Journal of Scientific Research & Reports

21(1): 1-9, 2018; Article no.JSRR.9004
ISSN: 2320-0227

Information Criteria's Performance in Finite Mixture Models with Mixed Features

Jaime R. S. Fonseca[1*]

[1]*Universidade de Lisboa, Instituto Superior de Ciências Sociais e Políticas, Rua Almerido Lessa, Lisboa, Portugal.*

Author's contribution

The sole author designed, analyzed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JSRR/2018/9004
Editor(s):
(1) Francisco Torrens, Institute for Molecular Science, University of Valencia, Paterna Institutes Building, Valencia, Spain.
Reviewers:
(1) Vasil Simeonov, University of Sofia, Bulgaria.
(2) Peter Trebuna, Technical university of Kosice, Slovak republic.
Complete Peer review History: http://www.sciencedomain.org/review-history/27051

Original Research Article

Received 11 January 2014
Accepted 10 February 2014
Published 05 November 2018

ABSTRACT

Aims: This study is intended to determine which information criterion is more appropriate for mixture model selection when considering data sets with both categorical and numerical clustering base variables (mixed case).
Study Design: In order to select among eleven information criteria which may support the selection of the correct number of clusters we conduct a simulation study. The generation of mixtures of both multinomial and multivariate normal data supports the proposed analysis.
Place and Duration of Study: Simulation: Instituto Superior de Ciências Sociais e Políticas (ISCSP), Universidade de Lisboa, 2012.
Methodology: The experimental design controls the number of normal (two and four) and multinomial (two and four) variables, the number of clusters (two, four and six), the level of clusters separation (ill and well), and for sample size we use three levels (400, 1200, 2000).
Thus, data sets were simulated with the following factors: two levels for the number of normal variables; two levels for the number of multinomial variables; two levels of segment separation, and three levels of number of clusters. Thus, the simulation plan uses a $2^3 \times 3^2$ factorial design with 72 cells. So with five replications (data sets) per cell, we generate a total of $2^3 \times 3^2 \times 5 = 360$ experimental data sets.
Results: The best overall performance goes to AIC3 (58%), followed by AICu (56%) and AICc (54%). About AIC_3, AICu and AICc, these criteria evidence a good compromise between *underfit*

**Corresponding author: E-mail: jaimefonseca@iscsp.ulisboa.pt;*

and *overfit*: AIC3, AIC and AICu underfit 11, 7 and 14%, and they overfit on 21, 18 and 18%, respectively. The most underfiting criterion is NEC, with 48%, and the most overfiting one is AIC, with 42%.
Conclusion: We run Friedman test for all the criteria, to test the null hypothesis that all the eleven populations distributions functions are identical We reject the null hypothesis and we accept the alternative (Monte Carlo p-value=0.000). Thus, we conclude that criteria performance was not identical for the eleven criteria, and we make multiple comparisons.
We concluded that AIC_3 and AICc have significantly different performances, but AIC_3 and AICu have similar performances. Thus we may conclude that AIC_3 and AICu are the best information criteria for selecting the true number of clusters when dealing with finite mixture models, mixed data and information criteria for model selection.

Keywords: Quantitative methods; cluster analysis; statistical and probabilistic modelling; finite mixture model; information theoretical criteria; simulation experiments; mixed variables.

1. INTRODUCTION

Finite mixture models (FMM) have proven to be powerful tools for clustering analysis, namely in the domain of social and behavioural science data [1]. There have been numerous proposals of information criteria for the model selection in FMM (model selection). Some of them traditionally proposed for regression models, others for finite mixture models.In the context of clustering, applications are common which consider not only categorical clustering base variables and/or numeric but, frequently, mixed (categorical, and numeric) clustering base variables.

The objective of this research is to address the performance of specific theoretical information criteria (for FMM selection) when dealing with mixed clustering variables. A simulation study is conducted for this purpose which results may help to support future analysts' decisions concerning the choice of particular information criteria when dealing with specific clustering applications.

This paper is organized as follows: in section 2, we define notation and review finite mixture models, clustering analysis through finite mixture models and we review previous work on the EM algorithm for the estimation of mixture models; in section 3, we review several model selection criteria proposed to estimate the number of components of a mixture (number of clusters); in section 4, we present the proposed simulation based approach to compare the performance of eleven information criteria; in section 5 we report on simulation results, and finally, in section 6 we present some concluding remarks.

2. CLUSTERING VIA FINITE MIXTURE MODELS

For illustratting the use of finite mixture models in the field of cluster analysis, see for instance [2]. FMM assume that parameters of a statistical model of interest differ across unobserved or finite mixture and they provide a useful means for clustering observations. In FMM, clustering base variables are assumed to be described by a different probability (density) distribution in each unobserved segment. These probability (density) functions typically belong to the same family and differ in the corresponding parameters' values.

This approach to clustering offers some advantages when compared with other techniques: provides unbiased segment memberships' estimates and consistent estimates for distributional parameters [3]; it provides means to select the number of clusters; it is able to deal with diverse types of data (different measurement levels [4]. In order to present FMM we give some notation below.

The mixture model approach to clustering assumes that data are from a mixture of an unknown number S of clusters in some unknown proportions, $\lambda_1, \cdots, \lambda_S$. The data $\underline{y} = (\underline{y}_1, ..., \underline{y}_n)$ are assumed to be a p-dimensional sample of size *n*, from a probability distribution with density

$$f(\underline{y}_i \mid \underline{\psi}) = \sum_{s=1}^{S} \lambda_s \, f_s(\underline{y}_i \mid \underline{\theta}_s) \tag{1}$$

where the mixing probabilities satisfy

$$\lambda_s > 0 \text{, s = 1, ..., S, and } \sum_{s=1}^{S} \lambda_s = 1 \tag{2}$$

The complete set of parameters we need to estimate, to specify the mixture model is

$$\underline{\psi} = \{\underline{\lambda}, \Theta\},\ \underline{\lambda} = \{\lambda_1, \cdots, \lambda_{S-1}\},\ \text{and}\ \ \Theta = \{\underline{\theta}_1, \cdots, \underline{\theta}_S\}$$

n	sample size
S	number of (unknown) clusters
$(Y_1, \cdots, Y_P)$	P clustering base variables
$(y_1, \cdots, y_n)$	measurements on variables $Y_1, \cdots, Y_P$
y_i	measurements of case *i* on $Y_1, \cdots, Y_P$
$\underline{z} = (\underline{z}_1, \ldots, \underline{z}_n)$	clusters-label vectors
z_i	vector indicating segment membership
$\underline{x} = (y, \underline{z})$	complete data
p(d)f	probability (density) function
θ_s	all p(d)f parameters of the s^{th} segment
$\Theta = (\theta_1 \ldots \theta_S)$	vector of parameters, without weights
λ	vector of weights (mixing proportions)
τ_{is}	Conditional probability
$\psi = (\lambda, \Theta)$	vector of all unknown parameters
$\hat{\psi} = (\hat{\lambda}, \hat{\Theta})$	estimate of all unknown parameters
L	likelihood function, L(ψ)
LL	log-likelihood function, log L(ψ)
LL_C	complete-data log-likelihood function
n_ψ	number of mixture model parameters

3. MODEL SELECTION

Selection of FMM solutions may rely on multiple Information Criteria, which turns opportune the specific issue concerning the selection among the criteria themselves Table 1.

On the other hand, applications are common in the clustering domain, which refer to base clustering variables of different types (different levels of measurement). This fact turns relevant the hypothesis of the existence of a relationship between information criteria's performance and the type of base variables' measurement level (categorical, numerical or mixed). In this study we propose an approach for evaluating several Information Criteria's performances, taking into account theirs relationship with base variables' measurement levels, for mixed case.

Information Criteria look for a trade-off between the precision of the ML estimate $\underline{\hat{\psi}}$, and the complexity of parameterization or model parsimony. They all balance fitness (trying to maximize the likelihood function) and parsimony (using penalties associated with measures of model complexity), trying to avoid overfit. Furthermore, fitting a model with a large number of clusters requires estimation of a very large number of parameters and a consequent loss of precision in these estimates [5].

The general form of information criteria is as follows

$$-2\log L(\hat{\psi}) + C, \tag{3}$$

where the first term is the negative logarithm of the maximum likelihood which decreases when the model complexity increases; the second term or penalty term penalizes too complex models, and increases with the model number of parameters. Thus, the selected mixture model should evidence a good trade-off between good description of the data and the model number of parameters.

AIC [6] and AIC_3 [7] are measures of model complexity associated with some criteria see Table 1 that only depend on the number of parameters; some other measures depend on both the number of parameters and the sample size, as AICc [8], AICu [9], CAIC [10], and BIC/MDL, [11,12]; others depend on entropy, as CLC [13], and NEC [14]; some of them depend on the number of parameters, sample size, and entropy, as ICL-BIC [15], and AWE [16]; L [17] depends on the number of parameters, sample size and mixing proportions, λ_S .

Table 1. Some information criteria for model selection on Finite Mixture Models

Criteria	Definition
AIC	$-2LL + 2n_\psi$
AIC3	$-2LL + 3n_\psi$
AICc	$AIC + (2n_\psi(n_\psi + 1))/(n - n_\psi - 1)$
AICu	$AICc + n\log(n/(n - n_\psi - 1))$
CAIC	$-2LL + n_\psi(1 + \log n)$
BIC/MDL	$-2LL + n_\psi \log n$
CLC	$-2LL + 2EN(S)$
ICL_BIC	$BIC + 2EN(S)$
NEC	$NEC(S) = EN(S)/(L(S) - L(1))$
AWE	$-2LL_c + 2n_\psi(3/2 + \log n)$
L	$-LL - (n_\psi/2)\sum \log(n\lambda_s/12)$

4. METHODOLOGY

4.1 Target Models

In the present work we specifically refer to information criteria presented in Table 1, which have been referred previously. All are currently in use for the estimation of FMM.

Several model selection criteria have been used in order to decide on the number of clusters that are present in data, when *a priori* knowledge does not exist. However, there is no indication concerning the selection of the selection criteria themselves.

In this paper we try to establish a relationship between type of clustering variables - mixed case - and the performance of information-based criteria. We also illustrate other factors that may influence the outcome, such as clusters' separation and sample size.

When some of the variables are continuous and some are categorical, we regard data as a random sample from the mixture model

$$f(\underline{y}_i \mid \underline{\psi}) = \sum_{s=1}^{S} \lambda_s \sum_{p=1}^{P} f_s(\underline{y}_{ip} \mid \underline{\theta}_{sp})$$

where $f_s(y_{ip} \mid \theta_{sp})$ is $N(\mu_{sp}, \sigma_{sp}^2)$, p = 1,…,k, for each one of the k continuous variables, and $\mathrm{Mult}_{C_p}(1; \theta_{sp1}, ..., \theta_{spc})$, for the (P-k) categorical variables, with C_p categories (see, *e.g.*, [18]).

4.2 Simulation Experiments

Thus, data sets were simulated with the following factors: two levels for the number of normal variables; two levels for the number of multinomial variables; two levels of segment separation, and three levels of number of clusters see Table 2.

Table 2. Factorial design for mixed variables

Factors		Number of levels
Normal variables	2; 4	2
Multinom. variables	2; 4	2
Multinom. categories	5	1
Separation levels	Well; ill	2
Number of clusters	2; 4; 6	3
Dimension	400; 1200; 2000	3
Factorial design		2^3*3^2

Thus, the simulation plan uses a $2^3 \times 3^2$ factorial design with 72 cells. So with five replications (data sets) per cell, we generate a total of $2^3 \times 3^2 \times 5 = 360$ experimental data sets.

In order to avoid local optima in the generated FMM estimation process, the EM algorithm is repeated 50 times with random starting centres, and the best solution for ML and model selection results are kept, with a tolerance level of 10^{-6} (the criterion for convergence of EM: difference between log-likelihood being smaller than 10^{-6}).

5. RESULTS

The results of the comparative experimental evaluation of the performance of eleven information criteria based on the proposed simulation study are presented below. They illustrate the relationship between the performance of information criteria and the clustering base variables' type.

Table 3 shows the percentage of cases (simulated experiments) each criterion determines the original (*true*) number of clusters (*fit*), across the used factors, and also the overall percentages *underfit* (percentage of times each criterion selects a model with a few number of clusters) and *overfit* (percentage of times each criterion selects a model with a high number of clusters).

The best overall performance goes to AIC3 (58%), followed by AICu (56%) and AICc (54%).

Concerning the number of clusters, we can see that NEC (98%) AWE (93%) and L (89%) outperform all the other criteria, for S=2; the same happens with AIC3 (70%), AICu (67%) and AICc (64%) for S=4; For S=6, AICc (30%) and AIC3 (26%), outperform the other criteria.

About AIC_3, AICu and AICc, these criteria evidence a good compromise between *underfit* and *overfit*: AIC3, AIC and AICu underfit 11, 7 and 14%, and they overfit on 21, 18 and 18%, respectively. The most underfiting criterion is NEC, with 48%, and the most overfiting one is AIC, with 42%.

As far as the type of clustering base variables is concerning, we can characterize the situation as follows: for clustering base variables with two normal and two categorical variables, the criteria with best performance are AIC3 (73%), AICu (68%) and AICc (66%); for clustering base variables with two normal and four categorical variables, the criteria with best performance are AIC3 (62%), AIC (60%) and AICc and AICu (with 58 and 57%, respectively); for clustering base variables with four normal and two categorical variables, the criteria with best performance are AWE (57%), L (50%) and AICu (49%); for clustering base variables with four normal and four categorical variables, the criteria with best performance are AIC3 and CAIC (exaquae with 54%), and L (52%).

Sample size doesn't show great influence on the information criteria performance, because AIC3, AICc and AICu show consistently good performance, for n=400, n=1200 and n=2000.

About clusters separation we noted that it is worthwhile to consider the three levels; it is enough the well-separated and ill-separated consideration, and empirically we consider well-separated clusters for $E_s \geq 0.95$. We see, in this study, that the classification criteria AWE with 73% and AIC3 with 68% outperform the other criteria for well-separated clusters; otherwise, for ill-separated clusters, AIC with 54% and AIC3 and AICc (exaquae, with 48%) was the criterion with the best performance.

Sample size doesn't show great influence on the information criteria performance, because AIC3, AICc and AICu show consistently good performance, for n=400, n=1200 and n=2000.

About clusters separation we noted that it is worthwhile to consider the three levels; it is enough the well-separated and ill-separated consideration, and empirically we consider well-separated clusters for $E_s \geq 0.95$.

We see, in this study, that the classification criteria AWE with 73% and AIC_3 with 68% outperform the other criteria for well-separated clusters; otherwise, for ill-separated clusters, AIC with 54% and AIC_3 and AICc (exaquae, with 48%) was the criterion with the best performance.

Table 3. Results of simulation study experiments

Factors		**BIC**	**AIC**	**AIC_3**	**AICc**	**AICu**	**CAIC**
	Fit	48	52	58	54	56	44
Overall	*Underfit*	26	2	11	7	14	30
	Overfit	15	42	21	18	18	13
	2	81	64	80	71	80	83
Number of Clusters	*4*	58	58	70	64	67	46
	6	8	41	26	30	20	4
2 Normals, 2 Categorical		53	61	73	66	68	44
2 Normals, 4 Categorical		47	60	62	58	57	47
4 Normals, 2 Categorical		48	44	48	44	49	44
4 Normals, 4 Categorical		47	46	54	47	46	54
Dimension	400	50	62	62	59	57	45
	1200	46	45	54	50	53	43
	2000	48	52	58	54	56	44
well-separated		66	50	68	61	66	61
Ill-separated		30	54	48	48	46	26

Table 3. Results of simulation study experiments (cont.)

		CLC	ICL-BIC	NEC	L	AWE
Overall	*Fit*	46	43	36	40	47
	Underfit	19	31	48	43	38
Number of Clusters	*Overfit*	28	31	4	11	2
	2	73	79	98	89	93
	4	46	44	8	0	41
	6	15	4	0	23	1
2 Norm. and 2 Categ.		53	46	31	33	38
2 Norm. and 4 Categ.		47	42	30	33	37
4 Norm. and 2 Categ.		44	42	42	50	57
4 Norm. and 4 Categ.		39	44	40	52	46
Dimension	400	49	42	35	36	38
	1200	43	41	38	39	48
	2000	46	43	36	40	47
well-separated		57	61	55	56	73
Ill-separated		35	25	16	24	21

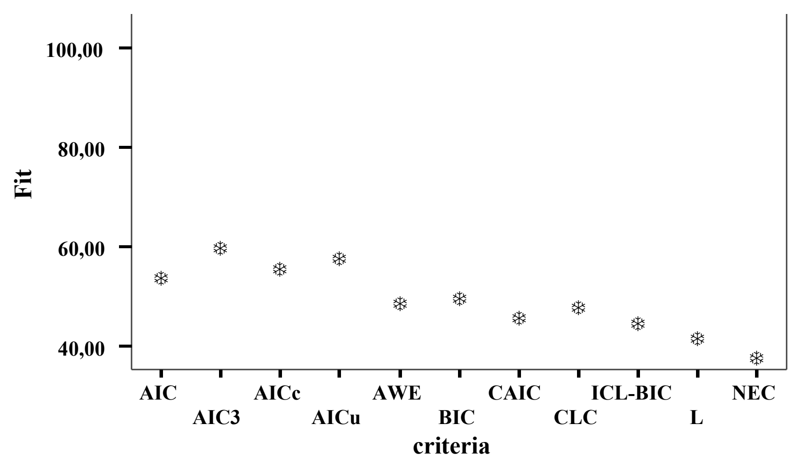

Fig. 1. The true number of clusters recovery (Fit), in percent

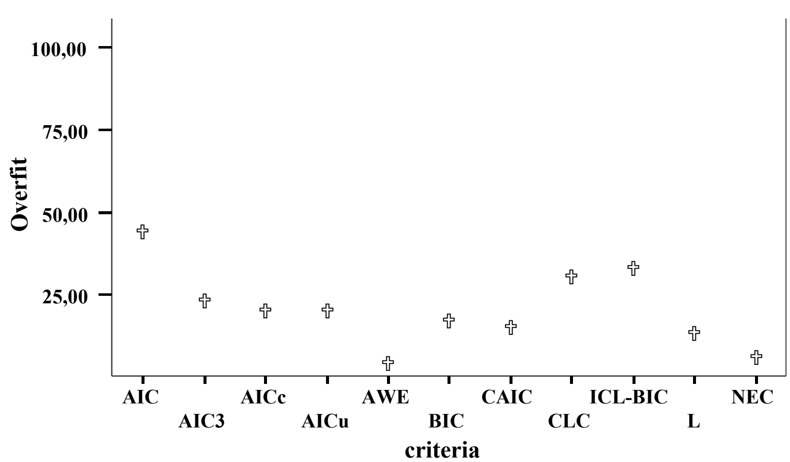

Fig. 2. Criteria selecting models with more clusters (overfit), in %)

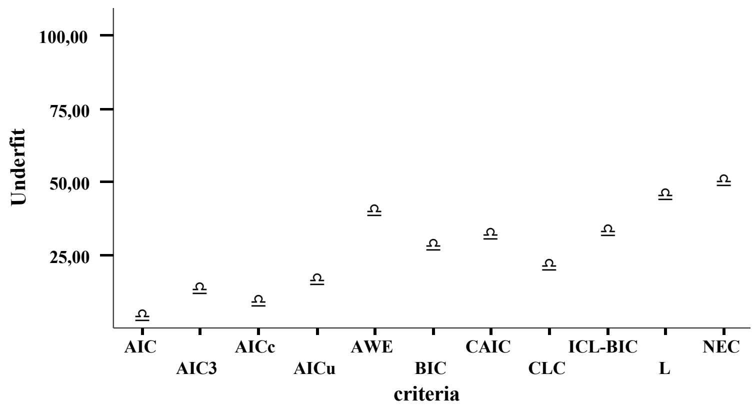

Fig. 3. Criteria selecting models with less clusters (underfit), in %)

Table 4. Matrix of rank order differences for multiple comparisons

Criteria		1	2	3	4	5	6	7	8	9	10	11
		90,5	89,5	128,5	102,5	109,0	70,5	65,0	48,0	26,5	55,0	73,0
1-BIC	90,5	0,0										
2-AIC	89,5	-1,0	0,0									
3-AIC_3	128,5	**38,0**	**39**	0,0								
4-AICc	102,5	12,0	13	**-26**	0,0							
5-AICu	109,0	18,5	19,5	-19.5	6,5	0,0						
6-CAIC	70,5	-20	-19	**-58**	**-32**	**-38,5**	0,0					
7-CLC	65,0	**-25,5**	**-24,5**	**-63,5**	**-37,5**	**-44**	-5,5	0,0				
8-ICL_BIC	48,0	**-42,5**	**-41,5**	**-80,5**	**-54,5**	**-61**	**-23**	-17	0,0			
9-NEC	26,5	**-64**	**-63**	**-102**	**-76**	**-82,5**	**-44**	**-39**	-22	0,0		
10-L	55,0	-35,5	**-34,5**	**-73,5**	**-47,5**	**-54**	-16	-10	7	**28,5**	0,0	
11-AWE	73,0	-17,5	-16,5	**-55,5**	**-29,5**	**-36**	2,5	8	**25**	**46,5**	18	0,0

6. DISCUSSION AND CONCLUSION

This study indicates the existence of a relationship between the performance of some information criteria and the mixed clustering variables which are considered for clustering with FMM.

In the present study, we conclude that AIC_3, AICu and AICc are preferable for data sets with both categorical and continuous clustering base variables (they select the right model in 62% of the cases). It is quite remarkable the good performance of AICu criterion, introduced in model selection for mixture model in [19].

From Fig. 1 which illustrates *fit* (percentage of the true structure recovery) we can see that AIC_3 have the best performance for mixed models.

Fig. 2 (criteria select models with less clusters, in %) shows that AIC almost never *underfits*; next, we have AICc, AIC_3 and AICu. Otherwise, we have NEC, L and AWE as the criteria with most underfitting.

As we can see from Fig. 3 (criteria select models with more clusters, in %), AIC is the criterion which *overfits* more, followed by ICL-BIC and CLC. On the other side we have criteria such as AWE and NEC, which almost never overfit.

Finally, in order to compare the criteria performances, we run Friedman tests, because the data consist of *b* mutually independent k-variate random variables (X_{i1},…,X_{ik}), called *b* blocks, *i*=1,…,*b*; the random variable X_{ij}.is in block *i* (the factors in analysis) and is associated with treatment *j* (the criteria we use).

We run Friedman test for all the criteria in Table 3, to test the null hypothesis that all the eleven populations distributions functions are identical, against the alternative which states that at least one of the populations tends to yield larger observations than at least one of the other populations. We reject the null hypothesis and we accept the alternative (Monte Carlo p-value = 0.000). Thus, we conclude that criteria performance was not identical for the eleven criteria in Table 3, and we make multiple comparisons.

Criteria *i* and *j* are considered to have different performance if the inequality

$$|S_i - S_j| > t_{(b-1)(k-1);1-\frac{\alpha}{2}} \left[\frac{2b(F_1 - F_2)}{(b-1)(k-1)} \right]^{\frac{1}{2}}$$

is satisfied, where $t_{(b-1)(k-1);1-\frac{\alpha}{2}}$ is the value of distribution *t* with (b-1)(k-1) degrees of freedom, and Rj, F_1 and F_2 are given by

$$F_1 = \sum_{i=1}^{b} \sum_{j=1}^{k} \left[R(X_{ij}) \right]^2 \text{ and } F_2 = \frac{1}{b} \sum_{j=1}^{k} R_j^2 \text{, with}$$

$$R_j = \sum_{i=1}^{b} R(X_{ij}),$$

where $R(X_{ij})$ is the rank, from 1 to k, assigned to X_{ij} within block *i*. Because we have

$$t_{(b-1)(k-1);1-\frac{\alpha}{2}} \left[\frac{2b(F_1 - F_2)}{(b-1)(k-1)} \right]^{\frac{1}{2}} = 22.6,$$

we can conclude, from Table 4 values, that because we have $|R_{AIC3}-R_{AICu}|$ =.19.5 < 22.6 and $|R_{AIC3}-R_{AICc}|$ =.26 > 22.6, we conclude that AIC_3 and AICc have significantly different performances, but AIC_3 and AICu have similar performances.

Thus we may conclude that AIC_3 and AICu are the best information criteria for selecting the *true* number of clusters when dealing with finite mixture models and information criteria for model selection.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Dias JG, Willekens F. Model-based clustering of sequential data with an application to contraceptive use dynamics. Mathematical Population Studies. 2005;12: 135-157.
2. Punj G, Stewart DW. Cluster analysis in marketing research: Review and suggestions for application. Journal of Marketing Research. 1983;134-148.
3. Dillon WR, Kumar A. Latent structure and other mixture models in marketing: An integrative survey and overview, chapter 9 in R.P. Bagozi (ed.). Advanced Methods of Marketing Research. 1994;352-388. Cambridge: Blackwell Publishers.
4. Vermunt JK, Magidson J. Latent class cluster analysis, in Hagenaars JA and McCutcheon AL (eds.), Applied Latent Class Analysis. 2002;89-106. Cambridge University Press.
5. Leroux BG, Puterman ML. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics. 1992;48:545-558.
6. Akaike H. Information theory and an extension of maximum likelihood principle, in Selected Papers of Hirotugu Akaike, in Proceedings of the Second International Symposium on Information Theory, Petrov BN and Caski F, eds., Akademiai Kiado, Budapest, 1973, 267-281, K. T. Emanuel Parzen, Genshiro Kitagawa, Ed., 1 ed. Texas: Springer-Verlag New York, Inc. 1973;434.
7. Bozdogan H. Proceedings of the first US/Japan conference on the Frontiers of Statistical Modeling: An Informational Approach. 1994;1:1 ed. Dordrecht: Kluwer Academic Publishers.
8. Hurvich CM, Tsai CL. Regression and Time Series Model Selection in Small Samples. Biometrika. 1989;76:297-307.
9. Mc Quarrie A, Shumway R, Tsai CL. The model selection criterion AICu. Statistics & Probability Letters. 1997;34:285-292.
10. Bozdogan H. Model selection and Akaikes's information criterion (AIC): The General Theory and its Analytical Extensions. Psycometrika. 1987;52:345-370.
11. Schwarz G. Estimating the dimenson of a model. The Annals of Statistics. 1978;6: 461-464.
12. Rissanen J. Modeling by shortest data description. Automatica. 1978;14:465-471.
13. Biernacki C. Choix de modéles en Classification, PhD Thesis. Compiègne University of Technology; 1997.
14. Biernacki C, Celeux G, Govaert G. An improvement of the NEC criterion for assessing the number of clusters in mixture model. Pattern Recognition Letters. 1999;20:267-272.
15. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering

with the integrated completed likelihood. IEEE Transactions on Pattern analysis and Machine Intelligence. 2000;22:719-725.
16. Banfield JD, Raftery AE. Model-based gaussian and Non-gaussian clustering. Biometrics. 1993;49:803-821.
17. Figueiredo MAT, Jain AK. Unsupervised Learning of Finite Mixture Models. IEEE Transactions on pattern analysis and Machine Intelligence. 2002;24:1-16.
18. Hunt LA, Basford KE. Fitting a mixture model to three-mode three-way data with categorical and continuous variables. Journal of Classification. 1999;16:283-296.
19. Fonseca JRS, Cardoso MGMS. Mixture-model cluster analysis using information theoretical criteria. Intelligent Data Analysis. 2007;11:173.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history/27051