INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

An entropy criterion for assessing the number of clusters in a mixture model

Gilles CELEUX
Gilda SOROMENHO

N° 1874
Avril 1993

PROGRAMME 3

Intelligence artificielle,
Systèmes cognitifs et
Interaction homme-machine

An entropy criterion for assessing the number of clusters in a mixture model

Gilles Celeux
INRIA Rocquencourt 78153 Le Chesnay

Gilda Soromenho
LEAD Lisbon University Alameda da Universidade 1600 Lisboa

Abstract

In this paper, we consider an entropy criterion to estimate the number of clusters arising from a mixture model. This criterion is derived from a relation linking the likelihood and the classification likelihood of a mixture. Its performances are investigated through Monte-Carlo numerical experiments and show favourable results as compared with other classical criteria.

Keywords: *Cluster Analysis, Gaussian Mixture, Entropy, Bayesian Criteria.*

Un critère d'entropie pour estimer le nombre de classes d'un modèle de mélange

Résumé

Nous proposons un critère d'entropie pour évaluer le nombre de classes d'une partition en nous fondant sur un modèle de mélange de lois de probabilité. Ce critère se déduit d'une relation liant la vraisemblance et la vraisemblance classifiante d'un mélange. Des simulations de Monte-Carlo illustrent ses qualités par rapport à des critères plus classiques.

Mots-clés : *classification, mélange de lois normales, entropie, critères bayésiens.*

1 Introduction

The choice of a "good" number of clusters in cluster analysis is a difficult problem. This question has to be considered in a precise framework to be addressed in a satisfactory manner. Basing cluster analysis on a probabilistic model provides a fruitful line of approach (see Bock 1985, 1989). Many authors have considered the problem of choosing the number of clusters within the context of multivariate mixture models (see Bock 1985, Celeux 1986, Bozdogan 1992, Banfield and Raftery 1992, McLachlan and Basford 1989, Windham and Cutler 1991, Bryant 1993 among others). These authors addressed the problem of assessing the number of components in a mixture in a clustering purpose. Some others authors have considered the problem of estimating the number of components in a mixture of distributions (see Wolfe 1970, Titterington, Smith and Makov 1985, Aitkin and Rubin 1985, McLachlan 1987 among others) without any reference to cluster analysis. Estimating the number of components is a difficult problem for which several approaches are in competition. The traditional likelihood ratio test can not be used since the classical regularity conditions which ensure a χ^2 distribution for the likelihood ratio statistic do not hold under any null hypothesis (see Aitkin and Rubin 1985 for a thorough discussion of this problem). The most efficient significance test is probably the procedure proposed by McLachlan (1987) which makes use of a parametric bootstrap to approximate the p-values of the generalized likelihood ratio (GLR) test. An other approach, that it will be discussed further in this paper, is to consider general techniques for model choice based on information criteria or Bayes factor (see, for instance Banfield and Raftery 1992 or Bozdogan 1992).

When cluster analysis is the main concern, it is implicitly assumed that each cluster can be approximately regarded as a sample from one of the mixture components. For instance, when a Gaussian mixture is considered for cluster analysis, the means of the mixture components are supposed to be significantly different. And, when the means are equal, there is only one component from the cluster analysis point of view even when the variance matrices of the mixture components are different. Then a point is worth noting. Even when they are concerned with cluster analysis, the afore mentioned authors - except Windham and Cutler 1991- proposed criteria, penalizing the log-likelihood statistic, devoted to select parsimonious mixture models. But those criteria are not devoted to measure directly the ability of the mixture to

provide a clustering structure. In this paper, we propose an entropy criterion which aims to choose the mixture model from which a clustering structure arises with the greatest evidence.

In Section 2, we present the mixture model and report briefly criteria proposed in the literature for approaching the problem of the number of components in a mixture. In Section 3, we analyze the relations between the mixture model and cluster analysis. Then we present and discuss our entropy criterion for assessing the number of clusters associated to a mixture model. In Section 4, we report numerical Monte-Carlo experiments in which the practical behaviour of the entropy criterion is investigated and compared with the behaviour of other criteria. A concluding section summarizes the main points of this paper.

2 Criteria for the number of components in a mixture

In the mixture model, the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are assumed to be a sample from a probability distribution with density

$$\phi(\mathbf{x}) = \sum_{k=1}^{K} p_k f(\mathbf{x}, \mathbf{a}_k) \tag{2.1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$) and the $f(\mathbf{x}, \mathbf{a_k})$ are densities from the same parametric family. For instance, $f(\mathbf{x}, \mathbf{a_k})$ denotes the d-dimensional Gaussian density with mean μ_k and variance matrix Σ_k and $\mathbf{a}_k = (\mu_k, \Sigma_k)$. The log-likelihood of the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$

$$L(K) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} p_k f(\mathbf{x}_i, \mathbf{a}_k)\right] \tag{2.2}$$

is an increasing function of K. Thus, $L(K)$ can not be used as a selection criterion which balances model fit and model complexity for choosing the number K of components in the mixture. Various criteria, to be minimized, have been proposed to measure a model's suitability taking this objective into account.

The *Akaike information criterion* (Akaike 1974) has been considered by Bozdogan and Sclove (1984) in the mixture context. It takes the form

$$\mathrm{AIC}(K) = -2L(K) + 2\nu(K) \tag{2.3}$$

where $\nu(K)$ is the number of free parameters in the mixture model with K-components. For instance for a d-dimensional Gaussian mixture, $\nu(K) = (K-1) + dK + dK(K+1)/2$.

The *Bayesian information criterion* as defined by Schwarz (1978) is an approximation of the exact Bayes solution to the problem of selecting the appropriate model. It is defined by

$$\mathrm{BIC}(K) = -2L(K) + \nu(K) \ln n. \tag{2.4}$$

Note that this criterion has also been proposed by Rissanen (1989) from another point of view based on coding theory.

Banfield and Raftery (1992) have also suggested an approximate Bayesian solution to the choice of the number of components in a mixture. Their approximation is somewhat different and lead to the so-called *approximate weight of evidence* which takes the form

$$\mathrm{AWE}(K) = -2L(K) + 2\nu(K)(\frac{3}{2} + \ln n). \tag{2.5}$$

Many authors (see for instance Koehler and Murphree 1988) observed that AIC criterion is order inconsistent (a criterion is order consistent if, as the sample size increases, it is minimised at the true order of the model with probability which approaches unity) and tends to overfit models. In the mixture context, it means that AIC tends to overestimate the correct number of components. In contrast, the BIC criterion has been proved to be order consistent in some contexts and under suitable conditions. In practical situations, BIC is expected to give an answer to the overparametrization of AIC. But, in the mixture context, there is no available consistency result for the BIC criterion. As pointed out in Titterington, Smith and Makov 1985 pp. 159, the reason is probably that theoretical justifications for criteria such as AIC or BIC rely on the same conditions as the usual asymptotic theory of the GLR test, and, as mentioned in the introduction, these standard regularity conditions do not hold in the mixture context. As for AWE, it is clear that it penalizes more drastically high order models than BIC.

Bozdogan (1990, 1992 and references therein) has proposed an *informational complexity criterion*, called ICOMP, for choosing parsimonious models. This criterion measures the complexity of a model by an approximation of the Fisher information of the model. In the context of a d-dimensional Gaussian mixture, it takes the form

$$\begin{aligned}
&\mathrm{ICOMP}(K) = -2L(K) \\
&+ \frac{dK + \frac{dK(K+1)}{2}}{2} \ln \frac{\sum_{k=1}^{K} \left[p_k^{-1}\mathrm{tr}\Sigma_k + 1/2\mathrm{tr}\Sigma_k^2 + (\mathrm{tr}\Sigma_k)^2 + \sum_{j=1}^{d}(\Sigma_k^{jj})^2\right]}{dK + dK(K+1)/2} \\
&- 1/2\left[(d+2)\sum_{k=1}^{K} \ln|\Sigma_k| - d\sum_{k=1}^{K} \ln(p_k n)\right] - Kd/2 \ln(2n). \qquad (2.6)
\end{aligned}$$

In this equation, Σ_k^{jj} denotes the jth diagonal term of the variance matrix Σ_k.

An interesting criterion, the *minimum information ratio*, has been proposed by Windham and Cutler (1991). This criterion, abbreviated MIR, is a measure of the proportion of information about the mixture parameters available without knowing the subpopulation memberships of the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This measure is the smallest eigenvalue of the information matrix $F_C^{-1}F$ where F is the Fisher information matrix for the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and F_C is the Fisher information matrix for the classified sample. The MIR can be approximated when using the EM algorithm (Dempster, Laird and Rubin 1977) for estimating the mixture parameters: MIR is one minus the EM rate of convergence. Then, if $(\theta^m, m > 0)$ is a sequence of parameter estimates produced by the iteration of EM, MIR is estimated by $1 - ||\theta^{m+1} - \theta^m||/||\theta^m - \theta^{m-1}||$, for m large enough, where $||.||$ is any convenient norm.

The MIR is an interesting criterion in a cluster analysis context since it may be interpreted as measuring the ability of the data to distinguish the component densities and a large MIR suggests a good clustering structure. However, the MIR presents some drawbacks. First, it is unable to compare the situations $K = 1$ versus $K > 1$. Moreover, in many circumstances, MIR revealed to be numerically difficult to calculate (when the ratio $||\theta^{m+1} - \theta^m||/||\theta^m - \theta^{m-1}||$ is near the form 0/0). On the other hand, Windham and Cutler have remarked that the modified criterion

$$\mathrm{AMIR}(K) = \mathrm{MIR}(K)(L(K) - L(1))$$

out performed MIR(K) in many practical situations since MIR(K) tends to underestimate the number of mixture components.

3 The entropy criterion

The entropy criterion that we propose in this section is aimed to measure the ability of a mixture model to provide well separated clusters. It is derived from a relation highlighting the differences between the maximum likelihood (ML) approach and the classification maximum likelihood (CML) approach of the mixture problem.

3.1 The two approaches of the mixture problem

The ML approach of the mixture problem consists in optimizing the log-likelihood $L(K)$ of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ defined in (2.2) using generally the EM algorithm (Dempster, Laird and Rubin 1977).

In the CML approach, the indicator vectors $\mathbf{z}_i = (z_{ij}, j = 1, \ldots, K)$ with $z_{ij} = 1$ or 0 according as $\mathbf{x}_i$ $(1 \leq i \leq n)$ has been drawn from the jth component or from another one, identifying the mixture component origin are treated as unknown parameters. The CML approach is employed only when cluster analysis is in order. The CML criterion to be optimized takes the form

$$CL(K) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \ln [p_k f(\mathbf{x}_i, \mathbf{a}_k)]. \tag{3.7}$$

Maximizing the CML criterion produces biased estimates of the mixture parameters (cf. for instance Bryant 1991). The source of the bias is essentially "all or nothing" classification. It can be severe, but it is supportable if the mixture components are well separated.

We turn now to an interesting simple relation between the log-likelihood $L(K)$ and the classification log-likelihood $CL(K)$. This relation was first exihibited by Hattaway (1986). Denoting

$$t_{ik} = \frac{p_k f(\mathbf{x}_i, \mathbf{a}_k)}{\sum_{j=1}^{K} p_j f(\mathbf{x}_i, \mathbf{a}_j)}$$

the posterior probability that x_i arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$), direct calculation show that

$$L(K) = C(K) + E(K) \tag{3.8}$$

with

$$C(K) = \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \ln [p_k f(\mathbf{x}_i, \mathbf{a}_k)].$$

and

$$E(K) = \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \ln t_{ik} \geq 0.$$

The relation (3.8) provides a decomposition of the log-likelihood $L(K)$ in a CML term $C(K)$ and the entropy term $E(K)$ which measures the overlapping of the mixture components. If the mixture components are very well separated the posterior probabilities t_{ik} tend to define a partition of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$: For each $\mathbf{x}_i$ there exists k ($1 \leq k \leq K$) such that $t_{ik} \simeq 1$. Then we have $t_{ik} \ln t_{ik} \simeq 0$ for all i ($1 \leq i \leq n$) and k ($1 \leq k \leq K$) (by convention $0 \ln 0 = 0$ since $\lim t \ln t = 0$ as $t \to 0$) and, as a consequence, $E(K) \simeq 0$. On the contrary, if the mixture components are intricated, $E(K)$ takes a large value. Thus, $E(K)$ can be regarded as a measure of the ability of the K-component mixture model to provide a relevant partition of the data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Moreover, from the relation (3.8), $E(K)$ appears to be a quantity measuring the difference between the ML and the CML approaches for the mixture problem: Both approaches can be thought of as equivalent as $E(K)$ is near 0. On the contrary, the bias of the CML approach can be expected to be severe for a large value of $E(K)$.

Remark 1: As $K = 1$ a natural way to extend $E(K)$ is to set $t_{i1} = 1$ for all i ($1 \leq i \leq n$) and then

$$E(1) = \sum_{i=1}^{n} t_{i1} \ln t_{i1} = 0.$$

The relation (3.8) holds with $K = 1$,

$$L(1) = C(1) + E(1) \tag{3.9}$$

with

$$C(1) = \sum_{i=1}^{n} t_{i1} \ln f(\mathbf{x}_i, \mathbf{a}_1).$$

Remark 2: An other measure proposed for assessing the number of clusters in a mixture makes use of the posterior probabilities t_{ik}. It is the so called partition coefficient PC $= \sum_{ik} t_{ik}^2$ (Bezdek 1981). Numerical experiments, reported in Windham and Cutler 1991, show clearly that the PC criterion tends to underestimate the correct number of clusters.

3.2 The criterion

The quantity $E(K)$ is the basis for a criterion assessing the number of clusters arising from a mixture. But, $E(K)$ need some modification to be effective.

The entropy $E(K)$ cannot be used directly as a criterion to assess the number of clusters in a mixture for two reasons. First, $E(K) \geq E(1)$ for any $K > 1$. Moreover, $L(K)$ is an increasing function of K and, as a consequence, the $E(K)$ are not comparable for different values of K. But, from (3.8) and (3.9), we can write

$$1 = \frac{C(K) - C(1)}{L(K) - L(1)} + \frac{E(K)}{L(K) - L(1)}, \quad K > 1 \tag{3.10}$$

and we propose

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

as a criterion to be minimized for assessing the number of clusters arising from a mixture.

However, there is still a problem for $K = 1$ since NEC(1) is of the form $\frac{0}{0}$ and is not defined. Thus, we are unable to compare the situations $K = 1$ versus $K > 1$ directly when using the criterion NEC(K). A solution is now proposed to overcome this difficulty for Gaussian mixtures. (Notice that Gaussian mixture is the only one mixture model considered in practical situations for analyzing multivariate quantitative data.) Before describing the way we proceed, it is important to recall that we aim to choose the mixture model which gives rise with the greatest evidence to a clustering structure. From this point of view, it may happen that an underlying mixture density is

not associated with any clustering model. For instance, a Gaussian mixture with equal means does not indicate that there is a relevant classification of the data. Our procedure works as follows:

- We estimate the parameters of the Gaussian mixture, using the maximum likelihod approach, for different values of K $(2 \leq K \leq K_{sup})$, K_{sup} being a reasonable upper bound of the number of mixture components. (See Bozdogan 1992, for a discussion concerning the choice of K_{sup}.) Then we determine $K*$ which minimizes $\mathrm{NEC}(K), (2 \leq K \leq K_{sup})$.

- To decide $K = K*$ or $K = 1$, we estimate the parameters of a $K*$ component Gaussian mixture with equal means $\mu_1 = \ldots = \mu_{K*} = \overline{\mathbf{x}}$, where $\overline{\mathbf{x}}$ represents the sample mean of the data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Denoting $\tilde{L}(1)$ and $\tilde{E}(1)$ the resulting log-likelihood and entropy, we compute the ratio

$$\mathrm{NEC}(1) = \frac{\tilde{E}(1)}{\tilde{L}(1) - L(1)}$$

  and we choose $K*$ clusters if $\mathrm{NEC}(K*) \leq \mathrm{NEC}(1)$, otherwise we declare for no clustering structure in the data.

Acting in such a way, we compare two rival models: a $K*$ component Gaussian mixture with different means, from which a partition of the data can easily be derived, and a $K*$ component Gaussian mixture with equal means, from which there is no evidence for partioning the data.

4 Numerical experiments

In this section, we compare the practical behaviour of the criterion NEC with the criteria AIC, BIC and AWE on the basis of Monte-Carlo experiments. We generated 20 samples from each type of simulated data set.

4.1 Experiment conditions

We simulated four univariate distributions for the two sample sizes $n = 200$ and $n = 50$: The first one was the standard Gaussian distribution with mean 0 and standard deviation 1. The second one was a two-component Gaussian mixture with means $\mu_1 = 0$ and $\mu_2 = 2$, with standard deviations

$\sigma_1 = \sigma_2 = 1$ and with equal proportions. The third one was a two-component Gaussian mixture with the same means and standard deviations than the second one and with proportions $p_1 = 0.7$ and $p_2 = 0.3$. The last one was a three-component Gaussian mixture with equal proportions, with means $\mu_1 = 0, \mu_2 = 2, \mu_3 = 4$ and standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

We also simulated three bivariate distributions with the sample sizes $n = 200$ and $n = 50$: The first one was a standard Gauusian distribution with mean vector 0 and variance matrix $\Sigma = I$. The second one was a two-component Gaussian mixture with equal proportions, mean vectors $\mu_1' = (0,0), \mu_2' = (2,2)$ and variance matrices $\Sigma_1 = \Sigma_2 = I$. The third one was a three-component Gaussian mixture with equal proportions, mean vectors $\mu_1' = (0,0), \mu_2' = (2,2), \mu_3' = (2,-2)$ and variance matrices $\Sigma_1 = \Sigma_2 = \Sigma_3 = I$.

The last example consists in simulating a sample of $n = 625$ observations from a four dimensional five-component Gaussian mixture with the following parameters:

$$\begin{aligned}
p_1 &= 0.12,\ \mu_1' = (10,12,10,12),\ \Sigma_1 = I \\
p_2 &= 0.16,\ \mu_2' = (8.5,10.5,8.5,10.5),\ \Sigma_2 = I \\
p_3 &= 0.20,\ \mu_3' = (12,14,12,14),\ \Sigma_3 = I \\
p_4 &= 0.24,\ \mu_4' = (13,15,7,9),\ \Sigma_4 = 4I \\
p_5 &= 0.28,\ \mu_5' = (7,9,13,15),\ \Sigma_5 = 9I.
\end{aligned}$$

This Gaussian mixture has been considered previously by Bozdogan (1992) to investigate the behaviour of the criterion ICOMP. As pointed out by Bozdogan, this five-component mixture gives rise to highly overlapping clusters.

The parameters of the Gaussian mixture models have been estimated using the EM algorithm. For the univariate and bivariate distributions the component variances or the component variance matrices where assumed to be equal whem running the EM algorithm. No restriction was placed on the mixture model for the four dimensional mixture: we ran the EM algorithm assuming different variance matrices for the mixture components. For each run, to attenuate its initial-position dependence, we initiated the EM algorithm with the true parameter values when it was possible, and otherwise we initiated it from a solution derived from the k-means algorithm. For each situation, we computed the normalized entropy NEC, as described

in Section 3, and the AIC, the BIC and the AWE criteria, as described in Section 2, from the estimated parameters.

4.2 Simulations results

Detailed results concerning the entropy criterion NEC are displayed in Table 1. In this table, the rows describe the underlying distribution: d represents the dimension of the sample space, n represents the sample size and the mixture parameters are described in the third column. Then, the mean values of the normalized entropy criterion NEC over the 20 trials and, into parentheses, their standard deviations are displayed. The results displayed in this table show a satisfactory behaviour of the criterion NEC. No surprisingly, NEC performs better for the largest sample size $n = 200$. Moreover, our procedure to compare $K = 1$ versus $K > 1$ seems to work well, despite a very slight tendency to favour the situation $K > 1$.

Table 2 displays the percent frequency of choosing K-component mixture for each of the criterion AIC, BIC, AWE and NEC for different values of K. Table 3 displays the same percent frequencies for the four dimensional mixture. From Table 2, it seems that the AWE criterion overpenalized high order models. The NEC criterion has an intermediate position between the AIC and the BIC criteria. As expected, AIC presents a slight tendency to overestimate the mixture model order. More surprisingly, the BIC criterion seems to underestimate the mixture model order. The entropy criterion has, in those experiments, the most satisfactory behaviour. It performs alike AIC without overestimating the correct number of clusters as AIC sometimes do. Table 3 corroborates the analogous behaviour for AIC and NEC criteria. Note that none of the criteria concluded to the presence of $K = 5$ clusters for this five-component mixture. But, as noted by Bozdogan (1992), choosing three clusters for those simulated data sets is a satisfactory solution. In this situation, observations arising from mixture components 1, 2 and 3 are merged into one cluster. But the 4 clusters solution, merging components 1 and 2 into one cluster is also quite reasonable. From those simulations, as shown in Table 4 which displayed the mean values and, into parentheses, standard deviation of criterion NEC, it appears that the entropy criterion NEC indicates clearly that two relevant partitions with three and four clusters can be derived from this five-component mixture.

			K=1	K=2	K=3	K=4	K=5
d=1	n=200	$p_1 = 1, \mu_1 = 0, \sigma_1 = 1$	16,41 (6.15)	17,93 (7.97)	22,83 (8.61)	29,32 (10.35)	
		$p_1 = 0.5$ $\sigma_1 = \sigma_2 = 1$ $\mu_1 = 0\ \mu_2 = 2$	13.83 (5.84)	10.98 (4.36)	15.72 (5.49)	22.10 (6.81)	47.86 (10.14)
		$p_1 = 0.7$ $\sigma_1 = \sigma_2 = 1$ $\mu_1 = 0\ \mu_2 = 2$	15.98 (6.12)	15.28 (5.95)	22.67 (7.31)	30.52 (10.11)	52.95 (16.82)
		$p_1 = p_2 = p_3 = 1/3$ $\sigma_1 = \sigma_2 = \sigma_3 = 1$ $\mu_1 = 0\ \mu_2 = 2\ \mu_3 = 4$	20.45 (12.88)	7.42 (2.35)	5.89 (0.97)	9.16 (3.41)	12.21 (4.18)
	n=50	$p_1 = 1, \mu_1 = 0, \sigma_1 = 1$	17.92 (7.21)	18.43 (8.91)	23.35 (10.72)	30.49 (12.74)	
		$p_1 = 0.5$ $\sigma_1 = \sigma_2 = 1$ $\mu_1 = 0\ \mu_2 = 2$	14.27 (7.42)	12.23 (5.82)	17.18 (7.79)	23.94 (9.61))	43.76 (12.45)
		$p_1 = 0.7$ $\sigma_1 = \sigma_2 = 1$ $\mu_1 = 0 \mu_2 = 2$	18.54 (10.17)	16.76 (7.64)	24.53 (9.86)	32.86 (13.22))	64.13 (18.12)
		$p_1 = p_2 = p_3 = 1/3$ $\sigma_1 = \sigma_2 = \sigma_3 = 1$ $\mu_1 = 0\ \mu_2 = 2\ \mu_3 = 4$	28.38 (11.82)	5.82 (2.10)	5.15 (0.62)	7.55 (2.84)	10.64 (3.58)
d=2	n=200	$p_1 = 1, \mu_1 = (0,0), \Sigma_1 = I$	3.41 (3.02)	16.84 (6.12)	18.69 (6.64)	24.47 (7.76)	
		$p_1 = p_2 = 0.5$ $\Sigma_1 = \Sigma_2 = I$ $\mu_1 = [0\ 0]\ \mu_2 = [2\ 2]$	8.86 (5.24)	6.24 (3.23)	8.12 (4.59)	10.44 (6.18)	
		$p_1 = p_2 = p_3 = 1/3$ $\Sigma_1 = \Sigma_2 = \Sigma_3 = I$ $\mu_1 = [0\ 0] \mu_2 = [2\ 2]$ $\mu_3 = [2 - 2]$	11.19 (4.78)	3.59 (2.44)	2.02 (0.49)	2.90 (0.88)	3.80 (1.57)
	n=50	$p_1 = 1, \mu_1 = (0,0), \Sigma_1 = I$	9.67 (4.15)	11.85 (6.99)	23.96 (7.44)	30.95 (8.94)	
		$p_1 = p_2 = 0.5$ $\Sigma_1 = \Sigma_2 = I$ $\mu_1 = [0\ 0]\ \mu_2 = [2\ 2]$	10.81 (6.68)	8.97 (4.12)	10.44 (5.83)	12.70 (7.56)	
		$p_1 = p_2 = p_3 = 1/3$ $\Sigma_1 = \Sigma_2 = \Sigma_3 = I$ $\mu_1 = [0\ 0]\ \mu_2 = [2\ 2]$ $\mu_3 = [2\ 2]$	11.97 (5.26)	4.43 (2.50)	2.35 (0.54)	2.92 (0.88)	3.96 (1.72)

Table 1: *Mean and standard deviation of the NEC criterion for the univariate and the bivariate distributions.*

				AIC	BIC	AWE	NEC
d=1	n=200	$p_1 = 1, \mu_1 = 0, \sigma_1 = 1$	$K = 1$	70	95	95	85
			$K = 2$	30	5	5	15
		$p_1 = 0.5$	$K = 1$	10	40	55	15
		$\sigma_1 = \sigma_2 = 1$	$K = 2$	80	60	45	80
		$\mu_1 = 0\ \mu_2 = 2$	$K = 3$	10	0	0	5
		$p_1 = 0.7$	$K = 1$	25	50	65	25
		$\sigma_1 = \sigma_2 = 1$	$K = 2$	65	50	35	70
		$\mu_1 = 0\ \mu_2 = 2$	$K = 3$	10	0	0	5
		$p_1 = p_2 = p_3 = 1/3$	$K = 1$	0	10	20	0
		$\sigma_1 = \sigma_2 = \sigma_3 = 1$	$K = 2$	5	15	50	10
		$\mu_1 = 0\ \mu_2 = 2\ \mu_3 = 4$	$K = 3$	80	75	30	80
			$K = 4$	15	0	0	10
	n=50	$p_1 = 1, \mu_1 = 0, \sigma_1 = 1$	$K = 1$	60	85	90	80
			$K = 2$	30	15	10	20
			$K = 3$	10	0	0	0
		$p_1 = 0.5$	$K = 1$	25	55	75	35
		$\sigma_1 = \sigma_2 = 1$	$K = 2$	60	45	25	55
		$\mu_1 = 0\ \mu_2 = 2$	$K = 3$	15	0	0	10
		$p_1 = 0.7$	$K = 1$	50	80	90	50
		$\sigma_1 = \sigma_2 = 1$	$K = 2$	35	20	10	40
		$\mu_1 = 0 \mu_2 = 2$	$K = 3$	15	0	0	10
		$p_1 = p_2 = p_3 = 1/3$	$K = 1$	0	20	50	0
		$\sigma_1 = \sigma_2 = \sigma_3 = 1$	$K = 2$	15	20	35	20
		$\mu_1 = 0\ \mu_2 = 2\ \mu_3 = 4$	$K = 3$	65	50	15	70
			$K = 4$	20	0	0	10
d=2	n=200	$p_1 = 1, \mu_1 = (0,0), \Sigma_1 = I$	$K = 1$	80	100	100	95
			$K = 2$	20	0	0	5
		$p_1 = p_2 = 0.5$	$K = 1$	5	35	50	10
		$\Sigma_1 = \Sigma_2 = I$	$K = 2$	80	60	50	80
		$\mu_1 = [0\ 0]\ \mu_2 = [2\ 2]$	$K = 3$	15	5	0	10
		$p_1 = p_2 = p_3 = 1/3$	$K = 1$	0	5	30	0
		$\Sigma_1 = \Sigma_2 = \Sigma_3 = I$	$K = 2$	5	25	50	5
		$\mu_1 = [0\ 0] \mu_2 = [2\ 2]$	$K = 3$	75	70	20	75
		$\mu_3 = [2\ -2]$	$K = 4$	15	0	0	20
			$K = 5$	5	0	0	0
	n=50	$p_1 = 1, \mu_1 = (0,0), \Sigma_1 = I$	$K = 1$	60	80	85	80
			$K = 2$	40	20	15	20
		$p_1 = p_2 = 0.5$	$K = 1$	10	60	75	10
		$\Sigma_1 = \Sigma_2 = I$	$K = 2$	65	40	25	70
		$\mu_1 = [0\ 0]\ \mu_2 = [2\ 2]$	$K = 3$	25	0	0	20
		$p_1 = p_2 = p_3 = 1/3$	$K = 1$	0	25	55	0
		$\Sigma_1 = \Sigma_2 = \Sigma_3 = I$	$K = 2$	20	30	40	10
		$\mu_1 = [0\ 0] \mu_2 = [2\ 2]$	$K = 3$	55	45	5	60
		$\mu_3 = [2\ -2]$	$K = 4$	20	0	0	30
			$K = 5$	5	0	0	0

Table 2: *Percent frequencies of choosing K clusters for the univariate and the bivariate distributions.*

K	AIC	BIC	AWE	NEC
2	0	5	30	0
3	60	65	60	55
4	40	30	10	45
5	0	0	0	0

Table 3: *Percent frequencies of choosing K clusters for the four dimensional mixture.*

K=1	K=2	K=3	K=4	K=5
3.15	0.59	0.41	0.38	0.84
(0.27)	(0.10)	(0.03)	(0.01)	(0.16)

Table 4: *Mean and standard deviation of the NEC criterion for the four dimensional mixture.*

5 Conclusion

We have proposed a criterion for assessing the number K of components in a mixture from a cluster analysis point of view. This criterion has been formed from the entropy of the posterior probabilities that the observations arose from one of the mixture components. This entropy measures the overlapping of the mixture components. The normalized entropy criterion NEC that we considered finally has been adapted for taking account the case $K = 1$. Numerical experiments show encouraging results compared with the performances of classical criteria. In particular, it seems that NEC does not suffer the overestimating tendency of AIC or an underestimating tendency as shown with BIC and AWE in the reported numerical simulations. Moreover, calculating the criterion NEC does not involve numerical difficulties such as those that can appear when using some criteria (ICOMP, MIR...).

Further experiments in various situations are needed to investigate more precisely the efficiency of NEC. And, our 'ad hoc' procedure to test $K = 1$ versus $K > 1$ can only be applied for Gaussian mixtures. Future research is needed to provide a general procedure to decide between $K = 1$ and $K > 1$ for any mixture model. But, henceforth, the entropy criterion NEC can be thought of as successful to propose a reasonable number of clusters arising from a mixture model.

References

Aitkin, M. and Rubin, D. M. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society*, Ser. B **47**, 67-75.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Trans. Auto. Control*, **19**, 716-723.

Banfield, J. D. and Raftery, A. E. (1992). Model-based Gaussian and non Gaussian clustering. *Biometrics,* **48**. (to appear).

Bezdek, J. C. (1981), *Pattern recognition with fuzzy objective function algorithms.* Plenum, New York.

Bock, H. H. (1985). On tests concerning the existence of a classification. *Journal of Classification*, **2**, 77-108.

Bock, H. H. (1989). Probabilistic aspects in cluster analysis. in *Conceptual and Numerical Analysis of Data.* O. Opitz (ed.) Springer-Verlag, Heidelberg. pp. 12-44.

Bozdogan, H. and Sclove, S. L. (1984). Multi-sample cluster analysis using Akaike 's information criterion. *Ann. Inst. Statist. Math.*, **36**, 163-180.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics, Theory and Methods* **19**, 221-278.

Bozdogan, H. (1992). Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix, in *Studies in Classification, Data Analysis, and Knowledge Organization*, O. Opitz, B. Lausen, and R. Klar (eds.), Springer-Verlag, Heidelberg. To appear.

Bryant, P. G. (1991). Large-sample results for optimization based clustering methods. *Journal of Classification*, **8**, 31-44.

Bryant, P. G. (1993). On detecting the nunbers of clusters using the MDL principle. in *Studies in Classification, Data Analysis, and Knowledge Organization* (Proceeding of IFCS 93, Paris), Springer-Verlag, Heidelberg. To appear.

Celeux, G. (1986). Validity tests in cluster analysis using a probabilistic teacher algorithm. *COMPSTAT 90*, Springer-Verlag, Heidelberg, pp. 163-169.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Ser. B **39**, 1-38.

Hattaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters*, **4**, 53-56.

Koehler, A. B. and Murphree, E. H. (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Applied Statistics*, **37**, 187-195.

McLachlan, G. J. (1987). On bootstraping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318-324.

Mclachlan, G. J. and Basford, K. E. (1989). *Mixture Models, Inference and Applications to Clustering.* Marcel Dekker, New York.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, World Scientific Publishing Company, Teaneck, New Jersey.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Titterington, D. M., Smith A. F. and Makov U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

Windham, M. P. and Cutler, A. (1991). Information ratios for validating cluster analyses. Working paper, Utah State University (Logan).

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.*, **5**, 329-350.

Unité de Recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LES NANCY Cedex (France)
Unité de Recherche INRIA Rennes IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Sophia Antipolis 2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399