Model selection in supervised classification

Guillaume Bouchard, Gilles Celeux

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Model selection in supervised classification

Guillaume Bouchard — Gilles Celeux

N° 5391

November 2004

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5391--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA FUTURS

Model selection in supervised classification

Guillaume Bouchard , Gilles Celeux

Thème COG — Systèmes cognitifs
Projets Select

Rapport de recherche n° 5391 — November 2004 — 22 pages

Abstract: This article is concerned with the selection of a generative model for supervised classification. Classical model selection criteria are assessing the fit of a model rather than its ability to produce a low classification error rate. A new criterion, the so called Bayesian Entropy Criterion (BEC) is proposed. This criterion is taking into account the decisional purpose of a model by minimizing the integrated classification entropy. It provides an interesting alternative to the cross validated error rate which is highly time consuming. The asymptotic behavior of BEC criterion is presented. Numerical experiments on both simulated and real data sets show that BEC is performing better than BIC criterion to select a model minimizing the classification error rate and is providing analogous performances than the cross validated error rate.

Key-words: Generative Classification, Integrated likelihood, Integrated conditional likelihood, Classification entropy, Cross validated error rate, AIC and BIC criteria.

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,
4, rue Jacques Monod, 91893 ORSAY Cedex (France)
Téléphone : +33 1 72 92 59 00 — Télécopie : +33 1 72 92 59 ??

Sélection de modèles en classification supervisée

Résumé : Le choix d'un modèle probabiliste pour l'analyse discriminante est l'objet de cet article. Les critères classiques de sélection de modèle privilégient l'adéquation du modèle à la distribution jointe des variables explicatives et de la variable de groupe plutôt que la minimisation du taux d'erreur du classifieur associé. Nous proposons un nouveau critère, le *Bayesian Entropy Criterion* (BEC), qui permet de sélectionner un classifieur prenant en compte l'objectif décisonnel par la minimisation de l'entropie intégrée de classification. Il représente une alternative intéressante à la validation croisée qui est très coûteuse. Les propriétés asymptotiques du critère BEC sont présentées et des expériences numériques sur des données simulées et des données réelles montrent que ce critère a un comportement meilleur que BIC pour choisir le modèle minimisant l'erreur de classification et analogue à celui de la validation croisée.

Mots-clés : Modèles d'analyse discriminante, vraisemblance intégrée, vraisemblance intégrée conditionnelle, entropie de classification, validation croisée, critères AIC et BIC.

1 Introduction

In statistical pattern recognition, the generative classification approach consists of modelling each class to be recognized with a probabilistic model. Many parametric or non parametric classification methods have been conceived or can be presented under this approach (see [23]). For many practical classification problems, it can be quite advantageous to consider many competing generative models in order to design a classification rule minimizing the future error rate. Examples of generative classification methods where a family of models is considered and for which the most efficient model is to be selected are [17] and [3]. Thus, in this perspective, an important task is to select a reliable model among a collection of generative models. A natural way to deal with this model selection problem is to assess the future performance of the model with its cross validated error rate. However, this type of criterion is painfully slow and alternative model selection criteria are desirable. But classical model selection criteria are not focusing on the classification task and can have a disappointing behavior. The aim of the present paper is to propose a new model selection criterion specifically suited to the supervised classification context. Before presenting this criterion, the points of view on which classical model selection criteria are based are recalled in this 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 [2] and [32]. 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 complex 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

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

where $\boldsymbol{p}(\boldsymbol{x}) = \prod_{1=1}^{n} \boldsymbol{p}(x_i)$ denotes the true distribution of the data $\boldsymbol{x} = (x_1, \ldots, x_n)$ (for simplicity, the $x_i's$ are supposed to be independent and identically distributed (iid)), $\boldsymbol{p}(\boldsymbol{x}|\theta_m) = \prod_{i=1}^{n} \boldsymbol{p}(x_i|\theta_m)$ is the distribution under the model m parameterized with θ_m, and $\hat{\theta}_m$ is the maximum likelihood 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 the parameters are chosen 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 $nD(X) - E(d(\boldsymbol{x}))$ where

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

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 proposed to estimate this difference with $2\nu_m$ where ν_m is the number of free parameters of the model m [2, 29]. This leads to the so called AIC criterion.

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

Relaxing this unrealistic assumption leads to alternative criteria such as the Network Information Criterion [25]. (Details can be found in [29], 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 [20]. This integrated likelihood is

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

$\pi(\theta_m)$ 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 [32]. It is

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

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 $\hat{\theta}_m$ [28]. 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 (see [4], chapter 6).

In recognition of the model selection uncertainty, there are more and more authors to think that it is unwise to separate the model selection process from the specific goal of inference. For instance, choosing a reliable number of components in a mixture model can highly depend of the modeller purpose. And, if BIC is working well at a practical level when the mixture model is considered in a density estimation purpose [30, 11], when the mixture model is considered in a cluster analysis perspective, some other criteria taking the clustering purpose into account as ICL (see [5] or [24], chapter 6) can appear to be more reliable. In the present paper, we are concerned with the problem of choosing a probabilistic model in a supervised classification context. Criteria as AIC and BIC are not taking the classification purpose into account and have fixed penalties. In this context, as said above, there exists however a reference criterion, the cross validated classification error rate, which is directly providing an estimate of the future error rate of the models in competition. But, this criterion is highly CPU time consuming. In this paper, we propose a penalized likelihood criterion which is taking into account the classification task. It is a BIC family criterion but it is approximating the integrated conditional likelihood of the generative classification models in competition instead of their integrated joint likelihood. It can be regarded as an efficient alternative to the cross validated classification error rate criterion.

The paper is organised as follows. Section 2 is devoted to the presentation of the model selection problem for generative models in supervised classification. Our criterion, the so called BEC criterion, is presented in Section 3. It asymptotic behavior is discussed in Section 4. Numerical experiments on both simulated and real data are presented in Section 5 to illustrate the practical behavior of BEC criterion. A short discussion section ends the paper.

2 Generative classifiers and model selection

Supervised classification is about guessing the unknown class, denoted by Y and taking value in $\{1,\ldots,K\}$ of an observation $\boldsymbol{X}$ taking value in $\mathbb{R}^d$. For that purpose, a decision function, called a classifier, $\delta(\boldsymbol{X}) : \mathbb{R}^d \to \{1,\ldots,K\}$ is designed from a learning sample $(\boldsymbol{x}_i, y_i), i = 1,\ldots,n$ (for simplicity, the $\boldsymbol{x}_i's$ are supposed to be iid). A classical approach to design a classifier is to represent the class conditional densities with a parametric density $\boldsymbol{p}(\boldsymbol{X}|Y = k, \theta_m)$ for $k = 1,\ldots,K$, m denoting the model with parameters $\theta_m \in \Theta_m$. In this work, we consider that Θ_m is a finite dimensional parameter space. Then an observation $\boldsymbol{X}$ is assigned to the class k maximizing the conditional probability of a class $\boldsymbol{p}(Y = k|\boldsymbol{X}, \theta_m)$. Using the Bayes rule, it leads to the classifier

$$\delta(\boldsymbol{X}) = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}} \boldsymbol{p}(\boldsymbol{X}, Y = k|\hat{\theta}_m), \tag{4}$$

$\hat{\theta}_m$ being a given estimator of the parameter θ_m based on the learning data. This approach is known as the generative classification approach [19, 31]. The maximum likelihood (ml) estimator based on the class-conditional distributions is a popular estimation procedure. In ml estimation, the joint likelihood of the input $\boldsymbol{x} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ and output $\boldsymbol{y} = (y_1,\ldots,y_n)$ is maximized:

$$\hat{\theta} = \arg\max_{\theta} \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\theta_m). \tag{5}$$

In supervised classification, it is often relevant to design different classification rules from a large collection of models and to choose the model leading, with the available learning data set, to the minimum classification error rate in the future. In the generative context, several proposed methods require the selection of a model in a collection of models. Recent examples that will be considered in Section 5 are multivariate Gaussian distributions with various variance decompositions [3] and Mixture Discriminant Analysis (MDA) [17]. For instance in the MDA approach where each class-conditional density is a mixture of Gaussian distributions, the number of mixture components *per* class are sensitive tuning parameters. They can either be supplied by the user [17], but it is clearly a sub-optimal solution, or they can be chosen to minimize the v-fold cross-validated error rate, as done in [12] or [3] for other tuning parameters. Despite the fact the choice of v can be sensitive, it can be regarded as a satisfactory solution, but it is highly CPU time consuming. Thus choosing such tuning parameter with a penalized loglikelihood criterion, as BIC, can be thought of as desirable in many situations. In such a classification context, denoting $\boldsymbol{y} = (y_1,\ldots,y_n)$ the classification of the learning sample, BIC takes the form

$$\mathrm{BIC}(m) = \log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\hat{\theta}_m) - \frac{\nu_m}{2}\log(n), \tag{6}$$

where ν_m is the dimension of θ_m. But, BIC measures the fit of the model m to the data $(\boldsymbol{x}, \boldsymbol{y})$ rather than its ability to produce a reliable classifier. Thus, in many situations, BIC can be disappointing to choose a model producing a low classification error rate. In order to answer this limitation, we propose a penalized likelihood criterion taking into account the classification task when evaluating the performance of a model.

3 The Bayesian Entropy Criterion

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

$$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, m) = \int \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, \theta_m)\pi(\theta_m|\boldsymbol{x})d\theta_m, \tag{7}$$

where

$$\pi(\theta_m|\boldsymbol{x}) \propto \pi(\theta_m)\boldsymbol{p}(\boldsymbol{x}|\theta_m)$$

is the posterior distribution of θ_m knowing $\boldsymbol{x}$. As for the integrated likelihood, this integral is generally difficult to calculate and has to be approximated. The approximation of $\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, m)$, we now present, leads to the so-called Bayesian Entropy Criterion (BEC). We have

$$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, m) = \frac{\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|m)}{\boldsymbol{p}(\boldsymbol{x}|m)} \tag{8}$$

with

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

and

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

The criterion that we now define is obtained through Laplace approximations applied on the two integrals (9) and (10).

It is valid to approximate logarithms of integrals (9) and (10) according to a line described in [28] to derive the BIC criterion. Denoting $\tilde{\theta}_m = \arg\max_\theta \boldsymbol{p}(\boldsymbol{x}|\theta_m)$ and assuming that the prior distribution $\pi(\theta_m)$ of θ_m may be approximated by a normal distribution with mean θ_m^0 and variance V_m^0, we can write [29]

$$\log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|m) \approx \log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\hat{\theta}_m) - \frac{\nu_m}{2}\log n - \frac{1}{2}\log|\hat{J}_J| - \frac{1}{2}(\hat{\theta}_m - \theta_m^0)^t V_0^{-1}(\hat{\theta}_m - \theta_m^0) \tag{11}$$

and

$$\log \boldsymbol{p}(\boldsymbol{x}|m) \approx \log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m) - \frac{\nu_m}{2}\log n - \frac{1}{2}\log|\tilde{J}_M| - \frac{1}{2}(\tilde{\theta}_m - \theta_m^0)^t V_0^{-1}(\tilde{\theta}_m - \theta_m^0), \tag{12}$$

where ν_m is the dimension of θ_m, $\hat{J}_J = J_J(\hat{\theta}_m)$ and $\tilde{J}_M = J_M(\tilde{\theta}_m)$ are the normalized Hessian of the negative joint and marginal log-likelihoods at $\hat{\theta}_m$ and $\tilde{\theta}_m$:

$$J_J(\theta_m) = -\frac{1}{n}\frac{\partial^2}{\partial\theta_m \partial\theta_m{}^T}\log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\theta_m), \quad J_M(\theta_m) = -\frac{1}{n}\frac{\partial^2}{\partial\theta_m \partial\theta_m{}^T}\log \boldsymbol{p}(\boldsymbol{x}|\theta_m).$$

Taking the difference of the two expressions (11) and (12) leads to

$$\begin{aligned}\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},m) \approx & \quad \log \boldsymbol{p}(\boldsymbol{x},\boldsymbol{y}|\hat{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m) \\ - & \quad \frac{1}{2}\log|\hat{J}_J \tilde{J}_M^{-1}| - \frac{1}{2}(\hat{\theta}_m - \tilde{\theta}_m)^t V_0^{-1}(\hat{\theta} + \tilde{\theta} - 2\theta_m^0)\end{aligned}$$

$$\begin{aligned}\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},m) \approx & \quad \log \boldsymbol{p}(\boldsymbol{x},\boldsymbol{y}|\hat{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m) \\ - & \quad \frac{1}{2}\log|I_d + \hat{J}_C \tilde{J}_M^{-1}| - \frac{1}{2}(\hat{\theta}_m - \tilde{\theta}_m)^t V_0^{-1}(\hat{\theta} + \tilde{\theta} - 2\theta_m^0).\end{aligned}$$

where $\hat{J}_C = J_C(\hat{\theta}_m)$ with

$$J_C(\theta_m) = -\frac{1}{n}\frac{\partial^2}{\partial\theta_m \partial{\theta_m}^T}\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},\theta_m).$$

Removing the terms of order O(1) gives

$$\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},m) \approx \log \boldsymbol{p}(\boldsymbol{x},\boldsymbol{y}|\hat{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m). \tag{13}$$

Thus the approximation of $\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},m)$ that we propose is

$$\text{BEC} = \log \boldsymbol{p}(\boldsymbol{x},\boldsymbol{y}|\hat{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m). \tag{14}$$

Some comments are in order.

- The conditional integrated likelihood can be interpreted as the Bayesian entropy of the classification derived from model m. This is the reason why we called this criterion *Bayesian Entropy Criterion* (BEC).
- Equation (13) is the approximation on which BEC is based. It is valid up to a constant term. It 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 $\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},m)$. However, the terms depending on n will dominate given enough data. 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 $(\boldsymbol{x},\boldsymbol{y})$ belongs to one of the models in competition. But, this is seldom the case.
- The BIC-like approximation of $\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})$

  $$\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}) \approx \log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},\theta_m^*) - \frac{\nu_m}{2}\log n, \tag{15}$$

  where

  $$\theta_m^* = \arg\max_{\theta_m} \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x},\theta_m),$$

  is not valid since, for any generative classification model, the posterior distribution $\pi(\theta_m|\boldsymbol{x})$ in (7) depends on n and cannot be neglected. It can be noticed that in a discriminative approach of supervised classification for which $\boldsymbol{x}$ is assumed to be not depending on θ this BIC-like approximation would be valid.

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

$$\boldsymbol{p}(\boldsymbol{x}_i|\theta_m) = \sum_{k=1}^{K} \boldsymbol{p}(y_i = k)\boldsymbol{p}(\boldsymbol{x}_i|y_i = k, , \theta_m), \tag{16}$$

  $\tilde{\theta}_m$ is the ml estimate of a finite mixture distribution. It can be derived from the EM algorithm [24]. Fortunately, in the present context, the well documented drawbacks of the EM algorithm (see for instance [24]) which are high dependence on initial position and slow convergence are easily avoided. Actually, the EM algorithm can be initialized in a quite natural way with $\hat{\theta}_m$. Thus the calculation of $\tilde{\theta}_m$ involves no difficulty. Despite the need to use the EM algorithm to estimate this parameter, it would be estimated in a stable and reliable way. It can also be noted that the mixing proportions $\boldsymbol{p}(y_i = k), k = 1, \ldots, K$ are not depending on the parameter θ_m of the generative model m. When the learning data set has been obtained through the diagnosis sampling scheme, the learning data set is the concatenation of K subsamples whose sizes are not random variables. Thus, the proportions in the mixture distribution (16) are fixed: $\boldsymbol{p}(y_i = k) = n_k/n$ where $n_k = \text{card}\{i \text{ such that } y_i = k\}$ for $k = 1, \ldots, K$. When the learning data set has been obtained through the mixture sampling scheme, they have to be estimated with θ_m to identify the mixture distribution (16) with the EM algorithm. But, again in a natural way, the initial proportions in EM can be chosen to be $\boldsymbol{p}(y_i = k) = n_k/n$ for $k = 1, \ldots, K$.

- In order to regard BEC as a penalized likelihood criterion, we can write

$$\begin{aligned} \text{BEC} &= \log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\hat{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\hat{\theta}_m) + \log \boldsymbol{p}(\boldsymbol{x}|\hat{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m) \\ \text{BEC} &= \log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, \hat{\theta}_m) - \left(\log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\hat{\theta}_m)\right). \end{aligned} \tag{17}$$

  The quantity $\text{pen} = \log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}|\hat{\theta}_m)$ is positive since $\tilde{\theta}$ is maximizing the marginal likelihood $\boldsymbol{p}(\boldsymbol{x}|\theta_m)$. It can be interpreted as a penalty applied on the conditional log-likelihood. This penalty is always non negative and is minimum when $\hat{\theta} = \tilde{\theta}$. It is implicitly dependent of the model complexity as illustrated in the toy example depicted in Figure 1. This is a two class problem. In the learning set, there are five points for each class, represented with a cross and a dot in the top graphics of Figure 1. Two Gaussian models have been considered for this data set: a "simple" model with spherical variance matrices and a "complex" model with free variance matrices. BEC is choosing the simplest one since the penalty term 'pen' for the complex model is dominating the increase it provides on the conditional log likelihood $\log(\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, \hat{\theta})$.

- Finally, it can be remarked from (17) that BEC is always smaller than 0 since $\log \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, m, \hat{\theta}) \leq 0$.

4 Asymptotic behavior of BEC criterion

Some theoretical properties of BEC are now highlighted. The behavior of BEC as the size of the learning set tends to infinity and when the sample distribution belongs to at least one of the model

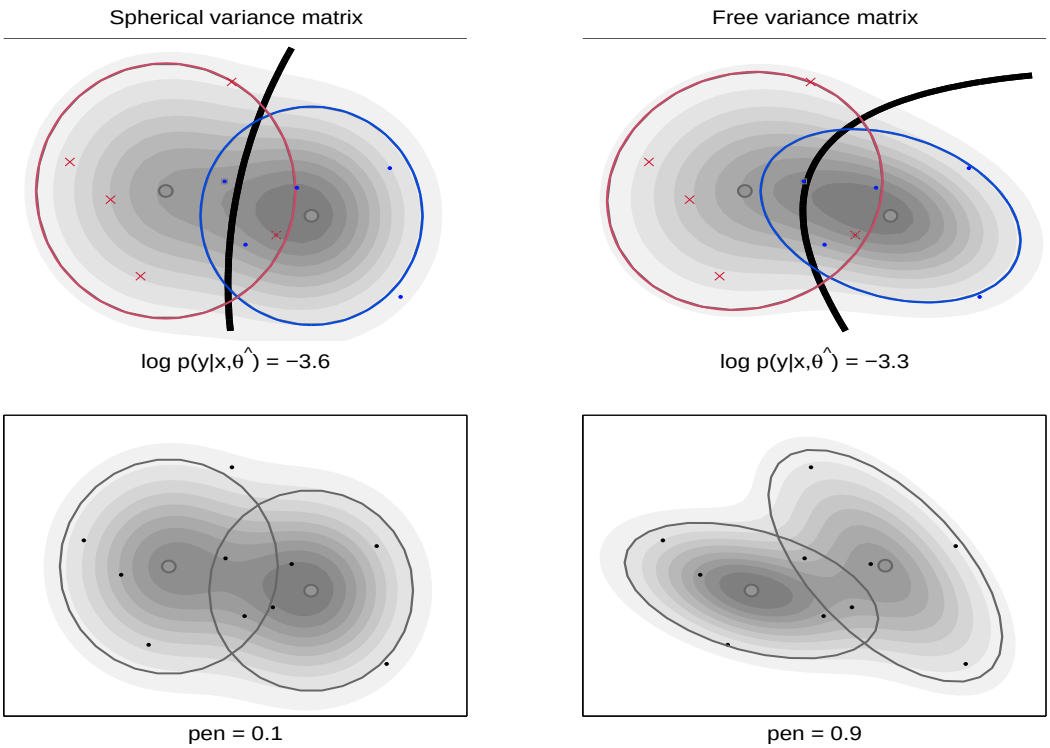

Figure 1: Illustration of the BEC model choice. A spherical Gaussian model (left graphics) is compared to a free variance matrices Gaussian model (right graphics) for a two class problem. At the top, the ml estimation of the joint distribution are shown, as well as the classification boundary of the corresponding generative classifier. In those graphics, the observations from a class are indicated with a cross and the observations from the other class are indicated with a dot. At the bottom, the ml estimation of the marginal distributions obtained with EM are shown. The density value of the distribution is proportional to the grey level. And 'pen' is the value of the penalty term isolated in (17).

in competition is studied. As previously written, assuming that the sample distribution belongs to one of the candidate models can be seen as an unrealistic assumption in most situations. However, it is a minimal requirement for a model selection criterion to behave as expected in such a situation. The BEC criterion has been conceived to find, in a collection of models, the model minimizing the classification error rate. If there is one and only one model m^* from which the sample distribution belongs, then BEC is expected to select this model m^*, since this is the unique model which attains

asymptotically the Bayes classification error rate. The following proposition proves actually that BEC chooses the unique true model if it exists.

Proposition 1 *If the sample joint distribution belongs to one and only one model m^* in the finite family of candidate models $\{m_1, \ldots, m_M\}$, and under standard regularity conditions on the family of candidate models, the BEC criterion would selects m^* with probability one as the sample size n of the learning set tends to infinity.*

PROOF. If the sample distribution $\boldsymbol{p}$ belongs to the model m^*, there exists a parameter value $\theta^0_{m^*}$ satisfying $\boldsymbol{p}(\boldsymbol{X}, Y) = \boldsymbol{p}(\boldsymbol{X}, Y|\theta^0_{m^*})$. The normalized criterion $\frac{1}{n}\text{BEC}(m^*)$ is the difference of the quantities $\frac{1}{n}\log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\hat{\theta}_{m^*})$ and $\frac{1}{n}\log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_{m^*})$. By the law of large numbers, $\hat{\theta}_{m^*} \to \theta^0_{m^*}$ and $\tilde{\theta}_{m^*} \to \theta^0_{m^*}$ as $n \to \infty$. Then, the regularity conditions implicitly assumed on the candidate models, ensure that the two quantities $\frac{1}{n}\log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\hat{\theta}_{m^*})$ and $\frac{1}{n}\log \boldsymbol{p}(\boldsymbol{x}|\tilde{\theta}_{m^*})$ tend almost surely to $E\left[\log \boldsymbol{p}(\boldsymbol{X}, Y|\theta^0_{m^*}\right]$ and $E\left[\log \boldsymbol{p}(\boldsymbol{X}|\theta^0_{m^*})\right]$, respectively. Hence,

$$\frac{1}{n}\text{BEC}(m^*) \to E\left[\log \boldsymbol{p}(Y|\boldsymbol{X}, \theta^0_{m^*})\right] \quad \text{as} \quad n \to \infty. \tag{18}$$

On the other hand, for any of the other models $m \neq m^*$ that does not contain the sample distribution, $\hat{\theta}_m \to \theta^1_m$ and $\tilde{\theta}_m \to \theta^2_m$ with $\theta^1_m \neq \theta^2_m$, so that for any model $m \neq m^*$

$$\begin{aligned} \frac{1}{n}\text{BEC}(m) &\to E\left[\log \boldsymbol{p}(\boldsymbol{X}, Y|\theta^1_m)\right] - E\left[\log \boldsymbol{p}(\boldsymbol{X}|\theta^2_m)\right] &(19)\\ &= \underbrace{E\left[\log \boldsymbol{p}(Y|\boldsymbol{X}; \theta^1_m)\right]}_{<E\left[\log \boldsymbol{p}(Y|\boldsymbol{X}; \theta^0_{m^*})\right]} - \underbrace{E\left[\log \boldsymbol{p}(\boldsymbol{X}|\theta^2_m) - \log \boldsymbol{p}(\boldsymbol{X}|\theta^1_m)\right]}_{>0}. &(20)\end{aligned}$$

The first inequality comes from the fact that the expected log-probability is maximized for the true parameter value $\theta^0_{m^*}$ (or equivalently the Kullback-Leibler divergence is minimum at $\theta^0_{m^*}$). The second inequality comes from the fact that θ^2_m maximizes the expected marginal likelihood in the parameter space of model m. □

Proposition 1 does not apply when the models in competition are nested models. In such a situation, the true distribution can belong to several candidate models.

Proposition 2 *Assuming that the true distribution $\boldsymbol{p}(\boldsymbol{X}, Y)$ belongs to two nested models m and m', with ν and ν' parameters, for any $\varepsilon > 0$, we have for n large enough*

$$E(BEC(m)) - E(BEC(m')) < \varepsilon.$$

PROOF. It is assumed that $\nu' > \nu$. The likelihood ratio statistic computed at the ml estimators of two nested models is asymptotically a χ^2 distribution with $\delta_\nu = \nu' - \nu$ degrees of freedom. When computing the difference of BEC on the two nested models m and m', two likelihood ratio (LR) statistics appear, corresponding to the joint and the marginal ml estimators:

$$\begin{aligned} \text{BEC}(m) - \text{BEC}(m') &= \log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\hat{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}|\hat{\theta}_{m'}) - \left(\log \boldsymbol{p}(\boldsymbol{x}; \tilde{\theta}_m) - \log \boldsymbol{p}(\boldsymbol{x}; \tilde{\theta}_{m'})\right)\\ &= \log\{\text{LR of } m \text{ vs. } m' \text{ for } \boldsymbol{p}(X, Y)\} - \log\{\text{LR of } m \text{ vs. } m' \text{ for } \boldsymbol{p}(X)\}\\ &\xrightarrow{\mathcal{D}} \frac{1}{2}\chi^{'2}_{\delta_\nu} - \frac{1}{2}\chi^2_{\delta_\nu} &(21)\end{aligned}$$

where $\chi'^2_{\delta_\nu}$ and $\chi^2_{\delta_\nu}$ are two dependent variables following χ^2 distributions with δ_ν degrees of freedom. The last approximation is valid for n sufficiently large and is $O_p(1)$. This proves that the random variable $\mathrm{BEC}(m) - \mathrm{BEC}(m')$ has asymptotically a zero mean.

□

It means that BEC criterion equally weights the two nested models. Thus, even asymptotically, we might find $\mathrm{BEC}(m') < BEC(m)$ and consequently choose the most complex model. In practical situations, this fact is rarely to occur since it needs two conditions, large sample size and a nearly exact collection of models. But, when nested models are in competition and the sample size can be regarded as large, it is of interest to display the BEC values in function of the model number of parameters. If a plateau appears on such a graph, it means that the collection of models is fitting well the sample distribution and we recommend to choose the simplest model on this plateau. Such a possible behavior of BEC is illustrated in Section 5.1, Figure 2.

Yet, the theory is still asymptotic and does not give any insight about the approximation qualities of BEC. The next section gives an answer to this question from numerical experiments on both simulated and real data sets.

5 Numerical experiments

In this section, some case studies for analyzing the practical ability of BEC to select a sensible classification model are reported and BEC is compared with criteria as the cross-validated error rate, AIC and BIC. First Monte Carlo numerical experiments are proposed in simple situations to highlight noticeable features of BEC behavior. Then the problem of selecting a reliable model in the context of *Eigenvalue Decomposition Discriminant Analysis* (EDDA) [3] and *Mixture Discriminant Analysis* (MDA) [17] is considered using Monte Carlo experiments on benchmark data sets. Finally, a study concerning a pattern recognition problem in computer vision is presented.

5.1 Monte Carlo numerical experiments

For the first experiment which is merely illustrative two simple models are compared. Five hundred samples of $n = 120$ observations in $\boldsymbol{R}^2$ from two classes with equal prior probabilities have been generated with the following class conditional densities:

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

and

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

In this experiment, the two models in competition are Gaussian class-conditional distributions with diagonal covariance matrices (DIAG) and with variance matrices proportional to the identity matrix (SPHE). The performances of criteria BEC and BIC are compared in Table 1. In this table, column $\overline{err}$ gives the error rate obtained with an independent test sample of size 50,000. It appears that most

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

Table 1: Comparison of criteria BEC and BIC for choosing between two models DIAG and SPHE. Column $\overline{err}$ gives the mean error rate evaluated on a test sample of size 50,000. Reported mean values are computed over 500 replications.

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.

The second Monte Carlo numerical experiment is aiming to illustrate the possibility for BEC to produce a plateau in function of the complexity of nested models. A two class problem has been considered in $\boldsymbol{R}^2$. Two data sets of size $n = 300$ were considered. For the first data set, each class-conditional distribution was a Gaussian distribution with variance matrix $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ and class-means $(0,0)^t$ and $(1.5,0)^t$. For the second data set, each class-conditional distribution was a three-component Gaussian mixture with identity variance matrix and component means $(0,0)^t, (3,0)^t$ and $(0,3)^t$ for class 1 and $(1.5,0)^t, (-1.5,0)^t$ and $(1.5,-3)^t$ for class 2. The generative models in competition for those two data sets were mixtures of Gaussian distributions with spherical variance matrices with the same volume, the number of mixture components varying from one to eight. Figure 2 provides the variations of BEC criterion regarding the number of spherical Gaussian components for both data sets. In this figure the BEC variations are given by the full line and the left scale provides the values of -BEC. Figure 2 provides also the variations of the error rate in dashed line, calculated from a test sample of size 50,000, and the right scale provides the classification error rate values. As expected, the right graphic of the figure shows a plateau from the true number of mixture components. Using the rule we recommend in Section 4 leads to choose the right number of mixture components (three) which provides the lowest classification error rate. For the first data set, no plateau is apparent and BEC chooses a four-component mixture. This is not the model which produces the lowest classification error rate, but in this case there is no sensitive differences between the classification error rates obtained with different number of mixture components (dashed line curve).

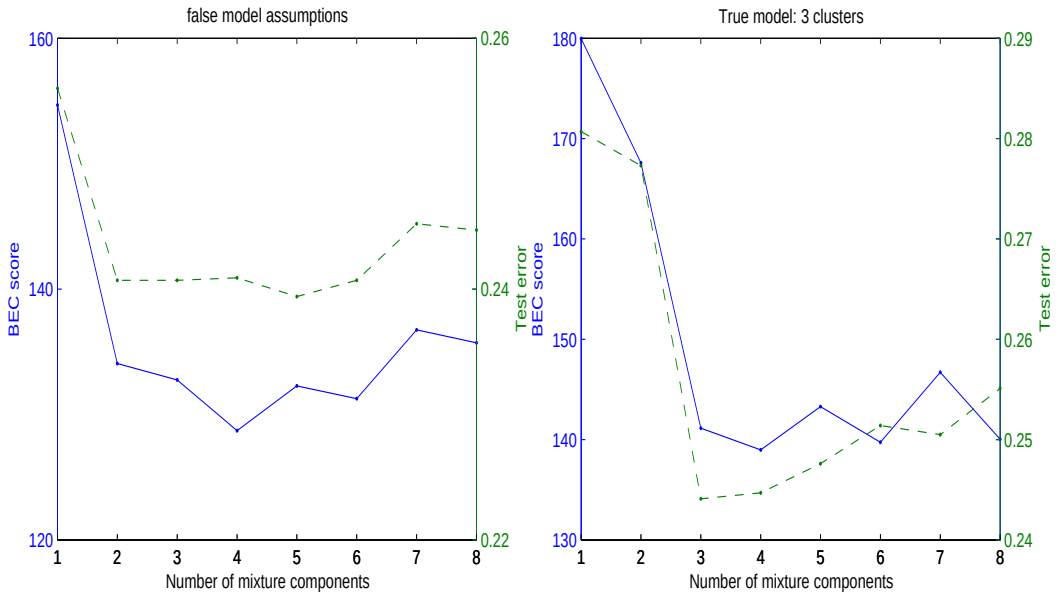

Figure 2: Illustration of the BEC behavior for nested model. In each graphic the full line gives the variations of -BEC whose values appears on the left scale and the dashed line gives the variations of the classification error rate for a test sample of size 50,000. The error rate are given on the right scale.

5.2 Choosing the variance matrix parametrization

Popular generative classifiers are Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Both methods assume Gaussian class-conditional densities with a common variance matrix for LDA and free variance matrices for QDA. Considering the eigenvalue decomposition of the class variance matrices lead to many alternative models [3]. Let $\Sigma_k = L_k D_k A_k D_k$ be the decomposition of the variance matrix of class k, where $L_k = |\Sigma_k|^{1/d}$ defines the volume of the distribution, D_k the matrix of eigenvectors of Σ_k defines its orientation, and A_k the diagonal matrix of normalized eigenvalues of Σ_k, defines its shape. Allowing some of those quantities to vary or not between classes leads to several models which can be of interest in a classification purpose. Moreover, assuming that Σ_k is a diagonal matrix or is proportional to the identity matrix lead to additional parsimonious models. In [3], 14 different parameterizations based on this eigenvalue decomposition has been considered and the model minimizing the cross validated error rate is selected. The corresponding method was called EDDA for *Eigenvalue Decomposition Discriminant Analysis*. In the present section, we examine the possibility to select one of the models in competition using AIC,

model	ν	BIC	AIC	BEC	CV3	test error
λI	10	0	0	0	0	0.293
$\lambda_k I$	13	0	0	0	0	0.289
λB	13	0	9	32	28	0.23
$\lambda_k B$	14	0	0	1	1	0.264
λB_k	16	0	0	0	0	0.287
$\lambda_k B_k$	17	93	0	0	0	0.276
$\lambda D^t A D$	19	0	23	38	36	**0.229**
$\lambda_k D^t A D$	20	0	0	0	1	0.261
$\lambda D_k^t A D_k$	25	0	68	25	34	0.23
$\lambda_k D_k^t A D_k$	26	0	0	3	0	0.258
$\lambda D_k^t A_k D_k$	28	0	0	0	0	0.291
$\lambda_k D_k^t A_k D_k$	29	7	0	1	0	0.274

Table 2: Comparison of the different model selection criteria on the *Australian credit* dataset. Two classes, 200 learning data, 490 test data. Ten continuous variables reduced in four dimensions. Each value represents the proportion of model choice among the 12 proposed models. In the first column, I stands for the identity matrix, B stands for a diagonal matrix. The presence of an index k indicates that the corresponding element is varying upon the classes.

BEC or BIC criteria instead of the cross validated error rate. Note that for simplicity the two cases with a common orientation and different shapes which require a specific algorithm are not been included in those experiments. Note also that the aim is not to assess the performance of EDDA methodology but to assess the ability of the considered model selection criteria to choose a reliable model from the models in competition.

For those numerical experiments, benchmark datasets from the UCI Machine Learning Database Repository available at `http://www.ics.uci.edu/~mlearn/` were used. Since reducing the dimension often improve the classification performances, the experiments were achieved on a space of dimension four generated by the $K-1$ canonical discriminant axes, eventually completed with the Principal Component Analysis (PCA) axes computed in the orthogonal space to the canonical discriminant sub-space. (For each considered data set, the original dimension d is given in Table 3.)

First, the behaviour of BEC is illustrated in some detail on the *Australian Credit Approval* dataset. This dataset contains discrete and continuous attributes, corresponding to various customer data. The index values of the four ordinal attributes were used as continuous variables, and the four binary variables were removed to avoid singularities in the estimation of the Gaussian distributions.

The following procedure has been applied 100 times: For each experiment, 200 learning data were randomly selected from the whole dataset of size 690 and used to learn the 12 models in competition and to calculate the values of criteria BIC, AIC, BEC, and the cross-validated error rate (CV3: it is the 3-fold cross validation procedure). The model optimizing each criterion was selected and its performance was assessed on the remaining test data set. Proportions of model choices are

Dataset	K	N	d	BIC	AIC	BEC	CV3	oracle
Abalone	3	4177	7	47.3	47.4	46.1	**45.9**	45.4
Bupa	2	345	6	37.5	38.3	**33.5**	34.6	31.6
Haberman	2	306	3	25.0	25.0	25.1	**24.9**	23.7
Pageblocks	5	5473	10	4.4	4.4	**2.8**	**2.8**	2.5
Teaching	3	151	5	63.8	63.3	63.8	**61.1**	56.9
Australian	2	690	14	26.3	26.4	**22.6**	22.8	21.9
Diabetes	2	768	8	26.0	25.6	**23.9**	24.2	23.0
German	2	1000	20	25.3	25.4	25.1	**24.9**	24.0
Heart	2	270	10	17.5	18.3	17.6	**17.3**	15.6

Table 3: Mean test error rate of the classifier selected with the four criteria. Those test error rates are averaged over 100 random learning/test splits. K is the number of classes, N the total number of samples and d the dimension of the description space.

given in Table 2. It shows that BEC selects a satisfactory model. The cross validation criterion CV3 and BEC have a quite similar behavior. On the contrary, as it often happens on other datasets, BIC chooses models that are strongly suboptimal in terms of mean test error rate. In this numerical experiment, BIC has selected most of the time a model with an error rate of 27.5%, far from the minimum, namely 22.9%, often selected with CV3 and BEC. It can be noticed that BEC and CV3 hesitate between the three models providing the smallest error rate. (AIC criterion has an analogous behavior despite a slight tendency to prefer the more complex model among the three models.) This suggests that BEC is also suitable for *model averaging* methods which weight a decision according to the posterior probabilities of candidate models. (See for instance [18].)

The datasets used to assess the performances of criteria AIC, BEC, BIC and CV3 were *Abalone* (classification between male, female and infant), *Bupa* (liver-disorders database), *Haberman* (survival data), *Pageblocks* (classifying the blocks of the page layout of a document), *teaching* (evaluations of teaching performance), *Australian* (credit approval), *Diabetes* (Pima indian diabete detection), *German* (credit risk evaluation), *Heart* (heart disease risk evaluation). A complete description of these dataset can be found in the UCI repository. In our experiments the binary variables were removed from *Abalone*, *Australian* and *Heart* datasets.

Experiments were similar to the one described for the *Australian credit* dataset. But this time, for each independent learning/test split, the test error rate corresponding to the chosen model is saved for each model selection criterion. The average error rate is plotted in Table 3 for each dataset and each criterion. We also add that we call the *oracle* performances, namely the test error rate that we would get by choosing at each experiment the model providing the smallest test error rate for one of the four considered criteria.

These experiments show that BEC clearly outperforms AIC and BIC in terms of error rate. BEC and BIC which approximate integrated likelihoods of the models in competition are of the same family of model selection criteria. But because BIC does not take into account the classification performance, it appears that it can choose a suboptimal model from the prediction point of view. It

is illustrated here for datasets *Bupa, pageblocks, Australian and diabete* where the positive difference between and BEC and CV3 error rates is higher than 2%. On the contrary, except for *Teaching dataset* for which EDDA performs quite poorly, the performance of BEC and CV3 are quite similar. It suggests that BEC is an interesting alternative to cross-validation for assessing the error rate of a classification model. However, those experiments on benchmark datasets remain somewhat superficial. Next, we present the performance of BEC in a more realistic setting.

5.3 Choosing of the number of mixture components in MDA

The model selection problem is now considered for the MDA classifiers [17]. In the present experimentation, attention is restricted to mixture of spherical Gaussian distributions, an attractive family of models for its simplicity and flexibility [6]. The class-conditional density is $p_k(\boldsymbol{x}|\theta_k) = \sum_{r=1}^{R_k} \pi_r \phi(\boldsymbol{x}|\boldsymbol{\mu}_r, \sigma_r^2 I_d)$ where R_k, $k = 1, \cdots, K$ denotes the number of mixture components for each class, and π_r, μ_r and σ_r are respectively the mixing proportion, 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 $\theta_k = (\pi_1, \ldots, \pi_{R_k-1}, \mu_1, \ldots, \mu_{R_k}, \sigma_1, \ldots, \sigma_{R_k})$. Obviously, the selection of the number of mixture components $\{R_k\}_{k=1,\cdots,K}$ is an important but difficult question to get good classification performances with this method. Evaluating the cross validation error rate is especially expensive in this context since the number of models to be compared is important. 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 to the learning data set rather than its ability to fulfil the classification task. Hence, it appears that it is difficult to guess the practical behavior of BIC and this criterion is rather disappointing to select a relevant number of components per class [6]. To illustrate this fact, we present a small Monte Carlo experiment before detailing an application of MDA for a pattern recognition problem in computer vision.

5.3.1 Monte Carlo numerical experiments

The same sample distribution as for the diagonal versus the spherical variance matrices case, presented in Section 5.1, was used. The considered models are here the spherical Gaussian mixture distributions described above, and the problem is to select the number of components $R_k, k = 1, 2$. For simplicity, we assume that $R_1 = R_2$. The behavior of criteria BEC and BIC are compared in Table 4. It can be remarked that BEC criterion selects the complexity suitable for the classification purpose. For instance, in the very well separated situation ($\Delta = 10$), the error rates of the different models are equivalent and BEC selects the simplest model most often. On the other side, BIC criterion selects always the same model without taking into account the separation between the classes.

separation	model	$\overline{err}$	-BIC	-BEC	BIC choice(%)	BEC choice(%)
$\Delta = 1$	1 components	0.266	1.6e+03	114	29	0
$\Delta = 1$	2 components	**0.25**	**1.6e+03**	98.5	70	12
$\Delta = 1$	3 components	0.251	1.62e+03	**94.6**	1	88
$\Delta = 3.5$	1 components	0.0748	1.6e+03	43.1	29	0
$\Delta = 3.5$	2 components	0.0625	**1.6e+03**	31.2	70	18
$\Delta = 3.5$	3 components	**0.0617**	1.62e+03	**28.8**	1	82
$\Delta = 5$	1 components	0.0227	1.6e+03	14.1	29	0
$\Delta = 5$	2 components	0.0151	**1.6e+03**	8.48	70	18.5
$\Delta = 5$	3 components	**0.0149**	1.62e+03	**7.16**	1	81.5
$\Delta = 7$	1 components	0.00357	1.6e+03	2.1	29	3.5
$\Delta = 7$	2 components	0.00174	**1.6e+03**	0.9	70	28
$\Delta = 7$	3 components	**0.0015**	1.62e+03	**0.702**	1	68.5
$\Delta = 10$	1 components	8.55e-05	1.6e+03	0.121	29	88
$\Delta = 10$	2 components	**1.8e-05**	**1.6e+03**	**0.0187**	70	5.5
$\Delta = 10$	3 components	**1.8e-05**	1.62e+03	0.033	1	6.5

Table 4: Comparison of criteria BEC and BIC for choosing the number of components in the spherical Gaussian mixture model. Column $\overline{err}$ gives the error rate evaluated on a test sample of size 50,000. Reported mean values are computed over 500 replications.

5.3.2 Model selection example in computer vision

Object categorization aims at classifying objects having common attributes. In this section, the problem of finding images containing a motorbike is considered. It is a typical example of object categorization since many different types of motorbikes exist, and the problem is to learn how to generalize the features specific to an object category. The motorbike and background datasets[1] considered here were originally studied by [10], but several authors compared on these databases object categorization methods based on interest point detection [27, 8, 9]. Those data sets contain respectively 826 and 900 images. Half of each data set was selected at random to learn the classifier and the remaining half data set has been used as test set.

To classify the data, a simple and computationally efficient "bag of features" method [8] was used. It is based on the quantization of scale-invariant descriptors of image patches. For each image, a k-dimensional vector is computed (k being the number of quantized vectors) and used as input to design a classifier. We briefly detail how these feature vectors have been generated, and then focus on the generative classifier selection problem.

- The images were rescaled to a maximum size of 320×160 pixels, preserving their initial aspect ratio.
- Then, a scale-invariant Harris-Laplace interest point detector extracts m location/scale points from the image $\mathcal{I}$. Depending of the image complexity, between 100 and 300 points are

[1] available at `http://www.vision.caltech.edu/html-files/archive.html`

detected. For each detected point, a 128-dimensional normalized vector is computed, called *Scale Invariant Feature Tranform* (SIFT). It encodes the visual appearance around the interest point [22].

- For each image, the set of appearance vectors is quantized into a 1000-dimensional vector from a clustering of images points into 1000 clusters: A fuzzy assignment of the vectors to the centers is computed using spherical Gaussian distributions with a variance equal to 0.36. Then each cluster is associated to a vector coordinate in the following way. The value given to each of the 1000 coordinates is the maximum probability of assignment to the corresponding cluster. The centers used in the quantization are learned by a k-means clustering on the learning set, so that the vectors approximately cover all possible appearances.

It is worth to note that the number of dimensions is much larger that the size of the learning data set. Some studies have shown that regularized discriminative classifiers are well suited for this type of situation [1, 27, 8], and, here, the question is to see if a generative classifier modelling the joint distribution of the input $(\boldsymbol{x}, \boldsymbol{y})$ can provide similar performances. In the present case, a generative classifier based on mixture of diagonal Gaussian distributions was used. This type of model can be expected to be suitable for the data we considered since several groups of motorbikes should be present, and the background images are expected to be associated to many different categories leading to a multimodal distribution.

The important problem to be solved here is to find a reliable number of components to describe each class. No prior information is available to help answering this question. Thus, it is solved using model selection criteria we considered in this paper. Mixtures were learned with one to five clusters for the motorbike images and with one to seven clusters for the background images. Criteria -BEC, -BIC, 10-fold cross validated error rate (CV10) were computed on the learning data and the error rate was evaluated on the test data. Table 5 gives the values of the different criteria for all possible models, up to 5 and 7 clusters.

Compared to other studies on this dataset, the error rate of 3.84% appears to be competitive. Some examples of misclassified images are given on Figure 3 The BIC tends clearly to select a too simple model, namely a 3×3 clusters mixture model. BEC is minimum for the same model as the test error rate for $R_1 = 3$ and $R_2 = 6$ clusters. Even for the other complexities, the values of BEC remarkably reproduces the behavior of test error rate. This illustrates the fact that BEC penalizes in a satisfactory manner the conditional likelihood so that the chosen classifier has nearly optimal performances. Now, for 20 independent random learning/test splits, we computed the relative performance improvement of BEC compared to BIC, that is to say

$$\alpha = \frac{\bar{err}_{\text{BIC}} - \bar{err}_{\text{BEC}}}{\min_{\{R_1, R_2\}} \bar{err}_{\text{test}}},$$

where $\bar{err}_c$ stands for the test error rate of a given criterion c. The mean value of α was 27.7 and the 95% confidence inteval was $[16.5, 38.9]$. This means that choosing a model with BEC improves on average the classification by 27.7% compared to a the classification based on BIC choice. A similar comparison between CV10 and BEC gives a confidence interval of $[-10.7, 8.8]$, which means that both criteria provide quite similar performances.

$-$-BIC ($\times 10^5$)

	R_1				
R_2	1	2	3	4	5
1	-9.111	-9.227	-9.255	-9.263	-9.264
2	-9.260	-9.257	-9.126	-9.243	-9.271
3	-9.279	**-9.281**	-9.275	-9.273	-9.126
4	-9.242	-9.270	-9.278	-9.279	-9.275
5	-9.272	-9.122	-9.239	-9.267	-9.275
6	-9.276	-9.271	-9.269	-9.115	-9.231
7	-9.259	-9.267	-9.268	-9.264	-9.261

$-$BEC ($\times 10^3$)

	R_1				
R_2	1	2	3	4	5
1	3.06	1.18	0.91	0.75	0.63
2	1.35	1.27	6.24	1.09	0.75
3	0.51	0.46	0.46	0.39	6.93
4	1.99	0.80	0.52	0.48	0.37
5	0.32	7.95	2.35	0.80	0.53
6	0.45	0.34	**0.29**	8.57	2.44
7	0.91	0.58	0.51	0.38	0.32

CV10 error rate ($\times 100$)

	R_1				
R_2	1	2	3	4	5
1	7.19	9.04	6.61	4.98	6.95
2	6.95	7.42	9.04	7.18	6.61
3	6.26	5.91	4.98	4.75	9.62
4	7.42	6.61	5.79	5.45	4.87
5	4.75	9.85	6.84	5.79	5.68
6	5.33	4.29	**4.09**	11.47	6.61
7	5.91	6.14	5.79	4.72	4.72

Test error rate ($\times 100$)

	R_1				
R_2	1	2	3	4	5
1	6.26	8.34	5.56	6.49	4.85
2	5.56	5.10	7.76	6.72	5.91
3	5.91	5.33	5.56	4.87	8.69
4	6.95	5.68	5.56	5.21	5.21
5	4.98	9.50	6.84	5.45	5.91
6	4.87	4.52	**3.84**	10.08	6.84
7	5.33	6.03	4.75	4.85	4.59

Table 5: Values of the different criteria for the mixture models with R_1 clusters in the distribution of the motorbike images and R_2 in the distribution of the background images. Last table provides the error rate computed on the independent test sample containing 863 images. Error rates have been computed considering that the positive and negative images have the same probability of occurrence.

Finally, on these 20 learning/test splits, we compared the classifier to a discriminative approach to see if the model fits well these high-dimensional data. A Gaussian kernel SVM classifier returned 3.46% error rate on these test data, where the kernel width and the slack-variable coefficient where chosen by 10-fold CV. The generative classifier based on BEC model choice gave 3.97% error rate. Since purely discriminative approaches give state-of-the-art performance in such context [8], the slight decrease of classification performance is acceptable since it open the door to more structured generative models, possibly using additional information such as the location and scale of the descriptors. Such model extensions are generally more difficult to introduce in discriminative approaches (see for instance [7]).

6 Discussion

We have proposed a promising model selection criterion which takes into account the classification task when selecting a generative model for supervised classification. It can be regarded as an efficient alternative to the cross-validated error rate when the collection of models in competition is large.

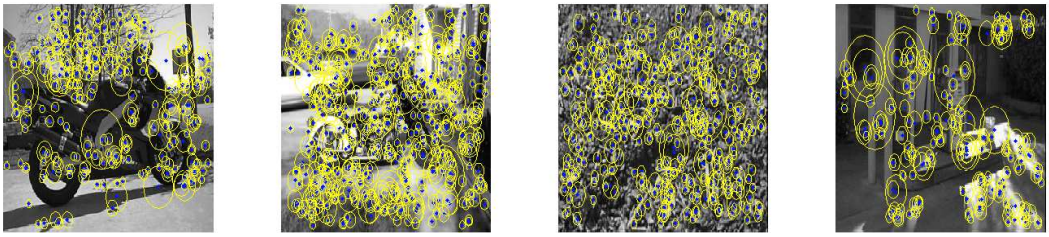

Figure 3: Examples of misclassified images with the corresponding scale-invariant Harris detectors.

This criterion is a BIC-like approximation of the classification entropy provided by a generative model. And, in many cases, it leads to select a model with a lower error rate than BIC criterion.

Now, it could be think of as desirable to estimate the parameter θ_m of a model m with $\theta_m^* = \arg\max_{\theta_m} \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, \theta_m)$ rather than with the ml estimate $\hat{\theta}_m$. However, we do not recommend taking into account the modelling purpose when estimating the model parameters because it could lead to quite unstable estimates and no performance improvement is to be expected for small and moderate training sample size [14, 26]. Moreover, as remarked in (15), $\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}, \theta_m^*)$ does not lead to a simple approximation of the classification entropy of a model m.

The BEC criterion could be applied to Bayesian Network Classifiers. Such classifiers can be very efficient in practice [13], but the choice of the graph structure is an open and difficult question. Studies focusing on discriminative parameter learning [15, 21] lead to much smaller improvements in classification error rate than methods based on discriminative model selection whose ideas lie on the same ground as BEC [16].

We think that estimating the parameter of a model and assessing its ability to fulfil the modelling purpose are two different problems that have to be treated separetaly. In a general perspective, when facing a collection of models, we recommend to estimate the model parameters by optimizing some contrast (as loglikelihood) measuring the fit of the model to the data. Then, when concerned with the model selection problem,we recommend to take into account the modelling purpose to choose a reliable, useful and stable model. And, in a supervised classification context, we think that the BEC criterion is doing the job in a satisfactory manner. Now, a possible further improvement would be to replace the mixture learning of $\tilde{\theta}$ with a simpler approximation.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 113–128, 2002.

[2] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[3] H. Bensmail and G. Celeux. Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:1743–48, 1996.

[4] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, 1st edition, 1994.

[5] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

[6] G. Bouchard and G. Celeux. Supervised classification with spherical Gaussian mixtures. In *Proceedings of CLADAG 2003*, pages 75–78, 2003.

[7] G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. Submitted to CVPR'05, October 2004.

[8] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Proceedings of the 8th European Conference on Computer Vision, Prague*, pages 59–74, 2004.

[9] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proceedings of the 9th International Conference on Computer Vision, Nice, France*, pages 634–640, 2003.

[10] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, June 2003.

[11] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[12] J. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[13] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[14] G. L. Goodman and D. W. McMichael. Objective functions for maximum likelihood classifier design. In Robin Evans, Lang White, Daniel McMichael, and Len Sciacca, editors, *Proceedings of Information Decision and Control 99*, pages 585–589, Adelaide, Australia, February 1999. Institute of Electrical and Electronic Engineers, Inc.

[15] R. Greiner and W. Zhou. Structural extension to logistic regression: Discriminant parameter learning of belief net classifiers. In *Proc. of the Eighteenth Annual National Conference on Artificial Intelligence*, pages 167–173, Edmonton, 2002.

[16] D. Grossman and P. Domingos. Learning bayesian network classifiers by maximizing conditional likelihood. In *Proc. of the 21st International Conference on Machine Learning*, 2004.

[17] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society series B*, 58:158–176, 1996.

[18] J. A. Hoeting, D. D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, 14:382–417, 1999.

[19] T. Jebara. *Discriminative, Generative and Imitative Learning*. PhD thesis, Media Laboratory, MIT, 2001.

[20] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[21] P. Kontkanen, P. Myllymäki, and H. Tirri. Classifier learning with supervised marginal likelihood. In J.Breese and D.Koller. Morgan Kaufmann Publishers, editors, *Proceedings of the 17th International Conference on Uncertainty in Artificial Intelligence*, pages 277–284, 2001.

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoint. *International Journal of Computer Vision*, 60:91–110, 2004.

[23] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[24] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[25] N. Murata, S. Yoshizawa, and S.-I. Amari. Network Information Criterion—determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5:865–872, November 1994.

[26] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In S. Becker T. Dietterich and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 609–616, Cambridge, MA, 2002.

[27] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proceedings of the 8th European Conference on Computer Vision, Prague*, volume 2, pages 71–84, 2004.

[28] A. E. Raftery. Bayesian model selection in social research (with discussion). *Sociological Methodology*, pages 111–196, 1995.

[29] B. D. Ripley. *Pattern Recognition and Neural Networks*. University Press, Cambridge, 1996.

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

[31] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

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

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifi que
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

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