مجلة المثنى للعلوم الادارية والاقتصادية

Modified information criteria for selecting a finite mixture model

Safaa K. Kadhem[*a] ، Hajem A. Daham[b]
College of Administration and Economics, Al Muthanna University, Iraq

Abstract

Recently, a paper published by Celeux et al. (2006) presented several forms for the deviation information criterion (DIC) for mixture models, each version is depended on the kind of probability function. However, no reliable version was adopted for those models. As an idea inspired by Brooks (2002, p. 617), we develop, in this paper, Bayesian deviations plugging into two known criteria: the Akaike information criterion (AIC) and Bayesian information criterion (BIC) for choosing best mix model. Due to unavailability the closed-form of the perceived likelihood of those models, we propose an algorithm for estimating the observed likelihood for mixture models via an Markov chain Monte Carlo (MCMC) approach. It is shown via recreation researches and examples include actual information applications that proposed AIC and BIC perform well.

Article information

Search history
The receipt:23/2/2020
Modification date:26/4/2020
Publishing acceptance:28/4/2020
Available online:30/6/2020

key words:
Finite mixture models
observed likelihood
Gibbs sampler
model selection

Introduction

Mixture models are type of latent variable models that have been developed to be important tool to accommodate the unobserved heterogeneity in data. A finite mixture model (FMM) is mostly employed in case a reflection is attributed to one of K groups (constituents), which has distinctive attributes and may be defined by diverse likelihood dispersals. In other words, these models are a weighted average of a finite number of distributions (mingling constituents). FMMs could be a finite combination of dispersals like Gaussian or Poisson dispersals (McLachlan & Peel, 2000; Fruhwirth-Schnatter, 2006). An important issue of the literature about the mix models is to identify the number of mixture constituents, which considers a specific part related with the process of the model estimation. Several applications of latent class models, the number of situation (components) of observed data is usually determined a significant, either by nature of problem or based on the technical perception (Scott, 2002; Celeux et al., 2006). However, determining of components on the basis of the perceived information denotes which increasing number of situations, and also the parameters, often improves fit of the model.

On other side, several applications could suppose that the number of implied situations is unknown and consider that number like arbitrary factor to be assessed together with the other model criterion. For instance, the fluctuating jump MCMC method was outlined by Richardson & Green (1997) to assess the constituents number independent combinations. However, this method is frequently computationally rigorous. Moreover, it needs attention while planning moves to confirm that Markov

* Corresponding author : E-mail addresses : safaakarem76@gmail.com.
2020 AL – Muthanna University . DOI:10.52113/6/2020-10-2/137-153

chains blend thoroughly both inside model gaps (similar situations, diverse factors) and amid model gaps (dissimilar situations). Moreover, it could face several approximating complexities. Additionally, that method has challenges on the prior selection for the number of latent states K (Fan & Sisson, 2011).

Two most common criteria, which are assumed to create stable model fit and parsimony, that are the Akaike Information Criterion (AIC; Akaike (1973)) and the Bayesian Information Criterion (BIC; Schwarz (1978)). Under frequentist context, the AIC and BIC are employed by Zucchini & MacDonald (2009) in many applications to choose the best FMM. However, these two criteria can lead to under-fitting or over- fitting problems due to irregular behavior the behavior of the likelihood function (Zucchini & MacDonald, 2009). Moreover, computing the over standards does not regard the ambiguity related to these measures when assessing models but are based only on point estimates of parameters. From this, the advantage of Bayesian inference comes to regard for different sources of ambiguity. Based on the principle, Brooks (2002, p. 617) referred to, through his remarks on the paper printed by Spiegel- halter et al. (2002), improving standards like AIC and BIC in a Bayesian framework. He illustrated the possibility of use the likelihood or deviation evaluated at an MCMC run and plug it into the AIC and BIC. This proposal was also proposed for different models by Ntzoufras (2009, pp.426-428), Carlin & Louis (2009, p.211) Congdon (2014, p.36) and Kadhem, Hewson & Kaimi (2018).

On this basis, the paper's goal is to produce original versions of AIC and BIC for FMMs that are evaluated in the Bayesian principle. Simply, the new forms of AIC and BIC are computed stand on the perceived-data probability estimated via subsequent draws. We implement a Monte Carlo simulation study to investigate all above proposed criteria. In this paper, it is not intended to make strong preceding expectations regarding the number of situations, but we deduce the model according to a definite number of components K. Consequently, that article considers the subject of model choice in the FMM perspective by supposing series of values of K between 1 and a pre-specified extreme value K_{max}, given an observed example of data. Consequently, we determine an appropriate number of components for a FMM, suitable for some models with maximizing number of components then choose the best model based on assumed model selection standards.

The paper is divided in the following manner: Section 2 introduces literature review for the estimation and selection of finite mixture models. Section 3 presents structure of the model. The model selection method is given in Section 4. Section 5 presents the simulation outcomes and two actual information application data example is introduced in. Section 6 introduces several deductions.

1. Bayesian Analysis of FMMs

Let $y = (y_1, y_2, \dots, y_T)$ refer to an example of perceived data of size T, the probability density function(pdf) of a combination model could be well-defined as a mixture of K constituent pdf:

$$P_r(y|\theta) = \sum_{k=1}^{k} \pi_k f_k(y|\theta_k) \qquad (2-1).$$

where $f_k(y|\theta_k)$ refers to the pdf of kth constituent, π_k is the population k weight that $0 \le \pi_k \le 1$, and $\sum_{k=1}^{K} \pi_k f_k(y|\theta_k) = 1, \Theta = (\pi; \theta) = (\pi_1 . \pi_2 . \dots . \pi_k; \theta_1 . \theta_2 . \dots . \theta_{1k})$ refers to a group of

unidentified weight and indexes of combination model. The key notion of combination model is that the remarks y are created by k distinctive arbitrary procedures so every procedure is shown by the density $f_k(y|\theta_k)$ and π_k rep- resents the equivalent ratio of remarks of this procedure. For instance, take FMM where $P_r(y|\Theta)$ is instituted from densities that are all Normal or Poisson distribution. Given an identically independent data (iid), $y = (y_1 . y_2 . \dots . y_T)$. produced from a k- constituent combination model shown in Equation (2-1), the probability function of those remarks, supposing that y_t is independently disseminated may be shown as

$$\Pr(y|\Theta) = L(y;\Theta) = \prod_{t=1}^{T}\{\pi_1 f_1(y_t|\theta_k) + \pi_2 f_2(y_t|\theta_2) + \cdots + \pi_k f_k(y_t|\theta_k)\}$$
$$= \prod_{t=1}^{T}\sum_{k=1}^{K} \pi_k f_k(y_t|\theta_k) . \qquad (2-2)$$

In the FMM in Equation (2-2), the unidentified parameter direction $\Theta = (\pi, \theta)$ requires assessed. To gain the subsequent distribution of Θ, it is required to integrate the data-dependent probability function $L(y;\Theta)$ of the combination model and the preceding dispersal of the unidentified indexes ;θ and π. The subsequent distribution may be shown as:

$$\Pr(y|\Theta) = L(\theta, \pi; y)\Pr(\theta)\Pr(\pi), \qquad (2\text{-}3)$$

where $L(\theta, \pi; y) = \prod_{t=1}^{T}\sum_{k=1}^{K} \pi_k f_k(y_t|\theta_k)$ is the probability, $\Pr(\theta)$ and $\Pr(\pi)$ represent the preceding dissemination of θ and π respectively. An effective way for simplifying the sampling from the subsequent dissemination is the date increase method suggested by Tanner & Wong (1987). This method depended on sampling from the whole data of subsequent dissemination $\Pr(\Theta, z|y)$ rather than $\Pr(\Theta|y)$ by suggesting secondary variables, named z, also called underlying indicator variables. If y and z are known, the analysis will be clearer.

It is proposed that there are distinct underlying indexes, z_t ;t = 1,2,...,T , related to every remark of the vector $y = (y_1 . y_2 . \dots . y_T)$. Since these indexes in actual life are unidentified parameters, the implication of a combination model needs approximating two unidentified measures: the constituent indicators, z, and the constituent parameters, $\Theta = (\pi, \theta)$. In the Bayesian standpoint, to acquire these quantities, they could be sampled from the next whole date posterior:

$$\Pr(z, \pi, \theta|y) = L_c(\theta, \pi; y, z)P_r(\theta)\Pr(\pi), \qquad (2-4)$$

where $L_c(z,\pi; y, z)$ is the whole data probability of a finite mix model, $\Pr(\theta)$ and $\Pr(\pi)$ are independent previous dissemination of the parameter θ and of the constituents weights π correspondingly. The whole-data probability is shown as:

$$L_c(\theta, \pi; y, z) = \prod_{t=1}^{T} \pi_{z_t} f(y_t|\theta_{z_t}) = \prod_{k=1}^{K} \prod_{t:z_t=k}^{T} \pi_k f(y_t|\theta_k),$$

$$= \prod_{k=1}^{K} \pi_k^{\sum_{t=1}^{T} I(z_t=k)} \prod_{t:z_t=k}^{T} f(y_t|\theta_k). \quad (2 - 5)$$

In order to complete the Bayesian requirement of the model, it is required to identify priors for the unidentified model indexes: π and θ. The prior on the constituent weights is denoted by a Dirichlet distribution as

$$\Pr(\pi) = \prod_{k=1}^{K} \pi_k \alpha \prod_{k=1}^{K} \pi_k^{\delta_{k-1}} = \text{Dirichlet}(\delta_1, \delta_2, \ldots, \delta_k), \qquad (2 - 6)$$

where $\delta k,\ k = 1,2,\ldots,K$ are positive ($\delta k > 0$) hyper-parameters of the Dirichlet distribution. The prior on the component-specific parameter, θ, depended on the shape of the parametric dissemination presumed for remarks, y. Generally speaking, to denote the prior on the component-specific parameter, θ, it should be written in the following form;

$$\theta \sim P_r(\theta|\varphi), \qquad (2 - 7)$$

where ϕ denotes a group of the hyper-parameters controlling the form of the prior dissemination of θ. Common MCMC methods are used. Gibbs sampler are used (Geman & Geman, 1984) to pretend from the full provisional succeeding disseminations of the FMM. The following dissemination in Equation (3) includes three full conditional distributions that are written as

$$\begin{aligned} z &\sim \Pr(z|y, \pi, \theta), \\ \pi &\sim \Pr(\pi|y, z), \qquad (2 - 8) \\ \theta &\sim \Pr(\theta|y, z). \end{aligned}$$

It is not that so difficult to achieve the Gibbs sampler to sample from these disseminations. In Bayesian inference for FMMs, the mingling ratio $\{\pi_1, \pi_2, \ldots, \pi_k\}$ could be seen as the preceding dissemination, which one remark relates to sub-population k. If the observations are, y_t , the complete provisional succeeding distribution of z_t could be gained as

$$\Pr(z_t = k|y_t, \pi, \theta)\ \alpha\ \pi_k \Pr(y_t|\theta_t) = \frac{\pi_k \Pr(y_t|\theta_t)}{\sum_{l=1}^{K} \pi_l \Pr(y_t|\theta_l)}\ . \qquad (2-9)$$

From Equation (2-9), the marginal distribution of the z_t is a multinomial distribution

$$z_t \sim \text{Multinomial}\{\Pr(z_t = 1), \Pr(z_t = 2), \dots, \Pr(z_t = K)\}. \ (2-10)$$

If the constituent indicators is z, the complete provisional succeeding of the constituent weights, π, can be sampled as

$$\Pr(\pi|y, z, \delta)\alpha\ L_c(\theta, \pi; y, z)$$

$$\Pr(\pi|\delta)\alpha \prod_{k=1}^{K} \pi_k^{\sum_{t=1}^{T} I(z_t=k)} \prod_{t:z_t} \Pr(y_t|\theta_k) \prod_{k=1}^{K} \pi_k^{\delta_{k-1}}\ \alpha \prod_{k=1}^{K} \pi_{t=1}^{I(z_t=k)+\delta_{k-1}}$$

$$\pi \sim \text{Dirichlet}(n_1 + \delta_1, n_2 + \delta_2, \dots., n_k + \delta_k), \qquad (2-11)$$

where $n_k = \sum_{\iota=1}^{T} I_{z_\iota}$, k = 1,2,...,K, refer to the sharing sizes. if z and y, the succeeding of θ is

$$P_r(\theta|y, z)\alpha\ L_c(\theta, \pi; y, z) P_r(\theta) \sim P_r(\theta) \prod_{t:z_t=k} P_r(y_t|\theta_k). \qquad (2\text{ -}12)$$

We follow the same algorithm introduced by Marin & Robert (2014, p.183) which defines the phases of sampling from the complete provisional subsequent distributions of a mix model.

2. New Bayesian Versions of AIC and BIC

In this section, we develop Bayesian amended forms for the Akaike Information

Criterion (AIC) (Akaike, 1973) and Bayesian Information Criterion (BIC) (Schwarz,

1978) for mixture models. Here, the criteria used are basically depended on a notion presented by Brooks (2002). Given an perceived probability function attained by assimilating the latent variables of a FMM, we calculate Bayesian forms for AIC and BIC which are estimated over the succeeding draws. To do that, the first is to define the deviation. If a model with parameters $\Theta = (\pi, \theta)$ and a sequence of perceived data, $y = (y_1 . y_2 . \dots . y_T)$, increased with a sequence of missing data, $z = (z_1 . z_2 . \dots . z_T)$,

the combined or comprehensive data distribution is shown as:

$$L_c(\Theta, y, z) = \Pr(y, z|\Theta), \qquad (3-1)$$

where Pr(y,z|Θ) refers to the complete data of probability. The perceived or combined probability function of observations, Pr(y|Θ) = L(Θ;y), is got by summing every probable situation sequences in the complete data probability. Thus

$$\Pr(y|\Theta) = \sum_{\forall z} \Pr(y,z|\Theta) = \sum_{\forall z} \Pr_r(y,z|\pi,\theta), \; = \prod_{t=1}^{T} \pi_{z_t} f(y_t|\theta_{z_t}),$$

$$= \prod_{k=1}^{K} \prod\nolimits_{t:z_t=k}^{T} \pi_k f(y_t|\theta_k), \qquad (3-2)$$

and the observed log-likelihood function $\ell(y|\Theta)$ can be given by:

$$\ell(y|\Theta) = \log\left\{\prod_{k=1}^{K} \prod_{t:z_t=k}^{T} \pi_k f(y_t|\theta_k)\right\} = \sum_{k=1}^{K}\sum_{t=1}^{T} \log[\pi_k \, f(y_t|\theta_k)]. \qquad (3-3)$$

So, the deviation then could be well-defined as:

$$D(\Theta) = -2[\ell(y|\Theta)] = -2 \sum_{k=1}^{K}\sum_{t=1}^{T} \log[\pi_k f(y_t|\theta_k)]. \qquad (3-4)$$

We indicate the suggested standards as AIC_B and BIC_B which are resulted from

$$AIC_B = E_{D_{Bayesian}(.)}\left[D_{Bayesian}(\Theta)\right] + 2h,$$

$$= -2 \, E_{\log P(.)}[\log L_c(\theta,\pi;y,z)] + 2h, \qquad (3-5)$$

and

$$BIC_B = E_{D_{Bayesian}(.)}\left[D_{Bayesian}(\Theta)\right] + h \, \log(T),$$

$$= -2 \, E_{\log Pr(.)}[\log L_c(\theta,\pi;y,z)] + h \, \log(T). \qquad (3-6)$$

This formula, $D_{Bayesian}(.)$ is a minimum Bayesian deviation, assessed by the subsequent samples of the model parameters simulated from an MCMC run. $\log L_c(\theta,\pi;y,z)$ is the complete log-likelihood function, and h is the number of free parameters, that is, $h = 3K - 1$ (Celeux et al., 2006). Given an observed log-likelihood in closed form into the AIC and BIC can be approximated as follows:

$$AIC_B = E_{\Theta|y}\left[D_{Bayesian}(\Theta|y)\right] + 2h$$

$$= -2 \, E_{\Theta|y}[\log \Pr(y|\pi,\theta)] + 2h$$

$$= -2 \int_{\pi}\int_{\theta}[\log \Pr(y|\pi,\theta)] \; \Pr(\pi,\theta|z) d_{\pi} d_{\theta} + 2h, \qquad (3\text{-}7)$$

and

$$BIC_B = E_{\Theta|y}\left[D_{Bayesian}(\Theta|y)\right] + h \log T,$$

$$= -2E_{\Theta|y}[\log \Pr(y|\pi,\theta)] + h \, \log T,$$

$$= -2 \int_{\pi}\int_{\theta}[\log \Pr(y|\pi,\theta)] \Pr(\pi,\theta|y,z) d_{\pi} d_{\theta} + h \, \log(T), \quad (3\text{-}8)$$

where

$E_{\Theta|y}[D_{Bayesian}(\Theta|y)]$= -2 $E_{\Theta|y}[\log Pr(y|\pi,\theta)]$

The two above standards, the anticipated perceived deviation assessed at draws from the subsequent dissemination of all model parameters, Pr(π,θ|y,z), perceived over an MCMC run.

3. Simulation study

We design, in this section, a simulation study to investigate the suggested standards; AIC and BIC under the Bayesian principle, for limited mix of regular distributions. This experiment, we try to select the best normal mixture model among several competing mixture models using proposed criteria. Before doing that, we first generate several data set from models with different complexities. Then, we implement the model estimation. Finally, we select the best model fitted to the generating data.

3.1 Generating synthetic data sets

We generated three data sets with size n=600 for each from normal mixture models with K_0=2,3 and 4 respectively, where K_0 denotes the model order:

First model $K_0 = 2$:

$0.3N(y_t|2,1) + 0.7N(y_t|10,1)$.

Second model $K_0 = 3$:

$0.3N(y_t|2,1)+0.5N(y_t|8,1)+0.2N(y_t|12,1)$.

Third model $K_0 = 4$:

$0.25N(y_t|2,1)+0.25N(y_t|8,1)+0.25N(y_t|12,1)+0.25N(y_t|20,1)$.

Figures (1), (2) and (3) show the histograms of the data sets simulated from the three models.

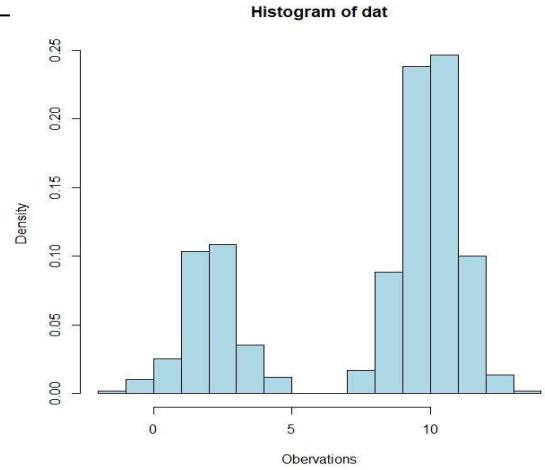

Figure 1: Histogram of the data generated from the first model (2- components).

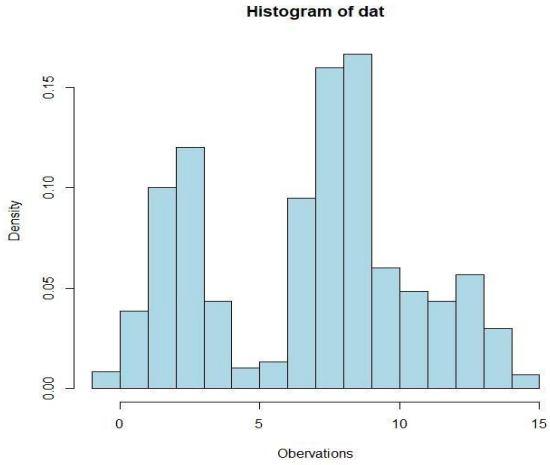

Figure 2: Histogram of the data generated from the second model (3- components).

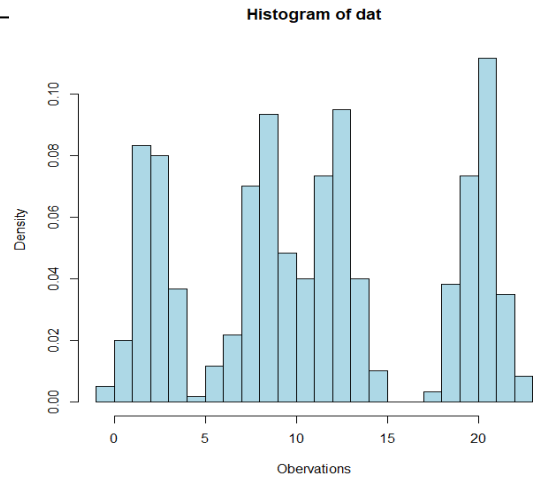

Figure 3: Histogram of the data generated from the third model (4- components).

3.2 Estimation and fitting the model

Throughout section, we estimate the model parameters employing the Gibbs sampler (Geman & Geman, 1984) of every model of the three models that generated the data sets. The parameters of the model are sampled as following:

$$\sigma_j^2 \sim \text{invGamma}(a_j, b_j),$$

$$\mu_j|\sigma_j^2 \sim \text{Normal}(\eta_j, \sigma_j^2|\xi_j),$$

$$\pi_j \sim \text{dirichlet}(\delta_j),$$

where η_j, ξ_j, a_j, b_j and δ_j are known hyper-parameters, $j = 1,2,\dots,K$. We utilize non-informative priors on the model indexes for the data plays the leading part in deducing the subsequent dissemination (Gelman et al., 2014). Therefore, Dirichlet distribution is used for π with $\delta_j = 1 , \forall j = \{1,2,\dots,k\}$,(Fruhwirth-Schnatter,2006).For difference parameter, σ^2 , we use the inverse Gamma with parameters a= 0.001 and b = 0.001 and so a mean of a/b = 1 and a variance of $a/b^2 = 1000$ could result diffuse values of this form. The prior of the mean parameter,μ , could be meant flat values from a Normal distribution with a shape parameter, η = 0, and a scale parameter, ζ = 0.001 that has great difference about 1000. We run the Gibbs sampler for 105,000 iterations, leaving the first 5,000 as a burn-in period and thinning the rest 100,000 iterations by keeping every 10th iteration. To avoid switching label problem, we put artificial restraints on the means parameter of the model, thus $\mu_1 < \mu_2 < \cdots < \mu_k$.

4. Results

4.1 Results of the model estimation

We show the outcomes of the valuation and choice of the all models fitted to the simulated datasets. Tables (1-3) show estimation result of each model respectively. Figures (4), (5), and (6) show the MCMC result of each model respectively. Note that the sampler accomplishes well in approximating the real parameters of three models.

Parameter	π_1	π_2
True	0.3	0.7
Estimated	0.281	0.719
95%CI	(0.196, 0.492)	(0.693, 0.791)
Parameter	μ_1	μ_2
True	2	10
Estimated	2.049	9.956
95%CI	(1.896, 3.292)	(9.593, 10.701)
Parameter	σ_1^2	σ_2^2
True	1	1
Estimated	1.096	1.026
95%CI	(0.996, 1.292)	(0.993, 1.188)

Table 1: Estimated parameters of a Normal mixture model with $K_0 = 2$.

Parameter	π_1	π_2	π_3
True	0.3	0.5	0.2
Estimated	0.312	0.489	0.213
95%CI	(0.296, 0.392)	(0.493, 0.5081)	(0.174,0.244)
Parameter	μ_1	μ_2	μ_3
True	2	8	12
Estimated	2.074	7.891	12.022
95%CI	(1.993, 2.292)	(7.644, 8.301)	(11.774,12.144)
Parameter	σ_1^2	σ_2^2	σ_3^2
True	1	1	1
Estimated	0.986	1.071	1.047

95%CI	(0.987, 1.116)	(0.903, 1.236)	(0.974,1.144)	

Table 2: Estimated parameters of a Normal mixture model with $K_0 = 3$.

Parameter	π_1	π_2	π_3	π_4
True	0.25	0.25	0.25	0.25
Estimated	0.262	0.277	0.248	0.228
95%CI	(0.183, 0.243)	(0.190, 0.297)	(0.188,0.276)	(0.174,0.284)
Parameter	μ_1	μ_2	μ_3	μ_4
True	2	8	12	20
Estimated	2.095	8.342	12.19	20.076
95%CI	(1.788, 2.731)	(7.312,8.664)	(11.444,12.219)	(19.201,12.326)
Parameter	σ_1^2	σ_2^2	σ_3^2	σ_4^2
True	1	1	1	1
Estimated	1.337	1.041	1.133	1.075
95%CI	(0.895,1.047)	(0.899,1.333)	(0.989,1.107)	(0.952,1.111)

Table 3: Estimated parameters of a Normal mixture model with $K_0 = 4$.

4.2 Result of the model selection

Having satisfactory estimates, by fitting three normal mixture models correctly, we now check the process of model selection. Given K=1, 2, ..., 7 of competitive models fitted for each one of the three synthetic datasets, the percentage of times out of 100 replications is reported that our criteria select the correct generating model in Table (5). In the example of producing data model with K0 =2 situations, it is clear that both AICB and AICB perform well in choosing the appropriate model (97%) with very slight overestimate the number of true components (3%).

In the second case, with increasing the complexity of generating model (K0=3), it can see that the AICB and AICB have a very satisfactory selection percentage for the correct model (92% and 93% respectively). In the third case, the generating model with K0=4, the AICB and AICB present successful in determining the four components in 100 simulations (90% and 91% respectively).

5. Application

In this section, we adopt two real application databases involving the acidity data and galaxy data, which were analyzed by Richardson & Green (1997), to evaluate our pro- posed criteria (figures 4 and 5).

K	K_0=2		K_0 = 3		K_0 = 4	
	AICB	BICB	AICB	BICB	AICB	BICB
K=1	0%	0%	0%	0%	0%	0%
K=2	97%	97%	2%	4%	0%	0%
K=3	3%	3%	92%	93%	3%	3%
K=4	0%	0%	6%	3%	91%	90%
K=5	0%	0%	0%	0%	0%	0%
K=6	0%	0%	0%	0%	0%	0%
K=7	0%	0%	0%	0%	0%	0%

Table 4: the number of times percentage in which the generating models with K_0=2, K_0=3 and K_0=4 are chosen by each criterion over 100 replications.

The Acidity data focus on the acidity index for a sample of 155 lakes in north- central Wisconsin. This parameter defines the lake's ability to absorb acid; decreased values may result in losing biological resources. It is supposed that ooze lakes that have no inlets or outlets, be apt to have minimum acidity parameter, and discharged lakes that have inlets and outlets have greater values. This information has been analyzed as a combination of regular disseminations on the log-scale by several authors, for example: Crawford et al. (1992) as well as Richardson & Green (1997).

According to several studies, the best model fitted for these data was its rank ranged between two and three components. For example, Richardson & Green (1997) used the reversible jump Markov chain Monte Carlo (RJMCMC) method and concluded that a normal mix model with two-components is the best for these data. Ishwaran, James & Sun (2001) suggested that two-components mixture model for the same data using the AIC and BIC. Also, Chen & Kalbfleisch (1996) pointed out that a normal mixture with K=2 is an appropriate model for the acidity data. On the other hand, McGrory & Titterington (2007) suggested a normal mixture with K=3 for these data.

The Galaxy data includes 82 speeds of remote galaxies differing from our own, sampled from 6 well-separated conic sections of the corona borealis. This data have been used by Crawford et al. (1992), Richardson & Green (1997), Celeux et al. (2006) and Papastamoulis & Iliopoulos (2010) for mixture modeling. The best fitting model appropriate for these data was a model with three mixture components as shown by Celeux et al. (2006).

By the same way implemented with simulated data, we fitted normal mixture models with different complexities ranged from K=2 until K=7 for the Galaxy and acidity data as shown in figures (4) and (5) respectively. The results of model selection are shown in Table (5) which show that the AICB and BICB select the model with K=2 along with the results obtained by most above authors. Note that fitting for the galaxy data after K=3 no longer important. This can be noted for the values of log-likelihood which were decreasing slightly. The same thing according to the acidity data.

K	Acidity data			Galaxy data		
	$\overline{\ell(y\|\Theta)}$	AICB	BICB	$\overline{\ell(y\|\Theta)}$	AICB	BICB
K=2	-188.710	387.421	402.638	-227.241	464.482	476.515
K=3	-187.760	391.521	415.868	-213.049	442.099	461.352
K=4	-187.149	397.298	430.776	-213.032	448.125	474.609
K=5	-186.207	402.415	445.023	-213.021	454.043	487.737
K=6	-186.101	410.201	461.941	-212.933	459.067	499.981
K=7	-186.004	416.008	476.877	-212.177	465.354	513.488

Table 5: Results of the model selection for two real data application.

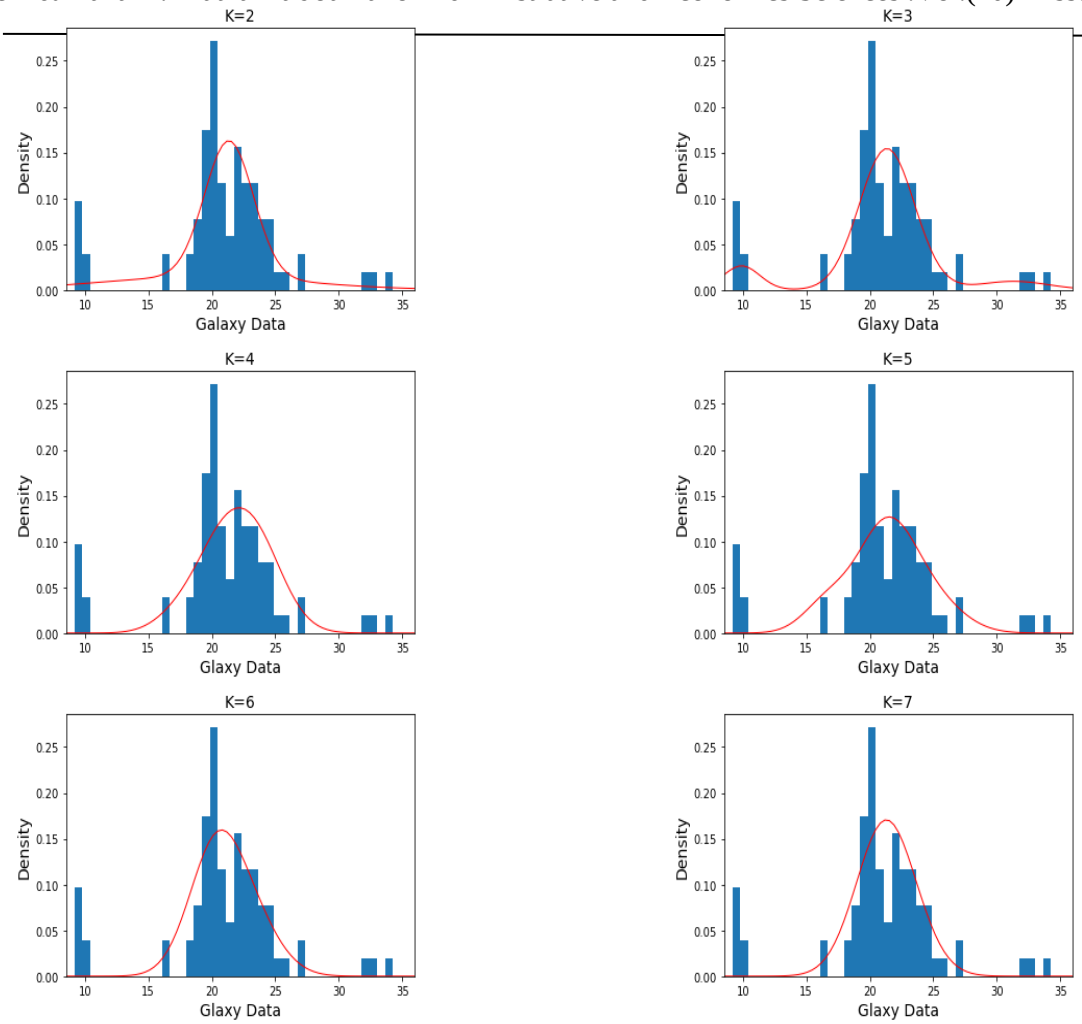

Figure 4: Fitting six test models for the galaxy data.

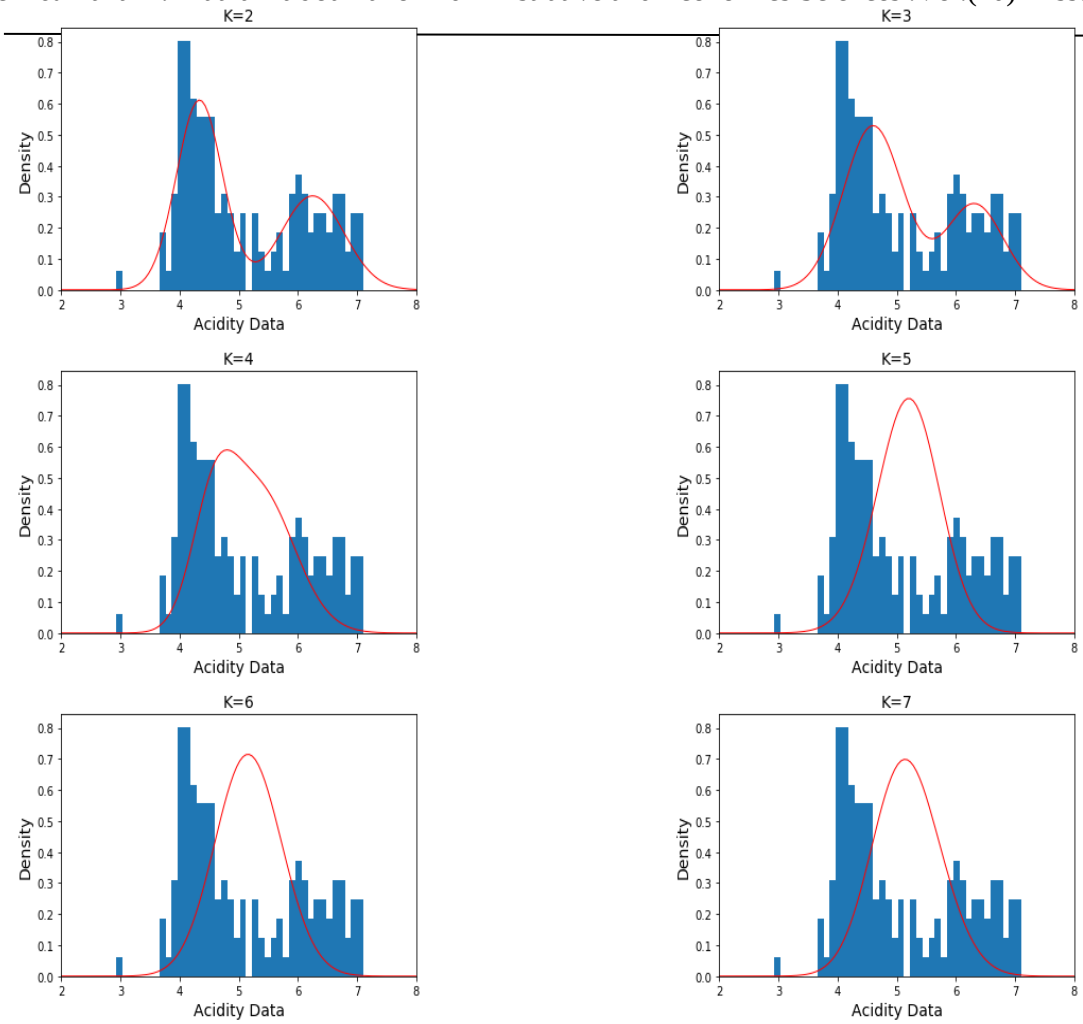

Figure 5: Fitting six test models for the acidity data.

6. Conclusion

This paper addressed the model selection issue for finite mixture model. We developed new modified information criteria for selecting the best finite mixture model, as an idea inspired by Brooks (2002, p. 617). We derived Bayesian deviations plugging into two known standards: the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to select the fittest mixture model. We showed via simulation studies and examples include real data applications that the these new criteria perform well. We recommend to extend the study to include more complicated models, for instance underlying Markov models, in which dependency among the hidden states may have important role in performance of the proposed criteria. In addition, we recommend in future to investigate other versions of our proposed criteria considering different forms of the likelihood, for instance, the complete and conditional likelihoods.

Acknowledgments

This work is a part from the scientific plan of the department of Banking and Financial science, College of administration and economics, Muthanna university, Iraq.

A. Computing the new modified AIC and BIC for a Normal mixture model

Given an observed log-likelihood in closed form, an approximate version of the AICB and BICB can be calculated using M simulated values from an MCMC run. For a finite mix of regular distributions the deviation can be given by:

$$\overline{D(\pi,\theta)} = \overline{D(\pi,\mu,\sigma^2)} \approx -\frac{2}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\log\left\{\sum_{k=1}^{K}\pi^{(m)}\,\phi_k\left(y_t\middle|\mu_k^{(m)},\sigma_k^{2(m)}\right),\right\}, \quad (7-1)$$

Where $\phi_k(.)$represent the density function of k-component normal distribution, M is the number of iteration and $(\pi^{(m)},\mu_k^{(m)},\sigma_k^{2(m)})$are the simulated parameters of the model. Therefor, the AIC_B and BIC_B can be approximated, respectively, as follows:

$$AIC_B \approx -\frac{2}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\log\left\{\sum_{k=1}^{K}\pi^{(m)}\,\phi_k\left(y_t\middle|\mu_k^{(m)},\sigma_k^{2(m)}\right)\right\} + 2h, \quad (7-2)$$

and

$$BIC_B \approx -\frac{2}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\log\left\{\sum_{k=1}^{K}\pi^{(m)}\,\phi_k\left(y_t\middle|\mu_k^{(m)},\sigma_k^{2(m)}\right)\right\} + h\log(T), \quad (7-3)$$

References

AKAIKE, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Proceedings of the Second International Symposium on Information, eds. B. Petrov & F. Csaki. Budapest, pp. 267–281.

B ROOKS, S. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde. Journal of the Royal Statistical Society, B 64, 616–618.

CARLIN, B.P. & LOUIS, T.A. (2009). Bayesian Methods for Data Analysis. Boca Raton, FL: Chapman and Hall/CRC Press, 3rd edn.

CELEUX, G., FORBES, F., ROBERT, C.P. & TITTERINGTON., D.M. (2006). Deviation information criteria for missing data models. Bayesian Analysis 1, 651–673.

CHEN, J. & KALBFLEISCH, J. (1996). Penalized minimum-distance estimates in finite mixture models. Canadian Journal of Statistics 24, 167–175.

CONGDON, P. (2014). Applied Bayesian Modelling. Mew York: John Wiley & Sons, 2nd edn.

CRAWFORD, S.L., DEGROOT, M.H., KADANE, J.B. & SMALL, M.J. (1992). Modeling lake-chemistry distributions: Approximate Bayesian an methods for estimating a finite-mixture model. Techno metrics 34, 441–453.

FAN, Y. & SISSON, S.A. (2011). Reversible jump MCMC. In Handbook of Markov Chain Monte Carlo, eds. S. Brooks, A. Gelman, G. Jones & X.L. Meng, chap. 3. CRC press, pp. 67–91.

FRUHWIRTH ¨ -SCHNATTER, S. (2006). Finite Mixture and Markov Switching Models. Springer Series in Statistics. Springer, New York.

10 GELMAN, A., CARLIN, J., STERN, H., DUNSON, D., VEHTARI, A. & RUBIN, D. (2014). Bayesian Data Analysis. Chapman and Hall/CRC, 3rd edn.

GEMAN, S. & GEMAN, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on pattern analysis and machine intelligence , 721–741.

ISHWARAN, H., JAMES, L.F. & SUN, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. Journal of the American Statistical Association 96, 1316–1332.

KADHEM, S.K., HEWSON, P. & KAIMI, I. (2018). Using hidden Markov models to model spatial dependence in a network. Australian & New Zealand Journal of Statistics 60, 423–446.

MARIN, J.M. & ROBERT, C. (2014). Bayesian Essentials with R. Springer New York Heidelberg Dordrecht London, 2nd edn.

MCGRORY, C.A. & TITTERINGTON, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. Computational Statistics & Data Analysis 51, 5352–5367.

MCLACHLAN, G. & PEEL, D. (2000). Finite Mixture Models. Wiley Series in Probability and Statistics, New York, USA.

NTZOUFRAS, I. (2009). Bayesian Modeling Using WinBUGS. Wiley Series in Computational Statistics, Hoboken, USA.

RICHARDSON, S. & GREEN, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society: series B (statistical methodology) 59, 731–792.

SCHWARZ, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464.

SCOTT, S.L. (2002). Bayesian methods for Hidden Markov models: Recursive computing in the 21th century. Journal of the American Statistical Association 97, 337– 351.

SPIEGELHALTER, D., BEST, N., CARLIN, B. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639.

TANNER, M.Y. & WONG, W.H. (1987). The Calculation of Subsequent Distributions by Data Augmentation. Journal of the American Statistical Association 82, 528–540.

ZUCCHINI, W. & MACDONALD, I.L. (2009). Hidden Markov Models for Time Series: An Introduction Using R. Chapman and Hall, London.