Choosing a model in a Classification purpose

Guillaume Bouchard, Gilles Celeux*

Abstract: We advocate the usefulness of taking into account the modelling purpose when selecting a model. Two situations are considered to support this idea: Choosing the number of components in a mixture model in a cluster analysis perspective, and choosing a probabilistic model in a supervised classification context. For this last situation we propose a new criterion, the Bayesian Entropy Criterion, and illustrate its behavior with numerical experiments.
Keywords: Integrated Likelihood, Integrated Complete Likelihood, Integrated conditional likelihood, AIC, BIC, ICL and BEC criteria, Mixture Discriminant Analysis.

1 Introduction

In statistical inference from data, selecting a parsimonious model among a collection of models is an important but difficult task. This general problem receives much attention since the seminal papers of Akaike (1974) and Schwarz (1978). A model selection problem consists essentially of solving the bias-variance dilemma: A too simple model will produce a large approximation error (underfitting) and a too complicated model will produce a large estimation error (overfitting).

A classical approach to the model assessing problem consists of penalizing the fit of a model by a measure of its complexity. A convenient measure of fit is the *deviance* of a model $m \in \mathcal{M}$, which is

$$d(\mathbf{x}) = 2[\log \mathbf{p}(\mathbf{x}) - \log \mathbf{p}(\mathbf{x}|m, \hat{\theta}_m)]$$

where $\mathbf{p}(\mathbf{x}) = \prod_{1=1}^{n} p(x_i)$ denotes the true distribution of the data $\mathbf{x} = (x_1, \ldots, x_n)$, (for simplicity, the $x_i's$ are supposed to be iid) $\mathbf{p}(\mathbf{x}|m, \theta_m) = \prod_{i=1}^{n} p(x_i|m, \theta_m)$ is the distribution under the model m parameterized with θ_m, and $\hat{\theta}_m$ is the maximum likelihood (ml) estimate of θ_m. Under the maximum likelihood approach and in a prediction perspective, a common way of penalization is based on the idea that the deviance will be smaller on a learning set than on a test set of comparable size, since we actually chose the parameters to minimize the deviance on the learning set. Thus, the problem when choosing a penalization term is to evaluate how large would be the difference on average over learning and test sets. That is the penalization would be an estimation of $nD(X) - E(d(\mathbf{x}))$ where

$$D(X) = 2E[\log p(X) - \log p(X|m, \hat{\theta}_m)]$$

is the expected deviance on a single test observation X. Assuming that the data arose from a distribution belonging to the collection of models in competition, Akaike (1974) proposes to estimate this difference with $2\nu_m$ where ν_m is the number of free parameters of the model m. This leads to the so called AIC criterion.

$$\mathrm{AIC}(m) = 2 \log \mathbf{p}(\mathbf{x}|m, \hat{\theta}_m) - 2\nu_m. \tag{1}$$

*The authors are with Inria. Address for correspondence: G. Celeux Dept. de mathmatiques, Btiment 425, Universit Paris-Sud F91405 Orsay Cedex, email: Gilles.Celeux@inria.fr.

Relaxing this unrealistic assumption leads to alternative criteria such as the NIC criterion of Murata *et al.* (1994). (Details can be found in Ripley 1996, pp. 32-34 and 61.)

An other point of view consists of basing the model selection on the integrated likelihood of the data in a Bayesian perspective (see Kass and Raftery 1995). This integrated likelihood is

$$\mathbf{p}(\mathbf{x}|m) = \int \mathbf{p}(\mathbf{x}|m, \theta_m)\pi(\theta_m)d\theta_m, \tag{2}$$

$\pi(\theta_m)$ being a prior distribution for parameter θ_m. The essential technical problem is to approximate this integrated likelihood in a right way. A classical asymptotic approximation of the logarithm of the integrated likelihood is the BIC criterion of Schwarz(1978). It is

$$\text{BIC}(m) = \log \mathbf{p}(\mathbf{x}|m, \hat{\theta}_m) - \frac{\nu_m}{2}\log(n). \tag{3}$$

This approximation needs regularity conditions on the likelihoods of the model collection $\mathcal{M}$ and is accurate when the prior distribution $\pi(\theta_m)$ is centered around the maximum likelihood estimate $\hat{\theta}_m$ (see Raftery 1995). Notice that it has been argued that this formulation may only be appropriate in circumstances where it was really believed that one and only one of the competing models is in fact true (Bernardo and Smith 1994, chapter 6).

Beyond technical difficulties which can occur when choosing a model, the scope of this paper is to show how it can be fruitful to take into account the purpose of the model user to get reliable and useful models for statistical description or decision tasks. In that viewpoint in mind, we focus on model-based cluster analysis and generative models for supervised classification.

The paper is organized as follows. In Section 2, the problem of assessing the number of components in a mixture model is considered from the cluster analysis point of view. In Section 3, a criterion for choosing a model in a supervised classification context is proposed and is experimented for choosing a model among mixture discriminant analysis models. A discussion section ends the paper.

2 Choosing the number of mixture components for clustering

Assessing the number K of components in a mixture model is a difficult question, from both theoretical and practical points of view, which had received much attention in the past two decades. In this paper, we do not propose a state of the art of this problem which has not been completely resolved. The reader is referred to the chapter 6 of the recent book of McLachlan and Peel (2000) for an excellent overview on this subject. We are essentially aiming to discuss elements of practical interest regarding the problem of choosing the number of mixture components when concerned with cluster analysis.

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

$$p(\mathbf{x}_i \mid K, \theta_K) = \sum_{k=1}^{K} p_k \phi(\mathbf{x}_i \mid \mathbf{a}_k) \tag{4}$$

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 a parameterized density (usually the d-dimensional Gaussian density) with parameter a_k, and $\theta_K = (p_1, \ldots, p_{K-1}, a_1, \ldots, a_K)$.

Assessing the number of components in the mixture model is known as a difficult problem from the theoretical point of view. As a matter of fact, even when K^* the right number of component

is assumed to exist, if $K^* < K_0$ then K^* is not identifiable in the parameter space Θ^{K_0} (see for instance McLachlan and Peel 2000, chapter 6).

But, here, we want to stress the importance of taking into account of the modelling context to select a reasonable and useful number of mixture components. Our opinion is that, behind the theoretical difficulties, assessing the number of components in a mixture model from data is a weakly identifiable statistical problem. Mixture densities with different number of components can lead to quite similar resulting densities. For instance, the galaxy velocities data of Roeder (1990) has became a benchmark data set and is used by many authors to illustrate procedures for choosing the number of mixture components. Now, according to those authors the answer lies from $K = 2$ to $K = 10$, and it is not exaggerating a lot to say that all the answers between 2 and 10 have been proposed as the good answer, at least one time, in the papers considering this particular data set. (An interesting and illuminating comparative study on this data set can be found in Aitkin, 2001.) Thus, we consider that it is highly desirable to choose K by keeping in mind what is expected from the mixture modelling to get a relevant answer to this question. Actually, mixture modelling can be used in quite different purposes. It can be regarded as a semi parametric tool for density estimation purpose or as a tool for cluster analysis.

In the first perspective, much considered by Bayesian statisticians, numerical experiments (see Roeder and Wasserman 1997 or Fraley and Raftery 1998, 2002) show that the BIC approximation of the integrated likelihood works well at a practical level. Moreover, under regularity conditions including the fact that the component densities are finite, Keribin (2000) proved that BIC provides a consistent estimator of K.

But, in the second perspective, the integrated likelihood does not take into account the clustering purpose at hand for selecting a mixture model in a model-based clustering setting. As a consequence, in the most current situations where the distribution from which the data arose is not in the collection of considered mixture models, BIC criterion will tend to overestimate the correct size regardless of the separation of the clusters (see Biernacki, Celeux and Govaert 2000 for illustrations).

To overcome this limitation, it can be advantageous to choose K in order to get the mixture giving rise to partitioning data with the greatest evidence. With that purpose in mind, Biernacki *et al.* (2000) considered the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z})$ (or integrated completed likelihood), $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ denoting the missing data such that $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ are binary K-dimensional vectors with $z_{ik} = 1$ if and only if $\mathbf{x}_i$ arises from component k. Those missing indicator vectors define a partition $P = (P_1, \ldots, P_K)$ of the observed data $\mathbf{x}$ with $P_k = \{\mathbf{x}_i \mid z_{ik} = 1\}$. Then, the integrated complete likelihood is

$$\mathbf{p}(\mathbf{x}, \mathbf{z} \mid K) = \int_{\Theta_K} \mathbf{p}(\mathbf{x}, \mathbf{z} \mid K, \theta)\pi(\theta \mid K)d\theta, \tag{5}$$

where

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

with

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

To approximate this integrated complete likelihood, those authors propose to use a BIC-like approximation leading to the criterion

$$\mathrm{ICL}(K) = \log \mathbf{p}(\mathbf{x}, \hat{\mathbf{z}} \mid K, \hat{\theta}) - \frac{\nu_K}{2} \log n, \tag{6}$$

where the missing data have been replaced by their most probable value for parameter estimate $\hat{\theta}$. (Details can be found in Biernacki *et al.* 2000.) Roughly speaking criterion ICL is the criterion BIC penalized by the estimated mean entropy

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

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

As a consequence, ICL favors K values giving rise to partitioning the data with the greatest evidence, as highlighted in the numerical experiments in Biernacki *et al.* (2000), because of this additional entropy term. Most generally, ICL appears to provide a stable and reliable estimate of K for real data sets and also for simulated data sets from mixtures when the components are not too much overlapping (see for instance McLachlan and Peel, 2000). But ICL, which is not aiming to discover the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components as illustrated in Figueiredo and Jain (2002).

On the contrary, BIC performs remarkably well to assess the true number of components from simulated data (see Biernacki *et al.* 2000, Fraley and Raftery, 1998, 2002, for instance). But, for real world data sets, BIC has a marked tendency to overestimate the numbers of components. The reason is that real data sets do not arise from the mixture densities at hand, and the penalty term of BIC is not strong enough to balance the tendency of the loglikelihood to increase with K to improve the fit of the mixture model.

3 Model Selection in Classification

Supervised classification is about guessing the unknown class, denoted by z and taking value in $\{1, \ldots, K\}$ of an observation $\mathbf{x}$. For that purpose, a decision function, called a classifier, $(\delta(\mathbf{x}) : \mathbf{R}^d \to \{1, \ldots, K\}$ is designed from a learning sample $(\mathbf{x}_i, z_i), i = 1, \ldots, n)$. A classical approach to design a classifier is to represent the class conditional densities with a parametric model $\mathbf{p}(\mathbf{x}|m, z = k, \theta_m)$ for $k = 1, \ldots, K$. Then the classifier is assigning an observation $\mathbf{x}$ to the class k maximizing the conditional probability of a class $p(z = k|m, \mathbf{x}, \theta_m)$. Using the Bayes rule, it leads to set $\delta(\mathbf{x}) = j$ if and only if $k = \arg\max_k p_k \mathbf{p}(\mathbf{x}|m, z = k, \hat{\theta}_m)$, $\hat{\theta}_m$ being the ml estimate of the class conditional parameters θ and p_k being the prior probability of class k. This approach is known as the generative discriminant analysis in the Machine Learning community (see for instance Schölkopf and Smola 2002).

In this context, it could be expected to improve the actual error rate by selecting a generative model m among a large collection of models $\mathcal{M}$ (see for instance Friedman 1989 or Bensmail and Celeux 1996). Recently Hastie and Tibshirani (1996) proposed to model each class density with a mixture of Gaussian distributions. In this approach the number of mixture components *per* class are sensitive tuning parameters. They can be supplied by the user, as in Hastie and Tibshirani (1996), but it is clearly a sub-optimal solution. They can be chosen to minimize the v-fold cross-validated error rate, as done in Friedman (1989) or Bensmail and Celeux (1996) for other tuning parameters. Despite the fact the choice of v can be sensitive, it can be regarded as a nearly optimal solution. But it is highly CPU time consuming. Thus choosing such tuning parameter with a penalized loglikelihood criterion, as BIC, can be expected to be much more efficient in many situations. In such a context, denoting $\mathbf{z} = (z_1, \ldots, z_n)$ the classification of the learning sample, BIC takes the form

$$\text{BIC}(m) = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \frac{\nu_m}{2} \log(n), \tag{7}$$

where ν_m is the dimension of θ_m. But, BIC measures the fit of the model m to the data $(\mathbf{x}, \mathbf{z})$ rather than its ability to produce a reliable classifier. Thus, in many situations, BIC can have a tendency to overestimate the complexity of the generative classification model to be chosen. In order to counter this tendency, we now propose a penalized likelihood criterion taking into account the classification task when evaluating the performance of a model.

3.1 The Bayesian Entropy Criterion

As stated above, a classifier deduced from model m is assigning an observation $\mathbf{x}$ to the class k maximizing $p(z = k|m, \mathbf{x}, \hat{\theta}_m)$. Thus, from the classification point of view, the conditional likelihood $\mathbf{p}(\mathbf{z}|m, \mathbf{x}, \theta_m)$ has a central position. For this very reason, we propose to make use of the integrated conditional likelihood

$$\mathbf{p}(\mathbf{z}|m, \mathbf{x}) = \int \mathbf{p}(\mathbf{z}|m, \mathbf{x}, \theta_m)\pi(\theta_m)d\theta_m, \tag{8}$$

where $\pi(\theta_m)$ is the prior distribution of θ_m, to select a relevant model m. As for the integrated likelihood, this integral is generally difficult to calculate and has to be approximated. The approximation we now present of $\log \mathbf{p}(\mathbf{z}|m, \mathbf{x})$ leads to the so-called Bayesian Entropy Criterion (BEC). In the following we denote

$$\hat{\theta}_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \theta_m),$$

$$\tilde{\theta}_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{x}|m, \theta_m)$$

and

$$\theta^\star_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{z}|m, \mathbf{x}, \theta_m).$$

We have

$$\mathbf{p}(\mathbf{z}|m, \mathbf{x}) = \frac{\mathbf{p}(\mathbf{x}, \mathbf{z}|m)}{\mathbf{p}(\mathbf{x}|m)}$$

with

$$\mathbf{p}(\mathbf{x}, \mathbf{z}|m) = \int \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \theta_m)\pi(\theta_m)d\theta_m, \tag{9}$$

and

$$\mathbf{p}(\mathbf{x}|m) = \int \mathbf{p}(\mathbf{x}|m, \theta_m)\pi(\theta_m)d\theta_m. \tag{10}$$

It is valid to approximate logarithms of integrals (**??**) and (**??**) with the BIC approximation according to a line described in Raftery (1995). It leads to

$$\log \mathbf{p}(\mathbf{x}, \mathbf{z}|m) = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \frac{\nu}{2}\log n + \mathrm{O}(1)$$

and

$$\log \mathbf{p}(\mathbf{x}|m) = \log \mathbf{p}(\mathbf{x}|m, \tilde{\theta}_m) - \frac{\nu}{2}\log n + \mathrm{O}(1),$$

ν being the dimension of the vector parameter θ. From which it follows that

$$\log \mathbf{p}(\mathbf{z}|m, \mathbf{x}) = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \log \mathbf{p}(\mathbf{x}|m, \tilde{\theta}_m) + \mathrm{O}(1). \tag{11}$$

Thus the approximation of $\log \mathbf{p}(\mathbf{z}|m, \mathbf{x})$ that we proposed is

$$\mathrm{BEC} = \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m) - \log \mathbf{p}(\mathbf{x}|m, \tilde{\theta}_m). \tag{12}$$

Some remarks are in order.

1. Denoting, as in Section 2, $t_{ik}(m, \hat{\theta}_m)$ the conditional probability that $\mathbf{x}_i$ arises from class k in model m with ml parameter estimate $\hat{\theta}_m$, we can write

$$\log \mathbf{p}(\mathbf{z}|m, \mathbf{x}, \hat{\theta}) = \sum_{i=1}^{n} \log t_{iz_i}(m, \hat{\theta}_m)$$

which can be regarded as the entropy of the classification $\mathbf{z}$. And, roughly speaking, the criterion defined in (**??**) is related to this term. This is the reason why we called this criterion Bayesian Entropy Criterion (BEC).

2. Equation (**??**) is the approximation on which BEC is based and its O(1) error means that, in general, the error in it does not vanish as n tends to infinity. Thus BEC can be thought of as a crude approximation of $\log \mathbf{p}(\mathbf{z}|m, \mathbf{x})$. The criterion BEC can be more accurate in practice when $\hat{\theta} \approx \tilde{\theta}$. Typically this fact occurs, for the true model, when the joint distribution of the data $\mathbf{x}, \mathbf{z}$ belongs to one of the models in competition. Thus, some more accurate criteria than the present BEC criterion are desirable. For instance, the *à la* BIC approximation

$$\log \mathbf{p}(\mathbf{z}|m, \mathbf{x}) \approx \log \mathbf{p}(\mathbf{z}|m, \mathbf{x}, \theta^*_m) - \frac{\nu}{2} \log n$$

is sensible, but θ^* is difficult to derive for most of models. An other track of approximation consists of starting from the equation

$$\mathbf{p}(\mathbf{z}|m, \mathbf{x}) = \int \frac{\mathbf{p}(\mathbf{x}, \mathbf{z}|m, \theta_m)}{\mathbf{p}(\mathbf{x}|m, \theta_m)} \pi(\theta_m) d\theta_m,$$

but the resulting approximation involves the difficult estimation of matrices

$$J = -\frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta^t} \log \mathbf{p}(\mathbf{x}, \mathbf{z}|m, \hat{\theta}_m)$$

and

$$K = -\frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta^t} \log \mathbf{p}(\mathbf{x}|m, \tilde{\theta}_m).$$

In the present moment, we are working to propose a more accurate and simple alternative to criterion BEC.

3. The criterion BEC needs to compute $\tilde{\theta}_m = \arg\max_{\theta_m} \mathbf{p}(\mathbf{x}|m, \theta_m)$. Since, for $i = 1, \ldots, n$,

$$p(\mathbf{x}_i|m, \theta_m) = \sum_{k=1}^{K} p(z_{ik} = 1|m, \theta_m) p(\mathbf{x}_i|z_{ik} = 1, m, \theta_m),$$

$\tilde{\theta}$ is the ml estimate of a finite mixture distribution. It can be derived from the EM algorithm (see McLachlan and Peel 2000). In the present circumstance, the mixing proportions are known : $p_k = \text{card}\{i \text{ such that } z_{ik} = 1\}/n$ for $k = 1, \ldots, K$ and the EM algorithm can be initiated in a quite natural way with $\hat{\theta}$. Thus the calculation of $\tilde{\theta}$ involves no difficulty.

3.2 Numerical experiments

In this section, we report on some case studies for analyzing the practical ability of BEC to select a reasonable classification model and compare BEC with criteria as the cross-validated error rate and BIC. In an illustrative purpose, we concentrate on a collection of models assuming that the

class densities are a mixture of Gaussian distributions. This kind of modelling has been called MDA and studied in Hastie and Tishirani (1996). More precisely, we restrict attention to mixture of spherical Gaussian distributions, an attractive family of models for its simplicity and flexibility (Bouchard and Celeux 2003). Note that the spherical Gaussian components wil be called "balls" in this section.

The model is the following: Assuming that the number of clusters in each class is fixed to R_k, $k = 1, \cdots, K$, the distribution of data in the k^{th} class is

$$p_k(\boldsymbol{x}; \theta_k) = \sum_{r=1}^{R_k} \pi_r \phi(\boldsymbol{x}; \boldsymbol{\mu}_r, \sigma_r^2 I_d) \tag{13}$$

where π_r, μ_r and σ_r are respectively the weight, mean and standard deviation of the r^{th} component, $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ denoting the density of a Gaussian distribution with mean μ and covariance Σ. The set of parameters of class k is denoted θ_k. Obviously, the selection of the number of mixture components $\{R_k\}_{k=1,\cdots,K}$ is an important and difficult question. Cross-validation error rate can be regarded as the reference criterion but it is time consuming especillay when the cardinality of the collection of models to be compared is large. For the time consuming point of view, BIC is attractive. But, assuming spherical Gaussian mixtures for the class conditional densities can be regarded in many situations as a rough model. Thus, BIC can be expected to perform poorly in such cases since this criterion measures the fit of the spherical Gaussian mixture rather than the fit to the classification task. In practice, it is difficult to guess the practical behavior of BIC (see Bouchard and Celeux 2003).

We first present some Monte Carlo numerical experiments on simulated data sets, then we present some numerical experiments on real data sets.

3.2.1 Monte Carlo numerical experiments

In the first experiment a couple of models are compared. Fifty samples of $n = 120$ points from two classes with equal prior probabilities have been generated with the following class conditional densities:

$$X|Z = 1 \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

and

$$X|Z = 2 \sim \mathcal{N}\left(\begin{bmatrix} \Delta \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}\right).$$

The first model (DIAG) is considering a Gaussian class conditional distribution with a diagonal variance matrix, while the second model (BALL) is considering a Gaussian class conditional distributions with a spherical variance matrix. The performances of criteria BEC and BIC are compared in Table **??**. In this table, the column $\overline{err}$ gives the error rate obtained with an independant test sample of size 50,000. It appears that most often BEC chooses the model giving the smallest error rate with an higher probability than BIC does. BIC often selects the spherical Gaussian distribution because it is more suitable as a density estimate. When the class separation increases, BEC tends to choose the most parsimonious model more often as expected.

In the second experiment, the same data sets are used, but the considered models are now the spherical Gaussian mixture distributions described above, and the problem is to select the number of balls $R_k, k = 1, 2$. For simplicity, we assume that $R_1 = R_2$. The behavior of criteria BEC and BIC are compared in table **??**. It can be remarked that BEC criterion selects the complexity suitable for the classification purpose. For instance, in the well separated situation, the error rates of the different models are equivalent and BEC selects the simplest model most often. On the

separation	model	$\overline{err}$	BIC	BEC	BIC choice(%)	BEC choice(%)
$\Delta = 1$	DIAG	**0.250**	502.331	**64.108**	24	98
$\Delta = 1$	BALL	0.268	**500.422**	69.665	76	2
$\Delta = 3.5$	DIAG	**0.070**	502.331	**22.067**	24	94
$\Delta = 3.5$	BALL	0.076	**500.422**	26.120	76	6
$\Delta = 5$	DIAG	**0.019**	502.331	**6.081**	24	84
$\Delta = 5$	BALL	0.023	**500.422**	8.310	76	16
$\Delta = 7$	DIAG	**0.002**	502.331	**0.458**	24	80
$\Delta = 7$	BALL	0.004	**500.422**	1.046	76	20
$\Delta = 10$	DIAG	**0.000**	502.331	**0.001**	24	60
$\Delta = 10$	BALL	0.000	**500.422**	0.002	76	40

Table 1: Comparison of criteria BEC and BIC for choosing between two models DIAG and BALL. Column $\overline{err}$ gives the error rate evaluated on a test sample of size 50,000. Means are computed over 50 replications.

other side, BIC criterion selects always the same model without taking into account the separation between the classes.

3.2.2 Real data sets

The first real dataset considered is the Pima Indian Diabete database[1]. It is described for instance in Ripley (1996), pp.14-15. It concerns a population of $n = 768$ women described by $d = 8$ variables and it is a two-class problem ($K = 2$) where class 2 is interpreted as "tested positive for diabete". For this dataset, two experimented where performed. The first experiment was achieved in the whole space of description, the second experiment was achieved on the plane generated by the first two axes of a Principal Component Analysis.(Actually dimension reduction is often relevant with Spherical Gaussian Mixture Discrimnant Analysis.)

Tables **??** and **??**) show that BEC selects a satisfactory number of balls. The behavior of BEC is here close to the cross-validated error rate criterion. On the contrary, as it happens often on real datasets, BIC is overestimating the number of balls that are needed to reach a small error rate.

The second example Contraceptive Method Choice. This data set (Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. 2000) is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics. The sample size is $n = 1473$ described with $d = 9$ variables and there $K = 3$ classes to be classified.

The results for this dataset are reported in Table **??**. The error rates are here rather poor and not very different. Probably the Spherical Gaussian Mixture Discriminant Analysis is not well adapted to this problem. But the interesting point is the difference of behaviour of BEC and BIC and this data set. BEC prefers parsimonious solution and its behaviour is similar to the behaviour of the cross-validated error rate. On the contrary BIC prefers the most complex model.

[1]It is available at `http://www.maths.lth.se/help/R/.R/library/mlbench/html/PimaIndiansDiabetes.html`

separation	model	$\overline{err}$	BIC	BEC	BIC choice(%)	BEC choice(%)
$\Delta = 1.5$	1 components	0.272	**501.745**	68.786	100	0
$\Delta = 1.5$	2 components	**0.259**	522.658	59.120	0	30
$\Delta = 1.5$	3 components	0.264	554.613	**57.390**	0	70
$\Delta = 3.5$	1 components	0.076	**501.745**	25.547	100	0
$\Delta = 3.5$	2 components	**0.064**	522.658	17.859	0	38
$\Delta = 3.5$	3 components	0.065	554.613	**17.124**	0	62
$\Delta = 5$	1 components	0.024	**501.745**	7.726	100	0
$\Delta = 5$	2 components	**0.016**	522.658	**4.269**	0	58
$\Delta = 5$	3 components	0.017	554.613	4.426	0	42
$\Delta = 7.5$	1 components	0.004	**501.745**	1.162	100	8
$\Delta = 7.5$	2 components	0.002	522.658	**0.510**	0	58
$\Delta = 7.55$	3 components	**0.002**	554.613	0.938	0	34
$\Delta = 10$	1 components	0.000	**501.745**	**0.062**	100	58
$\Delta = 10$	2 components	**0.000**	522.658	0.086	0	28
$\Delta = 10$	3 components	0.000	554.613	0.496	0	14

Table 2: Comparison of criteria BEC and BIC for choosing the number of component in the Spherical Gaussian mixture model. Column $\overline{err}$ gives the error rate evaluated on a test sample of size 50,000. Means are computed over 50 replications.

```
R1          R2      3-CV          BIC         BEC
1           1      0.2622         9119       719.3
1           2      0.2664         9125         783
1           3      0.2693         9140       946.1
2           1      0.2565         8855       513.1
2           2      0.2635         8861       662.2
2           3      0.2758         8876       761.3
3           1      0.2646         8779       523.2
3           2      0.2727         8784       669.6
3           3      0.2768         8800       684.5
4           1      0.2703         8770       655.2
4           2      0.2701         8776       665.5
4           3      0.2651         8778       663.5
Selected number of balls (3-fold CV)    : 2  1
Selected number of balls (BIC criterion): 4  1
Selected number of balls (BEC criterion): 2  1
```

Table 3: Results on Pima Indians Diabetes dataset.

R1	R2	3-CV	BIC	BEC
1	1	0.2409	9119	719.3
1	2	0.2427	9125	783
1	4	0.2424	9108	928.4
1	5	0.2464	9173	1056
3	1	0.2422	8779	523.2
3	2	0.2375	8784	669.6
3	4	0.2391	8768	608.8
3	5	0.2378	8833	712
5	1	0.2453	8796	678.1
5	2	0.2461	8798	677.1
5	4	0.2414	8784	638.3
5	5	0.2406	8851	701.9
7	1	0.2461	8762	616.4
7	2	0.2479	8768	638.4
7	4	0.2451	8751	584.3
7	5	0.2437	8817	610.1

Selected number of balls (3-fold CV) : 3 2
Selected number of balls (BIC criterion): 7 4
Selected number of balls (BEC criterion): 3 1

Table 4: Results on Pima Indians Diabetes data set with reduction in 2D by PCA.

4 Discussion

In this paper, we highlighted how it could be useful to take into account the model purpose to select a relevant and useful model. This point of view can lead to define different selection criteria than the classical BIC criterion. It has been illustrated in two situations: modelling in a clustering purpose and modelling in a supervised classification purpose. For this particular context, we have proposed a promising criterion, the so-called BEC criterion, which takes into account the classification task when selecting a model. It can be a efficient alternative to the cross-validated error rate when the collection of models in competition is large.

Now, it can be noticed that we do not considered the modelling purpose when estimating the model parameters. In both situations, we simply considered the ml estimator. Taking into account of the modelling purpose in the estimation process could be regarded as an interesting point of view. As a matter of fact, we do not think that this point of view is fruitful and, moreover, we think it can jeopardize the statistical analysis. For instance, in the cluster analysis context of Section 2, it could be thought of as more natural to compute the parameter value maximizing the complete loglikelihood $\log \mathbf{p}(\mathbf{x}, \mathbf{z}|\theta)$ rather than the observed loglikelihood $\log \mathbf{p}(\mathbf{x}|\theta)$. But as proved in Bryant and Williamson (1978), this strategy leads to asymptotically biased estimates of the mixture parameters. In the same manner, in the supervised classification context of Section 3, considering the parameter value maximizing directly the conditional likelihood $\log \mathbf{p}(\mathbf{z}|\mathbf{x}, \theta)$ could be regarded as an alternative to the classical ml estimation. But this would lead to a difficult optimization problem and would provide unstable estimate values. Finally, we do not recommend taking into account the modelling purpose when estimating the model parameters because it could lead to cumbersome algorithms or provoke undesirable biases in the estimation. On the contrary, we think that taking into account the model purpose when assessing a model could lead to choose reliable and stable models especially in unsupervised or supervised classification context.

R1	R2	R3	2-CV	BIC	BEC
1	1	1	0.5234	2.031e+004	2587
1	1	2	0.5249	2.017e+004	3050
1	1	3	0.5424	2.001e+004	3188
1	2	1	0.5223	2.025e+004	2818
1	2	2	0.5208	2.012e+004	3153
1	2	3	0.5431	1.995e+004	3197
1	3	1	0.5241	2e+004	3306
1	3	2	0.5256	1.986e+004	3462
1	3	3	0.5427	1.97e+004	3500
2	1	1	0.5065	1.999e+004	2738
2	1	2	0.5223	1.986e+004	3165
2	1	3	0.5183	1.969e+004	3301
2	2	1	0.509	1.993e+004	3079
2	2	2	0.5165	1.98e+004	3406
2	2	3	0.5229	1.963e+004	3446
2	3	1	0.5134	1.968e+004	3286
2	3	2	0.521	1.954e+004	3345
2	3	3	0.5259	1.938e+004	3293
3	1	1	0.5248	1.986e+004	3014
3	1	2	0.5261	1.973e+004	3339
3	1	3	0.5248	1.957e+004	3378
3	2	1	0.5229	1.981e+004	3144
3	2	2	0.5191	1.967e+004	3458
3	2	3	0.5229	1.951e+004	3564
3	3	1	0.5276	1.955e+004	3342
3	3	2	0.5248	1.942e+004	3321
3	3	3	0.529	1.925e+004	3464
4	1	1	0.5412	1.972e+004	3327
4	1	2	0.5286	1.959e+004	3375
4	1	3	0.528	1.942e+004	3348
4	2	1	0.5313	1.966e+004	3461
4	2	2	0.521	1.953e+004	3490
4	2	3	0.5183	1.936e+004	3606
4	3	1	0.5359	1.941e+004	3336
4	3	2	0.5221	1.927e+004	3467
4	3	3	0.5195	1.911e+004	3360

Selected number of balls (2-fold CV) : 2 1 1
Selected number of balls (BIC criterion): 4 3 3
Selected number of balls (BEC criterion): 1 1 1

Table 5: Results on Contraceptive data set.

References

Aitkin, M. (2001). Likelihood and Bayesian analysis of mixtures. *Statistical Modelling*, **1**, 287-304.

Akaike, H. (1974). A New Look at Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, **91**, 1743-48.

Bernardo, J. M. and Smith A. F. M. (1994). *Bayesian Theory*. Wiley.

Biernacki, C., Celeux., G. and Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Trans. on PAMI*, **22**, 719-725.

Bouchard, G. and Celeux, G. (2003). Supervised Classification with Spherical Gaussian Mixtures. *in CLADAG 2003*, University of Bologna, pp. 75-78.

Bryant, P. and Williamson, J. (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika*, **65**, 273-281.

Figueiredo, M. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transaction on PAMI*, **24**, 381-396.

Fraley, C. and Raftery, A. E. (1998). How Many Clusters ? Answers via Model-based Cluster Analysis. *The Computer Journal*, **41**, 578–588.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611-631.

Friedman, J. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, **84**, 165-175.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society* series B, **58**, 158-176.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Keribin, C. (2000). Consistent estimation of the order of mixture. *Sankhya*, **62**, 49-66.

Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning. *Machine Learning*, **40**, 203-228.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

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

Murata, N., Yoshizawa, S. and Amari, S. (1994) Network Information Criterion-Determining the Number of Hidden Units for Artificial neural network models. *IEEE Transactions on Neural Networks*, **5**, 865-872.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In *Sociological Methodology 1995*, (ed. Peter V. Marsden), Oxford, U.K.: Blackwells, pp. 111-196.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Roeder, K. (1990). Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in Galaxies. *Journal of the American Statistical Association*, **85**, 617-624.

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. *The Annals of Statistics*, **6**, 461-464.