null distribution of the LRTS. The bootstrap procedure is the following:

1. a bootstrap sample x^*_b is generated by simulating from the fitted model under the null hypothesis with K_0 components, i.e. from the GMM distribution with the vector of unknown parameters replaced by MLEs obtained from the original data under H_0;
2. the test statistic $LRTS^*_b$ is computed for the bootstrap sample x^*_b after fitting GMMs with K0 and K1 number of components;

3. steps 1. and 2. are replicated several times, say B = 999, to obtain the bootstrap null distribution of $LRTS^*$

A bootstrap-based approximation to the p-value may then be computed as:

$$p - value = \frac{(1 + \#\{LRTS*_b \geq LRT_{obs}\})}{(B+1)} \qquad (2.18)$$

Where B is the number of bootstrap samples, LRT_{obs} is LRTS computed on the observed data, and $LRTS*_b$ is LRTS computed on the b^{th} bootstrap sample.

3. Simulation Studies

3.1 Simulation Study 1

In this simulation study, we assessed how well different comparison criteria - AIC, BIC, and ICL - performed in various scenarios by estimating their values. We considered four configurations of true regression lines, which we named Model 1, Model 2, Model 3, and Model 4.

We simulated n= {100, 200,500, 1000} observations for Model 1, Model 2, Model 3, and Model 4 respectively. In the first scenario (Model 1), we used a finite mixture of two (K=2) parallel linear regression models which we call non-overlapping. In the second scenario (Model 2), we used a finite mixture of two (K=2) crossed-linear regression models which we call overlapping. The third scenario (Model 3) involved a finite mixture of three (K=3) linear regression models. Finally, in the fourth scenario (Model 4), we used a finite mixture of four (K=4) linear regression models. The vector of true parameters $\psi = (\alpha, \beta, \sigma^2)$ used to generate the mixture are reported in Table 1.

Table 1: True parameter values for Model 1, Model 2, Model 3, and Model 4

ψ	α_1	α_2	α_3	α_4	B_{01}	B_{02}	B_{03}	B_{04}	B_{11}	B_{12}	B_{13}	B_{14}	σ_1^2	σ_2^2	σ_3^2	σ_4^2
Model 1	0.3				-3	3			1	-1			0.1	0.1		
Model 2	0.3				-1	2			1	-2			0.1	0.1		
Model 3	0.3	0.4			-8	0	8		1	2	-1		0.1	0.1	0.1	
Model 4	0.3	0.2	0.3		-12	-6	0	6	-1	1	-1	1	0.1	0.1	0.1	0.1

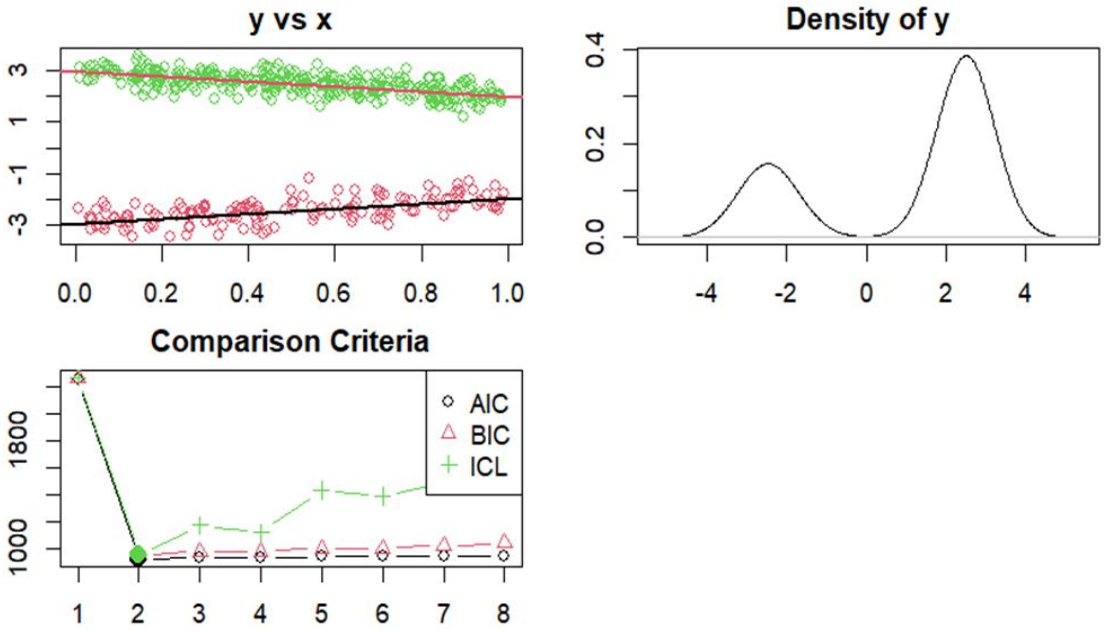

Figure 1: Scatter plot, the density function of sample size n=1000, and model selection criteria (AIC, BIC, and ICL) for the true model of two (K=2) non-overlapping regression lines (Model 1)

Table 2: Model selection criterion (BIC, AIC, and ICL) results for the chosen model١. (The True model is two (K=2) parallel regression lines). The best result is shown in boldface

Sample Size	K	BIC	AIC	ICL
100	1	468.0202	160.2047	468.0202
	2	**175.9907**	**154.7104**	**175.9907**
	3	186.4412	157.7844	197.5254
	4	186.4914	157.8345	192.9125
	5	186.4411	157.7843	197.5258
200	K	BIC	AIC	ICL
	1	921.8023	911.9074	921.8023
	2	**374.2062**	**351.1180**	**374.2062**
	3	374.2062	351.1180	374.2062
	4	391.1745	354.8930	517.2032
	5	374.2062	351.1180	374.2062
500	K	BIC	AIC	ICL
	1	2272.0864	2259.4425	2272.0864
	2	**2272.0864**	**918.7853**	**948.2876**
	3	973.1452	926.7845	1168.1854
	4	973.0063	926.6456	1116.2772
	5	996.8204	933.6013	1430.9980
1000	K	BIC	AIC	ICL
	1	4531.553	4516.830	4531.553
	2	**1637.894**	**1603.540**	**1637.894**
	3	1665.497	1611.511	1793.548
	4	1664.330	1610.344	1830.012
	5	1665.504	1611.519	1793.140

The results presented in Table 1 show the results of Model 1, results demonstrate this sensitivity of AIC to sample size, where it criterion fails to choose the correct model for a sample size of n=100 while BIC and ICL correctly identify the true models across different sample sizes. This highlights the importance of considering the characteristics of each model selection criterion, including their sensitivity to sample size when choosing the appropriate criterion for a given dataset and research question. Researchers should be mindful of the potential impact of sample size on model selection criteria and consider the trade-offs between model complexity and goodness of fit when interpreting the results of model selection analyses. The sensitivity of the Akaike Information Criterion (AIC) to sample size is a well-known characteristic of this model selection criterion. AIC tends to favor more complex models when the sample size is small, potentially leading to overfitting and selecting models that

are not the true underlying model. In contrast, the Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) criteria are generally more robust to sample size variations. These criteria penalize the model.

The following outputs are explained in Table 1, which shows the results of Model 1. The results demonstrate the sensitivity of AIC to sample size, where the criterion fails to choose the correct model for a sample size of n=100, while BIC and ICL correctly identify the true models across different sample sizes. This highlights the importance of considering the characteristics of each model selection criterion, including their sensitivity to sample size when choosing the appropriate criterion for a given dataset and research question. Researchers should be mindful of the potential impact of sample size on model selection criteria and consider the trade-offs between model complexity and goodness of fit when interpreting the results of model selection analyses.

The sensitivity of the Akaike Information Criterion (AIC) to sample size is a well-known characteristic of this model selection criterion. AIC tends to favor more complex models when the sample size is small, potentially leading to overfitting and selecting models that are not the true underlying model. In contrast, the Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) criteria are generally more robust to sample size variations. These criteria penalize model complexity more heavily than AIC, making them less prone to selecting overly complex models when the sample size is small.

Figure 1a, shows a scatter plot, the density function of sample sizes n=200, and model selection criteria (AIC, BIC, and ICL) for the true model of two (K=2) non-overlapping regression lines (Model 1). Through statistical analysis of data using information criteria methods to choose the best model, the BIC criterion is considered the best criterion at K=2 and sample size N=100. At the AIC criterion, we find that the best K=6, and the sample size N=100 are affected by the AIC criterion. It is also considered a very convenient standard, and the best standard is at K=2, and the sample size is n=100.

Figure 1b shows a scatter plot, the density function of sample size n=500, and model selection criteria (AIC, BIC, and ICL) for the true model of two (K=2) non-overlapping regression lines (Model 1). Through statistical analysis of data using information criteria methods to choose the best model, the BIC criterion is considered the best criterion at K=2, and sample size n=500.

Figure 1c shows a scatter plot, the density function of sample size n=1000, and model selection criteria (AIC, BIC, and ICL) for the true model of two (K=2) non-overlapping regression lines (Model 1). Through statistical analysis of data using information criteria methods to choose the best model, the BIC criterion is considered the best criterion at K=2, and sample size n=1000.

Figure 1d shows a scatter plot, the density function of sample size n=2000, and model selection criteria (AIC, BIC, and ICL) for the true model of two (K=2) non-overlapping regression lines (Model 1). Through statistical analysis of data using information criteria methods to choose the best model, the BIC criterion is considered the best criterion at K=2, and sample size n=2000.

Table 2: Model selection criterion (BIC, AIC, and ICL) results for the chosen model2. The True model is two (K=2) crossed regression lines). The best result is shown in boldface

Sample Size	K	BIC	AIC	ICL
100	1	468.0202	160.2047	468.0202
	2	**175.9907**	**154.7104**	**175.9907**
	3	186.4412	157.7844	197.5254
	4	186.4914	157.8345	192.9125
	5	186.4411	157.7843	197.5258
200	K	BIC	AIC	ICL
	1	921.8023	911.9074	921.8023
	2	**374.2062**	**351.1180**	**374.2062**
	3	374.2062	351.1180	374.2062
	4	391.1745	354.8930	517.2032
	5	374.2062	351.1180	374.2062
500	K	BIC	AIC	ICL
	1	2272.0864	2259.4425	2272.0864
	2	**2272.0864**	**918.7853**	**948.2876**
	3	973.1452	926.7845	1168.1854
	4	973.0063	926.6456	1116.2772
	5	996.8204	933.6013	1430.9980
1000	K	BIC	AIC	ICL
	1	4531.553	4516.830	4531.553
	2	**1637.894**	**1603.540**	**1637.894**
	3	1665.497	1611.511	1793.548
	4	1664.330	1610.344	1830.012
	5	1665.504	1611.519	1793.140

The results presented in Table 2 show the results of Model 2, results demonstrate this sensitivity of AIC to sample size, where it criterion fails to choose the correct model for a sample size of n=°00, n = 1000 when there are two crossed regression lines (k = 2) while BIC and ICL correctly identify the true models across different sample sizes. This highlights the importance of considering the characteristics of each model selection criterion, including their sensitivity to sample size when choosing the appropriate criterion for a given dataset and research question. Researchers should be mindful of the potential impact of sample size on model selection criteria and consider the trade-offs between model complexity and goodness of fit when interpreting the results of model selection analyses. The sensitivity of the Akaike Information Criterion (AIC) to sample size is a well-known characteristic of this model selection criterion. AIC tends to favor more complex models when the sample size is big the Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) criteria are generally more robust to sample size variations. These criteria penalize model complexity more heavily than AIC.

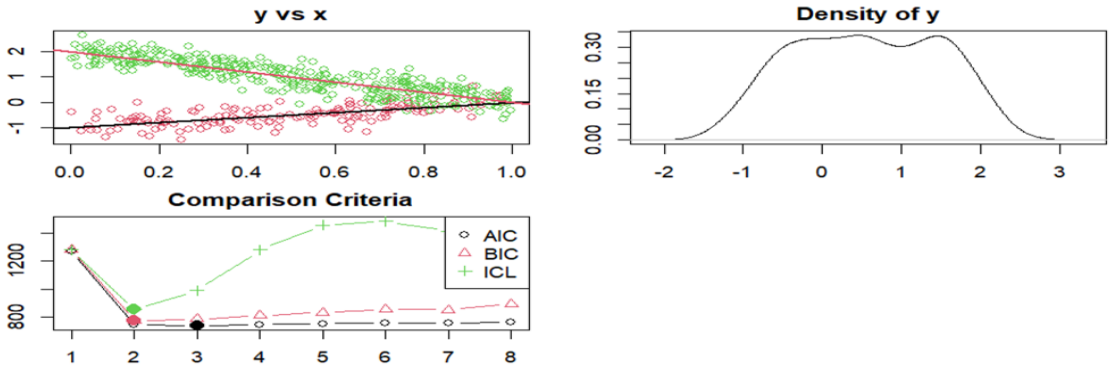

Figure 2: Scatter plot, the density function of sample size n=1000, and model selection criteria (AIC, BIC, and ICL) for the true model of two (K=2) overlapping regression lines (Model 2)

Table 3: Model selection criterion (BIC, AIC, and ICL) results for the chosen model. The True model is three (K=3) parallel regression lines. The best result is shown in boldface

Sample Size	K	BIC	AIC	ICL
100	1	648.6484	640.8329	648.6484
	2	657.1042	638.8680	704.3663
	3	**326.2399**	**297.5830**	**326.2399**
	4	591.8793	552.8018	615.6425
	5	326.2399	297.5830	326.2399
	6	337.2501	298.1725	348.2897
	7	337.2506	298.1730	348.3232
	8	337.2505	298.1729	348.3193
	K	BIC	AIC	ICL
200	1	1291.4399	1281.5450	1291.4399
	2	1017.0317	993.9435	1017.1534
	3	**590.1373**	**553.8558**	**590.1373**
	4	590.1373	553.8558	590.1373
	5	590.1373	553.8558	590.1373
	K	BIC	AIC	ICL
500	1	3203.355	3190.712	3203.355
	2	2566.235	2536.733	2566.698
	3	**1366.406**	**1320.045**	**1366.406**
	4	1366.406	1320.045	1366.406
	5	1366.406	1320.045	1366.406
	K	BIC	AIC	ICL
1000	1	6396.805	6382.082	6396.805
	2	5046.494	5012.139	5048.238
	3	**2797.032**	**2743.046**	**2797.032**
	4	2823.264	2749.648	3098.962
	5	2823.266	2749.650	3102.204

The results presented in Table 3 show the results of Model 3. The results for the model selection criteria AIC, BIC, and ICL across different sample sizes when the true model is Model 3. Changes in sample sizes do not affect the requirements, with a sample size of, n=100, n=200, n=500, and n=1000, the AIC, BIC, and ICL criteria correctly identify the true models for each sample size.

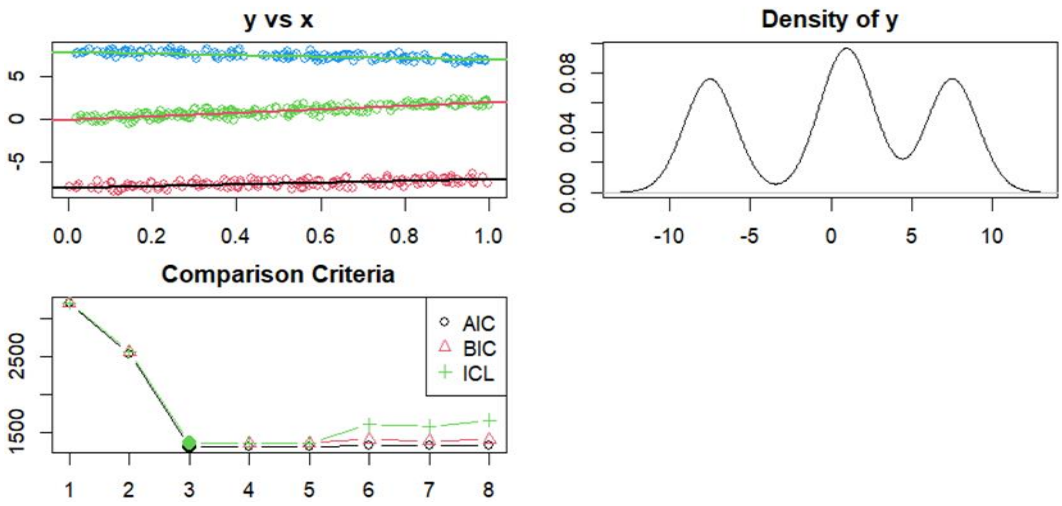

Figure 3: Scatter plot, the density function of sample size n=1000, and model selection criteria (AIC, BIC, and ICL) for the true model of three (K=3),(Model 3)

Table 4: Model selection criterion (BIC, AIC, and ICL) results for the chosen model (The True model is Four(K=4) parallel regression lines). The best result is shown in boldface

Sample Size	K	BIC	AIC	ICL
100	1	684.4862	676.6707	684.4862
	2	574.8336	556.5974	575.6727
	3	504.6746	476.0177	504.7525
	4	**395.0281**	355.9505	**395.0281**
	5	413.4477	363.9495	438.0130
	K	BIC	AIC	ICL
200	1	1357.6276	1347.7327	1357.6276
	2	1135.0639	1111.9756	1136.8952
	3	979.1034	942.8219	979.2235
	4	**735.7762**	**686.3015**	**735.7762**
	5	749.1888	686.5207	783.3479
	K	BIC	AIC	ICL
500	1	3373.512	3360.868	3373.512
	2	2796.424	2766.922	2800.059
	3	2409.976	2363.615	2410.279
	4	**1746.468**	**1683.249**	**1746.468**
	5	1746.468	1683.249	1746.468
	K	BIC	AIC	ICL
1000	1	6728.628	6713.904	6728.628
	2	6756.206	6721.852	8070.921
	3	4683.624	4629.639	4684.568
	4	**3289.217**	**3215.600**	**3289.217**
	5	3289.217	3215.600	3289.217

The results presented in Table 4 show the results of Model 4, results demonstrate this sensitivity of AIC to sample size, where it criterion fails to choose the correct model for a sample size of n=100, (k = 4) while BIC and ICL correctly identify the true models across different sample sizes. This highlights the importance of considering the

characteristics of each model selection criterion, including their sensitivity to sample size when choosing the appropriate criterion for a given dataset and research question. Researchers should be mindful of the potential impact of sample size on model selection criteria and consider the trade-offs between model complexity and goodness of fit when interpreting the results of model selection analyses. The sensitivity of the Akaike Information Criterion (AIC) to sample size is a well-known characteristic of this model selection criterion. AIC tends to favor more complex models when the sample size is small, potentially leading to overfitting and selecting models that are not the true underlying model. In contrast, the Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) criteria are generally more robust to sample size variations. These criteria penalize model complexity more heavily than AIC, making them less prone to selecting overly complex models when the sample size is small.

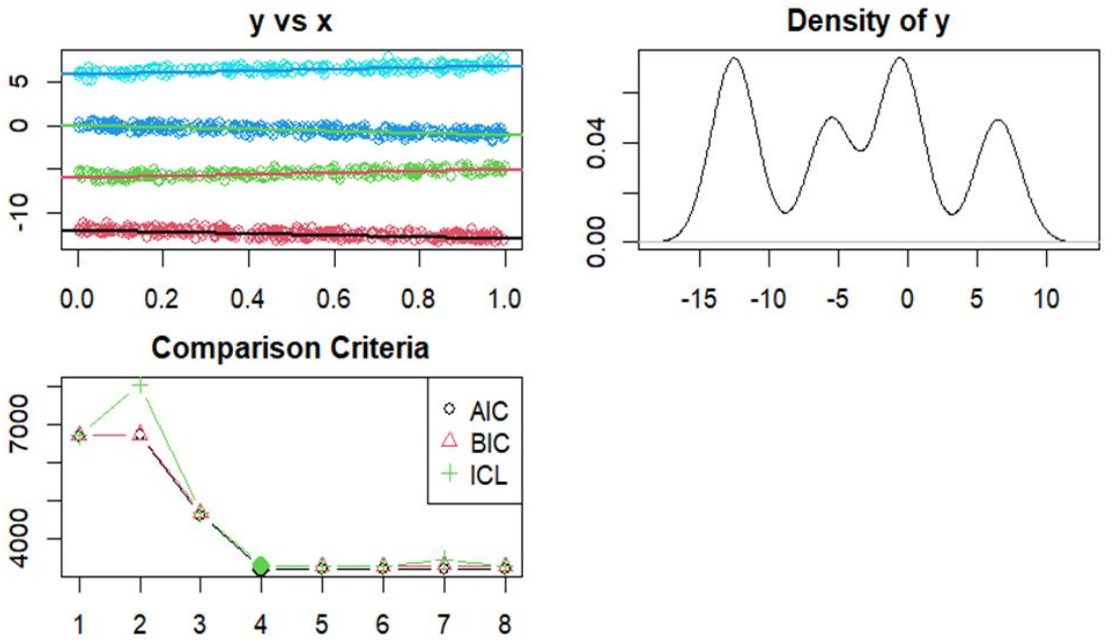

Figure 4: Scatter plot, the density function of sample size n=1000, and model selection criteria (AIC, BIC, and ICL) for the true model of Four(K=4) parallel regression lines

3.2 Simulation Study 2

In this simulation study, we aimed to evaluate the LRT bootstrap criterion by considering four different configurations of true regression lines. These configurations were also used in Simulation Study 1 and were referred to as Model 1, Model 3, and Model 4.

For each of these models, we simulated samples with 1000 observations. In the first scenario (Model 1), we used a finite mixture of two (K=2) linear regression models. The third scenario (Model 3) involved a finite mixture of three (K=3) linear regression models. Lastly, in the fourth scenario (Model 4), we used a finite mixture of four (K=4) linear regression models. The vector of true parameters used to generate the mixture is reported in Table 1.

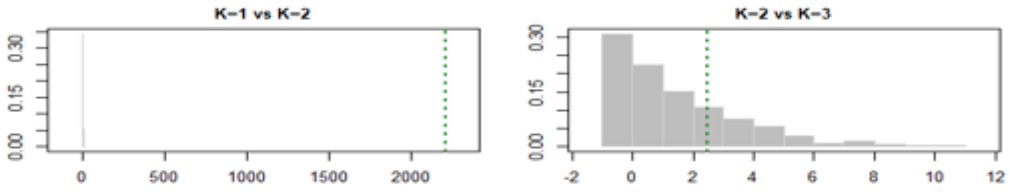

Figure 5: Histograms of LRTS bootstrap distributions for testing the number of mixture components of Model 1 (K=2). The dotted vertical lines refer to the sample values of LRTS

The Bootstrap Likelihood Ratio Test (BLRT; [17,23] is the third test used to compare fitted models with *K* components to similarly specified k-1 class models. The test records the difference in *-2LL* and then simulates data based on the parameter estimates from the *K-1* class model. Using the simulated data, the *K-1* and k components models are estimated to generate a sampling distribution for the difference in *-2LL* under the null hypothesis. The recorded difference in *-2LL* (from the empirical data) is then compared to the sampling distribution to estimate the p-value. Suppose the p-value is less than alpha (i.e., 0.05). In that case, the k- k-components model is preferred, and if the p-value is greater than alpha, then the fit of the two models is not statistically significant, and the *K*-components model is preferred.

In our analysis, we conducted a test using Model 1 with K=2 to compare the hypothetical number of components with K=1. We found that the associated P-value was 0.001, less than the significance level of 0.05, as per Table 1. This indicates that the difference we observed is statistically significant. We then tested to compare the fitted model with K=2 and K=3. In this case, we found that the associated p-value was 0.256 which is higher than the significance level of 0.05. This means that we cannot reject the null hypothesis, and the true number of components for the fitted Model 1 is K=2. In Table 5, we have presented the detailed results of our test. Additionally, Figure 5 supports our previous findings, providing us with a visual representation of the comparison between the fitted model with K=2 and K=3. Overall, our analysis suggests that Model 1 with K=2 is the best fit for our data.

Table 5: Results of Bootstrap Likelihood Ratio Test of Model 1 (K=2). The results reported in this table are obtained using 1000 Monte Carlo samples

Hypothesis	p-value
H_0: K=1 vs H_1: K=2	0.001
H0: K=2 vs H1:K=3	0.256

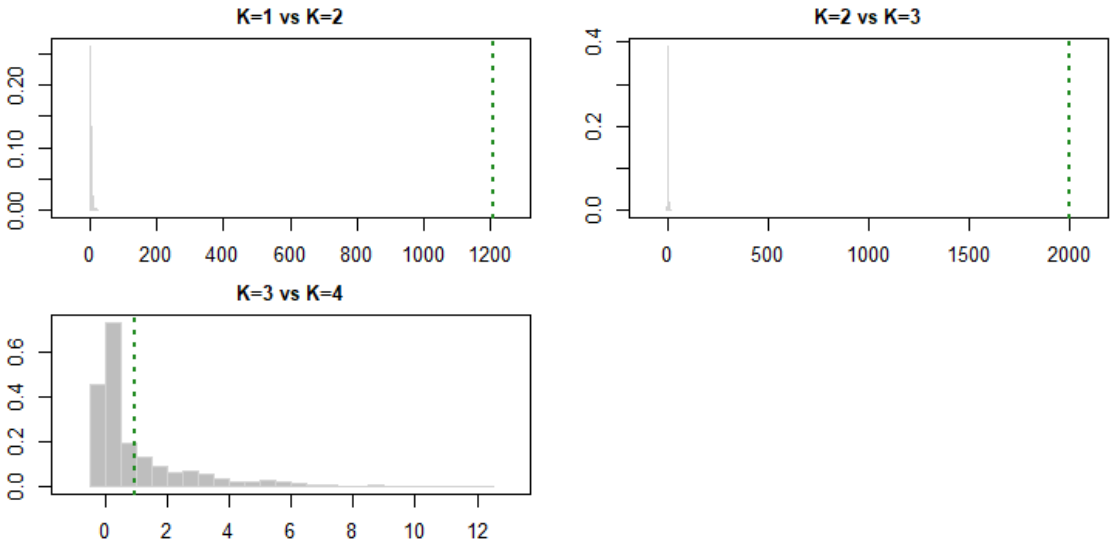

Figure 6: Histograms of LRTS bootstrap distributions for testing the number of mixture components of Model 3 (K=3). The dotted vertical lines refer to the sample values of LRTS

We tested Model 3 with K=3 and compared it with different numbers of components (K=1,2,4). The aim was to determine the hypothetical number of components that would work best. The results are presented in Table 6. We could not reject the null hypothesis when we compared the fitted model with K=3 and K =4, since the associated p-value was 0.317, which is higher than the level of significance at 0.05. Therefore, we cannot reject the null hypothesis, which means that the true number of components for the fitted Model 1 is K=3. Figure 6 further supports this conclusion.

Table 6: Results of Bootstrap Likelihood Ratio Test of Model 3 (K=3). The results reported in this table are obtained using 1000 Monte Carlo samples

Hypothesis	p-value
H_0: K=1 vs H_1: K=2	0.001
H0: K=2 vs H1:K=3	0.001
H_0: K=3 vs H_1: K=4	0.317

We conducted a thorough evaluation of Model 4 by varying the number of components (K=1,2,3,5) to determine the optimal hypothetical number. To achieve this, we employed various statistical techniques such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the Variational Bayesian Approximation (VBA). Our findings, presented in Table 7, reveal that the optimal hypothetical number of components for Model 4 is K=4. To verify our conclusion, we compared the fitted models with K=4 and K=5. Based on our analysis, we discovered that the associated p-value was 0.911, which is higher than the significance level of 0.05. Therefore,

we cannot reject the null hypothesis, indicating that the true number of components for the fitted Model 1 is K=4. Furthermore, we confirmed our findings through Figure 7, which illustrates the proportion of variance explained by each component. The figure clearly shows that the difference between K=4 and K=5 is minimal, and K=4 explains most of the variance. Hence, our conclusion is supported by both statistical analysis and visual representation.

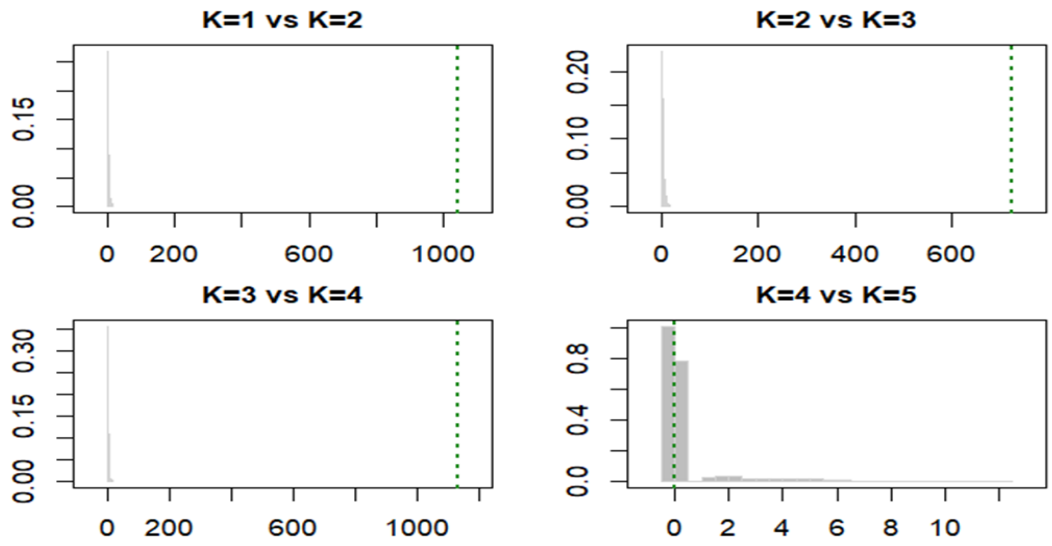

Figure 7: Histograms of LRTS bootstrap distributions for testing the number of mixture components of Model 4 (K=4). The dotted vertical lines refer to the sample values of LRTS

Table 7: Results of Bootstrap Likelihood Ratio Test of Model 4 (K=4). The results reported in this table are obtained using 1000 Monte Carlo samples

Hypothesis	p-value
H_0: K=1 vs H_1: K=2	0.001
H0: K=2 vs H1:K=3	0.001
H_0: K=3 vs H_1: K=4	0.001
H_0: K=4 vs H1: K=5	0.911

3.3 Summary of results

In Simulation Study 1, assessing the performance of model selection criteria (AIC, BIC, and ICL) across different scenarios with varying numbers of true regression lines (Model 1, Model 2, Model 3, and Model 4) and sample sizes (n=100, n=200, n=500, and n=1000), several key findings emerged: The study highlights the sensitivity of

model selection criteria, such as AIC, BIC, and ICL, to sample size variations in different models (Model 1 to Model 4). AIC tended to favor more complex models, potentially leading to overfitting, especially with small sample sizes, while BIC and ICL demonstrated greater robustness by penalizing model complexity more effectively. Across models, AIC showed inconsistencies in selecting the correct model with varying sample sizes, whereas BIC and ICL consistently identified the true models. Researchers are advised to carefully evaluate the characteristics of each criterion, particularly their sensitivity to sample size, to make informed choices when selecting the most appropriate criterion for model selection, emphasizing the balance between model complexity and goodness of fit for accurate interpretation of results.In Simulation Study 2, the Bootstrap Likelihood Ratio Test (BLRT) was employed to assess various configurations of true regression lines in models 1, 3, and 4 with different numbers of components (K=2, K=3, K=4). Findings revealed that for Model 1, K=2 was statistically favored over K=1 with a p-value of 0.001, and K=2 was deemed optimal compared to K=3 with a p-value of 0.256. Model 3's true component number was determined to be K=3 based on a p-value of 0.317 when compared to K =4. For Model 4, K=4 was identified as the optimal number of components through evaluations with AIC, BIC, and VBA, with a p-value of 0.911 supporting K=4 over K=5. Statistical analyses, bolstered by visual representations, validated the chosen component numbers for each model (K=2 for Model 1, K=3 for Model 3, and K=4 for Model 4) as best explaining the data variance. The BLRT and other statistical methods provided a robust framework for optimal component selection, emphasizing the crucial role of thorough model evaluation and selection in statistical analyses.

4. Summary and Conclusion

The article underscores the importance of carefully choosing the right model selection criterion based on the specific characteristics of the dataset and research goals. It points out that the Akaike Information Criterion (AIC) can be influenced by sample size, potentially favoring more complex models, especially when dealing with smaller sample sizes. On the other hand, the Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) criteria are noted for their ability to handle sample size variations better by effectively penalizing model complexity. The study also explores how sample size impacts model selection criteria, noting that AIC's sensitivity to sample size changes can lead to overfitting and the selection of models that may not accurately represent the true underlying model. Additionally, the article discusses the use of the Bootstrap Likelihood Ratio Test (BLRT) as a statistical tool for comparing models with different numbers of components, aiding in determining the most suitable model complexity. By utilizing statistical techniques like AIC, BIC, and ICL, the study identifies the optimal number of components for each model configuration through p-value analysis. Visual aids, such as scatter plots and density functions, are employed to complement the statistical findings, offering further insights into the performance and complexity of various models. Overall, the article stresses the importance of thoughtful model selection, taking into account sample size considerations, using statistical tests like BLRT, and incorporating visual representations to make well-informed decisions about the optimal model complexity in regression analysis.

References

[1] Abdalla, A., & Michael, S. (2019). A finite mixture of regression models for a stratified sample. *Journal of Statistical Computation and Simulation*, *89*(14), 2782-2800

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), Second International Symposium on Information Theory (pp. 267-281). Akademiai Kiado

[3] Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of computational and graphical statistics*, *19*(2), 332-353.

[4] 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.

[5] Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345–370.

[6] Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research, 33(2), 261-304.

[7] Celeux, G. (2015). On the different ways to compute the integrated completed likelihood criterion.

[8] Davison, A. and Hinkley, D. (1997) *Bootstrap Methods and Their Applications*. Cambridge University Press.

[9] De Veaux, R. D. (1989). A review of mixture models. Journal of the American Statistical Association, 84(406), 446-454.

[10] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1-22.

[11] Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal, 41(8), 578-588.

[12] Frühwirth-Schnatter, S. (2006). Finite mixture and Markov switching models. Springer.

[13] Grimm, K. J., An, Y., McArdle, J. J., Zonderman, A. B., & Resnick, S. M. (2017). Recent changes leading to subsequent changes: Extensions of multivariate latent difference score models. Structural Equation Modeling: A Multidisciplinary Journal, 24(3), 321-337.

[14] Jones, M. C., & McLachlan, G. J. (1992). Estimation in a family of linear regression models with mixtures of regressions. Journal of the American Statistical Association, 87(420), 958-967.

[15] Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

[16] Lindsay, B. G. (1995). Mixture models: Theory, geometry, and applications. IMS Lecture Notes-Monograph Series, 34, 1-70.

[17] McLachlan G.J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied Statistics, 36, 318-324.

[18] McLachlan, G. J., & Peel, D. (2000). Finite mixture models. New York, NY: Wiley.

[19] McLachlan, G. J., & Rathnayake, S. (2019). Finite mixture models. New York, NY: Wiley.

[20] McLachlan, G., & Basford, K. E. (1988). Mixture models: Inference and applications to clustering. New York, NY: Marcel Dekker.

[21] McLachlan, G., & Peel, D. (2000). Finite mixture models. New York, NY: Wiley.

[22] McLachlan, G., Lee, S. X., & Rathnayake, S. (2019). Finite mixture models. New York, NY: Wiley.

[23] McLachlan, G.J. and Peel, D. (2000) Finite Mixture Models. Wiley.

[24] McLachlan, G.J. and Rathnayake, S. (2014) On the number of components in a Gaussian mixture model. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,4(5), pp. 341-355.

[25] McNicholas, P. D. (2016). Mixture model-based classification. New York, NY: Chapman and Hall/CRC.

[26] Melnykov, V., Maitra, R., Ram, S., & Fu, M. (2015). Handbook of mixture analysis. New York, NY: CRC Press.

[27] Pearson, K. (1894). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London. Series A, 185, 71-110.

[28] Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and computing*, *10*, 339-348.

[29] Quandt, R. E., & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. Journal of the American Statistical Association, 73(364), 730-738.

[30] Rabe-Hesketh, S., & Skrondal, A. (2004). Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. New York, NY: CRC Press.

[31] Schlattmann, P. (2009). Medical applications of finite mixture models. New York, NY: Springer.

[32] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.

[33] Wedel, M., & Kamakura, W. A. (2012). Market segmentation: Conceptual and methodological foundations. New York, NY: Springer.

[34] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. The Annals of Statistics, 11(1), 95-103.

[35] Zucchini, W. (2000). An Introduction to Model Selection. *Journal of Mathematical Psychology*, 44, 41–61.