Journal of Systems Engineering and Electronics
Vol. 28, No. 4, August 2017, pp.801 – 809

A moment-based criterion for determining the number of components in a normal mixture model

Yimin Zhou[1], Liyan Han[1], Dan Wang[2], and Libo Yin[3,*]

1. School of Economics and Management, Beihang University, Beijing 100191, China;
2. School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100191, China;
3. School of Finance, Central University of Finance and Economics, Beijing 100081, China

Abstract: Determining the number of components is a crucial issue in a mixture model. A moment-based criterion is considered to estimate the number of components arising from a normal mixture model. This criterion is derived from an omnibus statistic involving the skewness and kurtosis of each component. The proposed criterion additionally provides a measurement for the model fit in an absolute sense. The performances of our criterion are satisfactory compared with other classical criteria through Monte-Carlo experiments.

Keywords: information criteria, Gaussian mixture, moment-based, number of components.

DOI: 10.21629/JSEE.2017.04.20

1. Introduction

Nowadays there is a widespread empirical evidence that finite mixture models have been powerful tools for analyzing data where observations originate from various components. The analyses of such finite mixture models are commonly carried out by using the maximum likelihood estimation with the known number of mixture components [1,2]. This naturally leads to the development of the selection criterion for determining the number of components in a mixture model.

A selecting criterion commonly deals with the trade-off between the quality of the model and the complexity of the model. Following this rule, there are various selection criteria to determine the number of components in a mixture model in literature. The Akaike's information criterion (AIC) [3] and the Bayesian information criterion (BIC) [4] are the most widely used criteria among them. In addition, numbers of studies provide targeted methods in accessing the number of components for a mixture model. These researches involve methods based on likelihood ratio test [5 – 9], Bayesian analysis [10 – 12], entropy criterion [13,14], homogeneity test [15,16] and graphical technique [17].

We provide a novel criterion to determine the appropriate number of components for a mixture model based on its moments. This criterion is derived from an omnibus statistic involving the skewness and kurtosis of each component. Our motivation for this study is twofold.

Firstly, our work is inspired by the simplicity of quality tests for normality. On the one hand, since mixtures of the normal distribution consist of more than one normal components, it is natural to wonder whether each normal component in a mixture model enjoys the similar statistic features with the normal distribution. On the other hand, the moment-based methods are widely used in testing normality [18 – 20].

In this sense, the criterion based on the moments is likely to be a neglected area of research in the mixture model. Based on our empirical results, the former four moments of each component, as defined in a Gaussian mixture model, share similar statistical characters with that of a normal distribution. Therefore, we find a supportive evidence for the feasibility of using the moment-based statistics to access the quality of our model.

Secondly, except for its comprehensibility and simplicity of calculation, our criterion additionally provides evaluation for the model quality in an absolute sense. We particularly compare our criterion with the AIC, BIC and normalized entropy criterion (NEC) which enjoy the computational convenience as well. It should be noticed that the AIC and BIC both provide evaluations of the models in a relative sense of testing its quality, which means that if all the candidate models fit poorly, these criteria will not give

Manuscript received February 12, 2016.
*Corresponding author.
This work was supported by the National Natural Sciences Foundation of China (71371022; 71401193; 71671193), the Program for Innovation Research in Central University of Finance and Economics, and the Innovation Foundation of BUAA for Ph.D. Graduates.

any warning of that. The work of Celeux and Soromenho [13] has approached this issue in a practical way. However, researchers have cast doubt that their procedure has shown a disappointing behavior [21]. In addition, compared with other criteria such as AIC, BIC, it suffers from the limitation that it cannot decide between one and more than one clusters [22].

Compared with the AIC and BIC, our criterion provides the testing of model quality in an absolute sense while those criteria are not devoted to measuring the performance of the mixture model. Besides, it provides more targeted criterion for the normal mixture model. Compared with the NEC, our criterion gets a better performance and is of no doubt with the decision between one and more than one clusters. The performances of our criterion are satisfactory compared with these classical criteria through Monte-Carlo experiments.

The rest of this paper is organized as follows. Section 2 introduces the classical criteria used for the model selection. Section 3 gives a heuristic discussion of our selection criterion, followed in Section 4, simulation procedures are designed to evaluate the performance of the criterion. Section 5 offers a summary.

2. Criteria for the number of components in a normal mixture model

A normal mixture model is a weighted sum of K component normal densities as given by

$$P(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_{j=1}^{K} \tau_j g(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{1}$$

where $\boldsymbol{x}$ is a d-dimensional vector, τ_j $(j = 1, 2, \ldots, K)$ is the mixture weights, and $g(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is the normal density

$$g(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}_j|^{-1/2} \exp[-(\boldsymbol{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)] \tag{2}$$

with the mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. The mixture weights satisfy the constraint that $\sum_{j=1}^{K} \tau_j = 1$. The mixture model assumes that each observed data point x_i has a corresponding unobserved data point, or latent variable τ_j^i $(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant K)$, specifying the mixture component that each data point belongs to. Let $x_1, x_2, \ldots, x_n$ denote a random training set of independent and identically distributed samples taken from the mixture distribution in (2). The log-likelihood function can be written as

$$L(K) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{K} \tau_j g(\boldsymbol{x}|\boldsymbol{\lambda}_j) \right) \tag{3}$$

where $\boldsymbol{\lambda}_j = \{\tau_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ $(j = 1, 2, \ldots, K)$. $L(K)$ is regarded to contain information about model fit. Meanwhile, $L(K)$ is an increasing function of K in general while a larger K indicates more model complexity. Various criteria to be minimized have been proposed to measure a model's suitability by balancing model fit and model complexity.

The Akaike information criterion [3], considered in [23] in the mixture context, takes the form

$$\mathrm{AIC}(K) = -2 \ln L(K) + 2N(K) \tag{4}$$

where $N(K)$ is the number of the unknown parameters.

The Bayesian information criterion defined by Schwarz [4] approximates the exact Bayes solution to the problem of selecting the appropriate model and is defined as

$$\mathrm{BIC}(K) = -2 \ln L(K) + \ln(n) N(K) \tag{5}$$

Closely related to the AIC, the BIC or Schwarz criterion is partly based on the likelihood function. When determining the number of mixture components, the AIC tends to overestimate and the BIC tends to underestimate the number of components K [22,24,25]. These criteria (especially the AIC) have been widely adopted in model selection, see for examples the works of Yamaoka et al. [26], Anderson et al. [27] and Posada and Crandall [28]. However, neither AIC nor BIC provides a test in an absolute sense of testing the model quality, which means that if all the candidate models fit poorly, these criteria will not give any warning of that.

The normalized entropy criterion proposed by Celeux and Soromenho [13] is derived from a relation linking the likelihood and the classification likelihood of a mixture and is expressed as

$$\mathrm{NEC}(K) = \frac{E(K)}{C(K) + E(K) - C(1)}. \tag{6}$$

This is a transform form of the entropy criterion of Celeux and Soromenho [13] where $E(K) = \sum_{j=1}^{K} \sum_{i=1}^{n} \tau_{ij} \ln(\tau_{ij})$ and $C(K) = \sum_{j=1}^{K} \sum_{i=1}^{n} \tau_j g(x_i; \boldsymbol{\lambda}_j)$. Celeux and Soromenho [13] concentrated on the view of the cluster analysis to estimate the number of components arising from a normal mixture model. Under their assumptions, the probability of each sample point to which component may belong should be one or zero in the best cases.

3. A moment-based criterion

We propose a criterion which aims to measure the ability of the model in separating mixtures of normals. It emphasizes the model with its goodness of fit in general, as well as its partial benefits. According to Akaike, the AIC criterion suggests the use of likelihood as a measure of fit of a

model. However, this measure can only make a comparison in the goodness of fit rather than an absolute judgment. The measure for partial benefits in the proposed criterion can make up for this advantage. Under the assumption of the accumulated normal density, each component is assumed to be normal. As for a normal distribution, each eigenvalue can be described by its mean and variance. It is unnecessary to consider its higher moments since that skewness and kurtosis are both equal to zero for the normal distribution. For this reason, the skewness and kurtosis of each component are to be penalized in our criterion.

3.1 Skewness and kurtosis of each component in a mixture of normals

3.1.1 Measuring the skewness and kurtosis of each component

In the probability theory and statistics, skewness and kurtosis both describe the shape of a probability distribution of a real-valued random variable. Skewness is a descriptor of the asymmetry while the kurtosis presents the "peakedness" [29]. There are different ways of quantifying them for a theoretical distribution and corresponding ways of estimating them from a sample. One common measure of skewness and kurtosis, originating with Karl Pearson [30], is based on a scaled version of the third and fourth moment of the data respectively. Pearson's moment coefficients of skewness and excess kurtosis are used to provide a comparison of the shape of a given distribution to that of the normal distribution. It has been argued that there was no emphasis in Pearson's original work on kurtosis as measuring (in part) tail heaviness which seems to be its more frequent contemporary usage. For this reason, an adjusted version of Pearson's kurtosis is commonly used instead. According to Pearson's moment coefficients of skewness and excess, the skewness of a random variable x is the third standardized moment, denoted by $\sqrt{b}_1$ and defined as

$$\sqrt{b}_1 = \frac{E[(x-u)^3]}{(E[(x-u)^2])^{3/2}} = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]. \quad (7)$$

Excess kurtosis is equal to the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3,

$$b_2 = \frac{E[(x-u)^4]}{(E[(x-u)^2])^2} - 3 = \frac{E[(x-u)^4]}{\sigma^4} - 3. \quad (8)$$

This subsection is designed to measure the skewness and kurtosis for each normal in the mixture model. Constraint that $\sum_{j=1}^{K} \tau_j = 1$. Coupled with the terms in (7) and (8), the coefficient of skewness for the kth normal distribution is computed as

$$\sqrt{b}_1^{(k)} = E_k\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] = \int_{-\infty}^{\infty}\left(\frac{x-u_k}{\sigma_k}\right)^3 g(x|\mu_k,\sigma_k)\mathrm{d}x. \quad (9)$$

In order to transform the term inside the integral into the production of $P(x)$ and some function $F(x|k)$, the term $g(x|\mu_k,\sigma_k)$ is transformed into form

$$g(x|\mu_k,\sigma_k) = g(x|k) = \frac{p(k|x)}{\tau_k}P(x). \quad (10)$$

By the Bayes' rule, and (9) turns to

$$\sqrt{b}_1^{(k)} = \int_{-\infty}^{\infty}\left(\frac{x-u_k}{\sigma_k}\right)^3 \frac{p(k|x)}{\tau_k}P(x)\mathrm{d}x. \quad (11)$$

Note that in (11), the term $\int_{-\infty}^{\infty}\left(\frac{x-u_k}{\sigma_k}\right)^3 \frac{p(k|x)}{\tau_k}P(x)\mathrm{d}x$ is the expectation of function $\left(\frac{x-u_k}{\sigma_k}\right)^3 \frac{p(k|x)}{\tau_k}$ under the probability distribution $P(x)$. The probability that a given point x_k belongs to the kth component is signed as τ_{ik}. Then, this expectation [31,32] can be approximated as

$$\sqrt{b}_1^{(k)} \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-u_k}{\sigma_k}\right)^3 \frac{\tau_{ik}}{\tau_k}. \quad (12)$$

In a similar way to the measure of the skewness $\sqrt{b}_1^{(k)}$, the coefficient of the kurtosis $b_2^{(k)}$ can be approximated as

$$b_2^{(k)} \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-u_k}{\sigma_k}\right)^4 \frac{\tau_{ik}}{\tau_k} - 3. \quad (13)$$

3.1.2 Asymmetric distributions of $\sqrt{b}_1^{(k)}$ and $b_2^{(k)}$

Pearson, over a long period, has studied the distribution of the sample skewness and kurtosis statistics $\sqrt{b}_1$ and b_2 with samples from a normal population, for example see [33]. Since then, $\sqrt{b}_1$ and b_2 have been used to test the null hypothesis that the distribution sampled is the normal [18,19,34]. D'Agostino and Pearson [18] have suggested an omnibus statistic in the form of $X_T^2 = X_1^2(\sqrt{b}_1) + X_2^2(b_2)$, where $X_1(\sqrt{b}_1)$ and $X_2(b_2)$ follow standardized normal distribution approximately, and the statistic X_T^2 is distributed as $\chi^2(2)$. $X_1(\sqrt{b}_1)$ and $X_2(b_2)$ can be written as

$$X_1(\sqrt{b}_1) = \sqrt{b}_1/\sqrt{6/n}, \quad X_2(b_2) = (b_2-3)/\sqrt{24/n}$$

while $6/n$ and $24/n$ are the asymptotic variances of $\sqrt{b}_1$ and b_2 [19]. A point worth noted is that the statistic $X_T^2 =$

$(\sqrt{b_1}/\sqrt{6/n})^2 + ((b_2-3)/\sqrt{24/n})^2$ requires large samples to be distributed as a χ^2 distribution.

Since a normal mixture distribution consists of more than one normal component, it is natural to wonder whether $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ in the kth component enjoy the similar statistic features with that of $\sqrt{b_1}$ and b_2. To answer this question, we generate Experiment 1 to assess the behavior of $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ in a mixture of two normal populations and compare with that of $\sqrt{b_1}$ and b_2 in a normal distribution.

Experiment In this experiment, 100 000 random samples are drawn from the distributions $N(2, 2^2)$ and $0.3N(2, 2^2) + 0.7N(4, 2^2)$ respectively. We carry out 10 000 replications and estimate the moments of $\sqrt{b_1}$, b_2, $\sqrt{b_1^{(k)}}$, $b_2^{(k)}$ and the related variables.

Table 1 Statistics of $\sqrt{b_1}$, b_2, $\sqrt{b_1^{(k)}}$, $b_2^{(k)}$ and related variables with sample size $n = 100\,000$

Statistics	Normal		Mixture of two normals ($\boldsymbol{\tau} = [0.3, 0.7]$)			
	$\sqrt{b_1}$	b_2	$\sqrt{b_1^{(1)}}$	$b_2^{(1)}$	$\sqrt{b_1^{(2)}}$	$b_2^{(2)}$
Mean	0.000 0	3.000 1	0.000 2	3.000 2	0.000 2	2.999 9
Std.	0.007 7	0.015 4	0.014 9	0.019 6	0.011 0	0.014 4
Jarque Bera	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.063 0
Std. of $X_k(\cdot)$	0.999 7	0.993 1	1.924 3	1.263 0	1.425 9	0.930 2

Note: The probability of rejecting the normality in Table 1 indicates that all the variables are considered normal statistically.

As seen in Table 1, it is not surprising that the distributions of $X_1(\sqrt{b_1})$ and $X_2(b_2)$ are approximated $N(0, 1^2)$ in the results of Monte Carlo simulations. Similarly, the results of Monte Carlo simulations on $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ indicate that the distributions of skewness and kurtosis in the mixture of normal condition are approximated normal. However, the distributions of $X_1(\sqrt{b_1^{(k)}})$ or $X_2(b_2^{(k)})$ do not approximate to a standard normal distribution as $X_1(\sqrt{b_1})$ or $X_2(b_2)$ does. Therefore, the similar transform functions cannot be used to generate the statistics for testing the local model fit. In this paper, we use the original forms of $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ which are approximated normal with limited means and variances.

3.2 The proposed criterion

Similar to the NEC, our criterion not only includes the model selecting function, but also rewards the model fit. The proposed criterion is derived from a statistic involving the skewness and kurtosis of each component. Our moment-based criterion (NMC) takes the form

$$\mathrm{NMC}(K) = \max\{|\sqrt{b_1^{(k)}}|, |b_2^{(k)}|\} \tag{14}$$

where $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ correspond to the measure of skewness and kurtosis of each normal component respectively.

For a well-fitting normal mixture model, the skewness $(\sqrt{b_1^{(k)}})$ and kurtosis $(b_2^{(k)})$ of each normal component are considered for the advantage that skewness and kurtosis are standards in test for normality and their value should be approximated to zero. On the contrary, if for some component j the distribution of the samples in its vicinity is non-normal, the associated value of $\sqrt{b_1^{(j)}}$ or kurtosis $b_2^{(j)}$ would deviate from zero to a positive or negative number. Vlassis and Likas [35] have proposed a total kurtosis measure $K_T = \sum_{j=1}^{K} \tau_j |b_2^{(j)}|$ to value how large this deviation is for the whole mixture. They also underlined the application of K_T as a measure of how well a normal mixture fits the data. It should reinforce the point that, considering the weighted sum of $|b_2^{(j)}|$ may ignore the information of kurtosis corresponding to a light weight τ_j. As an example, the maximum kurtosis and the total kurtosis K_T with different component number K are computed based on the method in [35]. As can be seen in Fig. 1, the value of the total kurtosis K_T (which corresponds to the red line) reflects that the goodness of fit is getting better while the value of maximum kurtosis indicates a worse model fit.

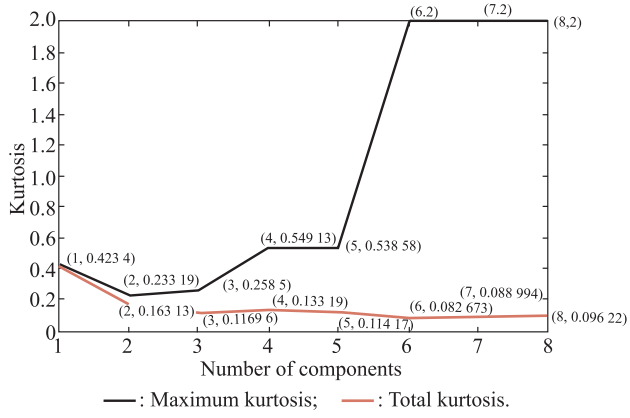

Fig. 1 The maximum kurtosis and the total kurtosis

In this paper, the maximum absolute value of skewness and kurtosis is used as a substitute indicator for the total value. The values of $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ as defined in (12) and (13) are the moment information of each component of the mixture model which should be both equal to zero under the normal assumption. In other words, for a normal mixture model the smaller the absolute value of $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ the better. Clearly, the upper bound of both $\sqrt{b_1^{(k)}}$ and $b_2^{(k)}$ can be described by their maximum absolute values as $\mathrm{Max}\{|\sqrt{b_1^{(k)}}|\}$ and $\mathrm{Max}\{|b_2^{(k)}|\}$. For the further step, the measure of the local goodness of the fit is defined as the maximum value of $\mathrm{Max}\{|\sqrt{b_1^{(k)}}|\}$ and $\mathrm{Max}\{|b_2^{(k)}|\}$.

As previously mentioned,

$$\mathrm{NMC}(K) = \max\{|\sqrt{b_1^{(k)}}|, |b_2^{(k)}|\}$$

is proposed as a criterion to be minimized for. This criterion $\mathrm{NMC}(K)$ is chosen for two reasons. Firstly, while the AIC can tell nothing about the quality of the model in an absolute sense, the moment-based criterion reflects the behavior of the model. Secondly, $\mathrm{NMC}(K)$ is not a monotone function of K. Given a set of candidate models for the data, the preferred model is the one with the minimum $\mathrm{NMC}(K)$ value.

4. Numerical experiments

Briefly, the simulation procedure involves three subroutines: with the first sub-routine generating data sets from the normal mixture models, whereas the second subroutine estimates the parameters of each competing model using the expectation maximization (EM) algorithm and completes the selection procedure and the third sub-routine evaluates the performance of the criterion. In the procedure of the parameter estimation, we run 10 times the EM algorithm which starts with random initial position with different values of K and select the solution with the maximum likelihood.

In order to illustrate the performances of the NMC in the problem of determining the number of components, the practical behavior of the NMC are compared with the well-known criteria AIC, AICc, BIC, NEC by Monte-Carlo experiments. AICc is an AIC with a greater penalty for extra parameters. Burnham and Anderson strongly recommend using AICc, rather than AIC [36], if n is small or $N(K)$ is large. The expressions for the model selection criteria are

$$\mathrm{AIC}(K) = -2\ln L(K) + 2N(K)$$

$$\mathrm{AICc}(K) = AIC(K) + \frac{2N(K)(N(K)+1)}{(n-N(K)-1)}$$

$$\mathrm{BIC}(K) = -2\ln L(K) + N(K)\ln(n)$$

$$\mathrm{NEC}(K) = \frac{E(K)}{(C(K)+E(K)-C(1))}$$

$$\mathrm{NMC}(K) = \max\{|\sqrt{b_1^{(k)}}|, |b_2^{(k)}|\}$$

where $N(K)$ is the number of the unknown parameters. When these criteria are applied, the model producing the lowest value is chosen. For the NEC, the model $\mathrm{NEC}(1)$ will not be considered since there remains controversy about this case.

4.1 Experiment conditions

(i) In this experiment, we take the same simulated data sets as [13]. Sample sizes $n = 50, 100, 200, 300$ are considered. The first data set is a two-component normal mixture distribution with parameters $\mu_1 = 0, \mu_2 = 2$ and standard deviations $\sigma_1 = \sigma_2 = 1$ and proportions $\tau_1 = \tau_2 = 0.5$. The second one is a two-component normal mixture with proportions $\tau_1 = 0.7$, with means $\mu_1 = 0, \mu_2 = 2$ and standard deviations $\sigma_1 = \sigma_2 = 1$. The third one is a three-component normal mixture with equal proportions, with means $\mu_1 = 0, \mu_2 = 2, \mu_3 = 4$ and standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

(ii) The main task of Experiment 2 is to compute the probability of each criterion in correctly estimating the true number of components. Note that this probability p takes a value range from zero to one inclusively, with $p = 0$ implying that the criterion chooses none of the correct model and $p = 1$ presents completely excellent selection ability of the criterion. In this experiment, series of simulated data are generated. The first 100 series of data are two-component normal mixture distributions with parameters $\mu_1 = 0 + 0.01t, \mu_2 = 2 + 0.01t$ where $t = 1, 2, \ldots, 100$ and standard deviations $\sigma_1 = \sigma_2 = 1$ and proportions $\tau_1 = \tau_2 = 0.5$. The second 100 series of data are two-component normal mixture with proportions $\tau_1 = 0.7$, with means $\mu_1 = 0 + 0.01t, \mu_2 = 2 + 0.01t$ and standard deviations $\sigma_1 = \sigma_2 = 1$. The third one is a three-component normal mixture with equal proportions, with means $\mu_1 = 0 + 0.01t, \mu_2 = 2 + 0.01t, \mu_3 = 4 + 0.01t$ and standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = 1$. Sample sizes are settled with $n = 50, 100, 150, 200, 300$.

4.2 Simulation results

Tables 2–4 show some details about the fitting results of the competing models. The maximum number of components is limited to be five, which means there are five competing models in total. Columns 2 and 3 display the maximum absolute value of skewness and kurtosis for the data points in each component from which the goodness of fit for each component can be investigated. The comparative performance of the five criteria in selecting the correct model is illustrated in columns 5 to 9. It is of interest to investigate the situation when the value of K is overestimated or underestimated. We will refer to the situation whereby a criterion selects a smaller number of components than the true ones as underestimate, whereas overestimate would mean the selection of a larger number of components than the true ones. As seen in Tables 2–4, the NEC presents a slight tendency to overestimate the number of normal components while AIC/AICc and BIC trend to underestimate the number of components. As for the moment-based criterion, there is no obvious regularity to be overestimate or underestimate. Taken together, the moment-based criterion has the most satisfactory behavior in those experiments. Compared with the AIC, the NMC seems to heavily penalize the local goodness of fit of all the components in the mixture model. For instance,

as exhibited in Table 2, the skewness and kurtosis play especially distinct roles in the selecting procedure since the model reaches the best local model fit when the number of components is 2.

Table 2 Sample 1 with data generated from $0.5N(0,1^2)+0.5N(2,1^2)$

Sample size	K	K_M	S_M	$\ln L$	Selection criteria				
					AIC	AICc	BIC	NEC	NMC
n=50	1	0.05	0.91	94.17	**192.33**	**192.59**	**196.15**	0.00	0.91
	2	0.41	0.37	91.54	193.08	194.45	202.64	1.52	**0.41**
	3	0.44	0.39	91.54	199.08	202.60	214.38	1.10	0.44
	4	0.25	0.44	91.53	205.06	212.00	226.09	**1.07**	0.44
	5	1.37	1.13	89.71	207.42	219.42	234.18	NaN	1.37
n=100	1	0.11	0.86	179.72	363.43	363.56	**368.64**	0.00	0.86
	2	0.58	0.14	175.56	**361.12**	**361.76**	374.15	1.33	**0.58**
	3	0.99	0.69	174.94	365.88	367.46	386.72	1.33	0.99
	4	1.00	0.66	174.97	371.95	374.95	400.60	1.09	1.00
	5	1.08	0.78	174.98	377.95	382.90	414.43	**1.06**	1.08
n=200	1	0.00	0.59	358.88	**721.75**	**721.81**	**728.35**	0.00	0.59
	2	0.21	0.51	355.99	721.98	722.29	738.47	**1.11**	**0.51**
	3	0.59	0.85	354.67	725.33	726.09	751.72	1.16	0.85
	4	0.46	1.71	352.11	726.22	727.62	762.50	NaN	1.71
	5	0.16	0.99	352.22	732.44	734.71	778.61	NaN	0.99
n=300	1	0.09	0.64	524.99	1 053.97	1 054.01	**1 061.38**	0.00	0.64
	2	0.35	0.16	519.70	**1 049.41**	**1 049.61**	1 067.93	**1.14**	**0.35**
	3	0.45	0.80	517.42	1 050.83	1 051.33	1 080.46	1.34	0.80
	4	0.70	0.90	513.86	1 049.73	1 050.64	1 090.47	NaN	0.90
	5	0.52	0.73	514.41	1 056.82	1 058.30	1 108.68	NaN	0.73

Note: The minimum value of AIC, AICc, BIC, NEC and NMC is marked in bold respectively. For the NEC, its value becomes NaN when the value of $C(K)+E(K)-C(1)$ is approximately equal to 0.

Table 3 Sample 2 with data generated from $0.7N(0,1^2)+0.3N(2,1^2)$

Sample size	K	K_M	S_M	$\ln L$	Selection criteria				
					AIC	AICc	BIC	NEC	NMC
n=50	1	0.22	0.91	91.85	187.71	**187.96**	**191.53**	0.00	0.91
	2	0.37	0.57	88.67	**187.34**	188.70	196.90	**1.70**	**0.57**
	3	0.81	0.27	88.38	192.75	196.26	208.05	2.41	0.81
	4	0.78	1.24	86.64	195.27	202.22	216.30	NaN	1.24
	5	0.81	1.70	86.44	200.88	212.88	227.65	NaN	1.70
n=100	1	0.29	0.14	164.92	**333.83**	**333.96**	**339.04**	0.00	0.29
	2	0.23	0.18	163.94	337.88	338.52	350.91	1.07	**0.23**
	3	0.35	0.25	163.94	343.88	345.47	364.73	1.03	0.35
	4	0.53	0.50	163.85	349.70	352.70	378.35	1.04	0.53
	5	0.27	0.29	163.94	355.87	360.82	392.35	**1.02**	0.29
n=300	1	0.18	0.55	342.63	689.27	689.33	**695.87**	0.00	0.55
	2	0.29	0.32	338.57	**687.14**	**687.45**	703.63	1.18	**0.32**
	3	0.17	0.44	338.44	692.88	693.64	719.27	1.05	0.44
	4	0.11	0.50	338.44	698.87	700.28	735.16	1.04	0.50
	5	0.22	0.61	338.36	704.72	706.99	750.90	**1.03**	0.61
n=300	1	0.12	0.40	523.49	**1050.97**	**1051.01**	**1058.38**	0.00	0.40
	2	0.14	0.19	521.45	1 052.90	1 053.10	1 071.42	1.09	**0.19**
	3	0.22	0.24	521.27	1 058.55	1 059.04	1 088.18	1.04	0.24
	4	0.29	0.29	521.24	1 064.49	1 065.40	1 105.23	**1.02**	0.29
	5	0.43	0.61	519.87	1 067.75	1 069.22	1 119.60	1.05	0.61

Note: The minimum value of AIC, AICc, BIC, NEC and NMC is marked in bold respectively. For the NEC, its value becomes NaN when the value of $C(K)+E(K)-C(1)$ is approximately equal to 0.

However, the evidence in Tables 2–4 is inadequate to prove the advantage of the NMC, as no criterion is found to consistently perform better than the rest in all cases. In order to answer the question that whether NMC is the most advantageous criterion for the selection issue of the normal mixture model, more experiments are taken to compute the probability of these five criteria to make the correct decision.

Table 4 Sample 3 with data generated from $1/3N(0,1^2)+1/3N(2,1^2)+1/3N(4,1^2)$

Sample size	K	K_M	S_M	$\ln L$	Selection criteria				
					AIC	AICc	BIC	NEC	NMC
n=50	1	0.23	0.83	105.25	214.51	214.76	**218.33**	0.00	0.83
	2	0.45	0.37	101.56	**213.12**	**214.48**	222.68	1.50	**0.45**
	3	0.87	0.93	100.70	217.40	220.91	232.69	1.19	0.93
	4	0.92	0.54	100.46	222.93	229.88	243.96	1.17	0.92
	5	0.90	0.89	100.27	228.53	240.53	255.30	**1.09**	0.90
n=100	1	0.05	0.81	199.92	403.84	**403.96**	**409.05**	0.00	0.81
	2	0.68	0.36	196.68	**403.37**	404.00	416.39	1.17	0.68
	3	0.35	0.28	195.25	406.51	408.09	427.35	1.54	**0.35**
	4	0.38	0.81	194.35	410.70	413.70	439.35	1.24	0.81
	5	0.19	0.91	194.40	416.80	421.74	453.27	**1.11**	0.91
n=200	1	0.08	0.88	411.21	826.41	826.47	833.01	0.00	0.88
	2	0.35	0.14	403.00	**816.00**	**816.31**	**832.49**	1.18	0.35
	3	0.20	0.34	402.77	821.55	822.30	847.94	1.08	**0.34**
	4	0.13	0.42	402.67	827.34	828.75	863.62	1.06	0.42
	5	0.14	0.45	402.60	833.21	835.48	879.38	**1.05**	0.45
n=300	1	0.02	0.75	629.48	1 262.96	1 263.00	**1 270.37**	0.00	0.75
	2	0.68	0.22	621.70	**1 253.39**	**1 253.60**	1 271.91	1.11	0.68
	3	0.34	0.37	621.03	1 258.05	1 258.55	1 287.68	1.07	**0.37**
	4	0.57	0.76	619.13	1 260.27	1 261.19	1 301.01	1.06	0.76
	5	0.70	0.73	619.10	1 266.19	1 267.67	1 318.04	**1.03**	0.73

Note: The minimum value of AIC, AICc, BIC, NEC and NMC is marked in bold respectively. For the NEC, its value becomes NaN when the value of $C(K)+E(K)-C(1)$ is approximately equal to 0.

The probability of these five criteria in correctly estimating the true number of components is tabulated in Table 5. Two major conclusions are procured. The first is that, the NEC performs better even if n is small or $N(K)$ is large. Less than half of the time, AIC/AICc, BIC and NEC correctly estimate the true number of components with a small sample size $n=50,100,150$. Particularly, in the case of the Sample 1 size equaling 150, the probability in correctly recovering the true number of components for each of the above criterion is 0.38, 0.4, 0.04 and 0.23. This means that out of 150 simulated data set, AIC/AICc, BIC and NEC respectively have correctly identified the true value of K 38, 40, 4 and 23 times. However, as regards the NMC, this quantity has reached 56. With a true value of K equaling 3, the NMC also impacts the advantages over other criteria. The results of Sample 3 in Table 5 shows that, with a sample size 150, the NMC has correctly identified the true value of K 24 times while the times of the correct selection of AIC/AICc, BIC and NEC are 20, 19, 0 and 12. It should be pointed out that the performance of the AIC is second only to our criterion.

The second finding revealed by Table 5 is that our

moment-based criterion performs better and better as the sample size grows. In fact, appearances are alike on this for these criteria. For example, with a sample size of 300, the probability concerned for each of the AIC, AICc, BIC and NMC in Sample 1 has reached a value of 0.64, 0.63, 0.07 and 0.51. However, as an exception, the NEC has failed to follow this rule.

Table 5 Probability of correctly estimate the true number of components K

	Sample size	Selection criteria				
		AIC	AICc	BIC	NEC	NMC
Sample 1	n=50	0.29	0.19	0.01	0.53	0.37
	n=100	0.32	0.33	0.05	0.33	0.46
	n=150	0.38	0.4	0.04	0.23	0.56
	n=200	0.53	0.53	0.06	0.2	0.54
	n=300	0.64	0.63	0.07	0.04	0.51
Sample 2	n=50	0.18	0.22	0.1	0.54	0.24
	n=100	0.36	0.33	0.05	0.3	0.42
	n=150	0.3	0.31	0.03	0.12	0.4
	n=200	0.5	0.48	0.05	0.09	0.51
	n=300	0.5	0.5	0.04	0.14	0.5
Sample 3	n=50	0.1	0.05	0	0.22	0.11
	n=100	0.15	0.11	0.01	0.1	0.19
	n=150	0.2	0.19	0	0.12	0.24
	n=200	0.17	0.17	0.01	0.06	0.22
	n=300	0.16	0.16	0.01	0.07	0.29

Note: In this experiment, 100 series of simulated data are generated. The probability of correctly estimate is calculated as times of correctly estimate/100. Each of the probability in this table takes a value range from zero to one inclusively, with 0 implying that the criterion chooses none of the correct model and 1 presents completely excellent selection ability of the criterion.

Table 6 Probability of underestimate of the true number of components K

	Sample size	Selection criteria				
		AIC	AICc	BIC	NEC	NMC
Sample 1	n=50	0.58	0.76	0.98	0.14	0.61
	n=100	0.53	0.59	0.95	0.06	0.47
	n=150	0.52	0.54	0.96	0.02	0.3
	n=200	0.4	0.41	0.94	0.01	0.26
	n=300	0.31	0.33	0.93	0	0.27
Sample 2	n=50	0.56	0.7	0.88	0.08	0.76
	n=100	0.55	0.59	0.95	0	0.37
	n=150	0.61	0.62	0.96	0	0.4
	n=200	0.44	0.46	0.95	0.02	0.38
	n=300	0.42	0.44	0.96	0	0.22
Sample 3	n=50	0.82	0.94	1	0.51	0.86
	n=100	0.81	0.86	0.99	0.27	0.56
	n=150	0.78	0.8	1	0.28	0.48
	n=200	0.81	0.81	0.99	0.13	0.44
	n=300	0.81	0.81	0.99	0.12	0.28

Note: The probability of underestimate is calculated as times of underestimate/100, and we refer to the situation whereby a criterion selects a smaller number of components than the true ones as under estimate.

Further investigation is regarding the under estimation and over estimation of these criteria. Table 6 and Table 7 reveal that the AIC/AICc, BIC and NEC tend to underestimate the true number of components while the NEC is diametrically opposed. For the NMC, the probability of under estimation falls in the range of 0.27 and 0.61 inclusively. The probability of the under estimation, on the other hand, reduces as the sample size grows. However, as researchers hardly have large samples, identifying the criterion that minimizes the probability of the under estimation may be a more practically effort. From this respect, it is observed from Table 5 that the NMC performs consistently better than other criteria, especially in small sample cases.

Table 7 Probability of overestimate of the true number of components K

	Sample size	Selection criteria				
		AIC	AICc	BIC	NEC	NMC
Sample 1	n=50	0.13	0.05	0.01	0.33	0.02
	n=100	0.15	0.08	0	0.61	0.07
	n=150	0.1	0.06	0	0.75	0.14
	n=200	0.07	0.06	0	0.79	0.2
	n=300	0.05	0.04	0	0.96	0.22
Sample 2	n=50	0.26	0.08	0.02	0.38	0
	n=100	0.09	0.08	0	0.7	0.21
	n=150	0.09	0.07	0.01	0.88	0.2
	n=200	0.06	0.06	0	0.89	0.11
	n=300	0.08	0.06	0	0.86	0.28
Sample 3	n=50	0.08	0.01	0	0.27	0.03
	n=100	0.04	0.03	0	0.63	0.25
	n=150	0.02	0.01	0	0.6	0.28
	n=200	0.02	0.02	0	0.81	0.34
	n=300	0.03	0.03	0	0.81	0.43

Note: The probability of underestimate is calculated as times of overestimate/100, and we refer to the situation whereby a criterion selects a larger number of components than the true ones as over estimate.

5. Conclusions

Mixture models, especially the mixture of normals, have got ever broader use in fitting asset returns. As assessing the number of components plays an important role in a mixture model, this paper attempts to provide a more targeted criterion which helps select the appropriate model for normal mixture models. Compared with the AIC, the NEC heavily penalizes the local goodness of fit of all the components in the mixture model. Our criterion additionally provides a test of a model in an absolute sense of testing its quality. Compared with the well-known criteria (AIC/AICc, BIC and NEC), the practical behavior of our moment-based criterion is superior to other criteria in the problem of determining the number of components. Moreover, the performance of the NMC becomes better and better as the sample size grows.

One limitation of our work is that the proposed method is applied only for univariate normal mixture models. For further study, the criterion can be extended to multivariate normal mixture models. Another possible extension of this work is to optimize the form for the measure of the local

goodness of fit. In this work, we simply make use of the maximum value of kurtosis and skewness which may not be suitable for data set with different sample sizes or sample shapes. One possible solution to overcome this problem is to discover the law of variances for the statistics of skewness and kurtosis.

References

[1] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* (*Methodological*), 1977, 39(1): 1 – 38.

[2] G. McLachlan, D. Peel. *Mixtures of factor analyzers*. USA: Finite Mixture Models, 2000: 238 – 256.

[3] H. Akaike. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 1973, 60(2): 255 – 265.

[4] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 1978, 6(2): 461 – 464.

[5] B. S. Everitt. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, 1981, 16(2): 171 – 180.

[6] G. J. McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, 1987, 36(3): 318 – 324.

[7] Z. D. Feng, C. E. McCulloch. On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variances. *Biometrics*, 1994, 50(4): 1158 – 1162.

[8] Y. Lo, N. R. Mendell, D. B. Rubin. Testing the number of components in a normal mixture. *Biometrika*, 2001, 88(3): 767 – 778.

[9] H. Chen, J. Chen, J. D. Kalbfleisch. Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society*, 2004, 66(1): 95 – 115.

[10] S. Richardson, P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B* (*Methodological*), 1997, 59(4): 731 – 792.

[11] M. Stephens. Bayesian Analysis of mixture models with an unknown number of components- an alternative to reversible jump methods. *Annals of Statistics*, 2000, 28(1): 40 – 74.

[12] A. Nobile. On the posterior distribution of the number of components in a finite mixture. *Annals of Statistics*, 2004, 32(5): 2044 – 2073.

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

[14] C. Biernacki, G. Govaert. Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 1997: 451 – 457.

[15] E. Susko. Weighted tests of homogeneity for testing the number of components in a mixture. *Computational Statistics & Data Analysis*, 2003, 41(3): 367 – 378.

[16] P. Schlattmann. Estimating the number of components in a finite mixture model: the special case of homogeneity. *Computational Statistics & Data Analysis*, 2003, 41(3/4): 441 – 451.

[17] K. Roeder. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 1994, 89(426): 487 – 495.

[18] S. M. Stigler. Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{\mathrm{b}}1$. *Biometrika*, 1973, 60(3): 613 – 622.

[19] K. O. Bowman, L. R. Shenton. Omnibus test contours for departures from normality based on $\sqrt{\mathrm{b}}1$ and b2. *Biometrika*, 1975, 62(2): 243 – 250.

[20] D. Seigmund. Tests for departure from normality: comparison of powers. *Biometrika*, 1977, 64(2): 231 – 246.

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

[22] A. Cutler, M. P. Windham. Information-based validity functionals for mixture analysis. *Proc. of the 1st US/Japan conference on the frontiers of statistical modeling: An informational approach*. Springer, 1994: 149 – 170.

[23] B. Hamparsum, S. L. Stanley. Multi-sample cluster analysis using Akaike's information criterion. *Annals of the Institute of Statistical Mathematics*, 1984, 36(1): 163 – 180.

[24] A. P. Liavas, P. Regalia. On the behavior of information theoretic criteria for model order selection. *IEEE Trans. on Signal Processing*, 2001, 49(8): 1689 – 1695.

[25] M. Miloslavsky, M. J. V. D. Laan. Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics & Data Analysis*, 2003, 41(3/4): 413 – 428.

[26] K. Yamaoka, T. Nakagawa, T. Uno. Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations. *Journal of Pharmacokinetics and Biopharmaceutics*, 1978, 6(2): 165 – 175.

[27] D. R. Anderson, K. P. Burnham, G. C. White. AIC model selection in overdispersed capture-recapture data. *Ecology*, 1994, 75(6): 1780 – 1793.

[28] D. Posada, K. A. Crandall. MODELTEST: testing the model of DNA substitution. *Bioinformatics*, 1998, 14(9): 817.

[29] Y. Dodge. *The dictionary of statistical terms*. Oxford: Oxford University Press, 2003.

[30] K. Pearson. Das Fehlergesetz und Seine Verallgemeinerungen Durch Fechner und Pearson. *Biometrika*, 1905, 4(1/2): 169 – 212.

[31] C. M. Bishop. *Pattern recognition and machine learning*. Berlin Heidelberg: Springer, 2006.

[32] A. Farag, A. S. El-Baz, G. Gimel'Farb. Precise segmentation of multimodal images. *IEEE Trans. on Image Processing*, 2006, 15(4): 952 – 968.

[33] E. S. Pearson. Tables of percentage points of $\sqrt{\mathrm{b}}1$ and b2 in normal samples: a rounding off. *Biometrika*, 1965, 52(1/2): 282 – 285.

[34] E. S. Pearson, H. O. Hartley. *Biometrika tables for statisticians*. 3rd ed. Cambridge, U.K.: Cambridge University Press, 1966.

[35] N. Vlassis, A. Likas. A kurtosis-based dynamic approach to Gaussian mixture modeling. *IEEE Trans. on Systems, Man and Cybernetics – Part A: Systems and Humans*, 1999, 29(4): 393 – 399.

[36] K. P. Burnham, D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Berlin Heidelberg: Springer Science & Business Media, 2002.

Biographies

Yimin Zhou was born in 1988. She is a Ph.D. student in School of Economics and Management, Beihang University. Her research interests are financial statistics, asset pricing, financial contagion and commodity market.
E-mail: zhouymmt@126.com

Liyan Han was born in 1955. He is a professor in School of Economics and Management, Beihang University. His research interests are financial engineering, financial statistics, investment management, financial market and intangible assets management.
E-mail: hanly@buaa.edu.cn

Dan Wang was born in 1993. He is a financial researcher in School of Economics and Management, Beihang University. He is taking in charge of programming development in current research group under the instruction of Professor Han Liyan. His research interests include asset pricing and asset allocation.
E-mail: william0423@hotmail.com

Libo Yin was born in 1988. She is an associate professor in School of Finance, Central University of Finance and Economics. Her research interests are financial engineering, asset pricing and commodity market.
E-mail: yinlibowsxbb@126.com