Using the classification likelihood to choose the number of clusters

C. Biernacki G. Govaert

URA CNRS 817
Université de Technologie de Compiègne
BP 529, 60205 Compiègne Cedex, France

Abstract

The problem to choose the number of clusters in clustering has already been dealt widely with, but still remains opened. We propose to use simply the classification likelihood of Symons [13] to solve this problem in the Gaussian mixture model approach, which is nowadays a standard approach in clustering. Beyond its simplicity, such a criterion is meaningful: it can be seen both a penalization of the standard likelihood by a term measuring overlapping of clusters and a penalization of the very famous k-means criterion by a function of the number of clusters. Several strategies to use the classification likelihood are presented and then compared with other well-known criteria by means of Monte-Carlo simulations and real data set. These new approaches show a very good behaviour for well-separated clusters in many situations.

1 Introduction

Clustering is helpful to improve the synthetic comprehension of a large data set, particularly in high dimension. To choose the number of clusters within a partition of these data is an old and difficult problem still opened. Solving it requires at first strict hypotheses about the clustering procedure.

A standard and powerful approach is available through a multivariate Gaussian mixture hypothesis. Data are assumed to be *independant* realizations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of $\mathbb{R}^d$ coming from a mixture of K subpopulations $P_1, \ldots, P_K$. Each cluster P_k $(1 \leq k \leq K)$ has a mixing proportion p_k $(0 < p_k < 1, \sum_{k=1}^K p_k = 1)$ and a Gaussian density $\phi_k(\boldsymbol{x}|\boldsymbol{a}_k)$ with $\boldsymbol{a}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ is the mean vector and $\boldsymbol{\Sigma}_k$ the variance matrix. So density is $f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^K p_k \phi_k(\boldsymbol{x}|\boldsymbol{a}_k)$, with $\boldsymbol{\theta} = (p_1, \ldots, p_{K-1}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_K)$.

When $\boldsymbol{\theta}$ is known, we can compute the probabilities t_{ik}'s that $\boldsymbol{x}_i$ arises from the cluster P_k defined by $t_{ik} = \frac{p_k \phi_k(\boldsymbol{x}|\boldsymbol{a}_k)}{f(\boldsymbol{x}|\boldsymbol{\theta})}$. The partition is obtained by assigning each point to the cluster which gives the largest *posterior* probability t_{ik}. This method is called MAP (*maximum a posteriori*). Many propositions start from these basic ideas to estimate the right number K of clusters in the mixture.

We will present standard estimation methods of $\boldsymbol{\theta}$ (Section 2), criteria to choose the number of components in mixture (Section 3), the classification likelihood as a new criterion and several strategies to use it (Section 4), and, at last, comparison of such criteria (Section 5).

2 Estimation: mixture and classification approaches

2.1 Maximum likelihood

In the mixture approach, the maximum likelihood is very used and consists in taking as estimator $\hat{\boldsymbol{\theta}}_L(K)$ of $\boldsymbol{\theta}$ the value which maximizes the log-likelihood:

$$\mathrm{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \ln \left(\sum_{k=1}^K p_k \phi_k(\boldsymbol{x}_i|\boldsymbol{a}_k) \right).$$

To obtain the optimum, we usually resort to iterative procedures as the EM algorithm (*Expectation Maximization*) (Dempster et al. [11]). Starting from parameters $\boldsymbol{\theta}^\circ$, this algorithm iterates in two steps:

1. E step (expectation): estimation of all probabilities t_{ik}'s when the parameters are known;

2. M step (maximization): choice of the estimator $\hat{\boldsymbol{\theta}}$ which maximizes $\mathrm{L}(\boldsymbol{\theta})$, when the t_{ik}'s are known.

The EM algorithm only converges to a *local maximum* (sometimes to a stationary value as a saddle point) of the likelihood function L: the result depends on the starting parameter.

2.2 Classification maximum likelihood

We introduce vectors $\boldsymbol{z}_i = (z_{ik}, k = 1, \ldots, K)$ which indicate the origin of the points: z_{ik} is equal to 1 or 0 according as $\boldsymbol{x}_i$ comes from the cluster P_k or not. These vectors are additional unknown parameters. The classification log-likelihood is given by (see Symons [13]):

$$\mathrm{CL}(\boldsymbol{\theta}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln(p_k \phi_k(\boldsymbol{x}_i | \boldsymbol{a}_k))$$

with $\boldsymbol{z} = (\boldsymbol{z}_i, i = 1, \ldots, n)$.

In such a situation, the search of maximum likelihood is done on the couple of parameters $(\boldsymbol{\theta}, \boldsymbol{z})$ to obtain finally estimators $\hat{\boldsymbol{\theta}}_{CL}(K)$ and $\hat{z}_{CL}(K)$. This procedure is typical in a clustering context since we want both estimators of $\boldsymbol{\theta}$ and $\boldsymbol{z}$.

In the same way that likelihood, an optimization iterative method is often used. For instance, the CEM algorithm (*Classification Expectation Maximization*) of Celeux and Govaert [8] is a classification variant of the EM algorithm which introduced a classification step (MAP) between the E and M steps.

The convergence of the CEM algorithm still remains local. It can produce also *biased estimators* of mixture parameters $\boldsymbol{\theta}$ if clusters are not well-separated or if mixing proportions are too extreme.

3 Choosing the number of clusters in mixture approach

The main methods to detect the number of clusters penalize the maximum likelihood by the measure of the model complexity given by the number of free parameters.

The AIC criterion (*Akaike information criterion*) of Akaike [1] was introduced by Bozdogan and Slove [7] in a mixture context: $\mathrm{AIC}_K = -2\mathrm{L}_K + 2q_K$, where $\mathrm{L}_K = \mathrm{L}(\hat{\boldsymbol{\theta}}_L(K))$ and q_K is the number of degrees of freedom for K clusters. Bozdogan [4] suggests a modified AIC criterion called AIC3

$$\mathrm{AIC3}_K = -2\mathrm{L}_K + 3q_K.$$

The Bayesian criterion BIC (*Bayesian information criterion*) of Schwarz [12] is an approximation of the exact Bayes solution to select the fitted number of clusters

$$\mathrm{BIC}_K = -2\mathrm{L}_K + \ln(n) q_K.$$

Finally Bozdogan [5] proposed ICOMP (*informational complexity criterion*) where the penalization contains the Fisher information matrix $\boldsymbol{F}_K$:

$$\mathrm{ICOMP}_K = -2\mathrm{L}_K + q_K \ln\left(\frac{\mathrm{tr}\boldsymbol{F}_K^{-1}}{q_K}\right) - \ln|\boldsymbol{F}_K^{-1}|.$$

.

Banfield and Raftery [3] obtained the AWE criterion by an approximation of the exact Bayes solution. In this case, it is the classification maximum likelihood $\mathrm{CL}_K = \mathrm{CL}(\hat{\boldsymbol{\theta}}_{CL}(K), \hat{z}_{CL}(K))$ that is penalized:

$$\mathrm{AWE}_K = -2\mathrm{CL}_K + 2q_K(\frac{3}{2} + \ln(n)).$$

The results obtained by intensive simulations of Cutler and Windham [10] show that AIC overestimates K, AIC3 underestimates it and BIC underestimates it much more. Such conclusions are coherent since we penalize complexity in an increasing order. ICOMP is the best criterion and AWE is well-known to underestimate K.

4 Choosing the number of clusters in clustering approach

4.1 The NEC criterion

Clustering criteria have the property of looking for well-separated clusters.

Celeux and Soromenho [9] suggested the NEC criterion, one of the most efficient clustering criterion, which is expressed by

$$\mathrm{NEC}_K = \frac{\mathrm{E}_K}{\mathrm{L}_K - \mathrm{L}_1}.$$

NEC must be small to obtain a compromise between a good classification feature and a good description of data respectively represented by the entropy $\mathrm{E}_K = -\sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \ln t_{ik}$ and $\mathrm{L}_K - \mathrm{L}_1$.

We note that this criterion is not defined in the monocluster situation as it is often the case for other clustering criteria. To solve this problem, an empirical version has been proposed in the Gaussian situation. NEC gives results like ICOMP (Celeux et Soromenho [9]).

4.2 Classification likelihood CL

4.2.1 CL: a penalized likelihood

We can make a link between the likelihood $\mathrm{L}(\boldsymbol{\theta})$ and the classification likelihood $\mathrm{CL}(\boldsymbol{\theta}, \boldsymbol{z})$:

$$\mathrm{CL}(\boldsymbol{\theta}, \boldsymbol{z}) = \mathrm{L}(\boldsymbol{\theta}) - \mathrm{LP}(\boldsymbol{\theta}, \boldsymbol{z})$$

where $\mathrm{LP}(\boldsymbol{\theta}, \boldsymbol{z}) = -\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln(t_{ik}) \quad (\geq 0)$. We can remark than $-\mathrm{LP}(\boldsymbol{\theta}, \boldsymbol{z})$ term corresponds to the logarithm of the probability of the partition $\boldsymbol{z}$ when the t_{ik}'s are known. So the classification likelihood is the standard likelihood penalized by a term which measures the quality of the partition.

4.2.2 CL: a penalization of the k-means

Celeux and Govaert [8] noted that optimizing $\mathrm{tr}(\boldsymbol{W}_K)$, the very famous Wald or k-means criterion in clustering ($\boldsymbol{W}_K = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'$), is equivalent to optimizing the classification likelihood criterion CL under the following hypotheses: equal mixing proportions, equal and "spherical" variance matrices ($\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \sigma^2 \boldsymbol{I}$ with $\boldsymbol{I}$ identity matrix).The link between these two criteria is

$$\mathrm{CL}_K = -\frac{nd}{2}\ln(\mathrm{tr}(\boldsymbol{W}_K)) - n\ln(K) + \mathrm{cst}.$$

4.2.3 CL: a criterion for the number of clusters?

It is well-known that both L_K and $\mathrm{tr}(\boldsymbol{W}_K)$ heavily overestimate K. The classification likelihood brings a penalization of these two terms and so it can be considered to detect the number of components.

4.2.4 Strategies to use the CL criterion

Now we suggest four strategies to use the classification likelihood to find the number of clusters.

The simplest one is to maximize directly $\mathrm{CL}(\boldsymbol{\theta}, \boldsymbol{z})$. However a theoretical drawback exits when we look for clusters with *free mixing proportions*. In such a case, noting Θ_K the space of parameters for K clusters, we have $\Theta_K \subset \Theta_{K+1} \Rightarrow \mathrm{CL}_K \leq \mathrm{CL}_{K+1}$. So we should always choose the upper bound of the number of clusters. Nevertheless, if we impose equality on mixing proportions, we have $\Theta_K \not\subset \Theta_{K+1}$ and therefore this systematic growth of CL disappears.

A second strategy consists in, first estimating the parameters $\boldsymbol{\theta}$ by maximum likelihood and then in penalyzing the maximum log-likelihood L_K by the term LP_K where $\mathrm{LP}_K = \min_{\boldsymbol{z}} \mathrm{LP}(\hat{\boldsymbol{\theta}}_L(K), \boldsymbol{z})$. The partition obtained in this way corresponds to the MAP of the t_{ik}'s given by maximum likelihood. We derive the CLM criterion (*classification likelihood in mixture*) $\mathrm{CLM}_K = \mathrm{L}_K - \mathrm{LP}_K$.

We also propose the strategy CL2 which is an extension of CL. CL2 refuses all number of clusters K where at least one of them contains a number of data equal or smaller than the number of free parameters. This approach will eliminate partitions with "too small" clusters which do not give a robust estimation of mixture parameters. We define in the same manner the strategy CLM2.

5 Numerical experiments

5.1 Simulated data

5.1.1 Data

We propose to compare the behaviour of our four criteria CL, CL2, CLM and CLM2 to standard criteria L, AIC, AIC3, BIC, ICOMP, NEC and AWE. In this aim, we will use intensive Monte-Carlo studies.

Uni- and bivariate parameters Each simulation set is considered with 200 points and three different separation degrees (except in the monocluster case). We use a separation measure derived from the theoretical misclassification rate but with the same weight to each cluster. It is defined by $s = 1 - \frac{1}{K}\sum_{k=1}^{K} \mathrm{P}(Z = k | Z^\circ = k)$, where the random variables Z° and Z respectively represent the true cluster and the obtained cluster.

We define the three following mixture degrees: $s_1 = 0.15$, $s_2 = 0.10$ and $s_3 = 0.05$ for respectively poor, middle and strong separation of clusters. The parameters are summarized in Table 1. Each of the 9 basic parameters sets (except the monocluster case) have in addition two types of mixing proportions (equal and different).

An example in dimension 4 We also propose a set of parameters in dimension 4 originally used by Bozdogan [6]. It is composed of 625 points in 5 clusters and the parameters of the model are:

$$\begin{array}{lll}
p_1 = 0.12 & \boldsymbol{\mu}_1 = (10, 12, 10, 12)' & \boldsymbol{\Sigma}_1 = \boldsymbol{I} \\
p_2 = 0.16 & \boldsymbol{\mu}_2 = (8.5, 10.5, 8.5, 10.5)' & \boldsymbol{\Sigma}_2 = \boldsymbol{I} \\
p_3 = 0.20 & \boldsymbol{\mu}_3 = (12, 14, 12, 14)' & \boldsymbol{\Sigma}_3 = \boldsymbol{I} \\
p_4 = 0.24 & \boldsymbol{\mu}_4 = (13, 15, 7, 9)' & \boldsymbol{\Sigma}_4 = 4\boldsymbol{I} \\
p_5 = 0.28 & \boldsymbol{\mu}_5 = (7, 9, 13, 15)' & \boldsymbol{\Sigma}_5 = 9\boldsymbol{I}.
\end{array}$$

5.1.2 The used models

For each set of parameters, we simulate 30 samples. For all of them, we start the CEM algorithm with the true parameters when the searched number of clusters is equal to the "true" one and otherwise we run 20 times the algorithm from random initial position and we select the best solution. Then the EM algorithm is started with parameters obtained by CEM if this last one is able to provide a result and else it is started 20 times at random.

The CEM and EM algorithms are successively used with equal and free mixing proportions p_k for each sample. The constraints about $\boldsymbol{\Sigma}_k$ are the closest to the structure of the true parameters: equal and spherical matrices ($\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \sigma^2\boldsymbol{I}$) for simulations 1 to 3 and 5 to 7, free spherical matrices ($\boldsymbol{\Sigma}_k = \sigma_k^2\boldsymbol{I}, 1 \leq k \leq$

N°	d	K	s	p_k	$\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$
1	1	1	s_3	$=$	
2	1	2	s_1	$=$	$\boldsymbol{\mu}_2 = 2.0$
				$\neq$	$\boldsymbol{\mu}_2 = 2.2$
			s_2	$=$	$\boldsymbol{\mu}_2 = 2.6$
				$\neq$	$\boldsymbol{\mu}_2 = 2.7$
			s_3	$=$	$\boldsymbol{\mu}_2 = 3.3$
				$\neq$	$\boldsymbol{\mu}_2 = 3.4$
3	1	3	s_1	$=$	$\boldsymbol{\mu}_2 = 2.4$, $\boldsymbol{\mu}_3 = 4.8$
				$\neq$	$\boldsymbol{\mu}_2 = 2.9$, $\boldsymbol{\mu}_3 = 5.2$
			s_2	$=$	$\boldsymbol{\mu}_2 = 2.9$, $\boldsymbol{\mu}_3 = 5.8$
				$\neq$	$\boldsymbol{\mu}_2 = 3.4$, $\boldsymbol{\mu}_3 = 6.0$
			s_3	$=$	$\boldsymbol{\mu}_2 = 3.6$, $\boldsymbol{\mu}_3 = 7.2$
				$\neq$	$\boldsymbol{\mu}_2 = 3.9$, $\boldsymbol{\mu}_3 = 7.35$
4	1	2	s_1	$=$	$\boldsymbol{\mu}_2 = 0.0$, $\boldsymbol{\Sigma}_2 = 0.0250$
				$\neq$	$\boldsymbol{\mu}_2 = 0.0$, $\boldsymbol{\Sigma}_2 = 0.0240$
			s_2	$=$	$\boldsymbol{\mu}_2 = 0.0$, $\boldsymbol{\Sigma}_2 = 0.0100$
				$\neq$	$\boldsymbol{\mu}_2 = 0.0$, $\boldsymbol{\Sigma}_2 = 0.0090$
			s_3	$=$	$\boldsymbol{\mu}_2 = 0.0$, $\boldsymbol{\Sigma}_2 = 0.0021$
				$\neq$	$\boldsymbol{\mu}_2 = 0.0$, $\boldsymbol{\Sigma}_2 = 0.0018$
5	2	1	s_3	$=$	

N°	d	K	s	p_k	$\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$
6	2	2	s_1	$=$	$\boldsymbol{\mu}_2 = (2.0, 0.0)')$
				$\neq$	$\boldsymbol{\mu}_2 = (2.2, 0.0)'$
			s_2	$=$	$\boldsymbol{\mu}_2 = (2.6, 0.0)'$
				$\neq$	$\boldsymbol{\mu}_2 = (2.7, 0.0)'$
			s_3	$=$	$\boldsymbol{\mu}_2 = (3.3, 0.0)'$
				$\neq$	$\boldsymbol{\mu}_2 = (3.4, 0.0)'$
7	2	3	s_1	$=$	$\boldsymbol{\mu}_2 = (1.8, 1.7)'$, $\boldsymbol{\mu}_3 = (1.8, -1.7)'$
				$\neq$	$\boldsymbol{\mu}_2 = (2.0, 1.9)'$, $\boldsymbol{\mu}_3 = (2.0, -1.9)'$
			s_2	$=$	$\boldsymbol{\mu}_2 = (2.0, 2.0)'$, $\boldsymbol{\mu}_3 = (2.0, -2.0)'$
				$\neq$	$\boldsymbol{\mu}_2 = (2.2, 2.2)'$, $\boldsymbol{\mu}_3 = (2.2, -2.2)'$
			s_3	$=$	$\boldsymbol{\mu}_2 = (2.6, 2.5)'$, $\boldsymbol{\mu}_3 = (2.6, -2.5)'$
				$\neq$	$\boldsymbol{\mu}_2 = (2.7, 2.7)'$, $\boldsymbol{\mu}_3 = (2.7, -2.7)'$
8	2	2	s_1	$=$	$\boldsymbol{\mu}_2 = (0.0, 0.0)'$, $\boldsymbol{\Sigma}_2 = 0.100\boldsymbol{I}$
				$\neq$	$\boldsymbol{\mu}_2 = (0.0, 0.0)'$, $\boldsymbol{\Sigma}_2 = 0.086\boldsymbol{I}$
			s_2	$=$	$\boldsymbol{\mu}_2 = (0.0, 0.0)'$, $\boldsymbol{\Sigma}_2 = 0.055\boldsymbol{I}$
				$\neq$	$\boldsymbol{\mu}_2 = (0.0, 0.0)'$, $\boldsymbol{\Sigma}_2 = 0.051\boldsymbol{I}$
			s_3	$=$	$\boldsymbol{\mu}_2 = (0.0, 0.0)'$, $\boldsymbol{\Sigma}_2 = 0.022\boldsymbol{I}$
				$\neq$	$\boldsymbol{\mu}_2 = (0.0, 0.0)'$, $\boldsymbol{\Sigma}_2 = 0.020\boldsymbol{I}$
9	2	2	s_1	$=$	$\boldsymbol{\mu}_2 = (2.6, 0.0)'$, $\boldsymbol{\Sigma}_1 = \boldsymbol{A}_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{A}_2$
				$\neq$	$\boldsymbol{\mu}_2 = (2.7, 0.0)'$, $\boldsymbol{\Sigma}_1 = \boldsymbol{A}_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{A}_2$
			s_2	$=$	$\boldsymbol{\mu}_2 = (3.3, 0.0)'$, $\boldsymbol{\Sigma}_1 = \boldsymbol{A}_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{A}_2$
				$\neq$	$\boldsymbol{\mu}_2 = (3.4, 0.0)'$, $\boldsymbol{\Sigma}_1 = \boldsymbol{A}_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{A}_2$
			s_3	$=$	$\boldsymbol{\mu}_2 = (4.6, 0.0)'$, $\boldsymbol{\Sigma}_1 = \boldsymbol{A}_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{A}_2$
				$\neq$	$\boldsymbol{\mu}_2 = (4.6, 0.0)'$, $\boldsymbol{\Sigma}_1 = \boldsymbol{A}_1$, $\boldsymbol{\Sigma}_2 = \boldsymbol{A}_2$

Table 1: Parameters of uni- and bivariate simulations. Strong, middle and poor mixture degrees are respectively called s_1, s_2 and s_3. Same mixing proportions between clusters are noted "=", different ones are noted "≠" and correspond to $p_1 = 0.3, p_2 = 0.7$ when $K = 2$ and to $p_1 = 0.1, p_2 = 0.3, p_3 = 0.6$ when $K = 3$. If the centers $\boldsymbol{\mu}_1$ are not explicit, they correspond to $\boldsymbol{\mu}_1 = 0.0$ or $\boldsymbol{\mu}_1 = (0.0, 0.0)'$ according to the space dimension d. In the same way, if any $\boldsymbol{\Sigma}_k$ is not explicit, it corresponds to $\boldsymbol{\Sigma}_k = \boldsymbol{I}$. We note also $\boldsymbol{A}_1 = \mathrm{diag}(4,1)$ and $\boldsymbol{A}_2 = \mathrm{diag}(1,9)$.

K) for simulations 4, 8 and dimension 4, and free diagonal matrices ($\boldsymbol{\Sigma}_k = \mathrm{diag}(\sigma^2_{1,k}, \ldots, \sigma^2_{d,k}), 1 \leq k \leq K$) for simulation 9.

At last, as regards the tested numbers of clusters, we successively use $K = 1$ to K_{sup} clusters, where K_{sup} is the smallest integer not less than $n^{0.3}$ (see Bozdogan [6]).

5.1.3 Simulations results

The results of uni- and bivariate simulations are summarized in Table 2, and in Table 3 for the dimension 4. We do not give the AIC and L criteria since both of them overestimate the right number of clusters (L more than AIC).

Model with equal mixing proportions

Data with equal mixing proportions The CL2 and CLM2 criteria are fully equivalent to the CL and CLM ones since there are never "too small" clusters. the CL and CLM criteria are often equivalent, and even better than, all other criteria *if a high separation* is present. A good separation of clusters is needed (s_2 or s_3) to detect the right number of clusters. The AWE criterion has the same behaviour. So the classification likelihood seems to detect clusters only if they are well-separated. On the other hand, the NEC criterion detects clusters less separated: although it is a classification criterion, it does not use any classification likelihood.

The effect of the dimension is important for all criteria but especially for the ones which use a CL term. In $d = 2$, clusters are detected with lower separation then in $d = 1$. These last criteria overestimate very rarely the good number of clusters. We also note that the CL criterion is better than the CLM and AWE ones: the CLM criterion underestimates the number of components more than CL and AWE much more.

Data with different mixing proportions For the same reasons than before, the CL2 and CLM2 criteria are respectively equivalent to the CL and CLM ones. The conclusion about dimension and separation of clusters are still good but all criteria decrease a lot: the mixture criteria AIC, AIC3, BIC and ICOMP become very bad since they overestimate highly the number of clus-

p_k		d	Sep.	K	AIC3	BIC	ICOMP	NEC	CLM	CLM2	CL	CL2	AWE
data	model												
equal	equal	1	s_1	=	69	58	59	61	24	24	27	27	17
				− / +	26 / 6	41 / 1	9 / 32	30 / 9	74 / 1	74 / 1	71 / 2	71 / 2	83 / 0
			s_2	=	92	89	77	82	42	42	43	43	36
				− / +	2 / 6	9 / 2	0 / 23	18 / 0	58 / 0	58 / 0	57 / 0	57 / 0	64 / 0
			s_3	=	92	97	88	79	93	93	94	94	81
				− / +	0 / 8	0 / 3	0 / 12	0 / 21	7 / 0	7 / 0	6 / 0	6 / 0	19 / 0
		2	s_1	=	96	100	44	88	43	43	51	51	13
				− / +	0 / 4	0 / 0	0 / 56	12 / 0	57 / 1	57 / 1	44 / 5	44 / 5	87 / 0
			s_2	=	93	99	43	90	86	86	93	93	57
				− / +	0 / 8	0 / 1	0 / 58	10 / 0	14 / 0	14 / 0	7 / 1	7 / 1	43 / 0
			s_3	=	97	99	57	87	100	100	100	100	100
				− / +	0 / 3	0 / 1	0 / 43	1 / 11	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
	free	1	s_1	=	54	40	58	22	7	14	0	0	13
				− / +	42 / 3	60 / 0	23 / 19	20 / 58	71 / 22	73 / 12	66 / 34	67 / 33	86 / 1
			s_2	=	87	71	87	60	22	39	6	14	33
				− / +	8 / 6	27 / 2	3 / 10	17 / 23	57 / 21	57 / 4	56 / 39	56 / 30	67 / 0
			s_3	=	92	98	87	67	81	93	83	92	64
				− / +	0 / 8	0 / 2	0 / 13	3 / 31	6 / 13	6 / 2	5 / 12	6 / 3	36 / 0
		2	s_1	=	93	98	86	31	4	42	4	41	13
				− / +	0 / 7	3 / 0	0 / 14	1 / 68	43 / 53	50 / 8	36 / 60	38 / 21	88 / 0
			s_2	=	91	100	84	34	19	88	5	77	51
				− / +	0 / 9	0 / 0	0 / 16	2 / 64	11 / 70	12 / 0	6 / 89	6 / 18	49 / 0
			s_3	=	94	100	81	62	49	97	39	95	99
				− / +	0 / 6	0 / 0	0 / 19	3 / 35	0 / 51	0 / 3	0 / 61	0 / 5	1 / 0
different	equal	1	s_1	=	7	14	9	50	31	31	19	19	31
				− / +	10 / 83	20 / 66	33 / 58	1 / 49	67 / 2	67 / 2	67 / 14	67 / 14	68 / 1
			s_2	=	2	4	7	44	32	32	32	32	33
				− / +	0 / 98	6 / 90	11 / 82	0 / 56	67 / 1	67 / 1	67 / 1	67 / 1	67 / 0
			s_3	=	2	3	4	66	48	48	50	50	38
				− / +	0 / 98	0 / 97	0 / 96	3 / 31	52 / 0	52 / 0	42 / 8	42 / 8	62 / 0
		2	s_1	=	33	64	15	51	30	30	19	19	17
				− / +	1 / 67	9 / 27	0 / 85	37 / 13	58 / 12	58 / 12	49 / 32	49 / 32	83 / 0
			s_2	=	18	41	13	64	50	50	43	43	38
				− / +	1 / 81	3 / 57	0 / 87	33 / 3	41 / 9	41 / 9	30 / 28	30 / 28	63 / 0
			s_3	=	2	18	3	72	79	79	78	78	78
				− / +	0 / 98	0 / 83	0 / 98	27 / 2	13 / 8	13 / 8	6 / 16	6 / 16	23 / 0
	free	1	s_1	=	68	56	69	22	6	23	6	7	32
				− / +	21 / 11	43 / 1	8 / 23	27 / 51	63 / 31	63 / 13	59 / 36	59 / 34	67 / 1
			s_2	=	90	84	83	37	12	33	13	18	36
				− / +	2 / 8	14 / 1	1 / 16	27 / 37	56 / 32	57 / 10	50 / 37	51 / 31	64 / 0
			s_3	=	90	100	91	86	72	98	89	99	64
				− / +	0 / 10	0 / 0	0 / 9	1 / 13	0 / 28	0 / 2	0 / 11	0 / 1	36 / 0
		2	s_1	=	88	88	78	26	16	62	19	63	39
				− / +	3 / 9	13 / 0	4 / 18	18 / 56	30 / 54	33 / 5	19 / 62	32 / 6	61 / 0
			s_2	=	91	98	78	38	30	88	22	87	57
				− / +	0 / 9	1 / 1	1 / 21	16 / 47	5 / 65	8 / 4	3 / 76	4 / 9	43 / 0
			s_3	=	92	100	88	71	36	98	22	98	100
				− / +	0 / 8	0 / 0	0 / 13	8 / 21	0 / 64	0 / 3	0 / 78	0 / 3	0 / 0

Table 2: Percentages of under-estimation (−), right estimation (=) and overestimation (+) of the number of clusters K in the uni- and bivariate simulations. Mixing proportions of data can be equal or free. Mixing proportions of the model are equal and free too. For instance, the criterion AIC3, with a middle mixture degree (s_2) in the univariate case, clusters having the same mixing proportions and the model imposing equal mixing proportions, detects in 92% of cases the right number of clusters, in 2% underestimates it and overestimates it in 6%.

ters, but the classification criteria CLM, CL, AWE and NEC seems better. So the classification criteria resist better an inappropriate model on mixing proportions.

Model with free mixing proportions In such a case (data have either equal or different mixing proportions), the CL and CLM criteria overestimate highly the number of clusters with a stronger overestimation for the CL criterion on average. However the CL and CLM criteria are pretty good with a slight advantage to the CL one in dimension 1 and if clusters are separated enough. This surprising good behaviour of the CL criterion (83 and even 89% of success in $d = 1$ with s_3) seems to contradict the theory for which K_{sup} ought to be the best value. On the other hand, in dimension 2, the CL and CLM criteria have poor success even if clusters are well-separated. The AWE criterion is often a better classification criterion in this case although it tends to underestimate a lot the number of clusters. The CL2 and CLM2 criteria give excellent results: they eliminate almost all overestimations of the CL and CLM criteria. They are improved with the space dimension and the mixture degree.

Results in dimension 4 Clusters are well-separated in this example. If we impose equal proportions, we obtain the same results than before: mixture criteria highly overestimate the number of clusters. However the BIC criterion has a good behaviour. The CL and CLM criteria are excellent (about 90% success) with a slight advantage to the CL one again. The CL2 and CLM2 criteria are still equivalent to the CL and CLM ones in this case. AWE can not decide between 4 and 5 clusters. NEC only detects 4 clusters.

If the mixing proportions are free, all mixture criteria get better. AWE also improves, NEC hesitates between $K = 4$ and $K > 4$ but CLM strongly overestimates the number of components. On the other hand, CL keeps a pretty good success rate with 73%. The reason is that CEM seldom converges when $K > 4$: some clusters were empty. The CL2 and CLM2 criteria work very well since they avoid overestimation of CL and CLM (100% success rate).

5.2 A real data set: the Iris

The Fisher's Iris (Anderson [2]) are a classical data set composed by 150 flowers measured on four variables and divided in three groups of same size (Iris Setosa, Iris Virginica, Iris Versicolor).

We use model with equal variance matrices ($\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_K = \mathbf{\Sigma}$) and mixing proportions equal and free. We compute all previous criteria by starting EM *and* CEM 20 times at random and by searching between 1 and 5 clusters (K_{sup} of Bozdogan).

Values of all criteria are reported in Table 6. We note than only CL2, CLM2, AWE and ICOMP detect three clusters with both equal and free mixing proportions. In the equal proportions case, NEC finds two clusters and BIC, CLM and CL find four clusters. Some evidence of hybridization between Iris Virginica and Versicolor is commonly accepted and we can divide this couple in one, two or even three groups. So, obtaining two, three or four clusters is coherent. In the case of the free mixing proportions model, ICOMP, NEC, CL2, CLM2, AWE and BIC stay in this good area. But criteria CL and CLM overestimate the number of components.

We note that, with the equal mixing proportions model, we have no longer the equivalence between CL, CL2 and CLM, CLM2 that we pointed out in the previous simulations. This may be the fact of the poor number of flowers (150) which allows quite easily to obtain clusters of size smaller than the number of free parameters.

6 Concluding remarks

In conclusion, our four strategies to use the classification likelihood may be recommended to search well-separated and equal mixing proportions clusters. Their simplicity of computation and the meaningful interpretation of the classification likelihood encourage to use them. In the other hand, if we let the mixing proportions free, only CL2 and CLM2 seem reliable enough.

As an extension of this work, investigations concerning behaviour of CL, CL2 and CLM, CLM2 would require to go further. We could study the dependence of these criteria on the sample size by using several values for this parameter ($n = 40, 100, 500$). We could also study their behaviour with non-Gaussian mixtures (Poisson, ...).

References

[1] H. Akaike. Information theory as an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, Akademiai Kiado, 1973.

[2] E. Anderson. The irises of the gaspé peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

model	K	AIC3	BIC	ICOMP	NEC	CLM	CLM2	CL	CL2	AWE
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
equal p_k	4	0	10	0	**93**	13	13	7	7	**53**
	5	**43**	**80**	**50**	7	**87**	**87**	**93**	**93**	47
	6	13	10	0	0	0	0	0	0	0
	7	**43**	0	**50**	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
free p_k	4	0	0	0	**67**	0	0	0	0	37
	5	**67**	**90**	**80**	17	13	**100**	**73**	**100**	**63**
	6	30	10	20	17	37	0	17	0	0
	7	3	0	0	0	**50**	0	10	0	0

Table 3: Percentage of choosing K clusters for the simulation in dimension 4.

model	K	AIC3	BIC	ICOMP	NEC	CLM	CLM2	CL	CL2	AWE
	1	801.8	829.9	804.6	8.2235	-379.9	-379.9	-379.9	-379.9	942.0
	2	663.8	700.0	688.1	**0.0001**	-304.9	-304.9	-304.9	-304.9	844.2
equal p_k	**3**	578.7	622.9	**650.0**	0.0517	-258.8	**-258.8**	-258.5	**-258.5**	**803.8**
	4	546.0	**598.2**	675.9	0.0458	**-236.2**	-	**-236.2**	-	811.2
	5	**545.4**	605.7	727.4	0.1587	-239.5	-	-237.7	-	866.0
	1	801.8	829.9	804.6	8.1804	-379.9	-379.9	-379.9	-379.9	942.0
	2	649.8	688.0	671.8	**0.0001**	-296.4	-296.4	-296.4	-296.4	840.2
free p_k	**3**	584.7	632.9	**649.9**	0.0514	-258.7	**-258.7**	-258.5	**-258.5**	**829.6**
	4	533.0	**591.4**	655.6	0.0460	-226.1	-	-254.2	-	886.0
	5	**527.5**	595.8	702.2	0.0736	**-217.9**	-	**-217.2**	-	877.0

Table 4: Values of criteria for the Iris with equal and free mixing proportions.

[3] J.D. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[4] H. Bozdogan. *Multi-Sample Cluster Analysis and Approaches to Validity Studies in Clustering Individuals.* PhD thesis, Department of Mathematics, University of Illinois, Chicago, IL 60680, 1981.

[5] H. Bozdogan. On the information-based measure of covariance complexity an its application to the evaluation of the multivariate linear models. *Communication in Statistics, Theory and Methods*, 19:221–278, 1990.

[6] H. Bozdogan. *Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix.* Information and Classification. Springer-Verlag, Heidelberg, 1993.

[7] H. Bozdogan and S.L. Slove. Multi-Sample Cluster Analysis using Akaike's Information Criterion. *Annals of Institute of Statistical Mathematics*, 36:163–180, 1984.

[8] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14:315–332, 1992.

[9] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 1994. to appear.

[10] A. Cutler and M.P. Windham. Information-Based Validity Functionals for Mixture Analysis. In H. Bozdogan, editor, *Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling*, pages 149–170, Amsterdam, 1993. Kluever (Academic Publishers).

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data with the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[12] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[13] M.J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.