On the Performance of Information Criteria in Latent Segment Models

Jaime R. S. Fonseca

Abstract**—**Nevertheless the widespread application of finite mixture models in segmentation, finite mixture model selection is still an important issue. In fact, the selection of an adequate number of segments is a key issue in deriving latent segments structures and it is desirable that the selection criteria used for this end are effective. In order to select among several information criteria, which may support the selection of the correct number of segments we conduct a simulation study. In particular, this study is intended to determine which information criteria are more appropriate for mixture model selection when considering data sets with only categorical segmentation base variables. The generation of mixtures of multinomial data supports the proposed analysis. As a result, we establish a relationship between the level of measurement of segmentation variables and some (eleven) information criteria's performance. The criterion AIC_3 shows better performance (it indicates the correct number of the simulated segments' structure more often) when referring to mixtures of multinomial segmentation base variables.

Keywords**—**Quantitative Methods, Multivariate Data Analysis, Clustering, Finite Mixture Models, Information Theoretical Criteria, Simulation experiments.

I. Introduction

CLUSTERING analysis' primary use in marketing has been for market segmentation, [37]. Finite mixture models (FMM) have proven to be powerful tools for clustering analysis, namely in the domain of social and behavioural science data, [16]. In this context they are commonly referred as Latent Segment Models (LSM).

There have been numerous proposals of information criteria for the selection of the number of segments of LSM (model selection).

In the context of market segmentation and social sciences in general, applications are common which consider basically categorical segmentation base variables.

The objective of this research is to address the performance of specific theoretical information criteria (for LSM selection) when dealing with the categorical segmentation base 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 segmentation applications.

This paper is organized as follows: in section II, we define notation and review finite mixture models, segmentation analysis through Latent Segments Models (mixture models) and we review previous work on the EM algorithm for the estimation of mixture models; in section III, we review several model selection criteria proposed to estimate the number of components of a mixture (number of segments); in section IV, we present the proposed simulation based approach to compare the performance of eleven information criteria; in section V we report on simulation results, and finally, in section VI we present some concluding remarks.

II. Segmentation via Latent Segments Models

[35] illustrated the use of mixture models (here referred as Latent Segments Models) in the field of cluster analysis. LSM assume that parameters of a statistical model of interest differ across unobserved or latent segments and they provide a useful means for clustering observations into segments. In LSM, segmentation 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. In order to present LSM we give some notation on Table I.

TABLE I
NOTATION

n	sample size
S	number of (unknown) segments
$(Y_1,\cdots,Y_P)$	P segmentation base variables (random variables)
$(\underline{y}_1,\cdots,\underline{y}_n)$	measurements on variables $Y_1,\cdots,Y_p$
$\underline{y}_i$	measurements vector of individual i on variables $Y_1,\cdots,Y_p$
$\underline{z}=(\underline{z}_1,...,\underline{z}_n)$	segments-label vectors
$\underline{z}_i=\left(z_{i1},...,z_{iS}\right)$	binary vector indicating segment membership
$\underline{x}=(\underline{y},\underline{z})$	complete data
p(d)f	probability (density) function
$\underline{\theta}_s$	vector of all unknown p(d)f parameters of the sth segment
$\Theta=\left(\underline{\theta}_1...\underline{\theta}_S\right)$	vector of mixture model parameters, without weights
$\underline{\lambda}=(\lambda_1,\cdots,\lambda_{S-1})$	vector of weights (mixing proportions)
τ_{is}	probability that an individual i belongs to the sth segment, given $\underline{y}_i$
$\underline{\psi}=(\underline{\lambda},\Theta)$	vector of all unknown mixture model parameters
$\underline{\hat{\psi}}=(\underline{\hat{\lambda}},\hat{\Theta})$	estimate of the vector of all unknown parameters
L	likelihood function, $L(\underline{\psi})$
LL	log-likelihood function, $\log L(\underline{\psi})$
LL_c	complete-data log-likelihood function
n_ψ	number of mixture model parameters

This approach to segmentation offers some advantages when compared with other techniques: provides unbiased segments memberships' estimates and consistent estimates for

Jaime R.S. Fonseca is with the CAPP - Centre for Public Administration and Policies – Quantitative Methods for Social and Health Sciences - ISCSP - Higher Institute of Social and Political Sciences, Technical University of Lisbon, Portugal (jaimefonseca@iscsp.utl.pt).

the distributional parameters, [17]; it provides means to select the number of segments, [34]; it is able to deal with diverse types of data (different measurement levels), [40].

The mixture model approach to segmentation assumes that data are from a mixture of an unknown number S of segments 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:

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

where the mixing probabilities satisfy:

$$\lambda_s > 0, \; s = 1, ..., S, \text{ and } \sum_{s=1}^{S} \lambda_s = 1 \qquad (2)$$

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

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

The log-likelihood function for the parameters is:

$$\log \; L(\underline{\psi}) = \sum_{i=1}^{n} \log \sum_{s=1}^{S} \lambda_s f_s(\underline{y}_i \mid \underline{\theta}_s) \qquad (3)$$

Developments in the area of multivariate mixtures are few and mostly concentrated on mixtures of multivariate normal, [7]. The particularization of mixture models for multinomial, normal multivariate and mixed models can be seen in works such as [22], [41], and [24], respectively.

When dealing with Mixture Models for segmentation purposes, we may define each complete data observation, $\underline{x}_i = (\underline{y}_i, \underline{z}_i)$, as having arise from one of the segments of the mixture (1). Values of segmentation base variables $\underline{y}_i$ are then regarded as being incomplete data, augmented by segment-label variables, z_{is}, that is, $\underline{z}_i = (z_{i1}, ..., z_{is})$ is the unobserved portion of the data; z_{is} are binary indicator latent variables, so that $z_{is} = (\underline{z}_i)s$ is 1 or 0, according as to whether $\underline{y}_i$ belongs or does not belong to the sth segment, for i = 1,…,n, and s = 1, …S.

Assuming that $\{\underline{Z}_i\}$ are independent and identically distributed, each one according to a multinomial distribution of S categories with probabilities $\lambda_1, \cdots, \lambda_S$, the complete-data log-likelihood to estimate $\underline{\psi}$, if the complete data $\underline{x}_i = (\underline{y}_i, \underline{z}_i)$ was observed, [33], is

$$\log L_c(\underline{\psi}) = \sum_{i=1}^{n} \sum_{s=1}^{S} z_{is} \{\log f_s(\underline{y}_i \mid \underline{\theta}_s) + \log \lambda_s\}$$

With the maximum likelihood approach to the estimation of $\underline{\psi}$, an estimate is provided by a suitable root of the likelihood equation:

$$\frac{\partial \log L(\underline{\psi})}{\partial \underline{\psi}} = \mathbf{O}$$

The maximum likelihood estimate (MLE) cannot be found analytically. The maximizations defining MLE are under the constraints in (2).

In order to derive meaningful results from segmentation, the mixture model must be identifiable, which simply means that an unique solution to the maximum likelihood problem must exist ([9].

Several researchers have studied consistency property of both the unknown number of segments (S) and of the other model parameters' MLE.

[27] presented one of the earliest works which refers to the consistency of MLE for the distributional parameters in a LSM and he stated that mixture model estimators are consistent and asymptotically normally distributed when segmentation variables are assumed to belong to the exponential family of distributions. Other researchers addressed this issue, like [14], [23], and [21].

The consistency of the number of segments' estimator ($\hat{S}$), was stated by [31], with a maximum-penalized-likelihood method for estimating a mixing distribution, considering information criteria AIC and BIC; they showed that this method produces a consistent estimator $\hat{S}$, in the sense of weak convergence.

This property of the number of segments' estimator was also discussed in [29], [26], [8], all concluding that the estimated number of segments to the true (despite unknown) number of segments is consistent.

Fitting finite mixture models (1) provides a probabilistic segmentation of the n entities in terms of their posterior probabilities of membership of the S segments of the mixture of distributions. Since the MLE of most of the latent segment model (1) cannot be found analytically, estimation of LSM iteratively computes the estimates of segments posterior probabilities and updates the estimates of the distributional parameters and mixing probabilities, [30].

Expectation-maximization (EM) algorithm, [15], is a widely used class of iterative algorithms for ML estimation in the context of incomplete data, e.g. fitting mixture models to observed data.

The EM algorithm proceeds by alternately applying two steps, until some convergence criterion is met.

The E-step, on the kth iteration, calculates the complete data expected log-likelihood function, given $\underline{y}$, defined by the so-called Q function where:

$$Q(\underline{\psi},\underline{\psi}^{(k)}) = \sum_{i=1}^{n}\sum_{s=1}^{S} \tau_s(\underline{y}_i;\underline{\psi}^{(k)})\{\log \lambda_s + \log f_s(\underline{y}_i;\underline{\theta}_s)\}$$

$$\hat{\tau}_{is} = \hat{\tau}_{\mathrm{s}}(\underline{y}_i \mid \underline{\psi}^{(k)}) = E(Z_{\mathrm{is}} \mid \underline{\mathrm{y}}_i;\underline{\psi}^{(k)}) = \frac{\hat{\lambda}_s^{(k)} f_s(\underline{\mathrm{y}}_i \mid \hat{\theta}_s^{(k)})}{\sum_{j=1}^{S} \hat{\lambda}_j^{(k)} f_j(\underline{\mathrm{y}}_i \mid \hat{\theta}_j^{(k)})}$$

is the membership probability of pattern $\underline{y}_i$ in segment s (posterior probability) (i=1,…,n, and s =1,…,S).

The M-step, on the (k+1)th iteration, demands the maximization of (3) with respect to $\underline{\psi}$, to update the parameter estimation, obtaining $\underline{\hat{\psi}}^{(k+1)}$.

Then, by the Bayesian rule, the ith pattern is probabilistically assigned into segment s, after algorithm EM convergence, if $\hat{\lambda}_s^{(k)} f_s(\underline{y}_i \mid \hat{\theta}_s^{(k)}) > \hat{\lambda}_{s'}^{(k)} f_{s'}(\underline{y}_i \mid \hat{\theta}_{s'}^{(k)})$, $\forall s \neq s' = 1,\ldots,S$.

Since the mixture likelihood L ($\underline{\psi}$) can never be decreased during the EM sequence,

$$L(\underline{\hat{\psi}}^{(k+1)}) \geq L(\underline{\hat{\psi}}^{(k)}),$$

it implies that $L(\underline{\hat{\psi}}^{(k)})$ converges to some L for a sequence of likelihood values bounded above. Since, typically with mixture model approach, the likelihood surface is known to have many local maxima the selection of suitable starting values for the EM algorithm is crucial, [5] or [28]. Therefore, it is usual to obtain several values of the maximized log-likelihood for each of the different sets of initial values applied to the given sample, and then consider the maximum value as the solution. Also, in order to prevent boundary solutions, the EM implementation may recur to maximum a posteriori estimates.

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

On the other hand, applications are common in the segmentation domain, which refer to segmentation base variables of different types (different levels of measurement). This fact turns relevant the issue concerning the existence of a relationship between information criteria's performance and the type of base variables' measurement level (categorical, continuous or mixed).

In the present study we propose an approach for evaluating several Information Criteria's performances, taking into account theirs relationship with base variables' measurement levels, categorical in case.

Information Criteria all balance fitness (trying to maximize the likelihood function) and parsimony (using penalties associated with measures of model complexity), trying to avoid overfit.

TABLE II
SOME INFORMATION CRITERIA FOR MODEL SELECTION ON LATENT SEGMENT MODELS

Criteria	Definition	Author
AIC	$-2LL + 2n_{\underline{\psi}}$	[1]
AIC$_3$	$-2LL + 3n_{\underline{\psi}}$	[11]
AICc	$AIC + (2n_{\underline{\psi}}(n_{\underline{\psi}}+1))/(n - n_{\underline{\psi}} - 1)$	[25]
AICu	$AICc + n\log(n/(n - n_{\underline{\psi}} - 1))$	[36]
CAIC	$-2LL + n_{\underline{\psi}}(1 + \log n)$	[10]
BIC/ MDL	$-2LL + n_{\underline{\psi}}\log n$	[39] / [38]
CLC	$-2LL + 2EN(S)$	[3]
ICL- BIC	$BIC + 2EN(S)$	[4]
NEC	$NEC(S) = EN(S)/(L(S) - L(1))$	[6]
AWE	$-2LL_c + 2n_{\underline{\psi}}(3/2 + \log n)$	[2]
L	$-LL + (n_{\underline{\psi}}/2)\Sigma \log(n\lambda_s/12) + S/2\log(n/12) + S(n_{\underline{\psi}} 1)/2$	[18]

Furthermore, fitting a model with a large number of segments requires estimation of a very large number of parameters and a consequent loss of precision in these estimates, [32].

Akaike's Information Criterion and Bayesian Information Criterion are, perhaps, the best known Information Criteria. These and some other criteria are presented in Table II.

The general form of information criteria is as follows:

$$-2\log L(\hat{\psi}) + C, \tag{4}$$

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 LSM should evidence a good trade-off between good description of the data and the model number of parameters.

The emphasis on information criteria begins with the pioneer work of [1], with the Akaike's information criterion; AIC chooses a model with S segments that minimizes (4) with $C = 2n_{\underline{\psi}}$.

Later, [10] suggested the modified AIC criterion (AIC$_3$) in the context of multivariate normal mixture models, using 3 instead of 2 as penalizing term, that is $C = 3n_{\underline{\psi}}$. When a vector parameter lies on the boundary of the parameter space (as in the case of the standard mixture problem), in comparing two models with n_ψ and n_ψ^* parameters, respectively, the likelihood ratio statistic has a non-central chi-square distribution with 2(n_ψ - n_ψ^*) degrees of freedom, instead of (n_ψ - n_ψ^*) considered in AIC. As a result, he obtained a penalization factor C = 2 $n_{\underline{\psi}}$ + $n_{\underline{\psi}}$.

Another variant of AIC, the corrected AIC (AICc), is proposed by [25], focusing on the small-sample bias adjustment (AIC may perform poorly if there are too many parameters in relation to the sample size); AICc thus selects a model with S segments that minimizes (4) with C = $2n_\psi n/(n-n_\psi-1)$.

Since AICc still tends to overfit as the sample size increases, [36] proposed a new criterion – AICu – which considers a greater penalty for overfitting, specially as the sample size increases.

The consistent AIC criterion (CAIC) with C = $n_{\underline{\psi}}$(1 + log n) was derived by [10]; it tends to select models with fewer parameters than AIC does.

The Bayesian Information Criterion (BIC) was proposed by [39]; initially proposed for linear models, it includes an adjustment for sample size and often favors a simpler model; it is intended to provide a measure of the weight of evidence favoring one model over another, [42]. It refers to C = $n_{\underline{\psi}}$ log n, and is equivalent to the MDL - Minimum Description Length, [38].

The CLC - Complete Likelihood Classification – criterion, [34], was originated from the link between the observed log-likelihood and log-classification likelihood, LLc = LL – EN(S). It considers C = 2EN(S), where the entropy term 2EN(S) penalizes poorly separated segments, with:

$$\mathrm{EN(S)} = -\sum_{i=1}^{n}\sum_{s=1}^{S}\tau_{is}\log\tau_{is}$$

In order to account for the ability of the latent segment model to give evidence for a segmentation structure of the data, [4] considered the integrated likelihood of the complete data ($\underline{x}, \underline{z}$) or Integrated Classification Likelihood criterion (ICL); an approximation, referred to as ICL-BIC by [34], chooses a model with S segments that minimizes (4) with C = EN(S) + $n_{\underline{\psi}}$ log n.

[6], suggested the improved NEC (originally introduced by [12]); because original NEC cannot be calculated for S = 1, they stated that NEC(1) = 1, and so we can choose a model with s segments for minimum NEC(s) $\leq$ 1, (2 $\leq$ s $\leq$ S); otherwise NEC declares there is no segmentation structure in the data.

[2] have suggested a Bayesian solution to the choice of the number of segments, based on an approximation of the classification likelihood, the so-called approximate weight of evidence – AWE – which penalizes more drastically complex models than BIC; so it will select more parsimonious models than BIC, except for well separated segments, and chooses a model with S segments that minimizes (4) with: $C = 2EN(S) + 2n_\psi(3/2+\log n)$.

Finally, [18] proposed the L criterion for any type of parametric mixture model for which it is possible to write an EM algorithm; this criterion chooses a model with S segments that minimizes:

$$-\mathrm{LL} + (n_{\underline{\psi}}/2)\Sigma\log(n\lambda_S/12) + S/2\log(n/12) + S(n_{\underline{\psi}}+1)/2\,.$$

AIC and AIC$_3$ are measures of model complexity associated with some criteria (see Table II) that only depend on the number of parameters; some other measures depend on both the number of parameters and the sample size, as AICc, AICu, CAIC and BIC/MDL; others depend on entropy, as CLC, and NEC; some of them depend on the number of parameters, sample size, and entropy, as ICL-BIC, and AWE; L depends on the number of parameters, sample size and mixing proportions, λ_s.

III. Methodology

In this section we present LSM referred to a mixture of categorical variables. When all the p (p = 1,...,P) segmentation base variables are categorical and the p-th variable has levels 1,..., C_p, indicator variables may be defined as follows:

$$y_{ipc} = \begin{cases} 1, \text{ if for entity } i, \text{attribute } p \text{ is at level } c \\ 0, \text{ if for entity } i, \text{attribute } p \text{ is not at level } c \end{cases}$$

for i = 1,...,n; c = 1,..., C_p ; p = 1,...,P

Let θ_{spc} be the probability that the p-th variable has level c in segment s (s = 1,...,S). The response for entity i, in level c may be considered distributed according to a multinomial model, consisting of one draw on C_p categories with probabilities $\theta_{sp1},\cdots,\theta_{spc}$, for c = 1,..., C_p, if it belongs to segment s; thus, conditional on entity i belonging to segment s,

$$\underline{y}_i \sim\triangleright \mathrm{Mult}_{C_p}(1;\theta_{sp1},...,\theta_{spc})$$

for each variable p.

Conditional on entity i belonging to segment s, the density function of an observation $\underline{y}_i$ is given by:

$$f(\underline{y}_i\,|\,\underline{\psi}) = \sum_{s=1}^{S}\lambda_s\, f_s(\underline{y}_i\,|\,\underline{\theta}_s) = \sum_{s=1}^{S}\lambda_s \prod_{p=1}^{P}\prod_{c=1}^{C_p}\theta_{spc}^{y_{ipc}}$$

where $\underline{\psi}$, the vector of unknown parameters is the $\{\theta_{spc}\}$, and $\underline{\lambda} = \{\lambda_1,...,\lambda_{s-1}\}$ for s = 1,...,S; p = 1,...,P; c = 1,..., C_p.

IV. Simulation Experiments

To evaluate the performance of the information criteria presented in Table I and robustness across experimental conditions, a simulation study is conducted. Because special care needs to be taken before arriving at conclusions based on simulations results, we performed some replications within each cell.

For testing the hypothesis that the segmentation base variables type influences the decision concerning the information criterion to use, we set up simulation studies for the three situations: categorical, continuous and mixed base variables.

As far as segmentation base variables with only categorical variables is concerned, the aim of this paper, the experimental design controls the number of categories for each variable, the number of segments, and the sample size; thus, data sets are simulated with two categories for each one of the three variables (this originates 2^3 data sets); we use two levels of segments' number (2, 3), and the sample size assumes the levels 400, 1200, and 2000. The simulation plan uses a 2^3×2×3 = 2^4×3 factorial design, with 48 cells (see table 3); For S = 2, we fixed the missing proportions at $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$; for S = 3 we fixed the missing proportions at $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, $\lambda_3 = 0.3$. Within each cell, five data sets are generated, so we work with 240 samples.

TABLE III
FACTORIAL DESIGN FOR CATEGORICAL SEGMENTATION BASE VARIABLES

	Variables			Number of segments	Sample size	Factorial design
	X1	X2	X3	2; 3	400; 1200; 2000	
Number of levels	2	2	2	2	3	2^4×3

In order to avoid local optima in the generated LSM estimation process, the EM algorithm is repeated 50 times with random starting centers, 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}).

V. 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 segmentation base variables' type.

Table IV summarizes the results of information criteria for categorical segmentation base variables. This table illustrate the percentage of cases when the original (true) number of segments is recovered (fit) and also the overall percentages corresponding to underfit (percentage of times each criterion selects a model with a few number of segments) and overfit (percentage of times each criterion selects a model with a high number of segments).

TABLE IV
SIMULATION RESULTS FOR CATEGORICAL EXPERIMENTS

		BIC	AIC	AIC_3	AICc	AICu	CAIC	L
Overall	Fit	79	90	94	90	92	77	56
	Underfit	21	6	6	6	8	23	44
	Overfit	-	4	-	4	-	-	-
Sample size	400	44	81	94	81	88	31	13
	1200	94	94	94	94	94	63	50
	2000	94	94	94	94	94	88	69
Number of Segments	2	88	92	100	92	100	75	75
	3	71	88	88	88	83	54	21

The performance of AIC_3 is very good; it consistently performs very well for the samples sizes and segment's number we use. Overall, it finds the correct number of segments in 94% of cases.

Other criteria perform very well, like AICu (overall 92%) and AIC and AICc (ex-aequo, overall 90%). The criteria which perform worst are excluded.

We also can see that AIC_3 (with AIC and AICu) only underfit 6% of the times; on the other side, L criterion underfit 44% of the times, followed by CAIC (23%) and BIC (21%). Thus AIC_3 is quite effective when considering categorical segmentation base variables.

VI. CONCLUSION AND DISCUSSION

We conduct a simulation study which aims to find an association between information criteria performance and the type of segmentation base variables (categorical) used in Latent Segments Models. This relationship is derived from the obtained results.

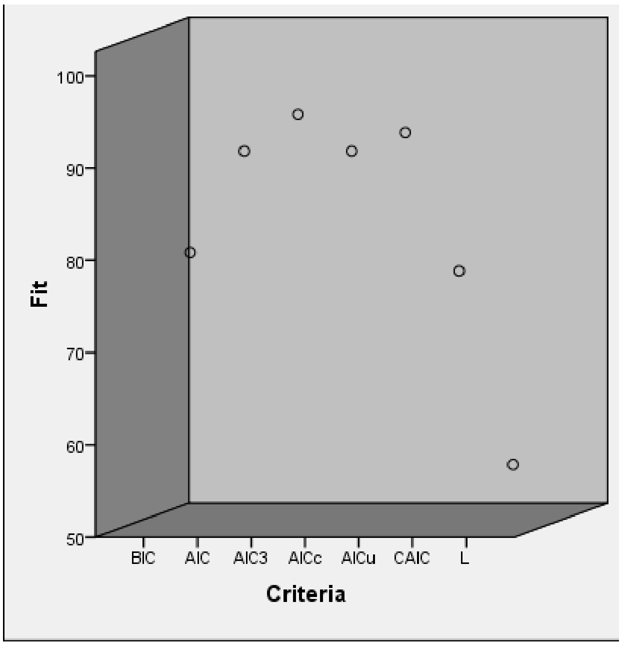

Fig. 1 Criteria fit performance in percentage

In the present study, we conclude that AIC$_3$, AICu and AICc show better performance when dealing with categorical segmentation base variables (AIC$_3$ selects the true number of segments in 94% of the simulated cases).

This study reinforces the conclusions from a previous study, [19] which was based on real data sets. Similar conclusions are achieved namely for segmentation base variables with only categorical variables (referred to 19 real data sets).

Fig. 1 illustrates fit (percentage of the true structure recovery) referred to all the experiments which are conducted in the present study, and the best of used criteria.

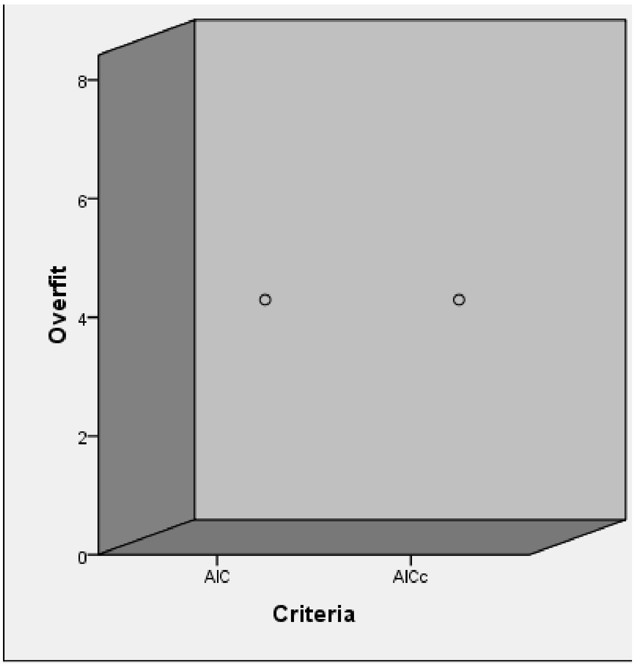

Fig. 2 Criteria overfit performance in percentage

As we can see from Fig. 2 (criteria select models with more segments, in %), AIC is the criterion which overfits more often, followed by AICc.

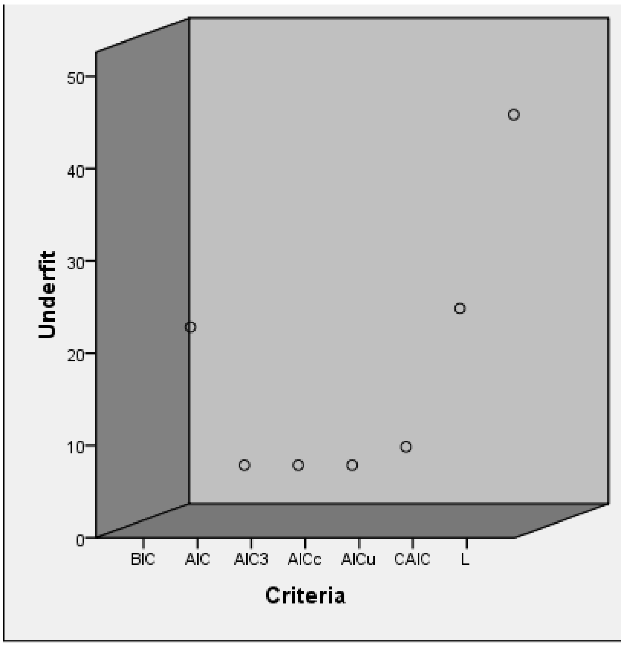

Fig. 3 Criteria underfit performance in percentage

Fig. 3 (criteria select models with less segments, in percent) shows that AIC almost never underfits; next, we have AIC$_3$, AICc and AICu.

Finally, in order to compare the criteria performances, we run Friedman tests ([20], because the data consist of b mutually independent k-variate random variables (X_{i1},…,X_{ik}), called b blocks (rows in table 4), i=1,…,b; the random variable X_{ij} is in block i (the factors in analysis: overall fit, sample size fit, and number of segments fit) and is associated with population j (criteria in the Table V' columns).

Concerning categorical experiments results (excluding underfit and overfit), we run Friedman test for all the criteria in Table IV (six observations corresponding to overall fit, different sample sizes and number of segments). We test the null hypothesis that the seven criteria perform identically (have similar medians) for recovery proportions. Since we reject the null hypothesis (Monte Carlo p-value of 0.001) we conclude that the criteria performance differs significantly (e.g. using $\alpha = 0,01$).

In order to seek for pairs significantly differences we conduct Friedman Multiple Comparisons, [13]. Criteria j and j' are considered to have different performance if the inequality:

$$| R_i - R_{j'} | > t_{(b-1)(k-1);1-\frac{\alpha}{2}} \left[\frac{2b(F_1 - F_2)}{(b-1)(k-1)} \right]^{\frac{1}{2}} \qquad (5)$$

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, F$_1$ and F$_2$ are given by:

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

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

where R(Xij) is the rank, from 1 to k, assigned to Xij within block i.

TABLE V
MATRIX FOR MULTIPLE COMPARISONS

Criteria		BIC	AIC	AIC3	AICc	AICu	CAIC	L
	Ri	22	29,5	36,5	29,5	32,5	11,5	6,5
BIC	22	0						
AIC	29,5	7,5	0					
AIC3	36,5	14,5	7	0				
AICc	29,5	7,5	0	-7	0			
AICu	32,5	10,5	3	-4	3	0		
CAIC	11,5	-10,5	-18	-25	-18	-21	0	
L	6,5	-15,5	-23	-30	-23	-26	-5	0

$$t_{(b-1)(k-1);1-\frac{\alpha}{2}} \left[\frac{2b(F_1 - F_2)}{(b-1)(k-1)} \right]^{\frac{1}{2}} = 6.26$$

Table V shows all the $R_{j'} - R_j$ values. The test yields significantly differences performances for AIC$_3$ and AICc (|RAIC$_3$-RAICc| = 7 is greater than 6.26), but there are no significant differences between AIC$_3$ and AICu (|RAIC$_3$-RAICu| = 4 is less than 6.26).

To sum up, considering the results obtained in the simulation study and tests, the information criteria to be used in latent segments model selection are:

- AIC$_3$ and AICu if all the segmentation base variables are categorical.

We think these simulation study results are particularly useful to help analysts selecting appropriates information criteria for LSM when dealing with specific segmentation problems. Further research should be conducted in order to provide new results in this area, namely considering additional information criteria.

REFERENCES

[1] H. Akaike, Information Theory and an Extension of Maximum Likelihood Principle, in K. T. Emanuel Parzen, Genshiro Kitagawa, ed., Selected Papers of Hirotugu Akaike, in Proceedings of the Second International Symposium on Information Theory, B.N. Petrov and F. caski, eds., Akademiai Kiado, Budapest, 1973, 267-281, Springer-Verlag New York, Inc, Texas, 1973, pp. 434.

[2] J. D. Banfield and A. E. Raftery, Model-Based Gaussian and Non-Gaussian Clustering, Biometrics, 49 (1993), pp. 803-821.

[3] C. Biernacki, Choix de modéles en Classification, PhD Thesis., Compiègne University of Technology, 1997.

[4] C. Biernacki, G. Celeux and G. Govaert, Assessing a Mixture model for Clustering with the integrated Completed Likelihood, IEEE Transactions on Pattern analysis and Machine Intelligence, 22 (2000), pp. 719-725.

[5] C. Biernacki, G. Celeux and G. Govaert, Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, Computational Statistics & Data Analysis, 41 (2003), pp. 561-575.

[6] C. Biernacki, G. Celeux and G. Govaert, An improvement of the NEC criterion for assessing the number of clusters in mixture model, Pattern Recognition Letters, 20 (1999), pp. 267-272.

[7] D. Böhning and W. Seidel, Editorial: recent developments in mixture models, Computational Statistics & Data Analysis, 41 (2003), pp. 349-357.

[8] S. Boucheron and E. Gassiat, Order Estimation and Model Selection, in e. O. C. A. T. Ryden, ed., Inference in Hidden Markov, 2002, pp. 25.

[9] H. Bozdogan, Mixture-Model Cluster Analysis using Model Selection criteria and a new Informational Measure of Complexity, in H. Bozdogan, ed., Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Approach, 69-113, Kluwer Academic Publishers, 1994, pp. 69-113.

[10] H. Bozdogan, Model Selection and Akaikes's Information Criterion (AIC): The General Theory and its Analytical Extensions, Psycometrika, 52 (1987), pp. 345-370.

[11] H. Bozdogan, Proceedings of the first US/Japan conference on the Frontiers of Statistical Modeling: An Informational Approach, Kluwer Academic Publishers, Dordrecht, 1994.

[12] G. Celeux and G. Soromenho, An entropy criterion for acessing the number of clusters in a mixture model, Journal of Classification, 13 (1996), pp. 195-212.

[13] W. J. Conover, Practical Nonparametric Statistics, John Wiley & Sons, Inc., New York, 1980.

[14] N. E. Day, Estimating the Components of a mixture of normal Distributions, Biometrika, 56 (1969), pp. 463-474.

[15] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from incomplete Data via EM algorithm, Journal of the Royal Statistics Society, B, 39 (1977), pp. 1-38.

[16] J. G. Dias and F. Willekens, Model-based Clustering of Sequential Data with an Application to Contraceptive Use Dynamics, Mathematical Population Studies, 12 (2005), pp. 135-157.

[17] W. R. Dillon and A. Kumar, 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, 352-388, Cambridge: Blackwell Publishers, 1994.

[18] M. A. T. Figueiredo and A. K. Jain, Unsupervised Learning of Finite Mixture Models, IEEE Transactions on pattern analysis and Machine Intelligence, 24 (2002), pp. 1-16.

[19] J. R. S. Fonseca and M. G. M. S. Cardoso, Mixture-Model Cluster Analysis using Information Theoretical Criteria, Intelligent Data Analysis, 11 (2007), pp. 155-173.

[20] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of American Statistical Association, 32 (1937), pp. 675-701.

[21] J. G. Fryer, and Robertson, C.A., A Comparision of Some methods for Estimating Mixed Normal Distributions, Biometrika, 59 (1972), pp. 639-648.

[22] P. Hall and D. M. Titterington, Efficient Nonparametric Estimation of Mixture Proportions, Journal of the Royal Statistical Society, Series B, 46 (1984), pp. 465-473.

[23] R. J. Hataway, A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions, The Annals of Statistics, 13 (1985), pp. 795-800.

[24] L. A. Hunt and K. E. Basford, Fitting a Mixture Model to Three-Mode Trhee-Way Data with Categorical and Continuous Variables, Journal of Classification, 16 (1999), pp. 283-296.

[25] C. M. Hurvich and C.-L. Tsai, Regression and Time Series Model Selection in Small Samples, Biometrika, 76 (1989), pp. 297-307.

[26] L. F. James, C. E. Priebe and D. J. Marchette, Consistency Estimation of Mixture Complexity, The Annals of Statistics, 29 (2001), pp. 1281-1296.

[27] A. B. M. L. Kabir, Estimation of Parameters of a finite Mixture of Distributions, Journal of the Royal Statistical Society, Series B, 30 (1968), pp. 472-482.

[28] D. Karlis and E. Xekalaki, Choosing initial values for the EM algorithm for finite mixtures, Computational Statistics & Data Analysis, 41 (2003), pp. 577-590.

[29] C. Keribin, Estimation consistante de l'orde de modèles de mélange, Comptes Rendues de l'Academie des Sciences, Paris, t. 326, Série I (1998), pp. 243-248.

[30] Y. Kim, W. N. Street and F. Menezer, Evolutionary model selection in unsupervised learning, Intelligent Data Analysis, 6 (2002), pp. 531-556.

[31] B. G. Leroux, Consistent Estimation of a Mixing Distribution, The Annals of Statistics, 20 (1992), pp. 1350-1360.

[32] B. G. Leroux and M. L. Puterman, Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models, Biometrics, 48 (1992), pp. 545-558.

[33] G. McLachlan and T. Krishnan, The EM Algorithm and Extensions, John Wiley & Sons, New York, 1997.

[34] G. F. McLachlan and D. Peel, Finite Mixture Models, John Wiley & Sons., 2000.

[35] G. J. McLachlan and K. E. Basford, Mixture Models: Inference and Applications to Clustering., Marcel Deckker, Inc., New York, 1988.

[36] A. McQuarrie, R. Shumway and C.-L. Tsai, The model selection criterion AICu, Statistics & Probability Letters, 34 (1997), pp. 285-292.

[37] G. Punj and D. W. Stewart, Cluster Analysis in Marketing Research: Review and Suggestions for Application, Journal of Marketing Research, XX (May 1983) (1983), pp. 134-148.

[38] J. Rissanen, Modeling by shortest data description, Automatica, 14 (1978), pp. 465-471.

[39] G. Schwarz, Estimating the Dimenson of a Model, The Annals of Statistics, 6 (1978), pp. 461-464.

[40] J. K. Vermunt and J. Magidson, Latent class cluster analysis., J.A. Hagenaars and A.L. McCutcheon (eds.), Applied Latent Class Analysis, 89-106., Cambridge University Press, 2002.

[41] H. x. Wang, Q. b. Zhang, B. Luo and S. Wei, Robust mixture modelling using multivariate t-distribution with missing information, Pattern Recognition Letters, 25 (2004), pp. 701-710.

[42] D. L. Weakliem, A critique of the Bayesian Criterion for Model Selection, Sociological Methodology & Research, 27 (1999), pp. 359-397.

Jaime R. S. Fonseca was born in Portugal, in 1952. He received a PhD degree in Quantitative Methods-Statistics and Data Analysis-from ISCTE Business School, Lisbon, Portugal, 2008, and a MSc degree in Computation and Data Analysis, from Sciences Faculty of Lisbon University, Lisbon, Portugal, 1988. He is currently Professor of Statistics/Data Analysis, and Segmentation Techniques, in Technical University of Lisbon, Portugal, Institute of Social and Political Sciences-ISCSP. He is an author of books such as Análise de Dados Univariados e Multivariados, Lisbon, Portugal, 2010,Editora Causa das Regras; Estatística Matemática, vol.II, Lisbon, Portugal, 2001,Edições Sílabo; Estatística Matemática, vol.I, Lisbon, Portugal, 2000,Edições Sílabo, and some published articles, such as Customer Satisfaction Study via a Latent Segment Model, Journal of Retailing and Consumer Services, 16, 352-359, 2009; Supermarket Customers Segments Stability, Journal of Targeting, Measurement and Analysis for Marketing, 15 (4), p. 210-221, 2007; Mixture-Model Cluster Analysis using Information Theoretical Criteria, Intelligent Data Analysis, 11 (2), p. 55-173, 2007. Current and previous research interests are multivariate data analysis, quantitative methods for social and health care sciences, latent class models, theoretical information criteria, market segmentation, customers' satisfaction.

Prof. Fonseca is membership of International Association of Statistical Computing, and reviewer of Journal of Applied Quantitative Methods (2009); 2009 IEEE Workshop on Statistical Signal Processing, paper 82906; 8th International Conference on Hybrid Intelligent Systems, 2008; IEEE Transactions on Knowledge and Data Engineering Reviewer, 2007.