Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood

Christophe Biernacki, Gilles Celeux

INRIA Rhône-Alpes

Gérard Govaert

UMR CNRS 6599 - UTC Compiègne

Abstract

We propose assessing a mixture model in a cluster analysis setting with the integrated completed likelihood. With this purpose, the observed data are assigned to unknown clusters using a maximum a posteriori operator. Then, the Integrated Completed Likelihood (ICL) is approximated using an *à la* Bayesian information criterion (BIC). Numerical experiments on simulated and real data of the resulting ICL criterion show that it performs well both for choosing a mixture model and a relevant number of clusters. In particular, ICL appears to be more robust than BIC to violation of some of the mixture model assumptions and it can select a number of clusters leading to a sensible partitioning of the data.

Keywords: mixture model, clustering, integrated likelihood, BIC, integrated completed likelihood, ICL criterion.

1 Introduction

Finite mixture models are commonly used as a basis for cluster analysis (see for instance [21]). One advantage of model-based clustering is that it provides a specific framework for assessing the resulting partitions of the data and especially for choosing a relevant number of clusters. A model-based clustering model is a parametric finite mixture model characterized by its form, denoted m in this article, (for instance m is a Gaussian mixture whose components have the same variance matrix) and the number K of the mixture components. Choosing a relevant model consists both of choosing its form m and the number of components K. In the Bayesian framework, a way of selecting a model among H models $M_1, \ldots, M_H$ is to choose the model of highest posterior probability. According to Bayes' theorem, the posterior probability of M_l given the data $\mathbf{x}$ is

$$P(M_l \mid \mathbf{x}) = \frac{\mathbf{f}(\mathbf{x} \mid M_l)P(M_l)}{\sum_{r=1}^{H} \mathbf{f}(\mathbf{x} \mid M_r)P(M_r)}$$

where $\mathbf{f}(\mathbf{x} \mid M_l)$ is the integrated or marginal likelihood of the model M_l and $P(M_l)$ is its prior probability. Thus, assuming that all models have equal prior probabilities, choosing the model with highest posterior probability is equivalent to select the model with the largest integrated likelihood. The Bayesian Information Criterion (BIC) of Schwarz [27] provides, under regularity conditions, a reliable approximation to the integrated likelihood. Although the regularity conditions for BIC do not hold for assessing the number of components K in a mixture model (see [1] for a precise insight), there is an increasing practical support for its use in this context (see for instance [17], [26]).

The point that we want to address here is the following. The integrated likelihood does not take into account the clustering purpose at hand for selecting a mixture model in a model-based clustering perspective. As a consequence, if the correct model is not in the family of considered models, BIC criterion will tend to overestimate the correct size regardless of the separation of the clusters (see [4] and Section 4 of the present article for illustrations).

In this article, we propose an Integrated Completed Likelihood (ICL) criterion which aims at answering this above mentioned limitation of BIC. In Section 2, the mixture model framework for clustering is reviewed and the differences between the likelihood and the completed likelihood are stressed. In Section 3, the ICL criterion is presented and discussed. Section 4 is devoted to numerical experiments on simulated and real data sets. A discussion section ends the paper.

2 Model-based clustering

In model-based clustering, observations are assumed to be a sample from a finite mixture of probability distributions. In a multivariate clustering context, we are mainly concerned with Gaussian distributions. For simplicity, we restrict attention to this situation. But the ICL criterion can be straightforwardly defined in other contexts such as the latent class model (see for instance [16]) in which a mixture of multivariate multinomial distributions is involved.

In the multivariate Gaussian mixture model, data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbf{R}^{nd}$ are assumed to be a sample from a probability distribution with density

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

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 $\phi(. \mid \mathbf{a}_k)$ denotes the d-dimensional Gaussian density with mean μ_k and variance matrix Σ_k with $\mathbf{a}_k = (\mu_k, \Sigma_k)$, and $\theta = (p_1, \ldots, p_K, \mathbf{a}_1, \ldots, \mathbf{a}_K)$ denotes the vector parameter of the mixture (m, K) at hand. The form m of a Gaussian mixture depends essentially on the assumptions concerning the variance matrices Σ_k (see [2] or [8] for a detailed presentation of some meaningful assumptions). In Section 4, most of those forms will be considered.

The mixture model is typically an incomplete data structure model (see [13]). The complete data are

$$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n))$$

where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ are binary K-dimensional vectors such that $z_{ik} = 1$ if and only if $\mathbf{x}_i$ arises from component k. Note that $\mathbf{z}$ defines a partition $P = (P_1, \ldots, P_K)$ of the observed data $\mathbf{x}$ with $P_k = \{\mathbf{x}_i \mid z_{ik} = 1\}$.

The observed log-likelihood of θ for the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

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

The complete log-likelihood of θ for the complete sample $\mathbf{y}$ is

$$CL(m, K) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(p_k \phi(\mathbf{x}_i \mid \mathbf{a}_k)). \tag{2.3}$$

In the clustering literature, it is also known as the classification log-likelihood [6]. In the mixture approach of model-based clustering the observed log-likelihood is maximized using generally the EM algorithm [23]. In the classification approach of model-based clustering, the completed log-likelihood is maximized using generally a Classification EM (CEM) algorithm (see [20] or [7]).

It is easily seen that the observed log-likelihood and the completed log-likelihood are linked by the following relation

$$CL(m, K) = L(m, K) - EC(m, K) \tag{2.4}$$

where

$$EC(m, K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log t_{ik} \geq 0,$$

with

$$t_{ik} = \frac{p_k \phi(\mathbf{x}_i \mid \mathbf{a}_k)}{\sum_{j=1}^{K} p_j \phi(\mathbf{x}_i \mid \mathbf{a}_j)} \tag{2.5}$$

denoting the conditional probability that $\mathbf{x}_i$ arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$).

The equation (2.4) shows that the completed log-likelihood can be regarded as a criterion penalizing the log-likelihood with $-EC(m, K)$. Moreover, $EC(m, K)$ is the realization of a random variable with mean $E(m, K)$, the entropy of the fuzzy classification matrix $\mathbf{t} = \{t_{ik}\}$,

$$E(m, K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \log t_{ik} \geq 0,$$

and with variance

$$\mathrm{Var}(EC(m, K)) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \log^2 t_{ik} - \sum_{i=1}^{n} \left[\sum_{k=1}^{K} t_{ik} \log t_{ik} \right]^2.$$

The entropy $E(m, K)$ (see [9]) is a measure of the ability of the K-component mixture model m to provide a relevant partition of the data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. If the mixture components are well separated, the classification matrix $\mathbf{t}$ tends to define a partition of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $E(m, K) \approx 0$. But if the mixture components are poorly separated, $E(m, K)$ has a large value. As a consequence, penalizing the log-likelihood with $-E(m, K)$, or $-EC(m, K)$, favors mixtures leading to a clustering of the data with the greatest evidence. In fact the random variable $CL(m, K)$ has been employed as a criterion for assessing the number of clusters arising from a Gaussian mixture model [4].

In practical situations, the criterion $CL(m, K)$ is computed in the following way. Let $\hat{\theta}$ be the maximum likelihood (m.l.) estimate of the mixture vector parameter and let $\mathbf{t}(\hat{\theta})$ be the corresponding estimate matrix of the classification matrix $\mathbf{t}$ where $\mathbf{t}(\hat{\theta})$ is derived from (2.5) by replacing $(p_k, \mathbf{a}_k)$ with $(\hat{p}_k, \hat{\mathbf{a}}_k)$. The missing cluster indicators z_{ik} are replaced with

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \arg\max_{\ell} t_{i\ell}(\hat{\theta}) = k \\ 0 & \text{otherwise.} \end{cases}$$

In the following, we will denote MAP (for Maximum A Posteriori) the function providing guessed values for the missing data from estimate value of θ:

$$\hat{\mathbf{z}} = \mathrm{MAP}(\hat{\theta}).$$

The completed likelihood criterion $CL(m, K)$ works well when the mixing proportions are restricted to be equal. But, it tends to overestimate the correct number of clusters when no restriction is placed on the mixing proportions (see [5]). The reason of this behavior is that the completed log-likelihood $CL(m, K)$ does not penalize the number of parameters in the mixture model. But, if a completed likelihood criterion would properly penalize the complexity of the model, it could be expected to provide a feasible estimate of the correct number of components in a mixture giving rise to partitioning the data with the greatest evidence. This penalized classification criterion is the integrated completed likelihood that we describe in the next section.

3 The integrated completed likelihood

A finite mixture model is characterized by the number of components K and the vector parameter $\theta = (p_1, \ldots, p_K, \mathbf{a}_1, \ldots, \mathbf{a}_K)$. We aim to find the mixture model leading to the greatest evidence for clustering the data $\mathbf{x}$. A classical way for choosing it is to select the model maximizing the integrated likelihood,

$$(\hat{m}, \hat{K}) = \arg\max_{m,K} \mathbf{f}(\mathbf{x} \mid m, K)$$

where

$$\mathbf{f}(\mathbf{x} \mid m, K) = \int_{\Theta_{m,K}} \mathbf{f}(\mathbf{x} \mid m, K, \theta)\pi(\theta \mid m, K)d\theta, \tag{3.6}$$

with

$$\mathbf{f}(\mathbf{x} \mid m, K, \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid m, K, \theta),$$

and $\Theta_{m,K}$ being the parameter space of the model m with K components and $\pi(\theta \mid m, K)$ a non informative or a weakly informative prior distribution on θ for the same model. A classical way to approximate (3.6) is to use BIC (see for instance [18])

$$\log \mathbf{f}(\mathbf{x} \mid m, K) \approx \log \mathbf{f}(\mathbf{x} \mid m, K, \hat{\theta}) - \frac{\nu_{m,K}}{2} \log(n), \tag{3.7}$$

where $\hat{\theta}$ is the m.l. estimate of θ

$$\hat{\theta} = \arg\max_{\theta} \mathbf{f}(\mathbf{x} \mid m, K, \theta)$$

and $\nu_{m,K}$ is the number of free parameters in the model m with K components. In the mixture context, the regularity conditions in [19] ensuring that

$$\log \mathbf{f}(\mathbf{x} \mid m, K) - \log \mathbf{f}(\mathbf{x} \mid m, K, \hat{\theta}) + \frac{\nu_{m,K}}{2} \log(n) = O_P(n^{1/2})$$

do not hold and there is a lack of theoretical justification for the BIC aproximation. But, simulations experiments (see [26] or [17]) show that the BIC approximation of the integrated likelihood works well at a practical level.

But, the use of the integrated likelihood (3.6) does not take into account the ability of the mixture model to give evidence for a clustering structure of the data. Instead we consider the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z})$ (or integrated completed likelihood)

$$\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) = \int_{\Theta_{m,K}} \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \theta)\pi(\theta \mid m, K)d\theta, \tag{3.8}$$

where

$$\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i, \mathbf{z}_i \mid m, K, \theta)$$

with

$$f(\mathbf{x}_i, \mathbf{z}_i \mid m, K, \theta) = \prod_{k=1}^{K} p_k^{z_{ik}} \left[\phi(\mathbf{x}_i \mid \mathbf{a}_k)\right]^{z_{ik}}.$$

To approximate this integrated completed likelihood we propose to use a BIC-like approximation. That is

$$\log \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) \approx \log \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \hat{\theta}^*) - \frac{\nu_{m,K}}{2} \log n \tag{3.9}$$

where

$$\hat{\theta}^* = \arg\max_{\theta} \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \theta). \tag{3.10}$$

But $\mathbf{z}$ is unknown. It means that the objective functions to be maximized in (3.8) and (3.10) are not available and so is $\hat{\theta}^*$. But, for n large enough, we have $\hat{\theta}^* \approx \theta$ and we approximate $\hat{\theta}^*$ with $\hat{\theta}$. Moreover, we replace the missing data $\mathbf{z}$ with $\hat{\mathbf{z}} = \mathrm{MAP}(\hat{\theta})$. Finally we propose the criterion

$$\mathrm{ICL}(m,K) = \log \mathbf{f}(\mathbf{x}, \hat{\mathbf{z}} \mid m, K, \hat{\theta}) - \frac{\nu_{m,K}}{2} \log n. \tag{3.11}$$

Some comments are in order.

1. The ICL criterion is essentially the ordinary BIC penalized by the substraction of the estimated mean entropy.

2. As for the BIC approximation of the integrated likelihood, and for the same reasons, there is a lack of theoretical justification of the *à la* BIC approximation of the integrated completed likelihood. Thus there is the need to consider simulations to see if the ICL criterion works well at a practical level. Such simulations are presented in the next section.

3. It is natural to replace the missing data $\mathbf{z}$ with $\hat{\mathbf{z}} = \mathrm{MAP}(\hat{\theta})$ since the ICL criterion is built with $\hat{\theta}$. In practical situations, replacing $\mathbf{z}$ using the MAP operator from an other consistent parameter estimate $\tilde{\theta}$ as the minimum Hellinger distance estimator, see [12], or a Bayesian estimator, see for instance [14], can be considered and will presumably do not affect the performance of the ICL criterion. On the contrary, we do not recommend replacing $\mathbf{z}$ from the classification m.l. estimator of θ since in general this estimator is inconsistent and can be highly biased (see [6]).

4 Numerical experiments

We compare the practical behavior of $\mathrm{BIC}(m,K)$ and $\mathrm{ICL}(m,K)$ for choosing the form m and the number K of components of a mixture model on the basis of numerical experiments on both simulated and real data. Since, in this article we are interested in model-based clustering, we restrict attention to multivariate Gaussian mixtures. The form m of the mixture model is defined by parameterizing the variance matrix Σ_k of a component in terms of its eigenvalue decomposition, as developed in [2] and [8],

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{4.12}$$

where $\lambda_k = |\Sigma_k|^{1/d}, d$ denoting the number of variables, D_k is the matrix of eigenvectors of Σ_k and A_k is a diagonal matrix, such that $|A_k| = 1$, with the normalized eigenvalues of Σ_k on the diagonal in a decreasing order. The parameter λ_k determines the volume of the kth group, D_k its orientation and A_k its shape. By allowing some but not all of these quantities to vary between groups, we obtain easily interpreted models which are appropriate to describe various clustering situations. Here, we considered 28 different models related to different assumptions on the group variance matrices and the proportions of the mixture model: 16 of these models are obtained by assuming equal or different volumes, shapes, orientations or proportions. We denote conventionally those models as exemplified now: $[p_k \lambda D_k A D_k']$ indicates the model with different proportions and orientations and equal volumes and shapes. Eight models assume diagonal variance matrices, we denote B a diagonal variance matrix and, for instance, $[p\lambda_k B]$ indicates the model with equal proportions, different volumes, equal shapes and diagonal orientations. Finally, four models assume spherical shapes: they are denoted $[p\lambda I]$, $[p\lambda_k I]$, $[p_k \lambda I]$, and $[p_k \lambda_k I]$.

In all experiments, the clustering have been derived from the m.l. estimate $\hat{\theta}$ of the mixture vector parameter at hand obtained with the EM algorithm. To get sensible maxima, for each considered situation, the EM algorithm is initiated $r = 20$ times with random centers and the solution providing the largest observed likelihood is selected.

4.1 Monte Carlo experiments

For each Monte Carlo experiment, we generate 50 samples from each type of simulated data.

4.1.1 Three clusters with different overlapping

We simulated two types of a three-component Gaussian mixture. Both types of Gaussian mixtures only differ by second component variance matrix Σ_2. The common characteristics of the simulated mixtures were the following

$$n = 400, d = 2, p_1 = 0.25, p_2 = 0.25, p_3 = 0.50$$

$$\mu_1 = \mu_2 = (0,0)', \mu_3 = (8,0)'$$

$$\Sigma_1 = \Sigma_3 = \begin{pmatrix} 0.11 & 0 \\ 0 & 9 \end{pmatrix}.$$

In both situations, we consider a variance matrix Σ_2 with the same volume and shape as Σ_1 but with a different orientation. In the first situation, we consider

$$\Sigma_2 = \begin{pmatrix} 0.96 & 2.61 \\ 2.61 & 8.15 \end{pmatrix}.$$

The angle between the first eigenvector of Σ_1 and Σ_2 is 30 degrees. (One of the 50 simulated data sets is displayed in Figure 1.)

In the second situation, we consider

$$\Sigma_2 = \begin{pmatrix} 7.33 & 2.64 \\ 2.64 & 1.67 \end{pmatrix}.$$

The angle between the first eigenvector of Σ_1 and Σ_2 is 18 degrees. Thus, the first two mixture components are quite overlapping. (One of the 50 simulated data sets is displayed in Figure 2.)

In this experiment all the 28 models were considered, with the number of clusters varying from one to seven.

For the first situation, BIC and ICL exhibit a similar behavior: both criteria dramatically favor the right model $[p_k \lambda D_k A D_k']$ with $K = 3$ components (92% for BIC and 88% for ICL). Note also that ICL (resp. BIC) chooses $K = 3$ in 96% (resp. 92%) and $K = 4$ in 4% (resp. 8%) of the simulations.

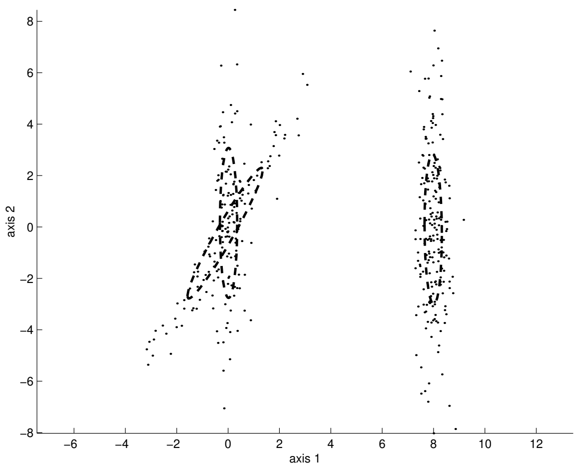
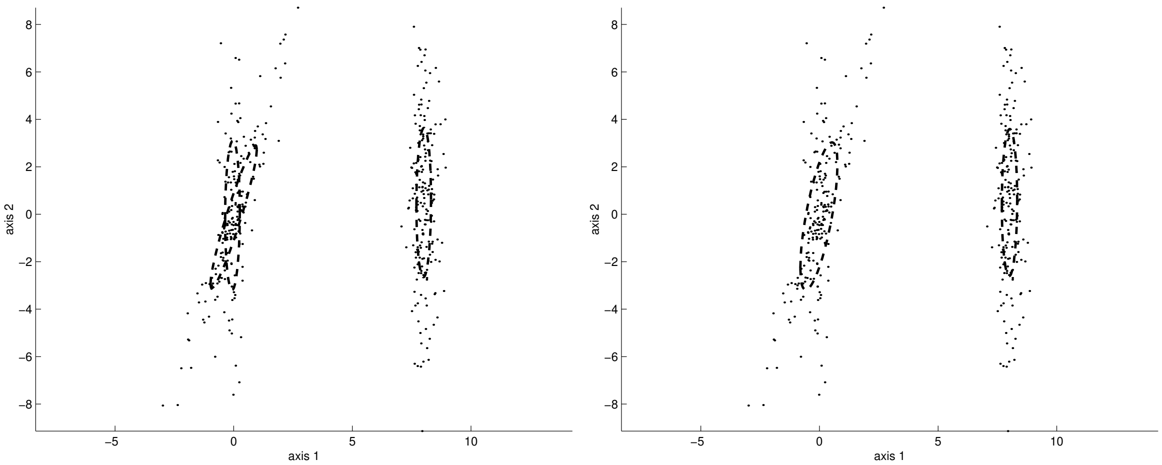


Figure 1: A typical BIC and ICL solution for an example of data in situation 1.

For the second "overlapping" situation, it is remarkable that BIC gives exactly the same answer : it chooses the right model $[p_k \lambda D_k A D_k']$ with $K = 3$ in 92% of the simulations and prefers a $K = 4$ solution otherwise. ICL highly prefers a $K = 2$ solution with the model $[p \lambda_k D_k A_k D_k']$ in 88% of the simulations and selects the right model with $K = 3$ in only 8% of the cases. But clearly, the solution selected with ICL makes sense from the clustering point of view. Figures 1 and 2 displayed an exemple of simulated data for situations 1 and 2 and depicted the BIC and ICL favorite couple (m, K) for those data sets.

4.1.2 A non-Gaussian cluster

We now consider experiment from a mixture of a uniform and a Gaussian cluster. One of the 50 simulated data sets is displayed in Figure 3 and the mixture characteristics are as



BIC solution ICL solution

Figure 2: Typical solutions for an example of data in situation 2.

follows:

$$n = 200, d = 2$$
$$f(\mathbf{x}) = 0.5 \underbrace{\left[0.25\ \mathbf{I}_{[-1,1]}(x^1)\ \mathbf{I}_{[-1,1]}(x^2)\right]}_{\text{non-Gaussian cluster}} + 0.5 \underbrace{\left[\phi(\mathbf{x} \mid (3.3, 0)', I)\right]}_{\text{Gaussian cluster}},$$

where $\mathbf{I}_{[-1,1]}$ denotes the indicator function of the interval $[-1, 1]$. When running the EM algorithm, only the model $[p\lambda I]$ is considered, K is varying from one to five. Percentage of choosing K is displayed in Table 1. In this case BIC has a disappointing behavior. This example highlights a tendency of this criterion, already mentioned in the introduction: When the clustering model at hand (here a Gaussian mixture model) does not fit well the data, BIC tends to overestimate the number of components. On the contrary, ICL includes a term $E(m, K)$, penalizing overlapping clusters, balancing the lack of fit of the data to the model at hand and can be thought of as more robust to violations of the model specifications than BIC, as it appears in this experiment.

4.2 Real data sets

4.2.1 The Old Faithful geyser

This first example on real data concerns the Old Faithful data (the version from [29]) which consists of data on 272 eruptions of the Old Faithful geyser in Yellowstone National Park. Each observation consists of two measurements: the duration (in minutes) of the eruption, and the waiting time (in minutes) before the next eruption. We consider the 28 models with $K = 1, \ldots, 6$ to compute BIC and ICL on the 272 eruptions. For almost all models, BIC prefers a $K = 3$ component mixture and selects the model $[p\lambda DAD']$. On the contrary, for almost all models, ICL prefers a $K = 2$ component mixture and selects the model $[p_k\lambda_k DA_kD']$. Figure (4) (a) (resp. (b)) depicts the favorite couple (m, K) of

BIC (resp. ICL). Those figures provide iso-density ellipses for each component. The ICL solution with $K = 2$ components clearly distinguishes two groups. The BIC solution with $K = 3$ components appears to model deviations from normality in the two obvious groups rather than a relevant additional group.

4.2.2 French departments

Figure 5(a) displays log-population versus log-density (in inhabitants/km^2) of 312 towns of three French departments: Two densely-populated departments in the suburbs of Paris, Seine-Saint-Denis and Hauts-de-Seine, and one rural department Haute Corse (source: census 1990 of the French population, INSEE Web site at `http://www.insee.fr/vf/-chifcles/rp90/index.htm`). For this bivariate data set, we consider all the 28 models with $K = 1, \ldots, 5$. Both criteria favor the model $[p_k \lambda_k DAD']$ and we restrict attention to this model. Table 2 gives BIC and ICL values for $K = 1, \ldots, 5$. BIC chooses the model $[p_k \lambda_k DAD']$ with $K = 3$, whereas ICL chooses $K = 2$, but its second choice $K = 3$ is not too far. The two-cluster solution has an interesting interpretation since one cluster is closely related to Haute Corse and the other cluster is closely related to the Paris area departments (the partitions are depicted in Figure 5 (b) (c). The three cluster solution splits Haute Corse into two clusters.

4.2.3 Acoustic emission control

This example is concerned with flaws detection on a pressurized vessel by acoustic emission. During a pressurization control, the vessel sounds (the *events*) are located on its surface. The first step of the flaw detection procedure consists of grouping those events in homogeneous clusters. Data at hand are 2061 event locations in a rectangle of $\mathbf{R}^2$ representing the vessel.

In this setting, a Gaussian mixture model with equal proportions, diagonal variance matrices with different volumes appears to be relevant. Moreover, the uniform background noise is taken into account with a uniform distribution on the rectangle where the sounds are located. It is worth noting that adding such a uniform distribution in the mixture is straightforward and simply leads to consider the proportion of the uniform component as an additional parameter.

For this example, the problem is to find a relevant number of mixture components leading to a clear grouping of the sound locations. In our experiments K is varying from 2 to 20 with the diagonal Gaussian mixture model with equal proportions and different volumes and we consider the additional uniform distribution. We ran the EM algorithm 50 times for each situation from random centers. We stopped EM each time the relative

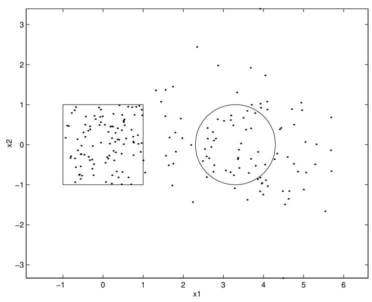
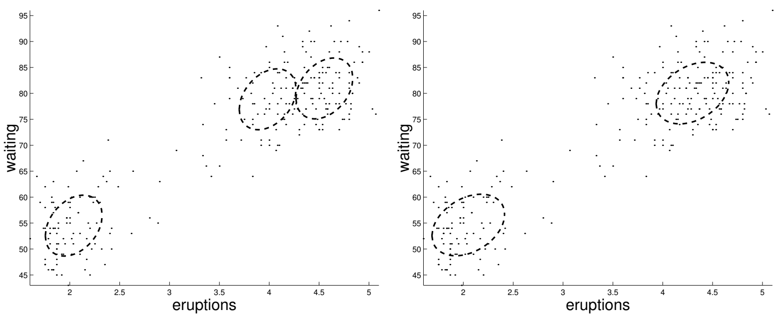


Figure 3: A uniform and a Gaussian cluster.

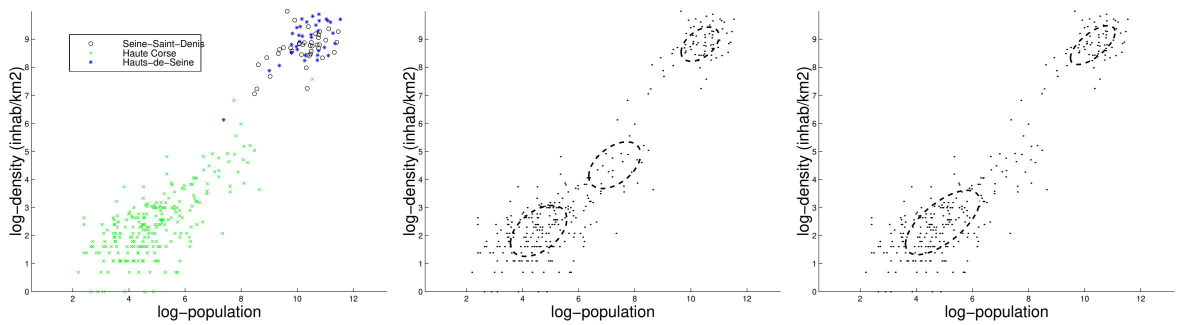
K	1	2	3	4	5
BIC	0	**60**	0	32	8
ICL	0	**100**	0	0	0

Table 1: Non-Gaussian cluster samples: percentage of choosing K with the model $[p\lambda I]$.



(a) BIC (b) ICL

Figure 4: Cluster ellipses for the Old Faithful geyser data.



(a) true partition (b) BIC partition (c) ICL partition

Figure 5: French departments.

criterion	1	2	3	4	5	$\hat{K}$
BIC	-1200	-1044	**-1035**	-1040	-1053	3
ICL	-1200	**-1044**	-1046	-1070	- 1105	2

Table 2: BIC and ICL values for the best model $[p_k\lambda_k DAD']$ on the French departments data.

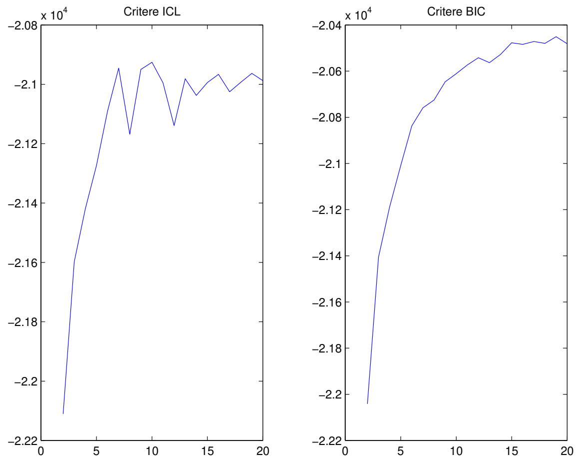


Figure 6: BIC and ICL values with the model applied to the vessel sound location data.

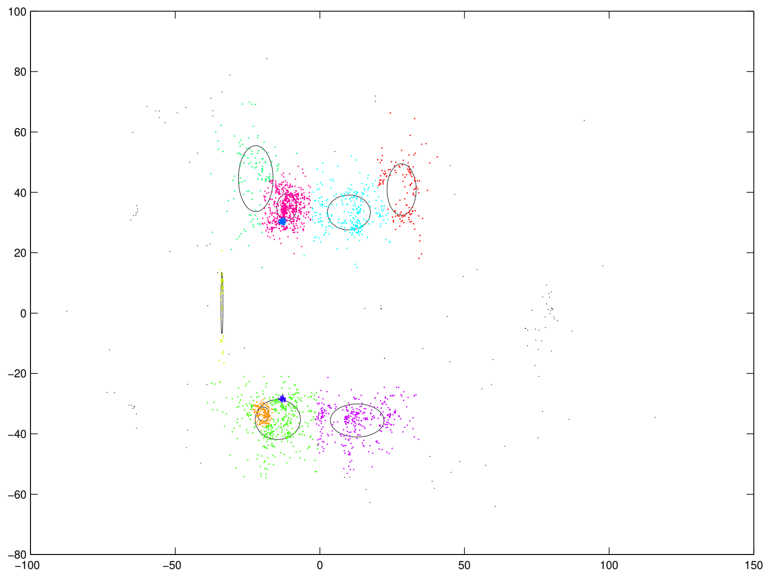


Figure 7: The ten-cluster partition of the vessel sound location data.

difference between two successive values of the observed log-likelihood was smaller than 10^{-16}. (We chose such a small convergence threshold in view of possible slow convergence of EM.) Figure 6 displays the values of BIC and ICL. BIC increases almost monotonically with K and does not provide evidence for any K value. On the contrary, ICL gives a preference for the ten-cluster partition which is depicted in Figure 7 by the iso-density of each of the ten components. Note that ICL also points out that the seven-cluster solution is of interest. But from the application in view a ten cluster solution seems more interesting. Actually, from Figure 7 it seems that the ten-cluster partition selected by ICL captures the high density regions appearing in this data set.

5 Discussion

Statistical analysis of finite mixtures are employed in statistical modeling with two different purposes. In one perspective, finite mixtures are essentially regarded as competitors to non parametric density estimation (see [15], [24] or [26]). In another view, finite mixtures are considered as a powerful modeling way in cluster analysis (see [17] or [21]). In both situations, choosing a relevant form m for the model and assessing a sensible number K of components is an important task.

When the concern of mixture modeling is density estimation, our numerical experiments confirm that the BIC approximation of the integrated observed likelihood can be regarded as a reasonable tool for comparing mixture models. Choosing the form of the model m and the number of components K by optimization of the BIC criterion will generally result in a good approximation of the density to be estimated. Experiments described in Section 4.1.1 and other experiments in [3], not reported here, highlight this satisfactory behavior of BIC in a spectacular way. Some other criteria, based on heuristic arguments, have been proposed to approximate the integrated observed likelihood in the mixture context. We can mention the Cheeseman-Stutz (CS) criterion [10] and [11], the MML criterion [22] and the Bayesian criterion in [25]. Numerical experiments showed that there is very little difference between those criteria and BIC. (Numerical experiments comparing BIC and CS criteria are in [3] and numerical experiments comparing BIC, MML and their Bayesian criterion are in [25].) An alternative promising approach for estimating the proper number of clusters is the cross-validated likelihood approach suggested by Smyth [28].

When the interest in mixture modeling is cluster analysis choosing a sensible number of clusters K is crucial. In this clustering context, the BIC criterion is less convincing. In particular, it tends to overestimate the number K of clusters when the fit of the data to the mixture model is not very good. (On the contrary when the fit is good, BIC tends to give too a few a number of clusters see [4] and [9] for illustrations.) The experiments in Section

4.1.2 and all the experiments in Section 4.2 are illustrations of such a behavior of BIC. In this context, we proposed maximizing the integrated completed likelihood rather than the integrated observed likelihood to select both a relevant form m of model and a relevant number of clusters K, the missing cluster indicators being replaced by their maximum a posteriori estimators. From a practical point of view, the ICL criterion which is basically the BIC approximation for the completed log-likelihood, seems to give an answer to the practical possible tendency of BIC to overestimate the number of clusters as it appears from numerical experiments in sections 4.1.2 and 4.2. Additionnal experiments supporting this assertion can be found in [3] where a slighlty different version of ICL is considered in a non informative Bayesian framework. Yet, it can be shown (through numerical experiments not reported here) that ICL outperforms heuristic criteria, developed for assessing mixture models in a clustering setting, as AWE [2] which tends to underestimate the number of clusters as shown in [9], or the entropy criterion NEC [9] which exhibits a disappointing behavior to choose a relevant form m of the mixture model as shown in [5].

As compared to the integrated observed likelihood, the integrated completed likelihood includes an additional entropy term $E(m, K)$ which favors well-separated clusters and which is the essential difference between BIC and ICL criteria.

Acknowledgements We are grateful to Catherine Hervé and Christel Rigault (CETIM) for allowing us to use the vessel sound location data and to Florence Forbes for helpful comments.

References

[1] M. Aitkin and D. B. Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, B*, 47:67–75, 1985.

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

[3] C. Biernacki, G. Celeux, and G Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. Technical Report 3521, Inria, 1998.

[4] C. Biernacki and G. Govaert. Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 29:451–457, 1997.

[5] C. Biernacki and G. Govaert. Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, 14:49–71, 1999.

[6] P. G. Bryant. Large sample results for optimization based clustering methods. *Journal of Classification*, 8:1–44, 1991.

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

[8] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

[9] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13:195–212, 1996.

[10] P. Cheeseman and J. Stutz. Bayesian classification (autoclass): Theory and results. In *Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA.*, pages 61–83, 1996.

[11] D. M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, 29:181–212, 1997.

[12] A. Cutler and O. Cordero-Braña. Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91:1716–1723, 1996.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

[14] J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, B*, 56:363–375, 1994.

[15] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[16] B. S. Everitt. *An Introduction to Latent Variables Models.* London: Chapman & Hall, 1984.

[17] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1998.

[18] R. E. Kass and A. E. Raftery. Bayes factor. *Journal of the American Statistical Association*, 90:733–795, 1995.

[19] R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90:928–934, 1995.

[20] G. J. McLachlan. *The classification and mixture maximum likelihood approaches to cluster analysis. I*, pages 199–208. Amsterdam: North-Holland, 1982.

[21] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[22] J. J. Oliver, R. A. Baxter, and C. S. Wallace. Unsupervised learning using mml. In *Proceedings of the 13th International Conference on Machine Learning*, pages 364–372. San Franscisco, 1996.

[23] R. A. Redner and Walker H. F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26:195–239, 1984.

[24] C. P. Robert. Mixtures of distributions: Inference and estimation. In Gilks W. R., Richardson S., and Spiegelhalter D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 441–464. London: Chapman & Hall, 1996.

[25] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Learning*, 20:1133–1142, 1999.

[26] K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902, 1997.

[27] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:61–464, 1978.

[28] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10:63–72, 2000.

[29] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag, 1994.