INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood

Christophe Biernacki, Gilles Celeux, Gérard Govaert

No 3521

THÈME 4

Rapport de recherche

ISSN 0249-6399

Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood

Christophe Biernacki, Gilles Celeux, Gérard Govaert

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet is2

Rapport de recherche n ° 3521 — — 27 pages

Abstract: We propose assessing a mixture model in a cluster analysis setting with the inegrated classification likelihood. With this purpose, the observed data are assigned to unknown clusters using a maximum a posteriori operator. The integrated completed likelihood approximation is derived without the theoretical difficulties encountered when approximating the integrated observed likelihood. 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.

Key-words: mixture model, clustering, integrated likelihood, BIC criterion, completed integrated likelihood, ICL criterion.

(Résumé : tsvp)

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
Téléphone : 04 76 61 52 00 - International: +33 4 76 61 52 00
Télécopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

Choix d'un modèle de mélange en classification à l'aide de la vraisemblance classifiante intégrée

Résumé : Nous proposons de choisir un modèle de mélange dans un but de classification par maximisation de la vraisemblance classifiante intégrée. Pour ce faire, nous affectons les points observés aux classes inconnues par un opérateur du maximum a posteriori. L'approximation de la vraisemblance complétée intégrée ainsi obtenue est ensuite réalisée sans rencontrer les problèmes théoriques d'approximation de la vraisemblance observée intégrée. Des expérimentation sur les données réelles et simulées illustrent le bon comportement du critère ICL en résultant pour, à la fois, choisir une forme de modèle et un nombre de classes pertinent. Il s'avère notamment que le critère ICL est beaucoup moins sensible que le critère BIC à un ajustement médiocre du modèle de mélange aux données et parvient à retenir les modèles donnant lieu à une classification pertinente des données.

Mots-clé : modèle de mélange, classification, vraisemblance intégrée, critère BIC, vraisemblance complétée intégrée, critère ICL.

1 Introduction

Finite mixture models are commonly used as a basis for cluster analysis (see for instance McLachlan and Basford 1988). One advantage of model-based clustering is that it provides a precise framework for assessing the resulting partitions of the data and especially for choosing a relevant number of clusters. A model-based cluster 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 in 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 choosing the model of highest posterior probability. By 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 the highest posterior probability is equivalent to select the model with the largest integrated likelihood. The Bayesian Information Criterion (BIC) of Schwarz (1978) 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, there is an increasing practical support for its use in this context (see for instance Fraley and Raftery 1998, Roeder and Wasserman 1997). However, using the BIC criterion for assessing a mixture model when grouping the data presents some drawbacks.

From a theoretical point of view, a condition of validity of the BIC approximation is that the estimated vector parameter of the model is well within the parameter space. For mixture models, if the true model has $K' < K$ components, then $K - K'$ of the mixing proportions will tend to zero as the sample size tends to infinity, thus the corresponding estimated proportions will be on the boundary of the parameter space (see Aitkin and Rubin 1985 for a more precise insight).

From a practical point of view, 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 the clusters separation (see Biernacki and Govaert 1997 and Section 4 of the present article for illustrations).

In this article, we propose an Integrated Classification Likelihood (ICL) criterion which aims answering the above mentioned limitations of BIC. In Section 2, the mixture model framework for clustering is reviewed and the differences between the likelihood and the classification 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 as, for instance, the latent class model (see for instance Everitt 1984) 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 $\mathbf{R}^d$ 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 at hand. The form m of a Gaussian mixture depends essentially of the assumptions concerning the component variance matrices Σ_k (see Banfield and Raftery 1993 or Celeux and Govaert 1995 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 Dempster, Laird and Rubin 1977). 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 with $z_{ik} = 1$ 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 $P_k = \{\mathbf{x}_i / z_{ik} = 1\}$.

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 (Bryant 1991). In the mixture approach of model-based clustering the observed log-likelihood is maximized using generally the EM algorithm (Redner and Walker 1984). In the classification approach of model-based clustering, the classification log-likelihood is maximized using generally a Classification EM (CEM) algorithm (Celeux and Govaert 1992).

It is easily seen that the observed log-likelihood and the classification log-likelihood are linked with 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

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

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 classification log-likelihood can be regarded as a criterion penalizing the log-likelihood with $-EC(m,K)$. And, $EC(m,K)$ 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 Celeux and Soromenho 1996) 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 (Biernacki and Govaert 1997).

In practical situations, they compute the criterion $CL(m,K)$ 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

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

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 classification 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 Biernacki 1997). The reason of this behavior is that the classification log-likelihood $CL(m, K)$ does not penalize the number of parameters in the mixture model. But, if a classification 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 classification likelihood that we describe in the next section.

3 The Integrated classification 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 the BIC criterion (see for instance Kass and Raftery 1995)

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

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. But, as seen above, this approximation is not valid in the mixture context and moreover the employment 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 classification 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

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

But, we are still faced with the problem that the BIC approximation for the logarithm of (3.8) is not valid since the estimates of some of the proportions will be on the parameter space boundary when estimating the mixture model with a too large number of components.

To circumvent this difficulty, it is of interest to isolate the contribution of the missing data $\mathbf{z}$ to the density $\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K)$ by conditioning on $\mathbf{z}$. The following lemma provides the justification for this. Let $\mathbf{p} = (p_1, \ldots, p_K)$, $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_K)$ and $\Theta_{m,K} = A_{m,K} \times P_{m,K}$.

Lemma 3.1 *If* $\pi(\theta \mid m, K) = \pi(\mathbf{a} \mid m, K)\pi(\mathbf{p} \mid m, K)$ *then*

$$\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) = \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K)\mathbf{f}(\mathbf{z} \mid m, K).$$

Proof:

$$\begin{aligned}
\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) &= \int_{\Theta_{m,K}} \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K, \theta)\pi(\theta \mid m, K) d\theta \\
&= \int_{\Theta_{m,K}} \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \theta)\mathbf{f}(\mathbf{z} \mid m, K, \theta)\pi(\theta \mid m, K) d\theta \\
&= \int_{\Theta_{m,K}} \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \mathbf{a})\mathbf{f}(\mathbf{z} \mid m, K, \mathbf{p})\pi(\mathbf{a} \mid m, K)\pi(\mathbf{p} \mid m, K) d\mathbf{a} d\mathbf{p} \\
&= \int_{A_{m,K}} \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \mathbf{a})\pi(\mathbf{a} \mid m, K) d\mathbf{a} \\
&\quad \int_{P_{m,K}} \mathbf{f}(\mathbf{z} \mid m, K, \mathbf{p})\pi(\mathbf{p} \mid m, K) d\mathbf{p}.
\end{aligned}$$

And, we have

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

$$= \int_{A_{m,K}} \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m.K, \mathbf{a})\pi(\mathbf{a} \mid m, K)d\mathbf{a} \int_{P_{m,K}} \pi(\mathbf{p} \mid m, K)d\mathbf{p}$$
$$= \int_{A_{m,K}} \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \mathbf{a})\pi(\mathbf{a} \mid m, K)d\mathbf{a}.$$

From which it follows that

$$\mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) = f(\mathbf{x} \mid \mathbf{z}, m, K)f(\mathbf{z} \mid m, K).$$

Remark 3.1 *The assumption $\pi(\theta \mid m, K) = \pi(\mathbf{a} \mid m, K)\pi(\mathbf{p} \mid m, K)$ is quite reasonable in many contexts and always made in full Bayesian inference for mixture models in a non informative or weakly informative context (see for instance Roeder and Wasserman 1997 or Richardson and Green 1997).*

Assuming, as in Lemma 3.1, that $\pi(\theta \mid m, K) = \pi(\mathbf{a} \mid m, K)\pi(\mathbf{p} \mid m, K)$, we have

$$\log \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K) = \log \int_{A_{m,K}} \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \mathbf{a})\pi(\mathbf{a} \mid m, K)d\mathbf{a}.$$

Now the BIC approximation is valid for $\mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K)$,

$$\log \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K) \approx \max_{\mathbf{a}} \log \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \mathbf{a}) - \frac{\lambda_{m,K}}{2} \log n \tag{3.9}$$

where $\lambda_{m,K}$ is the number of free components in $\mathbf{a}$. Note that the vector maximizing the conditional likelihood $\log \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \mathbf{a})$ with respect to $\mathbf{a}$ is not $\hat{\mathbf{a}}$. For instance, for a Gaussian mixture with no restriction placed on the variance matrix form, this vector is $((\bar{\mathbf{x}}_1, S_1), \ldots, (\bar{\mathbf{x}}_K, S_K))$ where

$$\bar{\mathbf{x}}_k = \frac{1}{n} \sum_{i=1}^{n} z_{ik} \mathbf{x}_i$$

and

$$S_k = \frac{1}{n} \sum_{i=1}^{n} z_{ik} \left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right) \left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right)'.$$

For $\log \mathbf{f}(\mathbf{z} \mid m, K)$ the BIC approximation is no longer applicable since the estimates of some mixing proportions can be on the boundary of the parameter space. But there is no need for any approximation for calculating $\mathbf{f}(\mathbf{z} \mid m, K)$. Actually, a flat prior distribution for $\mathbf{p}$ is a Dirichlet distribution $\mathcal{D}(\delta, \ldots, \delta)$ (see Diebolt and Robert 1994 or Richardson and Green 1997) and the Jeffreys non informative distribution is a $\mathcal{D}(1/2, \ldots, 1/2)$ (see Robert 1994). Then, we have

$$\mathbf{f}(\mathbf{z} \mid m, K) = \int p_1^{n_1} \ldots p_K^{n_K} \frac{\Gamma(K\delta)}{\Gamma(\delta) \ldots \Gamma(\delta)} \mathbf{I}_{\sum_k p_k = 1} d\mathbf{p},$$

where

$$n_k = \text{card}\{i, 1 \leq i \leq n, \text{such that } z_{ik} = 1\}(1 \leq k \leq K).$$

It follows that

$$\mathbf{f}(\mathbf{z} \mid m, K) = \frac{\Gamma(K\delta)}{\Gamma(\delta)^K} \frac{\Gamma(n_1 + \delta) \ldots \Gamma(n_K + \delta)}{\Gamma(n + \delta K)}.$$

In the particular case (of interest) $\delta = 1$, it leads to

$$\mathbf{f}(\mathbf{z} \mid m, K) = \frac{n_1! \ldots n_K!(K-1)!}{(n+K-1)!}.$$

Finally, if we opt for the Jeffreys non informative prior distribution for the proportions ($\delta = 1/2$), the integrated classification log-likelihood takes the form

$$\begin{aligned} \log \mathbf{f}(\mathbf{x}, \mathbf{z} \mid m, K) \quad &\approx \quad \max_{\mathbf{a}} \log \mathbf{f}(\mathbf{x} \mid \mathbf{z}, m, K, \mathbf{a}) - \frac{\lambda_{m,K}}{2} \log n \\ &\quad + \log \Gamma(\frac{K}{2}) + \sum_{k=1}^{K} \log \Gamma(n_k + \frac{1}{2}) \\ &\quad - K \log \Gamma(\frac{1}{2}) - \log \Gamma(n + \frac{K}{2}). \end{aligned} \tag{3.10}$$

Unfortunately $\mathbf{z}$ is unknown. It means that the objective function to be maximized in (3.8) is unavailable. We replace the missing data $\mathbf{z}$ with $\tilde{\mathbf{z}} = \text{MAP}(\tilde{\theta})$ where $\tilde{\theta}$ is an estimate of θ[1]. Finally we propose the criterion

$$\begin{aligned} \text{ICL}(m, K) \quad &= \quad \max_{\mathbf{a}} \log \mathbf{f}(\mathbf{x} \mid \tilde{\mathbf{z}}, m, K, \mathbf{a}) - \frac{\lambda_{m,K}}{2} \log n \\ &\quad + \log \Gamma(\frac{K}{2}) + \sum_{k=1}^{K} \log \Gamma(\tilde{n}_k + \frac{1}{2}) \\ &\quad - K \log \Gamma(\frac{1}{2}) - \log \Gamma(n + \frac{K}{2}), \end{aligned} \tag{3.11}$$

with

$$\tilde{n}_k = \text{card}\{i, 1 \leq i \leq n, \text{such that } \tilde{z}_{ik} = 1\}(1 \leq k \leq K).$$

Some comments are in order.

1. Despite a somewhat complicated expression, ICL is simple and easily computed.
2. Obviously other choices of δ (as $\delta = 1$) are possible when specifying the prior distribution $\pi(\mathbf{p} \mid m, K)$. But, we see no particular reason not to use Jeffreys prior distribution.

[1] In this article, we always consider the m.l. estimator $\tilde{\theta} = \hat{\theta}$, but other estimates as the classification m.l. estimator, the minimum Helliger distance estimator, see Cutler and Cordero-Brana 1996, or a Bayesian estimator, see for instance Diebolt and Robert 1994, can be considered.

3. In the ICL criterion, terms not depending on the sample size n appear. They can be considered negligible since the error mean of the BIC approximation in (3.9) is $O(1)$. Hoewever, Raftery (1995) mentioned that empirical experience has found the BIC approximation to be more accurate in practice than the $O(1)$ error term suggest. For instance, he mentions prior distributions leading to an $O(n^{-1/2})$ error rather than $O(1)$ for the BIC approximation. As a consequence the introduction of terms no depending on n in the ICL criterion makes sense.

4. When the $\tilde{n}_k$'s are large enough, namely as the $\tilde{p}_k$'s are far from 0, we can use the approximation of the Gamma function with the Stirling formula

$$\Gamma(t+1) \approx t^{t+1/2} \exp(-t)(2\pi)^{1/2}$$

for large values of t. Thus, calculating $\log \mathbf{f}(\tilde{\mathbf{z}} \mid m)$ using this approximation and removing the terms of order $O(1)$ (assuming that $K = o(n)$) leads to

$$\log \mathbf{f}(\tilde{\mathbf{z}} \mid m, K) \approx \sum_{k=1}^{K} \tilde{n}_k \log \tilde{n}_k - n \log n + (\delta - \frac{1}{2}) \sum_{k=1}^{K} \log \tilde{n}_k - (\delta K - \frac{1}{2}) \log n.$$

Since

$$\max_{\mathbf{p}} \log \mathbf{f}(\mathbf{z} \mid m, K, \mathbf{p}) = \sum_{k=1}^{K} \tilde{n}_k \log \frac{\tilde{n}_k}{n},$$

when $\delta = 1/2$, the ICL criterion would reduce to

$$\begin{aligned} \mathrm{ICL}(m, K) &= \max_{\mathbf{a}} \log \mathbf{f}(\mathbf{x} \mid \tilde{\mathbf{z}}, m, K, \mathbf{a}) + \max_{\mathbf{p}} \mathbf{f}(\tilde{\mathbf{z}} \mid m, K, \mathbf{p}) \\ &\quad - \frac{\lambda_{m,K}}{2} \log n - \frac{K-1}{2} \log n; \end{aligned}$$

that is,

$$\mathrm{ICL}(m, K) = \max_{\theta} \log \mathbf{f}(\mathbf{x}, \tilde{\mathbf{z}} \mid m, K, \theta) - \frac{\nu_{m,K}}{2} \log n. \tag{3.12}$$

It means that ICL would reduce to the *à la* BIC approximation of the logarithm of the complete integrated likelihood when each $\tilde{n}_k$ becomes large as n tends to infinity. (But, it is not true when K is greater than the correct number of mixture components since some of the p_k's tends to 0.)

5. The ICL criterion has some link with the approximation of the integrated observed likelihood, first proposed in Cheeseman and Stutz (1996), and more precisely detailed and discussed in Chickering and Heckerman (1997).

   This approximation for a general incomplete data model m is as follows. Let $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ be the complete data with density $\mathbf{g}(\mathbf{y} \mid \theta, m)$. The missing data $\mathbf{z}$ have the conditional density

$$k(\mathbf{z} \mid \mathbf{x}, m, \theta) = \frac{\mathbf{g}(\mathbf{y} \mid m, \theta)}{\mathbf{f}(\mathbf{x} \mid m, \theta)}, \tag{3.13}$$

and the Cheeseman-Stutz (C-S) approximation takes the form

$$\begin{aligned}\log \int \mathbf{f}(\mathbf{x} \mid m, \theta)\pi(\theta \mid m)d\theta &\approx \log \int \mathbf{g}(\mathbf{y} \mid m, \theta)\pi(\theta \mid m)d\theta \\ &\quad - \log \mathbf{g}(\mathbf{y} \mid m, \hat{\theta}) + \frac{v'}{2}\log n \\ &\quad + \log \mathbf{f}(\mathbf{x} \mid m, \hat{\theta}) - \frac{v}{2}\log n, \qquad (3.14)\end{aligned}$$

where $\hat{\theta}$ is the m.l. estimate of θ derived from the observed sample $\mathbf{x}$ of size n, v and v' denoting the number of parameters for respectively the incomplete data and the complete data models. As shown in Chickering and Heckerman (1997), this approximation is deduced from the identity

$$\mathbf{f}(\mathbf{x} \mid m) = \mathbf{g}(\mathbf{y} \mid m)\frac{\int \mathbf{f}(\mathbf{x} \mid \theta, m)\pi(\theta \mid m)d\theta}{\int \mathbf{g}(\mathbf{y} \mid \theta, m)\pi(\theta \mid m)d\theta}, \qquad (3.15)$$

$\mathbf{f}(\mathbf{x} \mid m)$ and $\mathbf{g}(\mathbf{y} \mid m)$ denoting respectively the marginal likelihood for the incomplete data and the complete data, and $p(\mathbf{x} \mid \theta, m)$ and $p(\mathbf{y} \mid \theta, m)$ denoting the corresponding joint distributions of the observed or complete data and the parameter. Applying the BIC approximation to the numerator and denominator of the second term of (3.15), and using the fact that the data $\mathbf{y}$ are completed so that its sufficient statistics match the expected sufficient statistics given $\mathbf{x}$ and m (As a consequence, the m.l. estimates of θ from the observed data $\mathbf{x}$ and from the completed data $\hat{\mathbf{y}}$ are the same.) leads to (3.14).

In the mixture context, we have $v = v' = \nu_{m,K}$ and the C-S approximation consists in replacing the missing data z_{iK} with $t_{iK}(\hat{\theta})$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ and it leads to

$$\begin{aligned}\log \int \mathbf{f}(\mathbf{x} \mid m, K, \theta)\pi(\theta \mid m, K)d\theta &\approx \log \int \mathbf{g}(\hat{\mathbf{y}} \mid m, K, \theta)\pi(\theta \mid m, K)d\theta \\ &\quad - \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}(\hat{\theta})\log(t_{ik}(\hat{\theta})).\end{aligned}$$

Finally, the C-S approximation of the logarithm of the integrated observed likelihood results in the following criterion

$$\begin{aligned}\mathrm{CS}(m, K) &= \max_{\mathbf{a}} \log \mathbf{f}(\mathbf{x} \mid \mathbf{t}(\tilde{\theta}), m, K, \mathbf{a}) - \frac{\lambda_{m,K}}{2}\log n \\ &\quad + \log \Gamma(\frac{K}{2}) + \sum_{k=1}^{K} \log \Gamma(\bar{n}_k + \frac{1}{2}) \\ &\quad - K \log \Gamma(\frac{1}{2}) - \log \Gamma(n + \frac{K}{2}) + E(m, K), \qquad (3.16)\end{aligned}$$

where

$$\bar{n}_k = \sum_{i=1}^{n} t_{ik}(\tilde{\theta}).$$

Here, we make two observations.

- First, in the mixture context, the BIC approximation is not valid for both the numerator $\int p(\mathbf{x}, \theta \mid m) d\theta$ and the denominator $\int p(\mathbf{y}, \theta \mid m) d\theta$ of (3.15). BIC is used heuristically as is often used in this context.
- Second it seems that there is little difference between $\mathrm{CS}(m, K)$ and the BIC criterion

$$\mathrm{BIC}(m, K) = \mathbf{f}(\mathbf{x} \mid m, K, \hat{\theta}) - \frac{\nu_{m,K}}{2} \log n. \tag{3.17}$$

  Both criteria are compared on the basis of numerical experiments in the next section.

4 Numerical experiments

We compare the practical behavior of criteria $\mathrm{BIC}(m, K)$, $\mathrm{CS}(m, K)$ 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 Gaussian mixtures. The form m of the mixture model is defined by parameterizing the covariance matrix Σ_k of a component in terms of its eigenvalue decomposition, as developed in Banfield and Raftery (1993) and Celeux and Govaert (1995),

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{4.18}$$

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. Eight models assume diagonal variance matrices, and four models assume spherical shapes. Table 1 provides the designation and the characteristics of the 28 models. 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, the EM algorithm is initiated in the following way: First, the CEM algorithm is ran r times from random centers. The chosen number r depends on the features of the data to be classified. Typically r increases with the sample size and the space dimension. It is specified for each experiment hereafter. Then, there are two possibilities: either the CEM algorithm provided

model	proportion	volume	shape	orientation
$[p\lambda I]$	fixed	fixed	spherical	NA
$[p\lambda_k I]$	fixed	variable	spherical	NA
$[p\lambda B]$	fixed	fixed	fixed	diagonal
$[p\lambda_k B]$	fixed	variable	fixed	diagonal
$[p\lambda B_k]$	fixed	fixed	variable	diagonal
$[p\lambda_k B_k]$	fixed	variable	variable	diagonal
$[p\lambda C]$	fixed	fixed	fixed	fixed
$[p\lambda_k C]$	fixed	variable	fixed	fixed
$[p\lambda C_k]$	fixed	fixed	variable	variable
$[p\lambda_k C_k]$	fixed	variable	variable	variable
$[p\lambda D A_k D']$	fixed	fixed	variable	fixed
$[p\lambda_k D A_k D']$	fixed	variable	variable	fixed
$[p\lambda D_k A D'_k]$	fixed	fixed	fixed	variable
$[p\lambda_k D_k A D'_k]$	fixed	variable	fixed	variable
$[p_k\lambda I]$	variable	fixed	spherical	NA
$[p_k\lambda_k I]$	variable	variable	spherical	NA
$[p_k\lambda B]$	variable	fixed	fixed	diagonal
$[p_k\lambda_k B]$	variable	variable	fixed	diagonal
$[p_k\lambda B_k]$	variable	fixed	variable	diagonal
$[p_k\lambda_k B_k]$	variable	variable	variable	diagonal
$[p_k\lambda C]$	variable	fixed	fixed	fixed
$[p_k\lambda_k C]$	variable	variable	fixed	fixed
$[p_k\lambda C_k]$	variable	fixed	variable	variable
$[p_k\lambda_k C_k]$	variable	variable	variable	variable
$[p_k\lambda D A_k D']$	variable	fixed	variable	fixed
$[p_k\lambda_k D A_k D']$	variable	variable	variable	fixed
$[p_k\lambda D_k A D'_k]$	variable	fixed	fixed	variable
$[p_k\lambda_k D_k A D'_k]$	variable	variable	fixed	variable

Table 1: Description of the 28 Gaussian mixture models.

a no empty cluster partition and the EM algorithm is initiated with the parameter values derived from this partition, or the CEM algorithm provided partitions with at least one empty cluster and the EM algorithm is initiated r times with random centers.

4.1 Monte Carlo experiments

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

4.1.1 Four clusters with different overlapping

We simulated a four-component Gaussian mixture with the following sample size, dimension and parameter values (one of the 50 simulated data sets is displayed in Figure 1(a)):

$$n = 200, d = 2 \quad \begin{array}{lll} p_1 = 0.25 & \mu_1 = (0,0)' & \Sigma_1 = I \\ p_2 = 0.25 & \mu_2 = (4,0)' & \Sigma_2 = I \\ p_3 = 0.25 & \mu_3 = (7,0)' & \Sigma_3 = I \\ p_4 = 0.25 & \mu_4 = (9,0)' & \Sigma_4 = I. \end{array}$$

In this experiment all the 28 models mentioned in Table 1 were considered, with the number of clusters varying from one to seven and r being set to 20. Percentage of choosing a couple (m, K) is displayed in Table 2 for BIC and CS, since both criteria give exactly the same answer (In fact, in all experiments we considered, there is very little difference between BIC and CS.), and in Table 3 for ICL. The favorite couple (m, K) of BIC and CS (resp. ICL) is depicted in Figure 1(b) (resp. (c)) for the data set displayed in Figure 1 (a). Those figures provide iso-density ellipses for each component. It is remarkable than BIC and CS select the exact model with the exact number of components. ICL prefers a two-component mixture of the form $[p_k\lambda_k B_k]$ which makes sense from the clustering point of view.

4.1.2 A four dimensional Gaussian mixture

The situation we consider here has been first proposed in Bozdogan (1993) and also considered in Celeux and Soromenho (1996). It is a five-component Gaussian mixture with the following characteristics:

$$n = 625, d = 4 \quad \begin{array}{lll} p_1 = 0.12 & \mu_1 = (10,12,10,12)' & \Sigma_1 = I \\ p_2 = 0.16 & \mu_2 = (8.5,10.5,8.5,10.5)' & \Sigma_2 = I \\ p_3 = 0.20 & \mu_3 = (12,14,12,14)' & \Sigma_3 = I \\ p_4 = 0.24 & \mu_4 = (13,15,7,9)' & \Sigma_4 = 4I \\ p_5 = 0.28 & \mu_5 = (7,9,13,15)' & \Sigma_5 = 9I. \end{array}$$

For this experiment, only the exact model $[p_k\lambda_k I]$ has been considered when running the EM algorithm, the number of clusters is varying from one to eight and $r = 50$. Percentage of choosing the number of components K is displayed in Table 4 for BIC, CS and ICL. We note that BIC and CS tend to overestimate the right number of components (five) whereas

$m\|K$	1	2	3	4	5	6	7	$\hat{m}$ (%)
$[p\lambda I]$	0	0	8	**78**	0	0	0	86
$[p\lambda_k I]$	0	0	0	0	0	0	0	0
$[p\lambda B]$	0	0	6	4	0	0	0	10
$[p\lambda_k B]$	0	2	0	0	0	0	0	2
$[p\lambda B_k]$	0	0	0	0	0	0	0	0
$[p\lambda_k B_k]$	0	0	0	0	0	0	0	0
$[p\lambda C]$	0	0	0	0	0	0	0	0
$[p\lambda_k C]$	0	0	0	0	0	0	0	0
$[p\lambda C_k]$	0	0	0	0	0	0	0	0
$[p\lambda_k C_k]$	0	0	0	0	0	0	0	0
$[p\lambda D A_k D']$	0	0	0	0	0	0	0	0
$[p\lambda_k D A_k D']$	0	0	0	0	0	0	0	0
$[p\lambda D_k A D'_k]$	0	0	0	0	0	0	0	0
$[p\lambda_k D_k A D'_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda I]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k I]$	0	0	0	0	0	0	0	0
$[p_k\lambda B]$	0	2	0	0	0	0	0	2
$[p_k\lambda_k B]$	0	0	0	0	0	0	0	0
$[p_k\lambda B_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k B_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda C]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k C]$	0	0	0	0	0	0	0	0
$[p_k\lambda C_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k C_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda D A_k D']$	0	0	0	0	0	0	0	0
$[p_k\lambda_k D A_k D']$	0	0	0	0	0	0	0	0
$[p_k\lambda D_k A D'_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k D_k A D'_k]$	0	0	0	0	0	0	0	0
$\hat{K}$ %	0	4	14	82	0	0	0	

Table 2: Four clusters: percentage of choosing (K, m) by BIC and CS.

$m \vert K$	1	2	3	4	5	6	7	$\hat{m}$ (%)
$[p\lambda I]$	0	0	20	0	0	0	0	20
$[p\lambda_k I]$	0	0	0	0	0	0	0	0
$[p\lambda B]$	16	0	2	2	0	0	0	20
$[p\lambda_k B]$	0	0	0	0	0	0	0	0
$[p\lambda B_k]$	0	0	0	0	0	0	0	0
$[p\lambda_k B_k]$	0	0	0	0	0	0	0	0
$[p\lambda C]$	0	2	2	0	0	0	0	4
$[p\lambda_k C]$	0	0	0	0	0	0	0	0
$[p\lambda C_k]$	0	0	0	0	0	0	0	0
$[p\lambda_k C_k]$	0	0	0	0	0	0	0	0
$[p\lambda D A_k D']$	0	0	0	0	0	0	0	0
$[p\lambda_k D A_k D']$	0	0	0	0	0	0	0	0
$[p\lambda D_k A D'_k]$	0	0	0	0	0	0	0	0
$[p\lambda_k D_k A D'_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda I]$	0	0	16	0	0	0	0	16
$[p_k\lambda_k I]$	0	0	0	0	0	0	0	0
$[p_k\lambda B]$	0	4	2	0	0	0	0	6
$[p_k\lambda_k B]$	0	0	0	0	0	0	0	0
$[p_k\lambda B_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k B_k]$	0	**28**	0	0	0	0	0	28
$[p_k\lambda C]$	0	0	2	0	0	0	0	2
$[p_k\lambda_k C]$	0	0	0	0	0	0	0	0
$[p_k\lambda C_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k C_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda D A_k D']$	0	0	0	0	0	0	0	0
$[p_k\lambda_k D A_k D']$	0	0	2	0	0	0	0	2
$[p_k\lambda D_k A D'_k]$	0	0	0	0	0	0	0	0
$[p_k\lambda_k D_k A D'_k]$	0	0	2	0	0	0	0	2
$\hat{K}$ %	16	34	48	2	0	0	0	

Table 3: Four clusters: percentage of choosing (K, m) by ICL.

ICL favors the four component solution. This four cluster solution, merging components 1 and 2 into one cluster, is a reasonable solution as it appears from Figure 2, which provides the density of the projections of each mixture component on the four coordinates axes. This solution has also been selected using the NEC criterion of Celeux and Soromenho (1996).

4.1.3 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 follows:

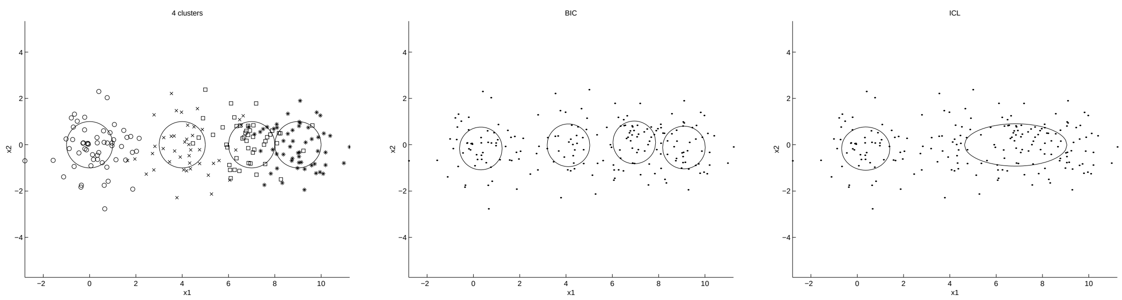
$$n = 200, d = 2$$
$$f(\mathbf{x}) = 0.5 \underbrace{\left[0.25\ \mathbf{I}_{[-1,1]}(x^1)\ \mathbf{I}_{[-1,1]}(x^2)\right]}_{\text{non-Gaussian cluster}} + 0.5 \underbrace{\left[\phi(\mathbf{x} \mid (3.3, 0)', I)\right]}_{\text{Gaussian cluster}},$$

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 and $r = 20$. Percentage of choosing K is displayed in Table 5. 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 unidimensional data. We consider the four models available in dimension one ($[p\lambda I]$, $[p_k\lambda I]$, $[p\lambda_k I]$, $[p_k\lambda_k I]$) with $K = 1, \ldots, 6$ and $r = 20$ to compute the BIC, CS and ICL criteria on the 299 eruption durations of the Old Faithful geyser (Azzalini and Bowman 1990). Numerical values of these criteria are respectively displayed in Tables 6, 7 and 8. In those tables, 'NA' indicates, in all cases but one, that the EM algorithm did not converge for the configuration at hand. For the solution selected by BIC (model $[p\lambda_k I]$ with $K = 4$), ICL is not available; but the reason for this is not the non convergence of EM, but it is due to the fact that the MAP operator leads to a degenerate partition (three clusters instead of four). BIC and CS select a couple (m, K) fitting well the second peak of the density (see Figure 4(a)) whereas ICL prefers a mixture with more separated components (see Figure 4(b)). Moreover, it is worth noting that ICL is the sole criterion choosing $K = 2$ for some of the models.



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

Figure 1: Four clusters with different overlapping.

K	1	2	3	4	5	6	7	8
BIC	0	0	0	22	18	**44**	16	0
CS	0	0	0	20	18	**44**	16	2
ICL	0	0	0	**52**	24	18	6	0

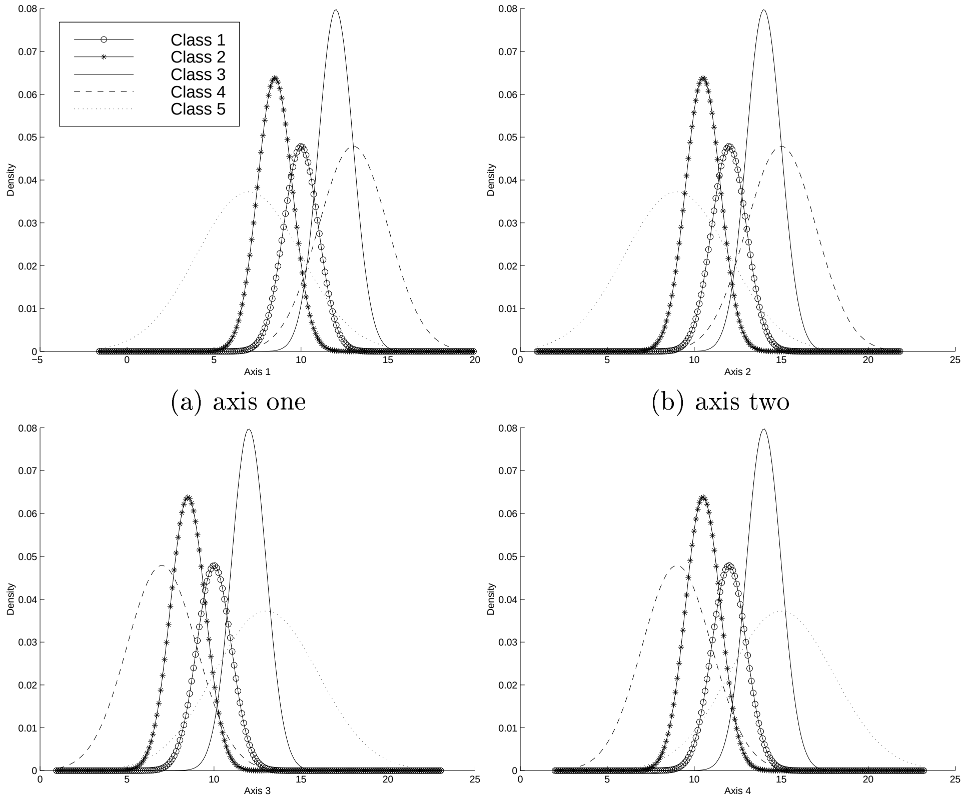
Table 4: Bozdogan's samples: percentage of choosing K with the model $[p_k \lambda_k I]$.

K	1	2	3	4	5
BIC	0	**66**	0	30	4
CS	0	**66**	0	30	4
ICL	0	**100**	0	0	0

Table 5: Non-Gaussian cluster samples: percentage of choosing K with the model $[p\lambda I]$.

$m \vert K$	1	2	3	4	5	6	$\hat{K}$
$[p\lambda I]$	-471	-326	-315	-324	-337	-322	3
$[p\lambda_k I]$	-471	-325	-298	**1302**	NA	NA	4
$[p_k\lambda I]$	-471	-476	-319	-324	-330	-336	3
$[p_k\lambda_k I]$	-471	-312	-288	NA	NA	NA	3
$\hat{m}$	$[p\lambda I]$	$[p_k\lambda_k I]$	$[p_k\lambda_k I]$	$[p\lambda_k I]$	$[p_k\lambda I]$	$[p\lambda I]$	

Table 6: BIC values for the Old Faithful geyser.



(a) axis one (b) axis two

(c) axis three (c) axis four

Figure 2: Bozdogan's class densities projected on the four coordinate axes.

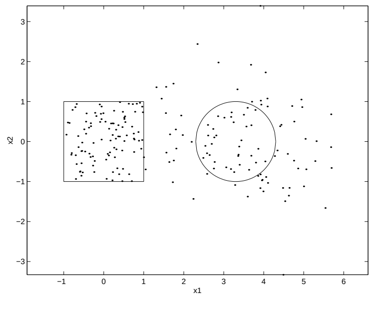


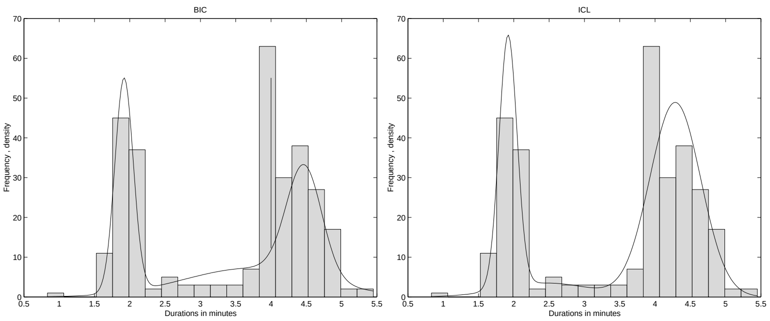
Figure 3: A uniform and a Gaussian cluster.

$m\|K$	1	2	3	4	5	6	$\hat{K}$
$[p\lambda I]$	-471	-326	-315	-324	-337	-322	3
$[p\lambda_k I]$	-471	-325	-298	**1283**	NA	NA	4
$[p_k\lambda I]$	-471	-477	-319	-324	-329	-334	3
$[p_k\lambda_k I]$	-471	-313	-288	NA	NA	NA	3
$\hat{m}$	$[p\lambda I]$	$[p_k\lambda_k I]$	$[p_k\lambda_k I]$	$[p\lambda_k I]$	$[p_k\lambda I]$	$[p\lambda I]$	

Table 7: CS values for the Old Faithful geyser.

$m\|K$	1	2	3	4	5	6	$\hat{K}$
$[p\lambda I]$	-471	-327	-349	-416	-476	-499	2
$[p\lambda_k I]$	-471	-326	-368	NA	NA	NA	2
$[p_k\lambda I]$	-471	-566	-347	-396	NA	NA	3
$[p_k\lambda_k I]$	-471	-313	**-297**	NA	NA	NA	3
$\hat{m}$	$[p\lambda I]$	$[p_k\lambda_k I]$	$[p_k\lambda_k I]$	$[p_k\lambda I]$	$[p\lambda I]$	$[p\lambda I]$	

Table 8: ICL values for the Old Faithful geyser.



(a) BIC (b) ICL

Figure 4: Density estimation for the Old Faithful geyser.

4.2.2 French departments

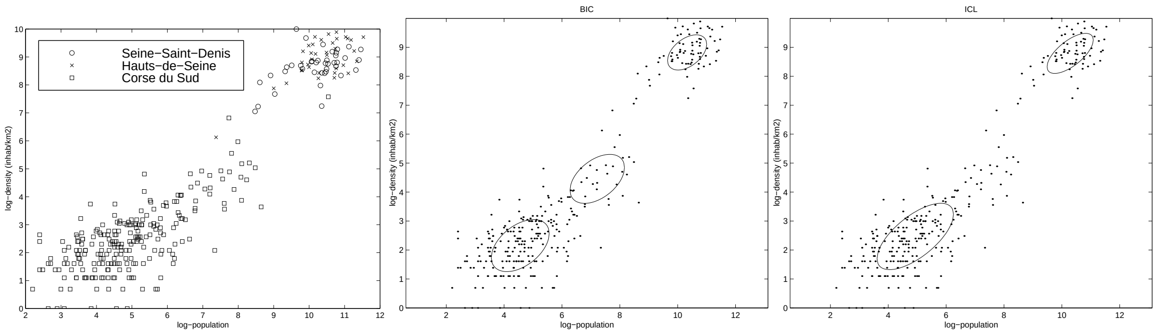
Figure 5(a) displays log-population versus log-density (in habitants/km2) of 312 towns of three French departments: Two densely-populated departments in the suburbs of Paris, Seine-Saint-Denis and Hauts-de-Seine, and one rural department Corse du Sud (source: census 1990 of the French population, World Wide Web of INSEE at `http://www.insee.fr/vf/-chifcles/rp90/index.htm`). For this bivariate data set, we consider all the 28 models described in Table 1 with $K = 1, \ldots, 5$ and $r = 20$. All the criteria favor the model $[p_k \lambda_k C]$ and we restrict attention to this model. Table 9 gives BIC and ICL values for $K = 1, \ldots, 5$. BIC and CS choose the model $[p_k \lambda_k C]$ with $K = 3$, whereas ICL hesitates between $K = 2$ and $K = 3$. The two-cluster solution has an interesting interpretation since one cluster is closely related to Corse du Sud 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 split Corse du Sud into two clusters.

4.2.3 Global precipitation climatology

We analyzed data for a global precipitation climatology that has been produced at the Joint Institute for the Study of the Atmosphere and Ocean. They are available on the World Wide Web at `tao.atmos.washington.edu/legates_msu`. The spatial resolution of this climatology is 2.5 degrees in latitude and longitude, which leads to a set of twelve 144 $\times$ 72 maps representing stations or points (pixels) at which monthly average precipitation (in mm) have been recorded or extrapolated, for each individual calendar month. Figure 6 shows such a map for the month of January. The land data for this climatology is taken from the Legates and Willmott (1990) climatology, which is based on the historical record of rain gauge measurements. The ocean precipitation estimates are from the Microwave Sounding Unit (MSU) (Spencer 1993), and the climatology is based on averages for the period 1979 to 1992.

Following Posse (1998), we pre-processed the data in the following way. As it is far from being normally distributed, a non-linear transform was first applied: the power 0.25 of each record was taken. Its dimension was then reduced via Principal Component Analysis and we experimented on the first two principal components depicted in Figure 8 (a).

We considered the most general model $[p_k \lambda_k C_k]$ and the number of clusters is varying from $K = 1$ to $K = 20$. Here, we did not initiate the EM algorithm from a CEM algorithm partition. Actually, it appeared that for this data set initiating from a random position leads to a larger likelihood than using a CEM initialization. Thus, we ran the EM algorithm $r = 20$ times from a random position and chose the solution providing the largest likelihood. Figure 7 displays values of BIC and ICL. BIC criterion increases monotically with K and does not provide evidence for any K value. On the contrary, ICL gives a preference for the seven-clusters partition displayed in Figure 8 for both the two first principal components and the corresponding map of the world. From Figure 8 (a) it seems that declaring for no clustering structure in this bidimensional data set is not realistic and that the seven-cluster partition selected by ICL captures the high density regions appearing in this data set.



(a) true partition (b) BIC-ICL 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**	**-1044**	-1049	- 1109	2

Table 9: BIC and ICL values for the best model $[p_k \lambda_k C]$ on the French department data.

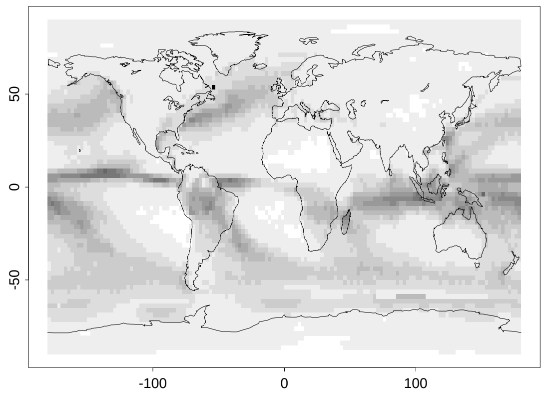


Figure 6: Monthly average precipitation (in mm) for the month of January. The spatial resolution is 2.5 degrees in latitude and longitude, which results in a 144 × 72 image.

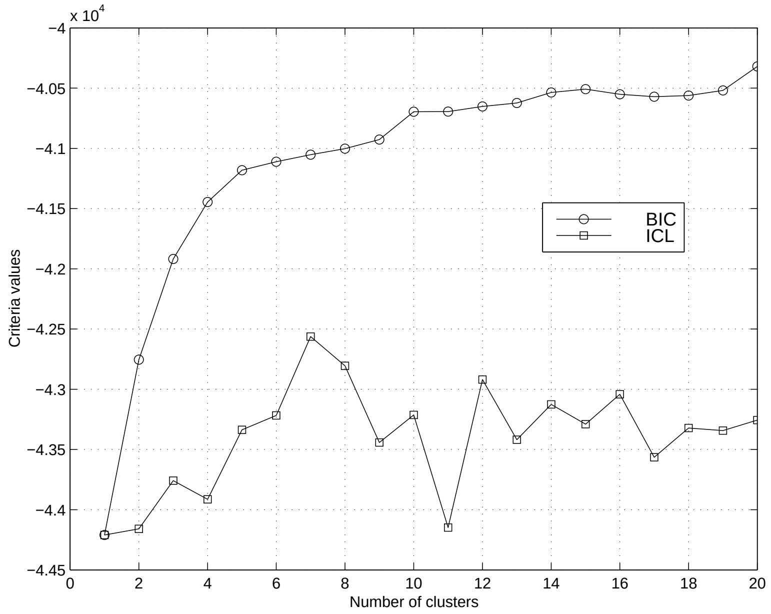
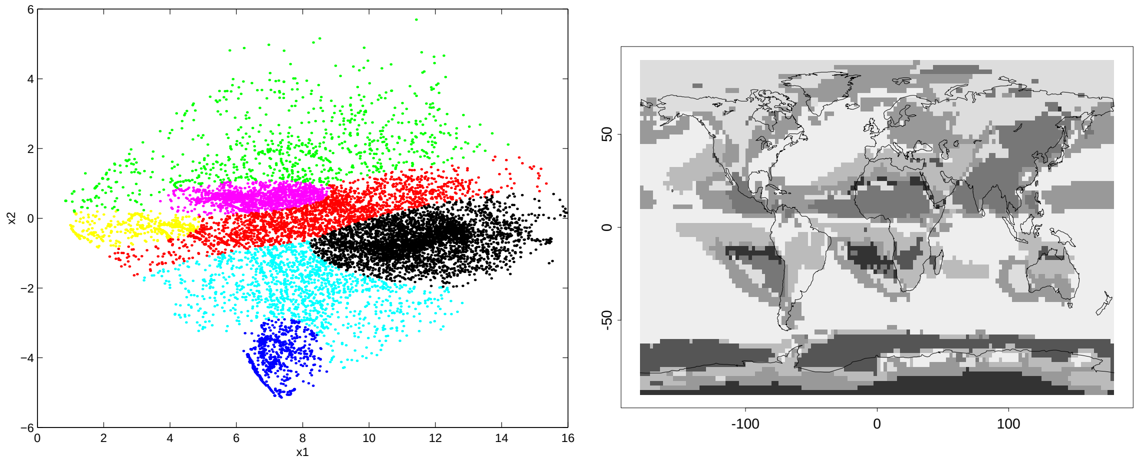


Figure 7: BIC and ICL values with the model $[p_k \lambda_k C_k]$ applied to the Legates/MSU data.



(a) the first two principal components (b) corresponding map of the world

Figure 8: Seven-cluster partition of the Legates/MSU precipitation data.

5 Discussion

Statistical analysis of finite mixtures are employed in statistical modelling with two different purposes. In one perspective, finite mixture are essentially regarded as competitors to non parametric density estimation (see Escobar and West 1995, Robert 1996 or Roeder and Wasserman 1997). In another view, finite mixture are considered as a powerful way of modelling in cluster analysis (see Fraley and Raftery 1998 or McLachlan and Basford 1988). 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 modelling 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 Section 4.2.1 highlight this satisfactory behavior of BIC in a spectacular way.

When the interest in mixture modelling is cluster analysis choosing a sensible number of clusters K is of crucial importance. 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. Experiment in Section 4.1.3 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. Firstly, the ICL criterion resulting from the approximation of this integrated completed likelihood does not suffer the lack of theoretical justification of the BIC approximation to the integrated observed likelihood for mixture model. Secondly, from a practical point of view, the ICL criterion 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 4.1.3 and 4.2. 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 (Banfield and Raftery 1993) which tends to underestimate the number of clusters as shown in Celeux and Soromenho (1996), or the entropy criterion NEC (Celeux and Soromenho 1996) which exhibits a deceptive behavior to choose a relevant form m of the mixture model as shown in Biernacki (1997).

As compared to the integrated observed likelihood, the integrated completed likelihood includes an additionnal entropy term $E(m, K)$ which favors well-separated clusters and which is the essential difference between BIC and ICL criteria. The fact that the approximation leading to ICL is well justified contrary to the BIC approximation seems to have a little practical impact. In our experiments, we also compute the *à la* BIC approximation of ICL (Eq. 3.12). This approximation did not give different answer than the ICL criterion both for choosing the form m of the mixture model and the number of clusters K in all the six experiments that we considered.

References

Aitkin, M. and Rubin, D. B. (1985). Estimation and Hypothesis Testing in Finite Mixture Models. *Journal of the Royal Statistical Society, B*, **47**, 67-75.

Azzalini, A. and Bowman A. W. (1990). A Look at some Data on the Old Faithful Geyser. *Applied Statistics*, **39**, 357-365.

Banfield, J. D. and Raftery A. E. (1993). Model-based Gaussian and non Gaussian clustering. *Biometrics*, **49**, 803-821.

Biernacki, C. (1997). Choix de modèles en classification. PhD. thesis, UTC Compiègne.

Biernacki, C. and Govaert, G. (1997). Using the Classification Likelihood to Choose the Number of Clusters. *Computing Science and Statistics*, **29**, 2, 451-457.

Bozdogan, H. (1993). 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.

Bryant, P. G. (1991). Large Sample Results for Optimization Based Clustering Methods. *Journal of Classification*, **8**, 31-44.

Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm and two Stochastic Versions. *Computational Statistics and Data Analysis*, **14**, 315-332.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781-793.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195-212.

Cheeseman, P., and Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and results. In Fayyad, U., Piatesky-Shapiro, G., and Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining*, pp. 61-83. AAAI Press, Menlo Park, CA.

Chickering, D. M. and Heckerman, D. (1997). Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. *Machine Learning*. **29**, 181-212.

Cutler, A. and Cordero-Braña, O. (1996). Minimum Hellinger Distance Estimation for Finite Mixture Models. *Journal of the American Statistical Association* **91**, 1716-1723.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood for Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, B*, **39**, 1-38.

Diebolt, J. and Robert, C. P. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society*, B, **56**, 363-375.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577-588.

Everitt, B. S. (1984). *An Introduction to Latent Variables Models.* London: Chapman & Hall.

Fraley, C. and Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. Technical Report No. 329, Department of Statistics, University of Washington. (To appear in *Computer Journal*.)

Kass, R. E. and Raftery, A. E. (1995). Bayes Factor. *Journal of the American Statistical Association*, **90**, 733-795.

Legates, D. R. and Willmott, C. J. (1990). Mean Seasonal and Spatial Variability Gauge-corrected, Global Precipitation. *International Journal of Climatology*, **10**, 111-127.

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

Posse, C. (1998). Hierarchical Model-based Clustering For Large Datasets. Technical Report, Department of Statistics, University of Minnesota.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research (with discussion). In *Sociological Methodology* (ed. P. V. Marsden), pp.111-195. Cambridge, Mass.: Blackwells.

Redner, R. A. and Walker H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195-239.

Richardson, S. and Green, P. J. (1997). Fully Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society, B*, **59**, 731-792.

Robert, C. P. (1994). *The Bayesian Choice : a Decision-Theoretic Motivation*. New York: Springer-Verlag.

Robert, C. P. (1996). Mixtures of Distributions: Inference and Estimation. *Markov Chain Monte Carlo in Practice*. (eds. Gilks W. R., Richardson S. and Spiegelhalter D. J.). London: Chapman & Hall.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association*, **92**, 894-902.

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

Spencer, R. W. (1993). Global Oceanic Precipitation from the MSU during 1979-91 and Comparisons to Other Climatologies. *Journal of Climate*, **6**, 1301-1326.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399