Adaptive mixture discriminant analysis for supervised learning with unobserved classes

Charles Bouveyron

Adaptive Mixture Discriminant Analysis for Supervised Learning with Unobserved Classes

Charles Bouveyron

SAMOS-MATISSE, CES, UMR CNRS 8174
Université Paris 1 (Panthéon-Sorbonne), Paris, France

Abstract

In supervised learning, an important issue usually not taken into account by classical methods is the possibility of having in the test set individuals belonging to a class which has not been observed during the learning phase. Classical supervised algorithms will automatically label such observations as belonging to one of the known classes in the training set and will not be able to detect new classes. This work introduces a model-based discriminant analysis method, called adaptive mixture discriminant analysis (AMDA), which is able to detect unobserved groups of points and to adapt the learned classifier to the new situation. Two EM-based procedures are proposed for parameter estimation and Bayesian model selection is used for unobserved class detection. Experiments on artificial and real data demonstrate the ability of the proposed method to deal with complex and real word problems. The proposed approach is also applied to the detection of novel species in DNA barcoding.

Key-words: supervised classification, unobserved classes, adaptive learning, novelty detection, model selection.

1 Introduction

The usual framework of supervised classification assumes that all existing classes in the data have been observed during the learning phase and does not take into account the possibility of having in the test set individuals belonging to a class which has not been observed. In particular, such a situation could occur in the case of rare classes or in the case of an evolving

population. For instance, an important problem in Biology is the detection of novel species which could appear at any time resulting from structural or physiological modifications. Unfortunately, classical supervised algorithms, like support vector machines or linear discriminant analysis, will automatically label observations from a novel class as belonging to one of the known classes in the training set and will not be able to detect new classes. It is therefore important to find a way to allow the supervised classification methods to detect unobserved situations and to adapt themselves to the new configurations.

In statistical learning, the problem of classification with unobserved classes is a problem which has received very few attention. Indeed, both supervised and unsupervised classification contexts have been widely studied but intermediate situations have received less attention. We would like however to mention two related topics in statistical learning called semi-supervised classification and novelty detection. Semi-supervised classification focuses on learning with partially labeled data whereas novelty detection tries to detect new or unknown data points in the test set. Unfortunately, both approaches are unable to detect unobserved groups of points in the test set and to adapt the classifier to the new situation.

To overcome this problem, this work introduces an approach based on the mixture model which combines unsupervised and supervised learning for detecting unobserved groups of observations in the test test and for adapting the supervised classifier to the new situation. The adapted classifier could be then used to correctly classify new observations in the future. Two EM-based approaches are proposed for parameter estimation: an inductive approach, which is made of a learning and a discovering phase, and a transductive approach which considers all available observations for learning in a unique step. The detection of the number of unobserved classes is done using Bayesian model selection criteria. Finally, once the classifier adapted, the classification of new observations can be then done through the classical *maximum a posteriori* rule.

The paper is organized as follows. A brief review on generative supervised classification is given in Section 2 as well as a presentation of related works in supervised learning with unobserved classes. Section 3 introduces an adaptive discriminant analysis method based on the mixture model which is able to detect unobserved classes and considers parameter estimation as

well. Experimental results highlighting the main features of the proposed method on simulated and real datasets are presented in Section 4. Section 5 is devoted to the application of the proposed approach to the detection of novel species in DNA barcoding. Finally, Section 6 proposes some concluding remarks and discusses further works.

2 Related works

This section briefly reviews the supervised classification problem and solutions based on the mixture model before to present related works on supervised learning with unobserved classes.

2.1 Generative supervised classification

Supervised classification, also known as discriminant analysis in the literature, aims to build a supervised classifier from a complete set of learning observations $\{(x_1, z_1), ..., (x_n, z_n)\}$ where x_i is an observation described by p variables and $z_i \in \{1, ..., K\}$ indicates the class of x_i. The learned classifier is then used for assigning a new observation x^* to one of the K known classes. Among all existing approaches, generative discriminant analysis is very popular because of its probabilistic background and its efficiency. Generative supervised classification assumes that the observations $\{x_1, ..., x_n\}$ and their labels $\{z_1, ..., z_n\}$ are respectively independent realizations of random vectors $X \in \mathbb{R}^p$ and $Z \in \{1, ..., K\}$ and that the conditional density of X given that $Z = k$ is a parametric density f_k parameterized by θ_k. Consequently, the marginal density of X is given by:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k),$$

where π_k is the prior probability of the kth class. The classification of a new observation x^* is done afterward using the *maximum a posteriori* (MAP) rule which assigns x^* to the class with the highest posterior probability $P(Z = k|X = x)$:

$$P(Z = k|X = x) = \frac{\pi_k f_k(x; \theta_k)}{\sum_{k=1}^{K} \pi_k f_k(x; \theta_k)}.$$

We refer the reader to [29] for more details on generative discriminant analysis. The following paragraphs review the most used parametric densities in generative supervised classification.

Mixture of Gaussians Among all parametric densities, the Gaussian model is probably the most used in classification. The Gaussian mixture model has been extensively studied in the last decades and used in many situations (see [3] and [30] for a review). Therefore, if the Gaussian model is chosen, $f_k(x;\theta_k)$ will denote the density of a multivariate Gaussian density parameterized by $\theta_k = \{\mu_k, \Sigma_k\}$ where μ_k and Σ_k are respectively the mean and covariance matrix of kth component of the mixture.

Mixture of parsimonious Gaussians In some situations, modelling the data with a full covariance matrix can be too expensive in terms of number of parameters to estimate. In such a case, it is possible to make additional assumptions on the structure of the covariance matrix. For example, in the well-known Linear Discriminant Analysis (LDA) method, the covariance matrices of the different components are supposed to be equal to a unique covariance matrix. It is also possible to assume that the covariance matrix of each mixture component is diagonal or proportional to the identity matrix. These models are known as parsimonious Gaussian models in the literature since they require to estimate less parameters than the classical Gaussian model. Celeux and Govaert proposed in [11] a family of parsimonious Gaussian models based on an eigenvalue decomposition of the covariance matrix including the previous models. These parsimonious Gaussian models were then applied in [5] to supervised classification.

Mixture of HD Gaussians Nowadays, many scientific domains produce high-dimensional data like medical research, image analysis or Biology (see an application to DNA barcoding in Section 5). Classifying such data is a challenging problem since the performance of classifiers suffers from the *curse of dimensionality* [4]. Classification methods based on Gaussian mixture models are directly penalized by the fact that the number of parameters to estimate grows up with the square of the dimension. Unfortunately, parsimonious models are usually too constrained to correctly fit the data in a high-dimensional space. To overcome this problem, Bouveyron *et al.* proposed

recently in [10] a family of Gaussian models adapted to high-dimensional data. This approach, based on the idea that high-dimensional data live in low-dimensional spaces, assumes that the covariance matrix of each mixture component has only $d_k + 1$ different eigenvalues where d_k is the dimension of the subspace of the kth mixture component. These Gaussian models were then used in [9] for high-dimensional data clustering.

Mixture with a noise component Banfield and Raftery have introduced in [3] a mixture model with a noise component in order to improve the robustness of the cluster analysis on noisy datasets. The original work proposed to add to the mixture model a uniform distribution over the convex hull of the data as an additional component. Good results of the use of this specific mixture model were observed in different situations. Hennig and Coretto [21] proposed recently to use an improper uniform distribution that does not depend on the data for improving the robustness and provide a better approximation of the likelihood than the one proposed in the original work. An application of noise detection is proposed in Section 4.

2.2 Semi-supervised classification

The first related topic to supervised classification with unobserved classes is semi-supervised classification. Semi-supervised classification is a topic which has been well studied for several years and which focuses on supervised classification with partially labeled data. Usually, unlabeled data are added to the learning data in order to improve the efficiency of the final classifier. Such an approach is particularly useful when only few labeled observations are available for learning (applications with a high supervision cost like DNA barcoding, cancer detection, ...). A good review on semi-supervised classification can be found in [37] and [24]. Earlier approaches [28, 32] used the EM algorithm to assign unlabeled observations to known classes. Most recent approaches include co-training algorithms [8] and graph-based techniques [24] which used prior informations on unlabeled observations. However, all semi-supervised classification methods are not able to detect unobserved groups of points. More importantly, they will use those new points to re-estimate the model parameters of known classes and the estimates of known classes parameters will be therefore deteriorated.

2.3 Novelty detection

The second and most related topic to supervised classification with unobserved classes is novelty detection. Novelty detection focuses on the identification of new or unknown data for which the learned classifier was not aware during the learning phase. This approach has become very popular in several application fields such as fault detection [14], medical imaging (mass detection in mammograms) [39] or e-commerce [25]. In the last years, many methods have been proposed to deal with this problem. An excellent review on novelty detection methods can be found in [26] and [27] which splits novelty detection methods into two main categories: statistical and neural network based approaches. Approaches based on statistical assumptions usually model the data on their statistical properties and use this information for deciding whether test data comes from the same distribution or not. Among parametric techniques, Chow [12] was the first to propose a threshold for outlier rejection which has been improved in [19] by introducing the classification confidence in the rejection. Gaussian densities were also used in [34] for modelling the learning data and detect outliers using a measure based on the Mahalanobis distance. Extreme value theory was also used in [33] for novelty detection by searching for low or high values in the tails of data distributions. Non-parametric approaches include k-NN based techniques [20, 31] or Parzen windows [41] for estimating the distribution of the data. Neural networks and kernel methods have been also widely used for novelty detection. Bishop [7] used parametric statistics by post-processing neural networks for detecting new data distribution whereas a probability density estimation of neural network outputs is used in [16] as a measure of novelty. Another approach based on neural networks was proposed in [35] which used a thresholding on the neural network output for detection new samples. Kohonen proposed also in [23] two types of novelty detectors based on self-organizing maps (SOM). More recently, Tax and Duin [40] as well as Schölkopf [38] used support vector machines (SVM) for distinguishing know and unknown objects. However, even though all these methods are able to detect new or unobserved data points, no one of them is able to recognize unobserved homogeneous groups of points and to adapt the classifier to the new situation for classifying future observations.

3 Adaptive mixture discriminant analysis

We introduce in this section an adaptive model-based classifier able to detect novel classes which have not been observed during the learning phase. Parameter estimation, model selection and classification of future observations will be discussed as well.

3.1 The mixture model

Let us consider a classical parametric mixture model of K components: the observations $\mathcal{X} = \{x_1, ..., x_n\} \in \mathbb{R}^p$ are assumed to be independent realizations of a random vector $X \in \mathbb{R}^p$ with density:

$$f(x; \Theta) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k), \tag{1}$$

where $\pi_k \geq 0$ for $k = 1, ..., K$ are the mixing proportions (with the constraint $\sum_{k=1}^{K} \pi_k = 1$), $f_k(x; \theta_k)$ is the density of the kth component of the mixture parameterized by θ_k and finally $\Theta = \{\pi_1, ..., \pi_K, \theta_1, ..., \theta_K\}$. We refer to the previous section regarding the choice of the mixture densities.

3.2 Parameter estimation: inductive approach

This paragraph focuses on the estimation of the mixture parameters in the specific situation where one or several classes have not be observed during the learning phase. In the mixture model framework, the maximum likelihood (ML) estimation method is usually used for estimating the model parameters. For the mixture model (1), the complete log-likelihood has the following form:

$$\ell(x_1, ..., x_n; \Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} s_{ik} \log \left(\pi_k f_k(x_i; \theta_k) \right),$$

where $s_{ik} = 1$ if x_i belongs to the kth class and $s_{ik} = 0$ otherwise. However, this work considers a specific learning situation in which one or several classes are not represented in the learning dataset. Therefore, the mixture parameter estimation can not be done using the classical way and a two step approach made of a learning phase and a discovery phase is proposed below.

The learning phase For this first phase of the parameter estimation, let us assume that only C classes are represented in the learning dataset $\mathcal{X} = \{x_1, ..., x_n\}$ with $1 \leq C \leq K$. Since the data of the learning set are complete, *i.e.* a label $z_i \in \{1, ..., C\}$ is associated to each observation x_i of the learning set ($i = 1, ..., n$), we fall into the classical estimation framework of model-based discriminant analysis. In such a case, the maximization of the likelihood reduces to separately estimate the parameters of each class density by maximizing the associated conditional log-likelihood ℓ_k, for $k = 1, ..., C$:

$$\ell_k(\mathcal{X}; \Theta) = \sum_{i=1}^{n} s_{ik} \log \left(\pi_k f_k(x_i; \theta_k) \right).$$

The maximization of the conditional log-likelihood $\ell_k(\mathcal{X}; \Theta)$, for $k = 1, ..., C$, conduces to an estimation of π_k by $\hat{\pi}_k = \frac{n_k}{n}$ where $n_k = \sum_{i=1}^{n} s_{ik}$ is the number of observations of the kth class and to an estimation of θ_k by $\hat{\theta}_k$ which depends on the chosen component density. For instance, in the case of a Gaussian density, the maximization of $\ell_k(\mathcal{X}; \Theta)$ conduces to an estimation of $\theta_k = \{\mu_k, \Sigma_k\}$ by:

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n_k} \sum_{i=1}^{n} s_{ik} x_k, \\ \hat{\Sigma}_k &= \frac{1}{n_k} \sum_{i=1}^{n} s_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^t, \end{aligned}$$

for $k = 1, ..., C$. We refer respectively to [11] and [10] for parameter estimation in the case of parsimonious and HD Gaussian models, and to [3] in the case of a mixture with a noise component.

The discovery phase Usually, in discriminant analysis, the classification phase consists only in assigning new unlabeled observations to one of known classes. However, in this work, it is assumed that all the classes have not been observed during the learning phase. It is therefore necessary to search for new classes before to classify the new observations for avoiding to misclassify observations of an unobserved class (by assigning them to one of the observed classes). Using the model and the notations introduced above, it remains to find $K - C$ new classes in the set of n^* new unlabeled observations $\mathcal{X}^* = \{x_1^*, ..., x_{n^*}^*\}$. Since these new observations are unlabeled, we have to fit the

mixture model in a partially unsupervised way. In this case, the completed log-likelihood has the following form:

$$\ell(\mathcal{X}^*;\Theta) = \sum_{i=1}^{n^*}\left(\sum_{k=1}^{C} s_{ik}^* \log\left(\pi_k f_k(x_i^*;\theta_k)\right) + \sum_{k=C+1}^{K} s_{ik}^* \log\left(\pi_k f_k(x_i^*;\theta_k)\right)\right),$$

where the parameters θ_k for $k = 1, ..., C$ have been estimated in the previous phase and the parameters θ_k for $k = C+1, ..., K$ remain to estimate. Due to the constraint $\sum_{k=1}^{K} \pi_k = 1$ on the parameters π_k, it is not possible to keep the proportions estimated during the learning and the mixture proportions have to be re-estimated on the new sample $\{x_1^*, ..., x_{n^*}^*\}$. However, the test set $\{x_1^*, ..., x_{n^*}^*\}$ is an incomplete dataset since the labels z_i^* are missing and the s_{ik}^* are consequently unknown for all observations of this set. In such a situation, the direct maximization of the likelihood is an intractable problem and the Expectation-Maximization (EM) algorithm [15] is usually used to estimate the mixture parameters by iteratively maximizing the likelihood through the maximization of the expectation of the completed log-likelihood conditionally to the posterior probabilities $t_{ik}^* = P(Z = k|X = x_i^*)$. We propose below a modified EM algorithm for estimating the parameters of the $K - C$ unobserved classes which alternates between the following E and M steps at each iteration q:

- **E step**: the posterior probabilities $t_{ik}^{*(q)} = P(Z = k|X = x_i^*)$, for $i = 1, ..., n^*$ and $k = 1, ..., K$, are updated according to the mixture parameters as follows:

$$t_{ik}^{*(q)} = \frac{\hat{\pi}_k^{(q-1)} f_k(x_i^*; \hat{\theta}_k^{(q-1)})}{f(x_i^*; \hat{\Theta}^{(q-1)})},$$

  where $\hat{\pi}_k^{(q-1)}$ and $\hat{\theta}_k^{(q-1)}$ are the mixture parameters estimated in the M step at the step $(q-1)$.

- **M step**: the parameters of the $K - C$ unobserved classes are estimated by maximizing the expectation of the completed log-likelihood conditionally to the posterior probabilities $t_{ik}^{*(q)}$ whereas the estimated parameters of the observed classes remain fixed to the values obtained in the learning phase except for the proportions which are re-estimated. Therefore, this step only updates the estimates of parameters π_k for

$k = 1, ..., K$ and θ_k for $k = C+1, ..., K$. In the case of the Gaussian mixture, for instance, the update formulas for the parameter estimates are, for $k = 1, ..., K$:

$$\hat{\pi}_k^{(q)} = \frac{n_k^{*(q)}}{n^*},$$

and for $k = C+1, ..., K$:

$$\begin{aligned} \hat{\mu}_k^{(q)} &= \frac{1}{n_k^{*(q)}} \sum_{i=1}^{n^*} t_{ik}^{*(q)} x_i^* \\ \hat{\Sigma}_k^{(q)} &= \frac{1}{n_k^{*(q)}} \sum_{i=1}^{n^*} t_{ik}^{*(q)} (x_i^* - \hat{\mu}_k^{(q)})(x_i^* - \hat{\mu}_k^{(q)})^t, \end{aligned}$$

where $n_k^{*(q)} = \sum_{i=1}^{n^*} t_{ik}^{*(q)}$.

Proofs of these results are given in Appendix A.1.

3.3 Parameter estimation: transductive approach

The previous paragraph proposed an EM-based algorithm which assumes that model parameters of the C observed classes have been estimated in the past (during the learning phase) and the modelling of new classes depends naturally on the quality of these estimates. However, in some situations, the cost of supervision is very high and only few labeled observations are available for learning. In such cases, the model parameters of observed classes could be badly estimated and the discovery phase could be consequently less efficient in detecting new classes. As the learning sample $\mathcal{X} = \{x_1, ..., x_n\}$ and the test sample $\mathcal{X}^* = \{x_1^*, ..., x_{n^*}^*\}$ are assumed to come from the same population, both samples could be used in the discovery phase to improve the model parameters while searching for unobserved classes in the test set. Such an approach should mainly benefit to the parameter estimation of the C observed classes and, consequently, should benefit as well to the detection of unobserved groups of observations. In this case, the completed log-likelihood has the following form:

$$\ell(\mathcal{X}, \mathcal{X}^*; \Theta) = \sum_{i=1}^{n} \sum_{k=1}^{C} s_{ik} \log\left(\pi_k f_k(x_i; \theta_k)\right) + \sum_{i=1}^{n^*} \sum_{k=1}^{K} s_{ik}^* \log\left(\pi_k f_k(x_i^*; \theta_k)\right),$$

and can be rewritten as follows:

$$\ell(\mathcal{X}, \mathcal{X}^*; \Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{s}_{ik} \log\left(\pi_k f_k(x_i; \theta_k)\right) + \sum_{i=1}^{n^*} \sum_{k=1}^{K} s^*_{ik} \log\left(\pi_k f_k(x^*_i; \theta_k)\right),$$

where $\tilde{s}_{ik} = s_{ik}$ if $k = 1, ..., C$ and $\tilde{s}_{ik} = 0$ for $k = C+1, ..., K$ and that for $i = 1, ..., n$. An alternative version of the EM-based algorithm proposed in the previous paragraph is presented below to jointly estimate model parameters while searching for new classes. The joint estimation procedure alternates between the following E and M steps at each iteration q:

- **E step**: on the one hand, the posterior probabilities $P(Z = k|X = x_i)$ remain fixed for the learning observations $\{x_1, ..., x_n\}$ and are equal to $\tilde{s}_{ik}$, for $i = 1, ..., n$ and $k = 1, ..., K$. On the other hand, the posterior probabilities $t^{*(q)}_{ik} = P(Z = k|X = x^*_i)$ are updated for the test sample $\{x^*_1, ..., x^*_{n^*}\}$, *i.e.* for $i = 1, ..., n^*$ and $k = 1, ..., K$, according to the mixture parameters as follows:

$$t^{*(q)}_{ik} = \frac{\hat{\pi}^{(q-1)}_k f_k(x^*_i; \hat{\theta}^{(q-1)}_k)}{f(x^*_i; \hat{\Theta}^{(q-1)})},$$

  where $\hat{\pi}^{(q-1)}_k$ and $\hat{\theta}^{(q-1)}_k$ are the mixture parameters estimated in the M step at the step $(q-1)$.

- **M step**: the parameters of the C observed classes and of the $K - C$ unobserved classes are estimated by maximizing the expectation of the completed log-likelihood conditionally to the posterior probabilities estimated in the E step. Therefore, this step updates now the estimates of parameters π_k and θ_k for $k = 1, ..., K$. In the case of the Gaussian mixture, for instance, the update formulas for the parameter estimates are, for $k = 1, ..., K$:

$$\begin{aligned}
\hat{\pi}^{(q)}_k &= \frac{n^{(q)}_k + n^{*(q)}_k}{n + n^*}, \\
\hat{\mu}^{(q)}_k &= \frac{1}{n^{(q)}_k + n^{*(q)}_k} \left(\sum_{i=1}^{n} \tilde{s}_{ik} x_i + \sum_{i=1}^{n^*} t^{*(q)}_{ik} x^*_i \right), \\
\hat{\Sigma}^{(q)}_k &= \frac{1}{n^{(q)}_k + n^{*(q)}_k} \left(S^{(q)}_k + S^{*(q)}_k \right).
\end{aligned}$$

where $S_k^{(q)} = \sum_{i=1}^{n} \tilde{s}_{ik}(x_i - \hat{\mu}_k^{(q)})(x_i - \hat{\mu}_k^{(q)})^t$, $S_k^{*(q)} = \sum_{i=1}^{n^*} t_{ik}^{*(q)}(x_i^* - \hat{\mu}_k^{(q)})(x_i^* - \hat{\mu}_k^{(q)})^t$, $n_k^{(q)} = \sum_{i=1}^{n} \tilde{s}_{ik}$ and $n_k^{*(q)} = \sum_{i=1}^{n^*} t_{ik}^{*(q)}$.

Proofs of these results are given in Appendix A.2.

3.4 Model selection: determining the number of components

In the usual case of supervised classification, the number of classes is known and the model selection consists only in choosing the most adapted densities for the considered dataset. In the context of the studied situation, the total number K of classes is assumed to be unknown and has to be chosen as well as the conditional densities of the mixture model. Classical tools for model selection in the mixture model framework are penalized likelihood criteria and include the AIC [1], BIC [36] and ICL [6] criteria. The Bayesian Information Criterion (BIC) is certainly the most popular and consists in selecting the model which maximizes the quantity:

$$BIC(\mathcal{M}) = \ell(x_1, ..., x_n; \Theta) - \frac{\nu(\mathcal{M})}{2} \log(n),$$

where $\nu(\mathcal{M})$ is the the number of parameters in model $\mathcal{M}$ and n is the number of observations. For instance, the number of parameters for the full Gaussian mixture model, *i.e.* full and different covariance matrices, is equal to $(k-1) + kp + kp(p+1)/2$. The AIC criterion penalizes the log-likelihood by $\nu(\mathcal{M})$ and the ICL criterion add to the BIC criterion the penalty $\sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \log(t_{ik})$ in order to favour well separated models. An evaluation of both criteria in the context of unobserved class detection is presented in the next section.

3.5 Classification with the adapted classifier

The previous paragraphs introduced a model-based discriminant analysis method which adapts its mixture model to a new situation including unobserved classes . Therefore, the adapted model can be used to classify new observations in the future. In the classical discriminant analysis framework, new observations are usually assigned to a class using the *maximum a posteriori* (MAP) rule. The MAP rule assigns a new observation $x \in \mathbb{R}^p$ to the class for which x has the highest posterior probability. Therefore, the classification step mainly consists in calculating the posterior probability

$P(Z = k|X = x)$ for each class $k = 1, ..., K$. In the case of the model described in this section, this posterior probability can be expressed classically using the Bayes' rule as follows:

$$P(Z = k|X = x) = \frac{\pi_k f_k(x; \theta_k)}{f(x; \Theta)},$$

where $f(x; \Theta) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k)$. Therefore, the posterior probabilities of the new observations depend on both the classes observed in the learning phase and the classes discovered in the test set.

4 Experimental results

This section presents experiments on toy and simulated datasets in order to highlight the main features of the method introduced in the previous section.

4.1 An introductory example: the iris dataset

The dataset considered in this first experiment is a classical one: the iris dataset made famous by its use by Fisher in [17] as an example for discriminant analysis. This dataset, in fact collected by Edgar Anderson [2] in the Gaspé Peninsula (Canada), is made of three classes corresponding to different species of iris (*setosa*, *versicolor* and *virginica*) among which the classes *versicolor* and *virginica* are difficult to discriminate (they are at least not linearly separable). The dataset consists of 50 samples from each of three species and four features were measured from each sample. The four measurements are the length and the width of sepal and petal. This dataset is used here as a toy dataset because of its popularity and its biological nature.

Detection of one unobserved class First, let suppose that botanists are studying iris species and have only observed the two species *setosa* and *versicolor*. For this experiment, the dataset has been randomly split into a learning dataset without *virginica* examples and a test dataset with several *virginica* examples. The top-left panel of Figure 1 shows what the botanists are supposed to have observed in the past. The top-center panel of the same figure presents a sample of new observations of iris for which the botanists are asked to classify. However, as the top-right panel indicates, this new sample contains individuals from a class which has not been observed by the

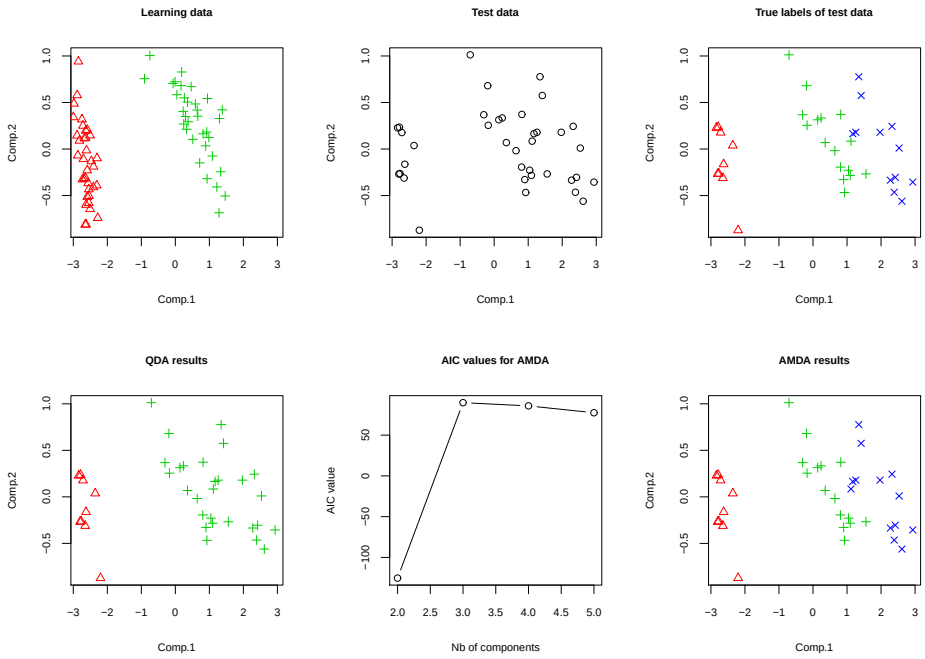

Figure 1: Detection of 1 unobserved class with AMDA on the Iris dataset: the classes "setosa" (red triangles) and "versicolor" (green plus-es) have been observed during the learning phase whereas the class "virginica" (blue crosses) has not.

botanists in the past and the iris experts will very likely classify all these new observations as belonging to either the class *setosa* or the class *versicolor*. The bottom-left panel of Figure 1 shows the result of such a *scenario,* using Quadratic Discriminant Analysis (QDA) in place of the iris experts, which yields to the classification of all *virginica* observations in the class *versicolor*. Remark that, even though this result is disappointing from our point of view, it is understandable both for an human expert and a classification method since the classes *versicolor* and *virginica* are indeed very difficult to discriminate. The strategy proposed in the previous section, hereafter referred to by Adaptive Model-based Discriminant Analysis (AMDA), was applied to this dataset. The bottom-center and right panels of Figure 1 presents the results provided by AMDA (with the inductive approach). On the one hand, it turns out that Bayesian model selection criteria (AIC here) succeed in identifying a new group of points in the test set. On the other hand, once the number K of mixture components chosen, AMDA classifies almost perfectly (only 2

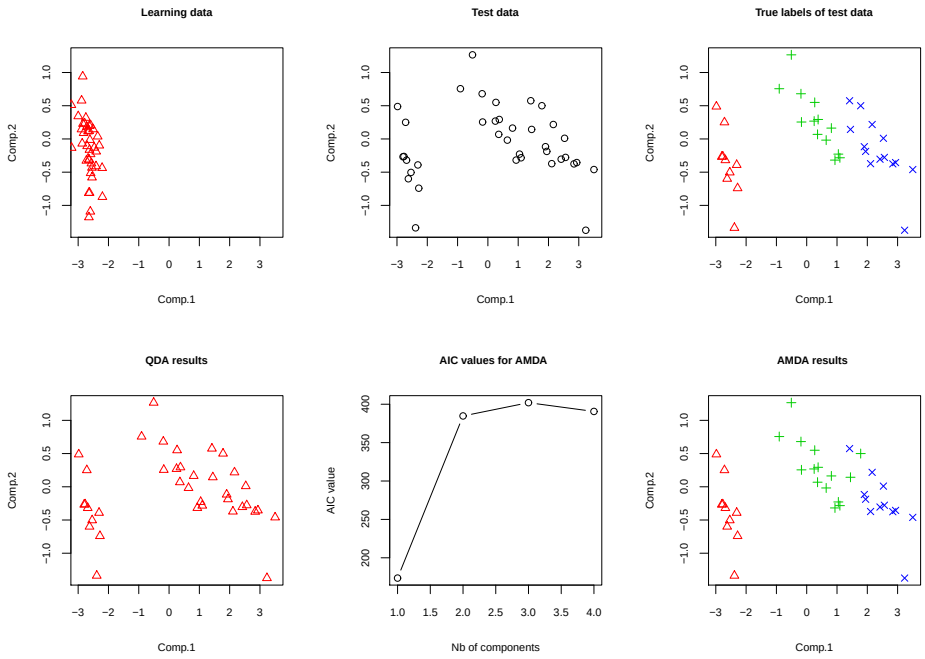

Figure 2: Detection of 2 unobserved classes with AMDA on the Iris dataset: the class "setosa" (red triangles) has been observed during the learning phase whereas the classes "versicolor" (green plus-es) and "virginica" (blue crosses) have not.

errors on this example) the observations of the unobserved class *virginica*.

Detection of two unobserved classes Here, the toy example turns to be a serious problem because the botanists are now assumed to have only observed one species, the species *setosa*, and will have therefore to discover two unobserved classes, the species *versicolor* and *virginica*. For this second experiment, the dataset has been randomly split into a learning dataset without *versicolor* and *virginica* examples and a test dataset with several *versicolor* and *virginica* examples. The top-left panel of Figure 2 shows what the botanists are supposed to have observed in the past whereas the center panel shows the new and unlabeled observations. As one can observe, the new observations are clearly different from the data observed in the past but it is actually not obvious to detect that these new observations come from two different iris species (*cf.* top-right panel of Figure 2). If a supervised classifier like QDA is used, the classifier will assign all the new observations

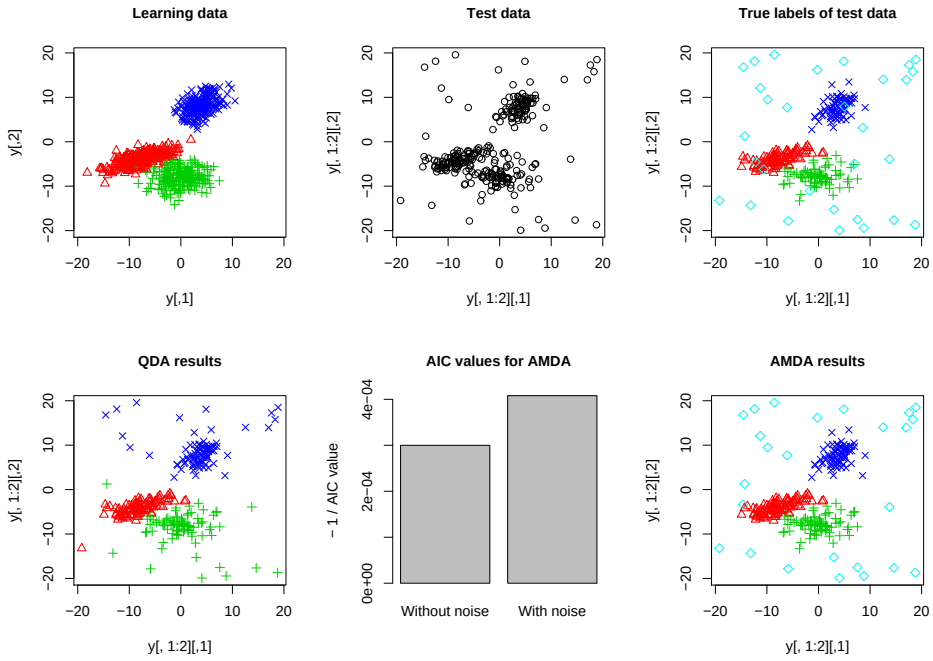

Figure 3: Detection of 1 unobserved noise class with AMDA on 2-dimensional simulated data: 3 observed classes and 1 unobserved noise class (light blue diamonds).

to the only one known class, the class *setosa*, and will make an important error (*cf.* bottom-left panel). In such a situation, there is no doubt that novelty detection methods presented in Section 2 are able to detect that the new observations do not belong to the species *setosa*. However, these techniques are not able to detect that the unlabeled observations are made of two homogeneous groups corresponding to two new iris species. The bottom-center and right panels of Figure 2 demonstrate that AMDA is actually able to detect the two unobserved iris species and can take this information into account to adapt the classifier for classifying future observations.

4.2 Detection of an unobserved noise class

This experiment aims to evaluate the ability of AMDA to detect an unobserved non Gaussian class of noise. For this, data were simulated in a 2-dimensional space according a mixture model made of 4 components: 3 Gaussian components and one uniform noise component. Means and co-

Truth / Classif.	1	2	3	4
1	75	**2**	0	0
2	0	78	0	**10**
3	0	0	65	**21**
-	-	-	-	-

(a) Confusion matrix for QDA

Truth / Classif.	1	2	3	4
1	75	**2**	0	0
2	0	77	0	0
3	0	0	64	0
4	0	**1**	**1**	**31**

(b) Confusion matrix for AMDA

Table 1: Confusion tables for QDA and AMDA on the test dataset for the simulated data with one unobserved noise class (class #4).

variance matrices of Gaussians were chosen in order to obtain separated enough groups. The learning set was made of 750 observations from the three Gaussian classes. The top-left panel of Figure 3 shows the observations of the learning set. The test set was made of 250 observations from the three Gaussian classes (observed during the learning) and 31 observations from the unobserved uniform noise class. The top-center panel of this figure presents the unlabeled test observations and the top-left panel indicates the true labels of these observations. The bottom-left panel of Figure 3 shows the classification of the test observations with the supervised classifier Quadratic Discriminant Analysis (QDA). Unsurprisingly, QDA classifies all the observations from the noise class to one of the three known Gaussian classes. Table 1 presents confusion tables for QDA and AMDA (inductive approach) on the test dataset and shows that all noise observations were classified into the classes #2 and #3. The bottom-center and right panels of Figure 3 show respectively the AIC values for AMDA with and without a noise component and the classification provided by AMDA with a detected noise component (as indicated by the largest AIC value). We can observe on this quite complex example that AMDA succeeds in both detecting the unobserved noise class and modelling it through a uniform component. Table 1.b confirms that AMDA recognizes all noise observations as belonging to one unobserved class in the past and makes only 2 false noise detections which is very satisfying. Naturally, it could be also possible to detect both unobserved classes and a noise component by comparing AIC curves with and without a noise component for different numbers of Gaussian components.

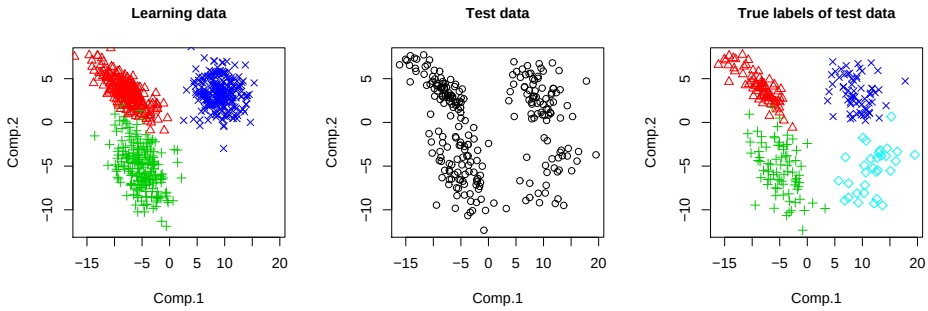

Figure 4: Classification with AMDA of simulated data: 3 observed classes and 1 unobserved class (light blue diamonds) in $\mathbb{R}^2$.

4.3 Monte Carlo simulations

This paragraph presents Monte Carlo experiments on simulated data in order to both compare inductive and transductive approaches, evaluate Bayesian model selection criteria in the context of unobserved class detection and determinate the breakdown group size for the detection of new classes. For the three following experiments, data were simulated according a Gaussian mixture model made of 4 groups (3 observed groups and one unobserved group) in a 2-dimensional space. Means and covariance matrices were chosen in order to obtain separated enough groups. Figure 4 presents the simulated learning dataset (left panel), the test dataset (center panel) and the true labels of the test observations (right panel). For each of the 50 replications of the Monte Carlo studies, 750 observations were simulated according to a Gaussian mixture model (250 obs. for each of the observed classes) in order to form the learning set and $250+\eta$ observations were simulated for the test set where η is the number of observations of the unobserved class. For each replication, the number η varied from 2 to 50.

Inductive vs. transductive approaches This first Monte-Carlo simulation aims to compare the inductive and transductive approaches proposed in Section 3. We choose to compare both approaches on modelling and classification criteria since supervised classification has two main objectives: efficiently classify new observations and correctly model the data for facilitating the understanding of classification results. For this simulation, the actual number of components was provided to the algorithms in order to focus on

modelling and classification abilities of both approaches. The left panel of Figure 5 shows the log-likelihood value of the whole dataset (training + test sets) divided by the number of observations for the inductive and transductive approaches according to the size of the unobserved class. In this figure, the information in which we are interested in is the relative behaviour of the inductive approach compared to the transductive one and not the the global behaviour of both curves. Indeed, for each situation, the approach with the highest log-likelihood value per point is the one which provides the best modelling of the data. On the one hand, it appears that the log-likelihood curve of the transductive approach is above the one of inductive approach for sizes of the unobserved class larger than 10. This indicates that, for large unobserved groups of points, the use of the all available observations allows to better model the data than using only the test observations. On the other hand, Figure 5 indicates as well that for small unobserved groups (smaller than 10) the inductive approach seems to better model the data than the transductive version of AMDA. This can be explained by the will of the transductive approach to consider small unobserved groups of points as extreme values of the observed classes. The right panel of Figure 5 shows the correct classification on a second test data set (different from the test set used for detecting new classes) for the two studied approaches according to the size of the unobserved class. A test set different from the test set used for detecting new classes is used here in order to evaluate the ability of both approaches to classify future unlabeled data with the adapted classifier including the discovered classes. One can observe that both classification rates are very good (between 0.97 and 0.99) and that the inductive version of AMDA appears to be slightly more efficient and stable than the transductive one to classify new data with the adapted classifier. In view of this results, we can recommend to use the transductive version for modelling purpose on large datasets and to use the inductive approach for classification purpose or modelling of small datasets.

Evaluation of model selection criteria This second Monte Carlo study aims to evaluate Bayesian model selection criteria in the context of unobserved class detection with AMDA. Figure 7 presents the rate of successful selection of the actual number of groups by the three selection model criteria AIC, BIC and ICL for both the inductive (left panel) and transductive (right

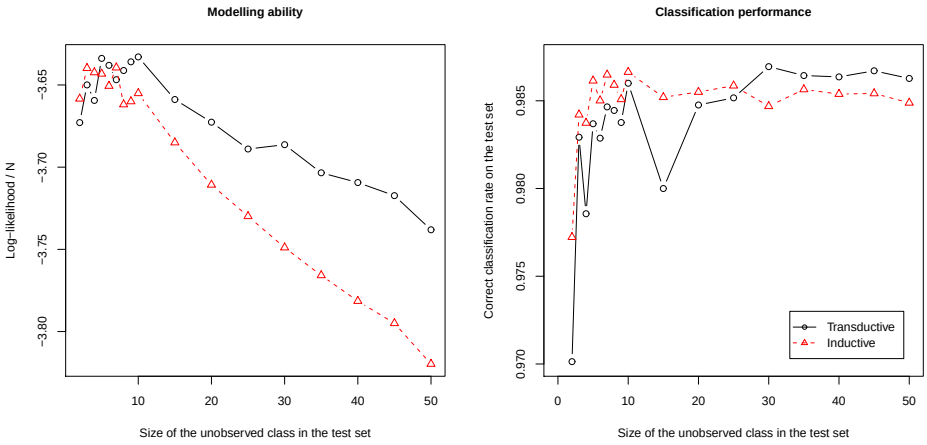

Figure 5: Modelling ability and classification performance of the inductive and transductive versions of AMDA according to the size of the unobserved class on simulated data: 3 observed classes and 1 unobserved class in $\mathbb{R}^2$.

panel) versions of AMDA. It appears that the three studied selection model criteria select always the correct number of groups when the unobserved group size is large (larger than 10 for the inductive approach and larger than 20 for the transductive one). For smaller sizes of the unobserved group, AIC turns out to be the more stable criterion since it selects the correct number of groups more frequently than the two other criteria. We therefore recommend the use of AIC as model selection criterion in the context of unobserved class detection.

Determination of the breakdown group size The two panels of Figure 7 shows three recognition rates averaged on the Monte Carlo replications for both the inductive (left panel) and transductive (right panel) versions of AMDA: total recognition rate, true positive rate and false positive rate. The total recognition rate measures the overall correct classification rate for the four classes (the three observed classes and the unobserved one). The true positive rate measures the correct classification rate for observations of the unobserved class (class #4). Conversely, the false positive rate evaluates how many observations of the three observed classes are classified as belonging to the new class. In a satisfying situation, the total recognition rate and the true positive rate should be close to 1 whereas the false positive rate should be close to 0. Both recognition rates were computed on a test dataset. Fig-

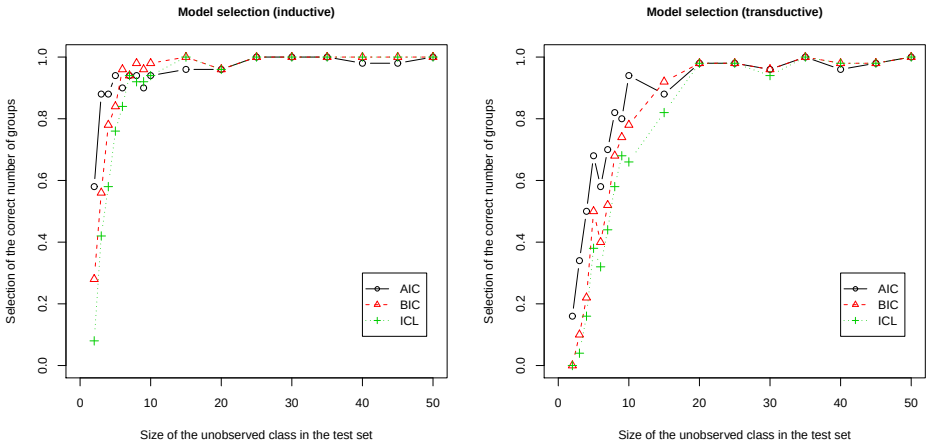

Figure 6: Successful selection of the actual number of groups using AIC, BIC and ICL with the inductive (left) and transductive (right) AMDA according to the size of the unobserved class on simulated data: 3 observed classes and 1 unobserved class in $\mathbb{R}^2$.

ure 7 shows that the three recognition rates are very good for sizes of the unobserved class larger than 10 in the case of inductive AMDA (left panel) and larger than 20 in the case of transductive AMDA (right panel). We observe as well that for sizes of the unobserved class smaller than 5–10 the true positive rate is very unstable and this means that the unobserved class is not well modeled. This confirms the observation made in the previous paragraph and the inductive approach seems more robust than transductive AMDA in the case of unobserved classes of small sizes.

To summarize, these Monte Carlo experiments have first demonstrated that the transductive version of AMDA must be used for modelling purpose on large dataset whereas inductive AMDA can be used for detecting and modeling small unobserved classes. They have also shown that AIC is the most efficient criterion for detecting unobserved classes and that the inductive version of AMDA is able to detect and model unobserved classes in the test set for unobserved classes as small as 5–10 observations.

5 Application to DNA barcoding

Determining to what species an organism belongs is probably the most common problem in Biology. The answer concerns many areas of practical im-

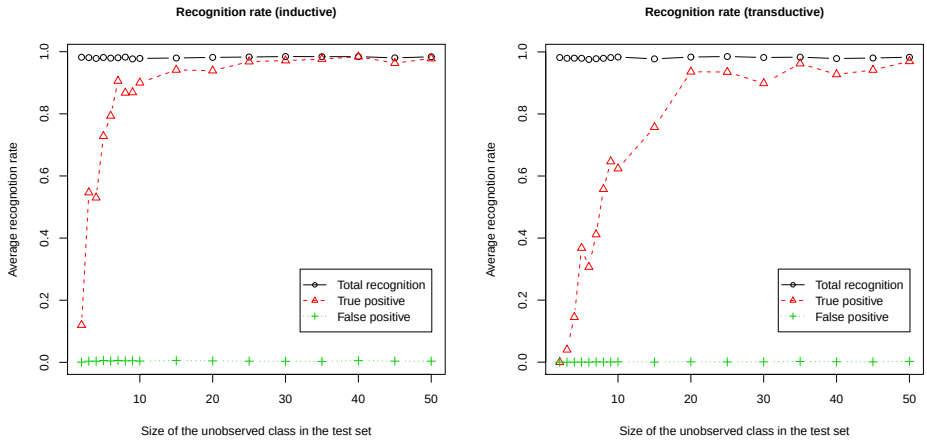

Figure 7: Recognition rates with the inductive (left) and transductive (right) AMDA according to the size of the unobserved class on simulated data: 3 observed classes and 1 unobserved class in $\mathbb{R}^2$.

portance such as protecting endangered species, sustaining natural resources, stopping disease vectors or monitoring environmental quality. Created in 2003, the Consortium for the Barcode of Life[1] is an international initiative devoted to developing DNA barcoding as a standard tool to identify species. Its purpose is to provide a simple and automatic method to correctly identify the species, with no or limited recourse to taxonomic expertise. The 5' half of the mtDNA gene COI has been chosen as the barcode locus for most animals, and gene markers with similar barcoding properties are investigated in plants, fungi, and protists. Traditionally, the barcoding procedure is based on an algorithm combining k-NN with neighbour-joining trees [22]. Several alternatives to this method were quite successfully applied to various kinds of organisms, although several major problems remain. One of them is: how to detect new or unobserved species? This section aims to demonstrate that AMDA can be used in such an application to efficiently detect new or unobserved species.

5.1 The data

The data considered for this application come from a study on neotropical bats within Guyana [13]. The original dataset contains DNA barcodes of 840

[1] *http://www.barcodingoflife.org*

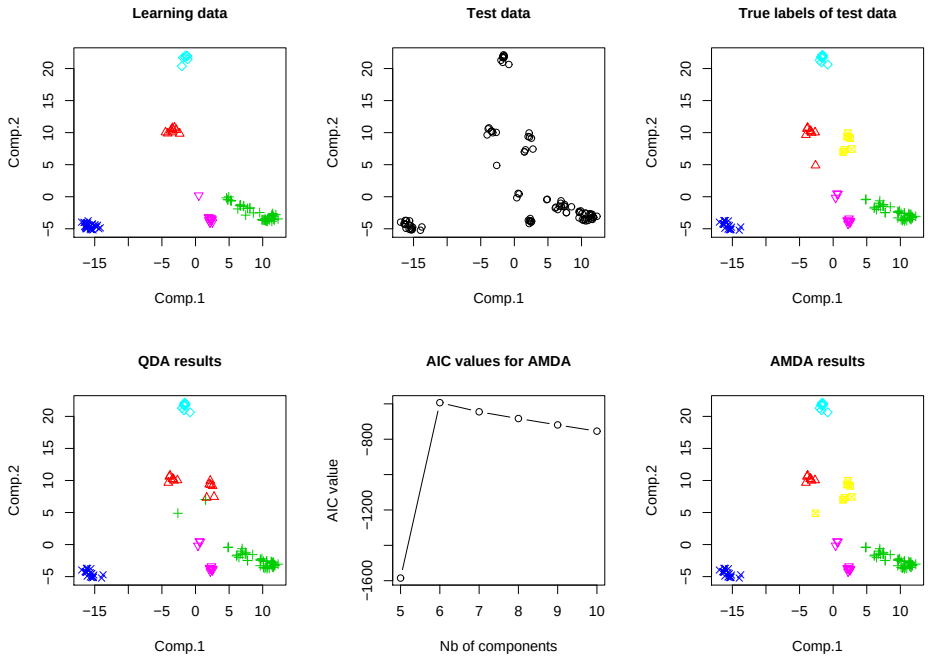

Figure 8: Classification with AMDA of the DNA barcode data: 5 observed bat species and 1 unobserved bat species represented on the two first principal axes.

bat specimens representing 87 species. Each of the 840 DNA sequences has a length of 657 variables and where each variable is coding for the "A", "C", "G", "T" nucleotides. Among the 840 observations of the original set, we only kept for our study the 471 observations belonging to the 6 most represented species (most of the original species only contain few observations). The 657 qualitative variables were also transformed to quantitative variables using Multiple Correspondence Analysis (see [18] for details) and the 176 axes associated to the largest eigenvalues (explaining more than 90% of the total variance) were kept for the remaining of the experiment. Finally, duplicate observations were removed from the dataset since many observations were not unique due to the nature of the data (DNA sequences). The final dataset contains 199 observations from 6 bat species described on 176 quantitative variables.

Truth / Classif.	1	2	3	4	5	6
1	8	0	0	0	0	**6**
2	**1**	42	0	0	0	**2**
3	0	0	20	0	0	0
4	0	0	0	9	0	0
5	0	0	0	0	11	0
-	-	-	-	-	-	-

(a) Confusion matrix for QDA

Truth / Classif.	1	2	3	4	5	6
1	8	2	0	0	0	0
2	0	42	0	0	0	0
3	0	0	20	0	0	0
4	0	0	0	9	0	0
5	0	0	0	0	11	0
6	**1**	0	0	0	0	**8**

(b) Confusion matrix for AMDA

Table 2: Confusion tables for QDA and AMDA on the the DNA barcode test dataset: 5 observed bat species and 1 unobserved bat species (class #6).

5.2 Experimental results

For this experiment, the preprocessed dataset was split into a learning set of 100 observations without observations of the 6th class and a test set of 99 observations containing 8 observations of the 6th class (assumed to be unobserved). The top panels of Figure 8 show the learning and test data on the two first principal axes. We can first observe that the different species are globally well separated and that some classes have inhomogeneous distribution. In particular, the classes #1 (red triangles) and #5 (purple inverse triangles) seem to be made of several sub-species. It appears as well that the observations of the unobserved class (yellow squares) are very close to observations of classes #1, #2 and #5 and this could complicate the discovery and classification tasks. Unsurprisingly, the supervised classifier QDA assigned the observations of the new class to the known classes #1, #2 and #5 as the bottom-left panel of Figure 8 shows. AMDA (inductive approach) was also applied to these data and AIC indicates (*cf.* bottom-center panel) that the most adapted model is a mixture model with 6 components which means that the test set contains 1 unobserved class. The bottom-right panel of Figure 8 shows the final classification of the test dataset provided by AMDA. As we can see, AMDA has correctly detected and classified the 8 observations of the unobserved species but misclassified 1 observations from class #1 (*cf.* Table 2). However, the belonging to the class #1 of this observation could be discussed regarding its DNA sequence. It is indeed natural to suspect a labelling or a sequencing error in this case. To summarize, this experiment has demonstrated the ability of AMDA to detect new species in

a complex real-world context and to adapt the supervised classifier to the new situation.

6 Conclusion and further works

This work has focused on the problem of learning a supervised classifier with unobserved classes. An adaptive model-based discriminant analysis method has been presented in this paper which is able to both detect unobserved groups of points in a new set of observations and to adapt the supervised classifier to the new situation. Two EM-based approaches have been proposed for parameter estimation: an inductive approach, which is made of a learning and a discovering phase, and a transductive approach which considers all available observations for learning in a unique step. The detection of the number of unobserved classes is done using Bayesian model selection criteria. Experiments on simulated and real datasets have shown that the proposed method is able to detect different kinds of unobserved classes (Gaussian, uniform noise, ...). The proposed strategy has been also applied with success to an important problem in Biology: the detection of novel species in DNA barcoding.

It remains however to deal in the future with the problem of label switching when $C - K > 1$. A way to solve this problem could be to ask domain experts to classify some observations of the new detected groups in order to associate a class name with the detected groups. Parsimonious Gaussian models could be used as well for modelling small groups in order to detect unobserved groups of points smaller than 5-10 observations. Finally, it could be very interesting to study the evolution of the proposed strategy in the context of dynamic classification.

A Appendix: proofs of parameter estimators

This ultimate section presents the proofs of parameter estimators given in Section 3 for both inductive and transductive approaches.

A.1 Inductive approach

At the iteration q of the M step, the expectation of the completed log-likelihood $Q(\mathcal{X}^*; \Theta)$ conditionally to the posterior probabilities t_{ik}^* has the

following form:

$$Q(\mathcal{X}^*;\Theta) = \sum_{i=1}^{n^*}\left(\sum_{k=1}^{C} t_{ik}^* \log\left(\pi_k f_k(x_i^*;\theta_k)\right) + \sum_{k=C+1}^{K} t_{ik}^* \log\left(\pi_k f_k(x_i^*;\theta_k)\right)\right),$$

where $\log\left(\pi_k f_k(x_i;\theta_k)\right)$ can be written as follows in the case of the multivariate Gaussian density:

$$\log\left(\pi_k f_k(x_i;\theta_k)\right) = -\frac{1}{2}\left(\log(\pi_k) + \log(|\Sigma_k|) + (x_i-\mu_k)^t\Sigma_k^{-1}(x_i-\mu_k)\right) + C,$$

where $C = -p\log(2\pi)/2$ is a constant which does not depend on mixture parameters. In the case of the discovery phase of the inductive approach, the maximization of $Q(\mathcal{X}^*;\Theta)$ according to the parameters π_k, μ_k and Σ_k can be done classically except that parameters μ_k and Σ_k have only to be estimated for $k = C+1,...,K$. We therefore refer to [30] for ML inference in finite mixture models.

A.2 Transductive approach

At the iteration q of the M step, the expectation of the completed log-likelihood $Q(\mathcal{X},\mathcal{X}^*;\Theta)$ conditionally to the posterior probabilities t_{ik}^* has the following form:

$$Q(\mathcal{X},\mathcal{X}^*;\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{s}_{ik}\log\left(\pi_k f_k(x_i;\theta_k)\right) + \sum_{i=1}^{n^*}\sum_{k=1}^{K} t_{ik}^*\log\left(\pi_k f_k(x_i^*;\theta_k)\right),$$

where $\log\left(\pi_k f_k(x_i;\theta_k)\right)$ is given above. We recall that $\tilde{s}_{ik} = s_{ik}$ if $k = 1,...,C$ and $\tilde{s}_{ik} = 0$ for $k = C+1,...,K$ and that for $i = 1,...,n$.

ML estimator for parameter π_k The maximization of $Q(\mathcal{X},\mathcal{X}^*;\Theta)$ according to the mixture proportion π_k under the constraint $\sum_{k=1}^{K}\pi_k = 1$ is equivalent to find a saddle point of the Lagrangian $\mathcal{L}(\Theta,\omega)$:

$$\mathcal{L}(\Theta,\omega) = Q(\Theta) - \omega\left(\sum_{k=1}^{K}\pi_k - 1\right),$$

where ω is the Lagrangian coefficient. The partial derivative of the Lagrangian $\mathcal{L}(\Theta, \omega)$ according to π_k is:

$$\frac{\partial}{\partial \pi_k}\mathcal{L}(\Theta, \omega) = \frac{(n_k + n_k^*)}{\pi_k} + \omega,$$

where $n_k = \sum_{i=1}^{n} \tilde{s}_{ik}$ and $n_k^* = \sum_{i=1}^{n^*} t_{ik}^*$. The relation $\frac{\partial}{\partial \pi_k}\mathcal{L}(\Theta, \omega) = 0$ implies that, for all $k = 1, ..., K$:

$$(n_k + n_k^*) + \omega \pi_k = 0, \tag{2}$$

and summing up this quantity over k provides the value of the Lagrangian coefficient ω:

$$\omega = n + n^*,$$

where $n = \sum_{k=1}^{K} n_k$ and $n^* = \sum_{k=1}^{K} n_k^*$. Finally, replacing ω by its value in (2) allows to find the ML estimate of π_k:

$$\hat{\pi}_k = \frac{(n_k + n_k^*)}{(n + n^*)}.$$

ML estimator for parameter μ_k The partial derivative of $Q(\mathcal{X}, \mathcal{X}^*; \Theta)$ according to μ_k has the following form:

$$\frac{\partial}{\partial \mu_k} Q(\mathcal{X}, \mathcal{X}^*; \Theta) = -\Sigma_k^{-1} \left(\sum_{i=1}^{n} \tilde{s}_{ik}(x_i - \mu_k) + \sum_{i=1}^{n^*} t_{ik}^*(x_i^* - \mu_k) \right).$$

The relation $\frac{\partial}{\partial \mu_k} Q(\mathcal{X}, \mathcal{X}^*; \Theta) = 0$ implies that:

$$\sum_{i=1}^{n} \tilde{s}_{ik}(x_i - \mu_k) + \sum_{i=1}^{n^*} t_{ik}^*(x_i^* - \mu_k) = 0,$$

which is equivalent to:

$$\mu_k \left(\sum_{i=1}^{n} \tilde{s}_{ik} + \sum_{i=1}^{n^*} t_{ik}^* \right) = \sum_{i=1}^{n} \tilde{s}_{ik} x_i + \sum_{i=1}^{n^*} t_{ik}^* x_i^*,$$

and this finally yields to the ML estimate of μ_k:

$$\hat{\mu}_k = \frac{1}{n_k + n_k^*} \left(\sum_{i=1}^{n} \tilde{s}_{ik} x_i + \sum_{i=1}^{n^*} t_{ik}^* x_i^* \right),$$

where $n_k = \sum_{i=1}^{n} \tilde{s}_{ik}$ and $n_k^* = \sum_{i=1}^{n^*} t_{ik}$.

ML estimator for parameter Σ_k At the optimum for parameter μ_k, the partial derivative of $Q(\mathcal{X}, \mathcal{X}^*; \Theta)$ according to Σ_k has the following form:

$$\begin{aligned} \frac{\partial}{\partial \Sigma_k} Q(\mathcal{X}, \mathcal{X}^*; \Theta) &= -\frac{1}{2} \frac{\partial}{\partial \Sigma_k} \left[\sum_{i=1}^{n} \tilde{s}_{ik} \left(\log(|\Sigma_k|) + (x_i - \hat{\mu}_k)^t \Sigma_k^{-1} (x_i - \hat{\mu}_k) \right) \right. \\ &\quad \left. + \sum_{i=1}^{n^*} t_{ik}^* \left(\log(|\Sigma_k|) + (x_i^* - \hat{\mu}_k)^t \Sigma_k^{-1} (x_i^* - \hat{\mu}_k) \right) \right]. \end{aligned}$$

Using the classical trick of the trace of the 1×1 matrix, we can write that $(x_i - \hat{\mu}_k)^t \Sigma_k^{-1} (x_i - \hat{\mu}_k) = \mathrm{tr}\left((x_i - \hat{\mu}_k)^t \Sigma_k^{-1} (x_i - \hat{\mu}_k) \right)$ and, using the identity $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we get:

$$\frac{\partial}{\partial \Sigma_k} Q(\mathcal{X}, \mathcal{X}^*; \Theta) = -\frac{1}{2} \frac{\partial}{\partial \Sigma_k} \left[(n_k + n_k^*) \log(|\Sigma_k|) + \mathrm{tr}\left(\Sigma_k^{-1} S_k \right) + \mathrm{tr}\left(\Sigma_k^{-1} S_k^* \right) \right],$$

where $S_k = \sum_{i=1}^{n} \tilde{s}_{ik} (x_i - \hat{\mu}_k)^t (x_i - \hat{\mu}_k)$ and $S_k = \sum_{i=1}^{n^*} t_{ik}^* (x_i^* - \hat{\mu}_k)^t (x_i^* - \hat{\mu}_k)$. Using the additivity property of the trace of square matrices, we end up with:

$$\frac{\partial}{\partial \Sigma_k} Q(\mathcal{X}, \mathcal{X}^*; \Theta) = -\frac{1}{2} \frac{\partial}{\partial \Sigma_k} \left[(n_k + n_k^*) \log(|\Sigma_k|) + \mathrm{tr}\left(\Sigma_k^{-1} (S_k + S_k^*) \right) \right].$$

Finally, using the matrix derivative formula of the logarithm of a determinant, $\frac{\partial}{\partial A} \log(|A|) = \left(A^{-1} \right)^t$, and of the trace of a product, $\frac{\partial}{\partial A} \mathrm{tr}(A^{-1} B) = -\left(A^{-1} B A^{-1} \right)^t$, the equality of $\frac{\partial}{\partial \Sigma_k} Q(\mathcal{X}, \mathcal{X}^*; \Theta)$ to the $p \times p$ zero matrix yields to the relation:

$$(n_k + n_k^*) \Sigma_k^{-1} = \Sigma_k^{-1} \left(S_k + S_k^* \right) \Sigma_k^{-1},$$

and, by multiplying on the left and on the right by Σ_k, we find out the ML estimate of Σ_k:

$$\hat{\Sigma}_k = \frac{1}{(n_k + n_k^*)} \left(S_k + S_k^* \right).$$

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] E. Anderson. The irises of the gaspé peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

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

[4] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[5] H. Bensmail and G. Celeux. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:1743–1748, 1996.

[6] 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.

[7] C. Bishop. Novelty detection and neural network validation. In *IEES Conference on Vision and Image Signal Processing*, pages 217–222, 1994.

[8] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory*, 1998.

[9] C. Bouveyron, S. Girard, and C. Schmid. High-Dimensional Data Clustering. *Computational Statistics and Data Analysis*, 52(1):502–519, 2007.

[10] C. Bouveyron, S. Girard, and C. Schmid. High Dimensional Discriminant Analysis. *Communications in Statistics: Theory and Methods*, 36(14):2607–2623, 2007.

[11] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[12] C. Chow. On optimum recognition error and reject tradeoff. In *IEEE Transactions on Information Theory*, pages 41–46, 1970.

[13] E. Clare, B. Lim, M. Engstrom, J. Eger, and P. Hebert. DNA barcoding of neotropical bats: species identification and discovery within guyana. *Molecular Ecology Notes*, 7:184–190, 2007.

[14] D. Dasgupta and F. Nino. A comparison of negative and positive selection algorithms in novel pattern detection. In *IEEE International Conference on Systems and Cybernetics*, pages 125–130, 2000.

[15] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[16] M. Desforges, P. Jacob, and J. Cooper. Applications of probability density estimation to the detection of abnormal conditions in engineering. In *Proc. Institute of Mechanical Engineers*, pages 687–703.

[17] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[18] M. Greenacre and J. Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall, 2006.

[19] L. Hansen, C. Liisberg, and P. Salamon. The error-reject tradeoff. *Open Systems and Information Dynamics*, 4:159–184, 1997.

[20] M. Hellman. The nearest neighbour classification with a reject option. In *IEEE Transactions on Systems Science and Cybernetics,*, pages 179–185, 1970.

[21] C. Hennig and P. Coretto. *The Noise Component in Model-based Cluster Analysi*, pages 127–138. Data Analysis, Machine Learning and Applications. Springer, 2008.

[22] R. Kelly, I. Sarkar, D. Eernisse, and R. Desalle. DNA barcoding using chitons (genus mopalia). *Molecular Ecology Notes*, 7:177–183, 2007.

[23] T. Kohonen. *Self-organisation and associative memory*. Springer-verlag, berlin edition, 1988.

[24] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo. On semi-supervised classification. In *NIPS*, 2004.

[25] C. Manikopoulos and S. Papavassiliou. Network intrusion and fault detection: a statistical anomaly approach. *rk intrusion and fault detection: a IEEE Communications Magazine*, 40(10):76–82, 2002.

[26] M. Markou and S. Singh. Novelty detection: A review - part 1: Statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[27] M. Markou and S. Singh. Novelty detection: A review - part 2: Neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003.

[28] G. McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, (70):365–369, 1975.

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

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

[31] T. Odin and D. Addison. Novelty detection using neural network technology. In *Proc. of COMADEN conference*, 2000.

[32] T. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, (73):821–826, 1978.

[33] S. Roberts. Novelty detection using extreme value statistics. In *IEE Proc. on Vision, Image and Signal Processing*, volume 146, pages 124–129, 1999.

[34] S. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6:270–284, 1994.

[35] J. Ryan, M. Lin, and R. Miikkulainen. Intrusion detection with neural networks. In *Advances in Neural Information Processing Systems*, 1998.

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

[37] M. Seeger. Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, 2001.

[38] B. Shölkopf, R. Williamson, A. Smola, J. Taylor, and J. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, pages 582–588, 2000.

[39] L. Tarassenko. Novelty detection for the identification of masses in mammograms. In *4th IEE International Conference on Artificial Neural Networks*, volume 4, pages 442–447, 1995.

[40] D. Tax and R. Duin. Outlier detection using classifier instability. In *Advances in Pattern Recognition*, pages 251–256, 1999.

[41] D. Yeung and C. Chow. Parzen window network intrusion detectors. In *Proc. of International Conference on Pattern Recognition*, 2002.