THE APPLICATION OF MIXTURE MODELING AND INFORMATION CRITERIA FOR DISCOVERING PATTERNS OF CORONARY HEART DISEASE

Jaime R. S. FONSECA[1]

PhD, Assistant Professor, ISCSP-Higher Institute of Social and Political Sciences, Technical University of Lisbon, Portugal

E-mail: jaimefonseca@iscsp.utl.pt

Abstract: This paper's purpose is twofold: first it addresses the adequacy of some theoretical information criteria when using finite mixture modelling (unsupervised learning) on discovering patterns in continuous data; second, we aim to apply these models and BIC to discover patterns of coronary heart disease. In order to select among several information criteria, which may support the selection of the correct number of clusters, we conduct a simulation study, in order to determine which information criteria are more appropriate for mixture model selection when considering data sets with only continuous clustering base variables. As a result, the criterion BIC shows a better performance, that is, it indicates the correct number of the simulated cluster structures more often. When applied to discover patterns of Coronary Heart Disease, it performed well, discovering the known pattern of data.

Key words: Quantitative Methods; Unsupervised Learning; Finite Mixture Models; Patterns in Continuous Data; Theoretical Information Criteria; Simulation experiments; Coronary Heart Disease

1. Introduction

As a technique of intelligent data mining, Finite mixture models (FMM) has proven to be a powerful tools for clustering analysis, namely in the domain of social, human and behavioural science data, (Dias and Willekens 2005), and in particular in segmentation, (Punj and Stewart 1983), (Fonseca and Cardoso 2007b). There have been numerous proposals of information criteria for the selection of the number of clusters (model selection) of FMM.

The main goal of this research is to address the performance of specific theoretical information criteria for mixture modelling selection, when dealing with the *continuous* clustering 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 clustering applications. Mainly, we want to know which criterion we should select in advance, knowing that clustering base variables are continuous.

This paper is organized as follows: in section 2, we define notation and review finite mixture models, and 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 clusters of a mixture structure; 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, about BIC and Coronary Heart Disease application.

2. Clustering via Mixture Models

For illustrating the use of mixture models in the field of cluster analysis, see for instance (McLachlan and Peel 2000), (McLachlan 1997), (Figueiredo and Jain 2002). FMM assume that parameters of a statistical model of interest differ across unobserved or latent clusters and they provide a useful means for clustering observations. In FMM, clustering base variables are assumed to be described by a different probability distribution in each latent cluster. These probability 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 clusters memberships' estimates and consistent estimates for the distributional parameters, (Dillon and Kumar 1994); it provides means to select the number of clusters, (McLachlan and Peel 2000); it is able to deal with diverse types of data (different measurement levels), (Vermunt and Magidson 2002). In order to present FMM we give some notation below (Table 1).

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

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

Table 1. 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
$\underline{z}_i = (z_{i1}, ..., z_{iS})$	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 s^{th} segment
$\Theta = (\underline{\theta}_1 ... \underline{\theta}_S)$	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 $\underline{s}^{th}$ segment, given
$\underline{\psi} = (\underline{\lambda}, \Theta)$	vector of all unknown mixture model parameters
$\hat{\underline{\psi}} = (\hat{\underline{\lambda}}, \hat{\Theta})$	estimate of the vector 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

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

$$\log\ L(\underline{\psi}) = \sum_{i=1}^{n} \log \sum_{s=1}^{S} \lambda_S \quad f_S(\underline{y}_i \mid \underline{\theta}_S) \tag{3}$$

When dealing with Mixture Models for clustering purposes, we may define each *complete data* observation, $\underline{x}_i = (\underline{y}_i, \underline{z}_i)$, as having arise from one of the clusters of the mixture (1). Values of clustering 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 $z_{iS} = (z_i)s$ is 1 or 0, according as to whether $\underline{y}_i$ belongs or does not belong to the s^{th} 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 ψ, if the complete data $\underline{x}_i = (\underline{y}_i, \underline{z}_i)$ was observed, (McLachlan and Krishnan 1997), is

$$\log L_c(\underline{\psi}) = \sum_{i=1}^{n} \sum_{s=1}^{S} z_{iS} \{\log f_S(\underline{y}_i \mid \underline{\theta}_S) + \log \lambda_S\} \tag{4}$$

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

$$\frac{\partial \log L(\underline{\psi})}{\partial \underline{\psi}} = \mathbf{O} \tag{5}$$

Fitting finite mixture models (1) provides a probabilistic clustering of the n entities in terms of their posterior probabilities of membership of the S clusters of the mixture of

distributions. Since the ML estimates of most of the latent segment model (1) cannot be found analytically, estimation of FMM iteratively computes the estimates of clusters posterior probabilities and updates the estimates of the distributional parameters and mixing probabilities, (Kim, Street, and Menezer 2002).

Expectation-maximization (EM) algorithm, (Dempster, Laird, and Rubin 1977), 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.

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, (Biernacki, Celeux, and Govaert 2003) or (Karlis and Xekalaki 2003). 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.

3. Model selection

Selecting FMM structures may rely on multiple information criteria, like, for instance, BIC, ICOMP, AIC, which turns opportune the specific issue concerning the selection among several criteria themselves.

Table 2. Some information criteria for model selection on Latent Segment Models

Criteria	Definition	Author
AIC	$-2LL + 2n_{\underline{\psi}}$	(Akaike 1973)
AIC3	$-2LL + 3n_{\underline{\psi}}$	(Bozdogan 1994)
AICc	$AIC + (2n_{\underline{\psi}}(n_{\underline{\psi}}+1))/(n-n_{\underline{\psi}}-1)$	(Hurvich and Tsai 1989)
AICu	$AICc + n\log(n/(n-n_{\underline{\psi}}-1))$	(McQuarrie, Shumway, and Tsai 1997)
CAIC	$-2LL + n_{\underline{\psi}}(1+\log n)$	(Bozdogan 1987)
BIC/MDL	$-2LL + n_{\underline{\psi}}\log n$	(Schwarz 1978) / (Rissanen 1978)
CLC	$-2LL + 2EN(S)$	(Biernacki 1997)
ICL_BIC	$BIC + 2EN(S)$	(Biernacki, Celeux, and Govaert 2000)
NEC	$NEC(S) = EN(S)/(L(S)-L(1))$	(Biernacki, Celeux, and Govaert 1999)
AWE	$-2LL_c + 2n_{\underline{\psi}}(3/2+\log n)$	(Banfield and Raftery 1993)
L	$-LL + (n_{\underline{\psi}}/2)\sum\log(n\lambda_S/12) + S/2\log(n/12) + S(n_{\underline{\psi}}+1)/2$	(Figueiredo and Jain 2002)

On the other hand, applications are common in the clustering domain, which refer to clustering base variables; also the criterion selection could be based on convergence property. In the present study we propose an approach for evaluating several (see table 2) information criteria's performances, taking into account theirs relationship with continuous

clustering base variables. Information criteria all balance fitness, trying to maximize the likelihood function, and parsimony, by using penalties associated with measures of model complexity, trying to avoid overfit. The general form of information criteria is as follows

$$-2\log L(\hat{\psi}) + C, \tag{6}$$

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

AIC (Akaike 1973) and AIC_3 (Bozdogan 1994) are measures of model complexity associated with some criteria (see table 2) that only depend on the number of parameters; some other measures depend on both the number of parameters and the sample size, as AICc (Hurvich and Tsai 1989), AICu (McQuarrie, Shumway, and Tsai 1997), CAIC (Bozdogan 1987), and BIC/MDL (Schwarz 1978) / (Rissanen 1978) ; others depend on entropy, as CLC (Biernacki 1997), and NEC (Biernacki, Celeux, and Govaert 1999); some of them depend on the number of parameters, sample size, and entropy, as ICL-BIC (Biernacki, Celeux, and Govaert 2000) , and AWE (Banfield and Raftery 1993) ; L (Figueiredo and Jain 2002) depends on the number of parameters, sample size and mixing proportions, λ_S .

4. Methodology

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, such as graphical techniques, likelihood ratio tests and theoretical information criteria. This work specifically refers to information criteria presented in table 2, which have been referred previously. This issue is in limelight, because there is no indication concerning the selection of the information criteria themselves, in a certain application, (Fonseca and Cardoso 2007). In this paper we try to establish a relationship between type of clustering variables - continuous - 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 we have a mixture of normal components ($1 \le s \le S$), the probability (density) function of an observation $\underline{y}_i$, conditional on entity *i* belonging to segment *s*, is given by

$$f_s(\underline{y}_i \mid \underline{\psi}) = \frac{1}{(2\pi)^{p/2} \mid \Sigma_s \mid^{1/2}} \exp\left(-\frac{1}{2}(\underline{y}_i - \underline{\mu}_s)^{\mathrm{T}} \Sigma_s^{-1} (\underline{y}_i - \underline{\mu}_s)\right) \tag{7}$$

Here, $\underline{\psi} = \{\underline{\lambda}, \underline{\theta}_S\}$, with $\underline{\theta}_S = (\underline{\mu}_S, \Sigma_S)$, the elements of components means, $\underline{\mu}_S$, and the distinct elements of the segment-covariance matrices Σ_S, s = 1,...S. To evaluate the performance of the information criteria presented in Table 2 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. The experimental design controls the number of variables, the number of clusters, the sample size, and the number of distributions; thus, data sets were simulated with two levels (*p* = 2 and *p* = 4) of clustering base variables, two levels of clusters (S = 2 and S = 4), three different distributions, and three levels of sample size (100,

500 and 2000); the simulation plan uses a $2^2 \times 3^2$ factorial design with 36 cells (see table 3). For S = 2, we fixed the missing proportions at $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$; for S = 4 we fixed the missing proportions at $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.25$. Within each cell 5 data sets were generated, so we work with 180 samples.

Table 3. Factorial design for continuous variables

Y_i	S	n	Number of Distributions	Factorial design
2; 4	2; 4	100; 500 2000	3	
2	2	3	3	2^2*3^2

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 of simulated experiments

Table 4 shows the percentage of cases (simulated experiments) each criterion determines the original (*true*) number of segments (*fit*), across the used factors, the overall percentages *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).

The best performance goes to BIC (overall 93%), followed by AIC_3 (overall 89%) and AICu (overall 88%). AIC_3 also performs very well, yielding the best performances when sample size decreases (85% for n = 100, against BIC 80%) and when the segment's number and variables' number increases (87% for S = 4 and p = 4, against BIC's 80%). Moreover, BIC only *overfits* and *underfits* on 1% and 6% of the times, respectively. As we could expect, other criteria, such as ICL-BIC, NEC, L, and AWE, almost never overfit; instead, they underfit a lot of time.

Concerning sample size BIC (80%) is outperformed by AIC_3 (85%), only when n = 100.

Table 4. Simulation results for continuous experiments

		BIC	AIC	AIC3	AICc	AICu	CAIC	CLC	ICL-	NEC	L	AWE
Overall	*Fit*	**93**	63	89	71	88	85	67	74	56	75	64
	Underfit	6	1	5	3	8	14	15	24	43	24	36
	Overfit	**1**	36	6	26	4	1	18	2	1	1	0
Sample size	*100*	80	72	85	83	77	69	45	65	51	55	49
	500	100	61	99	67	99	99	83	79	57	75	69
	2000	87	63	81	60	87	83	76	71	52	84	71

Number of seg./variables	2 P=2	98	70	93	78	83	93	76	80	81	90	83
	2 P=4	100	53	98	67	100	100	78	98	93	100	93
	4 P=2	73	67	73	64	67	62	49	31	4	29	13
	4 P=4	80	40	87	58	73	76	42	76	18	49	44

As far as the number of segments and variables number is concerned, BIC (80%) is only outperformed by AIC_3 (87%). Nevertheless the number of variables and sample size, the simulation experiment results show that information criteria BIC is quite effective for FMM with continuous clustering base variables, in order to select the *true* model.

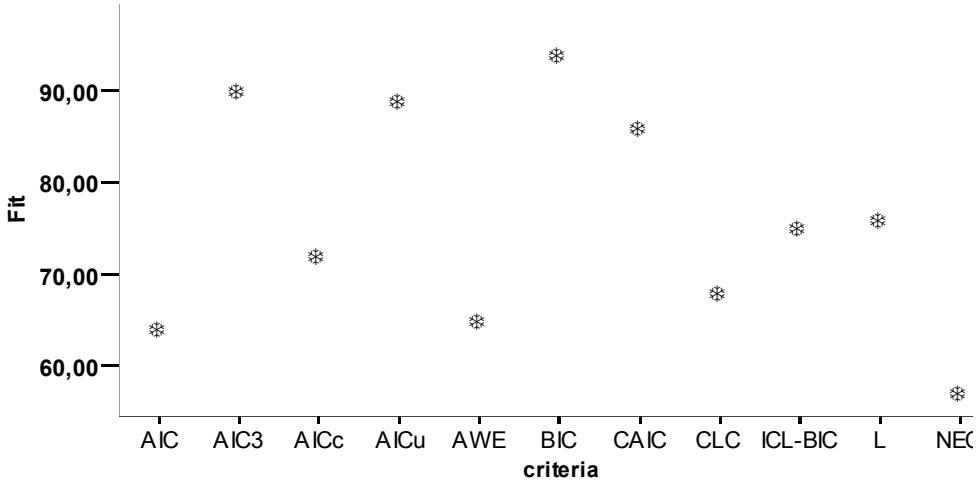

Figure 1. The true number of segments recovery (Fit), in percent

Figures 1, 2, and 3 show the percentage of cases (simulated experiments) each criterion determines the original (*true*) number of segments (*fit*), across the used factors, and also the overall percentages *overfit* ,and *underfit* respectively.

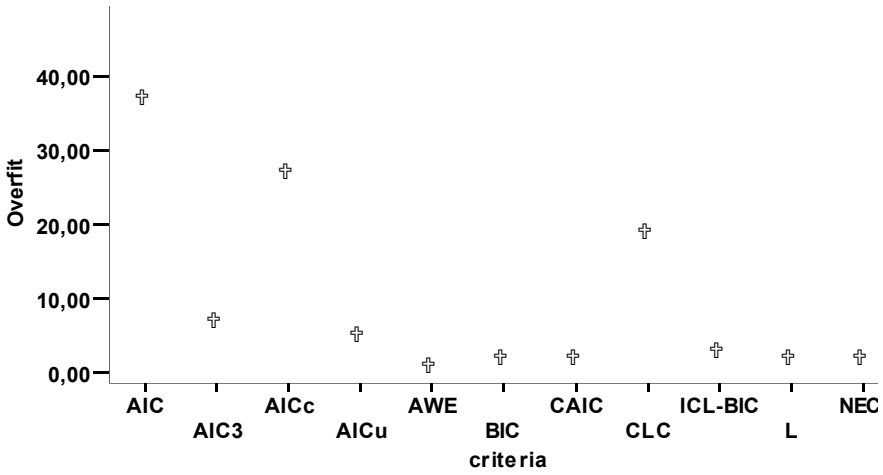

Figure 2. Criteria selecting models with more segments (overfit), in percent

As we can see from figure 2 (criteria select models with more segments, in %), AIC is the criterion which *overfits* more often, followed by AICc and CLC. Figure 3 (criteria select models with less segments, in %) shows that AIC almost never *underfits*; next, we have AIC_3, AICu and AICc; we also can see that BIC almost never *underfits* on normal multivariate models.

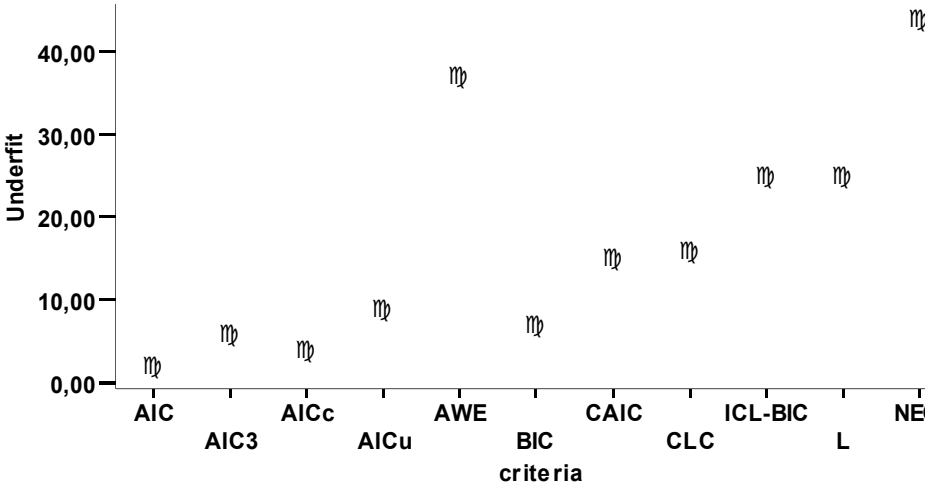

Figure 3. Criteria selecting models with less segments (underfit), in percent

6. Coronary Heart Disease Application

In order to see the performance of these models and information criterion BIC, we analyze a dataset (n = 231) with known diagnostic classification (normal, premature, serious and permanent), and five continuous variables: NAOHDLCC, CHOLESTEROL, LDLC, HDLC, TG.

In order to "guess" the diagnostic classification, we apply FMM approach, with information criterion BIC, and we display in table 5 the results for model selection. Because information criterion BIC presents an elbow for S = 4, we selected a model with four clusters, the true diagnostic classification, with relative sizes: 28, 23, 20, and 11 percent, respectively.

Thus, we can conclude that these models, finite mixture modeling, with information criterion BIC for model selection, are good for discovering patterns in continuous data, in particular for guessing true diagnostic classification for coronary heart disease.

Table 5. Model Selection (Information criterion BIC)

Model	LogL	BIC
1-Cluster	-5570,87	11196,08
2-Cluster	-5309,55	10733,21
3-Cluster	-5175,43	10524,74
4-Cluster	-5080,44	10394,53
5-Cluster	-5027,73	10348,89

As we can see from figure 4, the items NAOHDLCC, CHOLESTEROL, and LDLC are the most important ones, in order to discriminate between the four clusters.

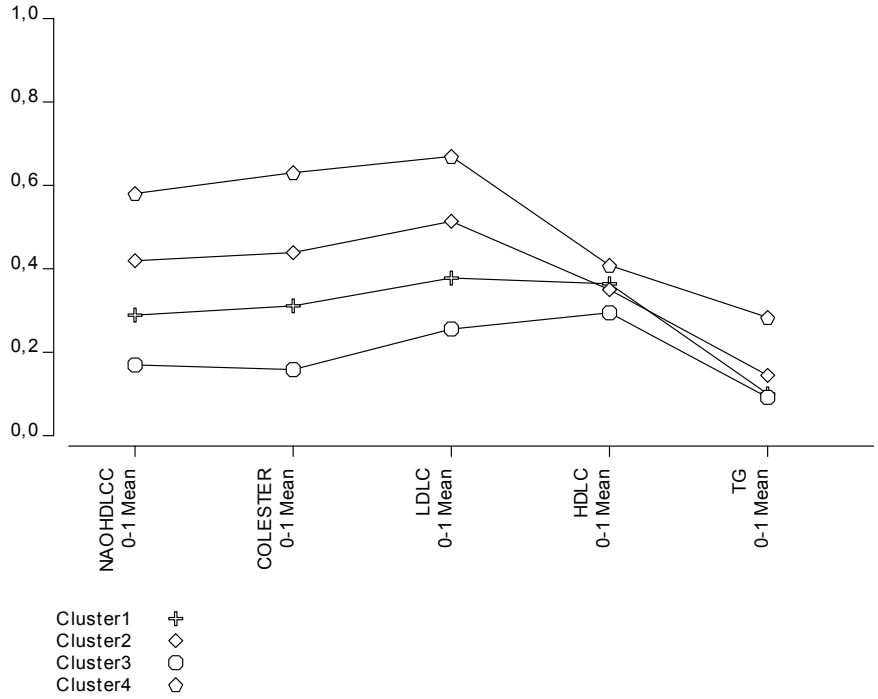

Figure 4. Conditional probabilities of four clusters

7. Conclusions

The results of this study help on developing a consistent way of selecting an appropriate information criterion for model selection when dealing with finite mixture modelling and continuous clustering base variables.

As a result of the simulation study, BIC and AIC3 (followed by AICu) are the best performing criteria when dealing with continuous segmentation base variables; moreover, BIC selects the right model in 93% of the time (Figure 1 and table 4). We also can see that BIC almost never *overfits* (Figure 2), and rarely *underfits* (Figure 3). Thus we conclude that BIC is a good criterion to select the best model and so to discover patterns in continuous data.

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

Thus we run Friedman test for all the criteria in table 2, to test the null hypothesis that all the eleven means performances are identical. We reject the null hypothesis (Monte Carlo p-value of 0.000). Thus, we conclude that criteria performance was not identical for the eleven criteria in table 2, 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

$$F1 = \sum_{i=1}^{b}\sum_{j=1}^{k}\left[R(X_{ij})\right]^2 \text{ and } F_2 = \frac{1}{b}\sum_{j=1}^{k} R_j^2 \text{, with } R_j = \sum_{i=1}^{b} R(X_{ij}),$$

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

Table 10 Matrix for multiple comparisons

Criteria		BIC	AIC	AIC3	AICc	AICu	CAIC	CLC	ICL-BIC	NEC	L	AWE
	R_i	82,5	26,5	74,5	38	68	62,5	31	44,5	17,5	50,5	30,5
BIC	82,5	0,0										
AIC	26,5	-56,0	0,0									
AIC3	74,5	-8,0	48,0	0,0								
AICc	38	-44,5	11,5	-36,5	0,0							
AICu	68	-14,5	41,5	-6,5	30,0	0,0						
CAIC	62,5	-20,0	36,0	-12,0	24,5	-5,5	0,0					
CLC	31	-51,5	4,5	-43,5	-7,0	-37,0	-31,5	0,0				
ICL-BIC	44,5	-38,0	18,0	-30,0	6,5	-23,5	-18,0	13,5	0,0			
NEC	17,5	-65,0	-9,0	-57,0	-20,5	-50,5	-45,0	-13,5	-27,0	0,0		
L	50,5	-32,0	24,0	-24,0	12,5	-17,5	-12,0	19,5	6,0	33,0	0,0	
AWE	30,5	-52,0	4,0	-44,0	-7,5	-37,5	-32,0	-0,5	-14,0	13,0	-20,0	0,0

$$(t_{(b-1)(k-1);1-\frac{\alpha}{2}}\left[\frac{2b(F_1-F_2)}{(b-1)(k-1)}\right]^{\frac{1}{2}} = 18.4)$$

Because we have

$$t_{(b-1)(k-1);1-\frac{\alpha}{2}}\left[\frac{2b(F_1-F_2)}{(b-1)(k-1)}\right]^{\frac{1}{2}} = 18.4,$$

as we can see, we have $|R_{BIC}-R_{AICu}| = 14.5$, $|R_{BIC}-R_{AIC3}| = .8$, and $|R_{AIC3}-R_{AICu}| = 6.5$, all less than 18.4; then, we can conclude that BIC, AIC_3 and AICu have similar performances. They differ from all the others information criteria with relation with performance.

To sum up, we conclude that for determining the number of segments, BIC, AIC_3 and AICu, with 93, 89 e 88 percent, respectively, perform very well when using FMM for discovering patterns in continuous data. Moreover, they perform well for several sample sizes and *true* number of segments, and they almost never overfit and underfit.

Then we apply this criterion, with mixture models, in order to discover the patterns *of coronary heart disease*, and the results are very good, because this approach selects a model with four clusters, which was the known pattern of data.

References

1. Akaike, H. **Information Theory and an Extension of Maximum Likelihood Principle,** in Petrov, B.N. and Caski, F. (eds.),"Selected Papers of Hirotugu Akaike, in Proceedings of the Second International Symposium on Information Theory", Akademiai Kiado, Budapest, 1973, pp. 267-281, edited by Parzen, K. T. E. and Kitagawa, G., Texas, Springer-Verlag New York, Inc., 1973
2. Banfield, J. D., and Raftery, A. E. **Model-Based Gaussian and Non-Gaussian Clustering,** Biometrics, 49 (3), 1993, pp. 803-821
3. Biernacki, C. **Choix de modéles en Classification, PhD Thesis.,** Compiègne University of Technology, 1997
4. Biernacki, C., Celeux, G. and Govaert, G. **An improvement of the NEC criterion for assessing the number of clusters in mixture model,** Pattern Recognition Letters 20, 1999, pp. 267-272
5. Biernacki, C., Celeux, G. and Govaert, G. **Assessing a Mixture model for Clustering with the integrated Completed Likelihood,** IEEE Transactions on Pattern analysis and Machine Intelligence 22 (7), 2000, pp. 719-725
6. Biernacki, C., Celeux, G. and Govaert, G. **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
7. Bozdogan, H. **Model Selection and Akaikes's Information Criterion (AIC): The General Theory and its Analytical Extensions,** Psycometrika 52 (3), 1987, pp. 345-370
8. Bozdogan, H. **Proceedings of the first US/Japan conference on the Frontiers of Statistical Modeling: An Informational Approach, 1 ed. 3 vols. Vol. 1.,** Dordrecht, Kluwer Academic Publishers, 1994
9. Dempster, A.P., Laird, N. M. and Rubin, D.B. **Maximum Likelihood from incomplete Data via EM algorithm,** Journal of the Royal Statistics Society, *B* 39, 1977, pp. 1-38
10. Dias, J. G., and Willekens, F. **Model-based Clustering of Sequential Data with an Application to Contraceptive Use Dynamics,** Mathematical Population Studies 12, 2005, pp. 135-157
11. Dillon, W.R., and Kumar, A. **Latent structure and other mixture models in marketing: An integrative survey and overview,** chapter 9 in Bagozi, R. P. (ed.), "Advanced methods of Marketing Research", pp. 352-388, Cambridge: Blackwell Publishers, 1994
12. Figueiredo, M.A.T., and Jain, A.K. **Unsupervised Learning of Finite Mixture Models,** IEEE Transactions on pattern analysis and Machine Intelligence 24 (3), 2002, pp. 1-16
13. Fonseca, J. R.S., and Cardoso, M.G.M.S. **Mixture-Model Cluster Analysis using Information Theoretical Criteria,** Intelligent Data Analysis 11 (2), 2007, pp. 155-173
14. Fonseca, J. R.S., and Cardoso, M.G.M.S. **Supermarket Customers Segments Stability.** Journal of Targeting, Measurement and Analysis for Marketing 15 (4), 2007b, pp. 210-221
15. Hurvich, C.M., and Tsai, C.-L. **Regression and Time Series Model Selection in Small Samples,** Biometrika 76 (2), 1989, pp. 297-307
16. Karlis, D., and E. Xekalaki. 2003. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis* 41:577-590.
17. Kim, Y., Street, W.N. and Menezer, F. **Evolutionary model selection in unsupervised learning,** Intelligent Data Analysis 6, 2002, pp. 531-556
18. McLachlan, G., and Krishnan, T. **The EM Algorithm and Extensions,** New York: John Wiley & Sons, 1997
19. McLachlan, G.F. and Peel, D. **Finite Mixture Models, first ed. 1 vols,** John Wiley & Sons, 2000
20. McLachlan, G.J., Peel, D. and Prado, P. **Clustering via Normal Mixture models,** 1997
21. McQuarrie, A., Shumway, R. and Tsai, C.-L. **The model selection criterion AICu,** Statistics & Probability Letters 34, 1997, pp. 285-292

22. Punj, G., and David W. S. **Cluster Analysis in Marketing Research: Review and Suggestions for Application,** Journal of Marketing Research XX, May 1983, pp. 134-148
23. Rissanen, J. **Modeling by shortest data description,** Automatica 14, 1978, pp. 465-471
24. Schwarz, G. **Estimating the Dimenson of a Model,** The Annals of Statistics 6 (2), 1978, pp. 461-464
25. Vermunt, J.K. and Magidson, J. **Latent class cluster analysis,** in Hagenaars, J.A. and McCutcheon, A.L. (eds.), "Applied Latent Class Analysis", Cambridge University Press, 2002, pp. 89-106

---

[1] Jaime R. S. Fonseca is currently an Assistant Professor in the ISCSP-Higher Institute of Social and Political Sciences, Technical University of Lisbon, Ph.D in Quantitative Methods - Statistics and Data Analysis - from ISCTE - Business School, Department of Quantitative Methods, Lisbon, Portugal, teaching Data Analysis with computer in the classroom, and researcher in the CAPP - Centre for Public Administration and Policies.