Neurocomputing 112 (2013) 153–163

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Model-based functional mixture discriminant analysis with hidden process regression for curve classification

F. Chamroukhi [a,b,*], H. Glotin [a,b,c], A. Samé [d]

[a] Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
[b] Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
[c] Institut Universitaire de France,[1] France
[d] UPE, IFSTTAR, GRETTIA, France

ARTICLE INFO

Available online 5 March 2013

Keywords:
Functional mixture discriminant analysis
Model-based approaches
Curve classification
Hidden process regression
Unsupervised learning
Clustering
EM algorithm

ABSTRACT

In this paper, we study the modeling and the classification of functional data presenting regime changes over time. We propose a new model-based functional mixture discriminant analysis approach based on a specific hidden process regression model that governs the regime changes over time. Our approach is particularly adapted to handle the problem of complex-shaped classes of curves, where each class is potentially composed of several sub-classes, and to deal with the regime changes within each homogeneous sub-class. The proposed model explicitly integrates the heterogeneity of each class of curves via a mixture model formulation, and the regime changes within each sub-class through a hidden logistic process. Each class of complex-shaped curves is modeled by a finite number of homogeneous clusters, each of them being decomposed into several regimes. The model parameters of each class are learned by maximizing the observed-data log-likelihood by using a dedicated expectation–maximization (EM) algorithm. Comparisons are performed with alternative curve classification approaches, including functional linear discriminant analysis and functional mixture discriminant analysis with polynomial regression mixtures and spline regression mixtures. Results obtained on simulated data and real data show that the proposed approach outperforms the alternative approaches in terms of discrimination, and significantly improves the curves approximation.

1. Introduction

Most statistical analyses involve vectorial data when the observations are finite dimensional vectors. However, in many application domains, such as diagnosis of complex systems [8,27], electrical engineering [19], speech recognition (e.g., the phoneme data studied in [11]), radar waveform [9], etc., the data are functions (i.e curves) rather than finite dimensional vectors. The paradigm of analyzing such data is known as functional data analysis (FDA) [25]. The statistical approaches for FDA concern the analysis of data for which the individuals are entire functions or curves rather than finite dimensional vectors. The goals of FDA include data representation for further analysis, data visualization, exploratory analysis by performing clustering or projections, regression, classification, etc. Additional background on FDA, examples and analysis techniques can be found in Ramsay and Silverman [25]. The analysis task leads in general to learning a statistical model namely in a supervised context for classification, regression, or in an unsupervised way for a clustering or a segmentation task [15,8,27,14,4,17,21,11], etc. The challenge is therefore to build adapted models to be learned from such data living in a very high or an infinite dimensional space. In this paper, we consider the problem of supervised functional data classification (discrimination) where the observations are temporal curves presenting several regime changes over time. However, while the global task is supervised, as we shall present it later, this global discrimination task includes two unsupervised tasks. The first one is to automatically cluster possibly dispersed classes into several homogeneous clusters (i.e., sub-classes), and the second one is to automatically determine the temporal regimes of each sub-class which can be seen as a temporal segmentation task.

Indeed, concerning the first point of class dispersion, that is the need of sub-classses, in many areas of application of classification, a class itself may be composed of several unknown (unobserved) sub-classes. The learning has therefore to be treated in an unsupervised way within each class, since no labels of sub-groups are available. For example, in handwritten digit recognition, there are several characteristic ways to write a digit,

* Corresponding author at: Université de Toulon, LSIS, UMR CNRS 7296, Bâtiment R, BP 20132, 83957 La Garde Cedex, France. Tel.: +33 4 94 14 20 06; fax: +33 4 94 14 28 97.
E-mail address: faicel.chamroukhi@univ-tln.fr (F. Chamroukhi).
[1] iuf.amue.fr.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.10.030

and therefore a creation of several sub-classes within the class of digit itself, which may be modeled using a mixture density as in Hastie and Tibshirani [18]. In complex systems diagnosis application, where we have to decide between two classes: without defect/with defect, one would have only the class labels as either with or without defect, however no labels according to how a defect would happen, namely the type of defect, the degree of defect, means minor, critical, etc. Thus, providing an automatic tool that decomposes the class into sub-classes would be very helpful in making accurate decisions as well as for well interpretation. Another example is the one of gene function classification based on time course gene expression data. As stated in Gu and Li [17] when considering the complexity of the gene functions, one functional class may includes genes which involve one or more biological profiles. Describing each class as a combination of sub-classes, which unfortunately are very often unknown, is necessary to provide realistic description, rather than providing a rough representation.

We mainly focus on generative approaches, in particular latent variable models, which are dedicated to explain the underlying processes generating the data. As it should be explicitly described later, this can be achieved by explicitly integrating the problem of class dispersion and the one of the regimes changes over time into a two-level latent data model. The generative approaches for functional data related to this work are essentially based on regression analysis, including polynomial regression, splines and B-splines [14,4,17,21], or also generative polynomial piecewise regression as in Chamroukhi [4] and Chamroukhi et al. [8]. Non-parametric statistical approaches have also been proposed for functional data discrimination as in Ferraty and Vieu [13], Delaigle et al. [11] and clustering as in Delaigle et al. [11]. Another possible curve modeling can be the Gaussian processes approach [26] which is a non-parametric approach that has been used in functional data analysis [20,29,30], one can site in particular recent Gaussian Processes for functional regression [30]. While they are used as a non-parametric approach, inference in such models requires performing MCMC sampling and direct implementation is computationally expensive. In this paper, we focused on parametric approaches where the computation of conditional expectations is analytic. The model parameters can be further used for summarizing a set of curves into a parameter vector. This is useful for example for a feature extraction prospective [6]. Furthermore, while Gaussian process approaches are well adapted to approximate and to cluster non-linear functions or curves as in Hierarchical Gaussian process mixtures for regression [30,20], the problem of regime changes within each set of curves is still not taken into account in such approaches; only a non-linear approximation is provided, without segmentation. In this paper, we propose a new generative approach for modeling classes of complex-shaped curves where each class is itself composed of unknown homogeneous sub-classes. In addition, the model is particularly dedicated to address the problem when each homogeneous sub-class presents regime changes over time. Here we extend the functional discriminant analysis approach presented in Chamroukhi et al. [8], which relates modeling each class of curves presenting regime changes with a single mean curve, to a mixture formulation which leads to a functional mixture-model based discriminant analysis. More specifically, this approach uses a mixture of regression models with hidden logistic processes (RHLP) [4,27] for each class of functional data, and derives a functional mixture discriminant analysis framework for functional data classification. The resulting discrimination approach is therefore a model-based functional discriminant analysis in which learning the parameters of each class of curves is achieved through an unsupervised estimation of a mixture of RHLP (MixRHLP) models. A first idea of this approach was presented in Chamroukhi [4] and Chamroukhi et al. [5].

In the next section we give a brief background on discriminant analysis approaches for functional data classification including functional linear and mixture discriminant analysis, and then we present the proposed model-based functional mixture discriminant analysis with hidden process regression for curve classification, which we will abbreviate as FMDA-MixRHLP, and the corresponding parameter estimation procedure using a dedicated expectation–maximization (EM) algorithm. Then, we will present the model selection using the Bayesian Information Criterion (BIC) [28].

In the following we denote by $((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n))$ a given labeled training set of curves issued from G classes where $y_i \in \{1,\ldots,G\}$ is the class label of the ith curve $\mathbf{x}_i$. We assume that $\mathbf{x}_i$ consists of m observations $(x_{i1},\ldots,x_{im})$, regularly observed at the time points $(t_1,\ldots,t_m)$ with $t_1 < \cdots < t_m$.

2. Background on functional discriminant analysis

In this section, we give a background on generative discriminant analysis approaches for functional data classification based on functional regression. Functional discriminant analysis approaches extend discriminant analysis approaches for vectorial data to functional data or curves. The functional discriminant analysis principle is as follows. Assume that we have a labeled training set of curves and the classes' parameter vectors $(\boldsymbol{\Psi}_1,\ldots,\boldsymbol{\Psi}_G)$ where $\boldsymbol{\Psi}_g$ is the parameter vector of the density of class g $(g=1,\ldots,G)$ (e.g., provided by an estimation procedure from a training set). In functional discriminant analysis, a new curve $\mathbf{x}_i$ is assigned to the class $\hat{y}_i$ using the maximum a posteriori (MAP) rule, that is

$$\hat{y}_i = \arg\max_{1\le g\le G} \frac{w_g p(\mathbf{x}_i|y_i=g,\mathbf{t};\boldsymbol{\Psi}_g)}{\sum_{g'=1}^{G} w_{g'} p(\mathbf{x}_i|y_i=g',\mathbf{t};\boldsymbol{\Psi}_{g'})}, \quad (1)$$

where $w_g = p(y_i=g)$ is the prior probability of class g, which can be computed as the proportion of the class g in the training set, and $p(\mathbf{x}_i|y_i=g,\mathbf{t};\boldsymbol{\Psi}_g)$ its conditional density.

Different ways are possible to model this conditional density. By analogy to linear or quadratic discriminant analysis for vectorial data, the class conditional density for each class of curves can be defined as a density of a single model, e.g., a polynomial regression model, spline, including B-spline [21], or a generative piecewise regression model with a hidden logistic process (RHLP) [8] when the curves further present regime changes over time. These approaches lead to Functional Linear (or quadratic) Discriminant Analysis which we will abbreviate as (FLDA).

In the next section, we briefly recall the FLDA based on polynomial or spline regression.

2.1. Functional linear discriminant analysis

Functional linear discriminant analysis (FLDA), firstly proposed in James and Hastie [21] for irregularly sampled curves, arises when we model each class conditional density of curves with a single model. More specifically, the conditional density $p(\mathbf{x}_i|y=g,\mathbf{t};\boldsymbol{\Psi}_g)$ in Eq. (1) can for example be the one of a polynomial, spline or B-spline regression model with parameters $\boldsymbol{\Psi}_g$, that is:

$$p(\mathbf{x}_i|y_i=g,\mathbf{t};\boldsymbol{\Psi}_g) = \mathcal{N}(\mathbf{x}_i;\mathbf{T}\boldsymbol{\beta}_g,\sigma_g^2\mathbf{I}_m), \quad (2)$$

where $\boldsymbol{\beta}_g$ is the coefficient vector of the polynomial or spline regression model representing class g and σ_g^2 the associated noise variance, the matrix $\mathbf{T}$ is the matrix of design which depends on the adopted model (e.g., for polynomial regression, $\mathbf{T}$ is the

$m \times (p+1)$ Vandermonde matrix with rows $(1, t_j, t_j^2, \ldots, t_j^p)$ for $j = 1, \ldots, m$, p being the polynomial degree), and $\mathcal{N}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In this case, estimating the model for each class consists therefore in estimating the regression model parameters $\boldsymbol{\Psi}_g$, namely by maximum likelihood which is equivalent to performing least squares estimation in this Gaussian case. A similar FLDA approach that fits a specific generative piecewise regression model governed by a hidden logistic process to homogeneous classes of curves presenting regime changes has been presented in Chamroukhi et al. [8] and Chamroukhi [4].

However, all these approaches, as they involve a single component density model for each class, are only suitable for homogeneous classes of curves. For complex-shaped classes, when one or more classes are dispersed, the hypothesis of a single component density model description for the whole class of curves becomes restrictive. This problem can be handled, by analogy to mixture discriminant analysis for vectorial data [18], by adopting a mixture model formulation [24,32] in the functional space for each class of curves. The functional mixture can for example be a polynomial regression mixture or a spline regression mixture [14,4,17]. This leads to functional mixture discriminant analysis (FMDA) [5,4,17].

The next section describes the previous work on FMDA which uses polynomial regression and spline regression mixtures.

2.2. Functional mixture discriminant Analysis with polynomial regression and spline regression mixtures

A first idea on functional mixture discriminant analysis (FMDA), motivated by the complexity of the time course gene expression functional data for which modeling each class with a single function using FLDA is not adapted, was proposed by Gui and Li [17] and is based on B-spline regression mixtures. In the approach of Gui and Li [17], each class g of functions is modeled as a mixture of K_g sub-classes, each sub-class k ($k = 1, \ldots, K_g$) is a noisy B-spline function (can also be a polynomial or a spline function) with parameters $\boldsymbol{\Psi}_{gk}$. The model is therefore defined by the following conditional mixture density:

$$p(\mathbf{x}_i | y_i = g, \mathbf{t}; \boldsymbol{\Psi}_g) = \sum_{k=1}^{K_g} \alpha_{gk}\, p(\mathbf{x}_i | y_i = g, z_{gi} = k, \mathbf{t}; \boldsymbol{\Psi}_{gk}) = \sum_{k=1}^{K_g} \alpha_{gk} \mathcal{N}(\mathbf{x}_i; \mathbf{T}\boldsymbol{\beta}_{gk}, \sigma_{gk}^2 \mathbf{I}_m), \quad (3)$$

where the α_{gk}'s are the non-negative mixing proportions that sum to 1 such that $\alpha_{gk} = p(z_i = k | y_i = g)$ (α_{gk} represents the prior probability of the sub-class k of class g), z_i is a hidden discrete variable in $\{1, \ldots, K_g\}$ representing the labels of the sub-classes for each class, and $\mathbf{I}_m$ is the m dimensional identity matrix. The parameters of this functional mixture density for each class g (Eq. (3)) denoted by

$$\boldsymbol{\Psi}_g = (\alpha_{g1}, \ldots, \alpha_{gK_g}, \boldsymbol{\Psi}_{g1}, \ldots, \boldsymbol{\Psi}_{gK_g})$$

can be estimated by maximizing the observed-data log-likelihood by using the expectation–maximization (EM) algorithm [12,23] as in Gui and Li [17].

However, using polynomial or spline regression for class representation, as studied in Chamroukhi [4] and Chamroukhi et al. [8], is more adapted for curves presenting smooth regime changes and for the splines the knots have to be fixed in advance. When the regime changes are abrupt, capturing the regime transition points needs to relax the regularity constraints on splines, since a spline is a smooth function [10], which leads to piecewise regression [22] for which the knots can be optimized using a dynamic programming procedure [1,31]. On the other hand, the regression model with a hidden logistic process (RHLP) presented in Chamroukhi et al. [8] and used to model each homogeneous set of curves with regime changes is flexible and explicitly integrates the smooth and/or abrupt regime changes via a logistic process. As pointed in Chamroukhi et al. [8], this approach however has limitations in the case of complex-shaped classes of curves since each class is only approximated by a single RHLP model.

In this paper, we extend the discrimination approach proposed in Chamroukhi et al. [8], which is based on functional linear discriminant analysis (FLDA) using a single density model (RHLP) for each class, to a functional mixture discriminant analysis framework (FMDA), where each class conditional density model is assumed to be a specific mixture density. This density is a mixture of regression models with hidden logistic processes, which we will abbreviate as MixRHLP. Thus, by using this functional mixture discriminant analysis approach, we may therefore overcome the limitation of FLDA (and FQDA) for modeling complex-shaped classes of curves, via the mixture formulation. Furthermore, thanks to the flexibility of the RHLP model that approximates each sub-class, as studied in Chamroukhi et al. [7,8], we will also be able to automatically and flexibly approximate the underlying hidden regimes for each sub-class.

The proposed functional mixture discriminant analysis with hidden process regression and the unsupervised learning procedure for each class through the EM algorithm are presented in the next section.

3. Proposed functional mixture discriminant analysis with hidden process regression mixture

Let us assume as previously that each class g has a complex shape so that it is composed of K_g homogeneous sub-classes. Furthermore, now let us suppose that each sub-class k of class g is itself governed by L_{gk} unknown regimes.

3.1. Modeling the classes of curves with a mixture of regression models with hidden logistic processes

In the proposed functional mixture discriminant analysis (FMDA) approach, we model each class of curves by a specific mixture of regression models with hidden logistic processes (MixRHLP) as in Chamroukhi [4] and Samé et al. [27]. The approach will thus be abbreviated as FMDA-MixRHLP. According to the MixRHLP model, each class of curves g is assumed to be composed of K_g homogeneous sub-groups with prior probabilities $\alpha_{g1}, \ldots, \alpha_{gK_g}$. Each of the K_g sub-groups is governed by L_{gk} hidden polynomial regimes. Thus, for the ith curve $\mathbf{x}_i$ issued from sub-class k of class g, the observation x_{ij} may switch from one regime to another at each time point t_j.

We let z_{gi} denote the variable representing the unobserved (hidden) sub-class (cluster) label of the ith curve $\mathbf{x}_i$ for class g. We have therefore $z_{gi} = k \in \{1, \ldots, K_g\}$ for sub-class k. We will thus denote by $\mathbf{z}_g = (z_{g1}, \ldots, z_{gn})$ the hidden cluster labels for class g.

Furthermore, we let r_{gkj} denote the unobserved regime label for sub-class k of class g at time t_j. Thus, we have $r_{gkj} = \ell \in \{1, \ldots, L_{gk}\}$ for regime ℓ. The variable r_{gkj} allows for switching from one regime to another among L_{gk} regimes over time. We let therefore $\mathbf{h}_{gk} = (h_{gk1}, \ldots, h_{gkm})$ denote the labels vector governing each sub-class k of class g.

In the following, we will encode each of the random variables z and r in a binary manner as follows. The variable z will be indexed by the three indexes

- g: the group (class)
- k: the sub-class (cluster)
- i: the observation, which is the ith curve in this case,

so that we have z_{gki} equals 1 if and only if the sub-class label of the curve $\mathbf{x}_i$ belonging to class g is k, that is $z_{gi}=k$. Similarly, we will use a binary coding for the variable r which will be indexed by the four indexes

- g: the group (class)
- k: the sub-class (cluster)
- ℓ: the regime
- j: the observation at time t_j,

so that we have $r_{gk\ell j}$ equals 1 if and only if, for sub-class k of class g, the regime at time t_j is ℓ, that is $r_{gkj}=\ell$.

The distribution of each configuration of the discrete variable r_{gkj} is assumed to be logistic, thus $\mathbf{h}_{gk}$ governing each sub-class is therefore assumed to be a logistic process. This choice is due to the flexibility of the logistic function in both determining the regime transition points and accurately modeling abrupt and/or smooth regime changes. Indeed, as it has been well detailed in Chamroukhi et al. [7,8], the logistic function (4) controls through its parameters the regime transition points and the quality of regime (smooth or abrupt) via the parameters $\{w_{gk\ell 0}, w_{gk\ell 1}\}$. The probability of each regime ℓ is given by

$$\pi_{gk\ell}(t_j;\mathbf{w}_{gk}) = p(r_{gkj}=\ell|t_j;\mathbf{w}_{gk}) = \frac{\exp(w_{gk\ell 0}+w_{gk\ell 1}t_j)}{\sum_{u=1}^{L_{gk}}\exp(w_{gk\ell u0}+w_{gk\ell u1}t_j)}, \tag{4}$$

where $\mathbf{w}_{gk}=(\boldsymbol{w}_{gk1},\ldots,\boldsymbol{w}_{gkL_{gk}})$ is its parameter vector, $\boldsymbol{w}_{gk\ell}=(w_{gk\ell 0}, w_{gk\ell 1})$ being the two-dimensional coefficient vector for the ℓth logistic component. Furthermore, the regimes are assumed to be noisy polynomial functions and the resulting model for each sub-class is the regression model with hidden logistic process (RHLP) [7,8]. The RHLP model indeed assumes that the curves of each sub-class (or cluster) k of class g are generated by K_g polynomial regression models governed by a hidden logistic process $\mathbf{h}_{gk}$. Thus, as stated in Chamroukhi et al. [7,8], the distribution of a curve $\mathbf{x}_i$ belonging to sub-class k of class g is defined by

$$p(\mathbf{x}_i|y_i=g,z_{gi}=k,\mathbf{t};\boldsymbol{\Psi}_{gk}) = \prod_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\pi_{gk\ell}(t_j;\mathbf{w}_{gk})\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk\ell}^T\mathbf{t}_j,\sigma_{gk\ell}^2) \tag{5}$$

where $\boldsymbol{\Psi}_{gk}=(\mathbf{w}_{gk},\boldsymbol{\beta}_{gk1},\ldots,\boldsymbol{\beta}_{gkL_{gk}},\sigma_{gk1}^2,\ldots,\sigma_{gkL_{gk}}^2)$ for $(g=1,\ldots,G;\ k=1,\ldots,K_g)$ is its parameter vector. Hence, the resulting conditional distribution of a curve $\mathbf{x}_i$ issued from class g is given by the following conditional mixture density:

$$\begin{aligned} p(\mathbf{x}_i|y_i=g,\mathbf{t};\boldsymbol{\Psi}_g) &= \sum_{k=1}^{K_g} p(z_i=k|y_i=g)p(\mathbf{x}_i|y_i=g,z_i=k,\mathbf{t};\boldsymbol{\Psi}_{gk}) \\ &= \sum_{k=1}^{K_g}\alpha_{gk}\prod_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\pi_{gk\ell}(t_j;\mathbf{w}_{gk})\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk\ell}^T\mathbf{t}_j,\sigma_{gk\ell}^2) \end{aligned} \tag{6}$$

where $\boldsymbol{\Psi}_g=(\alpha_{g1},\ldots,\alpha_{gK_g},\boldsymbol{\Psi}_{g1},\ldots,\boldsymbol{\Psi}_{gK_g})$ is the parameter vector for class g, $\boldsymbol{\Psi}_{gk}$ being the parameters of each of its RHLP component density, that is $\prod_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\pi_{gk\ell}(t_j;\mathbf{w}_{gk})\ \mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk\ell}^T\mathbf{t}_j, \sigma_{gk\ell}^2)$ as given by Eq. (5). Notice that the key difference between the proposed FMDA with hidden process regression mixture and the FMDA proposed in Gui and Li [17] is that the proposed approach uses a generative hidden process regression model (RHLP) for each sub-class rather than a spline; the RHLP is itself based on a dynamic mixture formulation as it can be seen in Eq. (5). Thus, the proposed approach is more adapted for capturing the regime changes within curves during time.

Now, once we have defined the model for each class of curves g, we have to estimate its parameters $\boldsymbol{\Psi}_g$. The next section presents the unsupervised learning of the model parameters $\boldsymbol{\Psi}_g$ for each class of curves by maximizing the observed-data log-likelihood through the EM algorithm.

3.2. Maximum likelihood estimation via the EM algorithm

Given an independent training set of labeled curves $((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n))$, the parameter vector $\boldsymbol{\Psi}_g$ of the mixture density of class g given by Eq. (6) is estimated by maximizing the following observed-data log-likelihood:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Psi}_g) &= \log\prod_{i|y_i=g} p(\mathbf{x}_i|y_i=g,\mathbf{t};\boldsymbol{\Psi}_g) \\ &= \sum_{i|y_i=g}\log\sum_{k=1}^{K_g}\alpha_{gk}\prod_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\pi_{gk\ell}(t_j;\mathbf{w}_{gk})\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk\ell}^T\mathbf{t}_j,\sigma_{gk\ell}^2). \end{aligned}$$

The maximization of this log-likelihood cannot be performed in a closed form. We maximize it iteratively by using a dedicated EM algorithm. The EM scheme requires the definition of the complete-data log-likelihood. The complete-data log-likelihood for the proposed MixRHLP model for each class, given the observed data which we denote by $\mathcal{D}_g=(\{\mathbf{x}_i|y_i=g\},\mathbf{t})$, the hidden cluster labels $\mathbf{z}_g$, and the hidden processes $\{\mathbf{h}_{gk}\}$ governing the K_g clusters, is given by

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\Psi}_g) = \sum_{i|y_i=g}\sum_{k=1}^{K_g} z_{gki}\Bigg[&\log\alpha_{gk}+\sum_{j=1}^{m}\sum_{\ell=1}^{L_{gk}} r_{gk\ell j}\log\pi_{gk\ell}(t_j;\mathbf{w}_{gk}) \\ &+\sum_{j=1}^{m}\sum_{\ell=1}^{L_{gk}} r_{gk\ell j}\log\mathcal{N}(y_{ij};\boldsymbol{\beta}_{gk\ell}^T\mathbf{t}_j,\sigma_{gk\ell}^2)\Bigg]. \end{aligned} \tag{7}$$

The next paragraph shows how the observed-data log-likelihood $\mathcal{L}(\boldsymbol{\Psi}_g)$ is maximized by the EM algorithm.

3.3. The dedicated EM algorithm for the unsupervised learning of the parameters of the MixRHLP model for each class

For each class g, the EM algorithm starts with an initial parameter $\boldsymbol{\Psi}_g^{(0)}$ and alternates between the two following steps until convergence:

3.3.1. E-step

This step computes the expected complete-data log-likelihood, given the observations $\mathcal{D}_g$, and the current parameter estimation $\boldsymbol{\Psi}_g^{(q)}$, q being the current iteration number:

$$Q(\boldsymbol{\Psi}_g,\boldsymbol{\Psi}_g^{(q)}) = \mathbb{E}[\mathcal{L}_c(\boldsymbol{\Psi}_g)|\mathcal{D}_g;\boldsymbol{\Psi}_g^{(q)}].$$

As it can be seen in the expression of $\mathcal{L}_c(\boldsymbol{\Psi}_g)$, this step simply requires the calculation of conditional expectations of the variables z_{gki} and $r_{gk\ell j}$. More specifically, the expected complete-data log-likelihood is given by

$$\begin{aligned} Q(\boldsymbol{\Psi}_g,\boldsymbol{\Psi}_g^{(q)}) = \mathbb{E}[\mathcal{L}_c(\boldsymbol{\Psi}_g)|\mathcal{D}_g;\boldsymbol{\Psi}_g^{(q)}] &= \sum_{i|y_i=g}\sum_{k=1}^{K}\gamma_{gki}^{(q)}\log\alpha_{gk} \\ &+\sum_{i|y_i=g}\sum_{k=1}^{K_g}\sum_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}\log\pi_{gk\ell}(t_j;\mathbf{w}_{gk}) \\ &+\sum_{i|y_i=g}\sum_{k=1}^{K_g}\sum_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}\log\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk\ell}^T\mathbf{t}_j,\sigma_{gk\ell}^2). \end{aligned} \tag{8}$$

As shown in the expression of $Q(\boldsymbol{\Psi}_g,\boldsymbol{\Psi}_g^{(q)})$, this step simply requires the calculation of the posterior sub-class probabilities, i.e., the probability that the observed curve $\mathbf{x}_i$ originates from sub-class (cluster) k for class g, which we index in a similar way as z_{gki} and are given by

$$\begin{aligned}\gamma_{gki}^{(q)} &= p(z_{gi}=k|\mathbf{x}_i,y_i=g,\mathbf{t};\boldsymbol{\Psi}_{gk}^{(q)})\\ &= \frac{\alpha_{gk}^{(q)}p(\mathbf{x}_i|y_i=g,z_{gi}=k,\mathbf{t};\boldsymbol{\Psi}_{gk}^{(q)})}{\sum_{k'=1}^{K_g}\alpha_{gk'}^{(q)}p(\mathbf{x}_i|y_i=g,z_{gi}=k',\mathbf{t};\boldsymbol{\Psi}_{gl}^{(q)})}\\ &= \frac{\alpha_{gk}^{(q)}\prod_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\pi_{gk\ell}(t_j;\mathbf{w}_{gk}^{(q)})\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk\ell}^{T(q)}\mathbf{t}_j,\sigma_{gk\ell}^{2(q)})}{\sum_{k'=1}^{K_g}\alpha_{gk'}^{(q)}\prod_{j=1}^{m}\sum_{\ell=1}^{R_{gk'}}\pi_{gk'r}(t_j;\mathbf{w}_{gk'}^{(q)})\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk'\ell}^{T(q)}\mathbf{t}_j,\sigma_{gk'\ell}^{2(q)})}\end{aligned} \tag{9}$$

and the posterior regime probabilities for each sub-class, i.e., the probability that the observed data point x_{ij} at time t_j originates from the ℓth regime of sub-class k for class g, which we index in a similar way as $r_{gk\ell j}$ and are given by

$$\begin{aligned}\tau_{gk\ell ij}^{(q)} &= p(r_{gkj}=\ell|x_{ij},y_i=g,z_{gi}=k,t_j;\boldsymbol{\Psi}_{gk}^{(q)})\\ &= \frac{\pi_{gk\ell}(t_j;\mathbf{w}_{gk}^{(q)})\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gk\ell}^{T(q)}\mathbf{t}_j,\sigma_{gk\ell}^{2(q)})}{\sum_{l=1}^{L_{gk}}\pi_{gkl}(t_j;\mathbf{w}_{gk}^{(q)})\mathcal{N}(x_{ij};\boldsymbol{\beta}_{gkl}^{T(q)}\mathbf{t}_j,\sigma_{gkl}^{2(q)})}.\end{aligned} \tag{10}$$

3.3.2. M-step

This step updates the value of the parameter $\boldsymbol{\Psi}_g$ by maximizing the function $Q(\boldsymbol{\Psi}_g,\boldsymbol{\Psi}_g^{(q)})$ given by Eq. (8) with respect to $\boldsymbol{\Psi}_g$, that is

$$\boldsymbol{\Psi}_g^{(q+1)} = \arg\max_{\boldsymbol{\Psi}_g} Q(\boldsymbol{\Psi}_g,\boldsymbol{\Psi}_g^{(q)}).$$

It can be shown that this maximization can be performed by separate maximizations w.r.t. the mixing proportions $(\alpha_{g1},\ldots,\alpha_{gK_g})$ subject to the constraint $\sum_{k=1}^{K_g}\alpha_{gk}=1$, and w.r.t. the regression parameters $\{\boldsymbol{\beta}_{gk\ell},\sigma_{gk\ell}^2\}$ and the hidden logistic processes' parameters $\{\mathbf{w}_{gk}\}$.

The mixing proportions' updates are given, as in the case of standard mixtures, by

$$\alpha_{gk}^{(q+1)} = \frac{1}{n_g}\sum_{i|y_i=g}\gamma_{gki}^{(q)} \quad (k=1,\ldots,K_g), \tag{11}$$

n_g being the cardinal number of class g. The maximization w.r.t. the regression parameters consists in performing separate analytic solutions of weighted least-squares problems where the weights are the product of the posterior probability $\gamma_{gki}^{(q)}$ of sub-class k and the posterior probability $\tau_{gk\ell ij}^{(q)}$ of regime ℓ of sub-class k. Thus, the regression coefficient updates are given by

$$\boldsymbol{\beta}_{gk\ell}^{(q+1)} = \left[\sum_{i|y_i=g}\sum_{j=1}^{m}\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}\mathbf{t}_j\mathbf{t}_j^T\right]^{-1}\sum_{i|y_i=g}\sum_{j=1}^{m}\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}x_{ij}\mathbf{t}_j \tag{12}$$

and the updates for the variances are given by

$$\sigma_{gk\ell}^{2(q+1)} = \frac{\sum_{i|y_i=g}\sum_{j=1}^{m}\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}(x_{ij}-\boldsymbol{\beta}_{gk\ell}^{T(q+1)}\mathbf{t}_j)^2}{\sum_{i|y_i=g}\sum_{j=1}^{m}\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}}. \tag{13}$$

Finally, the maximization w.r.t. the logistic process parameters $\{\mathbf{w}_{gk}\}$ consists in solving multinomial logistic regression problems weighted by $\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}$ which we solve with a multi-class IRLS algorithm. The IRLS algorithm (e.g., see [16,4]) is an iterative algorithm which consists of starting with an initial parameter vector $\mathbf{w}_{gk}^{(0)}$, and updating the estimation until convergence. A single update of the IRLS algorithm at iteration l is given by

$$\mathbf{w}_{gk}^{(l+1)} = \mathbf{w}_{gk}^{(l)} - \left[\frac{\partial^2 Q_{\mathbf{w}_{gk}}^{(q)})}{\partial\mathbf{w}_{gk}\partial\mathbf{w}_{gk}^T}\right]_{\mathbf{w}_{gk}=\mathbf{w}_{gk}^{(l)}}^{-1}\left.\frac{\partial Q_{\mathbf{w}_{gk}}^{(q)}}{\partial\mathbf{w}_{gk}}\right|_{\mathbf{w}_{gk}=\mathbf{w}_{gk}^{(l)}}, \tag{14}$$

where $Q_{\mathbf{w}_{gk}}^{(q)}$ denotes the term in the Q-function (8) that depend on $\mathbf{w}_{gk}$, that is $\sum_{i|y_i=g}\sum_{k=1}^{K_g}\sum_{j=1}^{m}\sum_{\ell=1}^{L_{gk}}\gamma_{gki}^{(q)}\tau_{gk\ell ij}^{(q)}\log\pi_{gk\ell}(t_j;\mathbf{w}_{gk})$. The parameter update $\mathbf{w}_{gk}^{(q+1)}$ is then taken at convergence of the IRLS algorithm (14).

The pseudo-code in Algorithm 1 summarizes the EM algorithm for the proposed MixRHLP model for each class.

Algorithm 1. Pseudo-code of the proposed algorithm for the MixRHLP model for a set of curves.

Inputs: Labeled training set of curves $((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n))$ sampled at the time points $\mathbf{t}=(t_1,\ldots,t_m)$, the number of sub-classes K_g, the number of regimes L_{gk} and the polynomial degree p.

Initialize: $\boldsymbol{\Psi}_g^{(0)}=(\alpha_{g1}^{(0)},\ldots,\alpha_{gK_g}^{(0)},\boldsymbol{\Psi}_{g1}^{(0)},\ldots,\boldsymbol{\Psi}_{gK_g}^{(0)})$
fix a threshold $\epsilon>0$ (e.g., $\epsilon=10^{-6}$),
set $q\leftarrow 0$ (EM iteration)
while increment in log-likelihood $>\epsilon$ **do**
  // E-Step
  for $k=1,\ldots,K_g$ **do**
    compute $\gamma_{gki}^{(q)}$ for $i=1,\ldots,n$ using Eq. (9)
    for $r=1,\ldots,L_{gk}$ **do**
      compute $\tau_{gk\ell ij}^{(q)}$ for $i=1,\ldots,n$; $j=1,\ldots,m$ using Eq. (10)
    end for
  end for
  // M-Step
  for $k=1,\ldots,K_g$ **do**
    compute the update $\alpha_{gk}^{(q+1)}$ using Eq. (11)
    for $r=1,\ldots,L_{gk}$ **do**
      compute the update $\boldsymbol{\beta}_{gk\ell}^{(q+1)}$ using Eq. (12)
      compute the update $\sigma_{gk\ell}^{2(q+1)}$ using Eq. (13)
    end for
    //IRLS updating loop (Eq. (14))
    Initialize: set $\mathbf{w}_{gk}^{(l)}=\mathbf{w}_{gk}^{(q)}$
    set a threshold $\zeta>0$
    $l\leftarrow 0$ (IRLS iteration)
    while increment in $Q_{\mathbf{w}_{gk}}>\zeta$ **do**
      compute $\mathbf{w}_{gk}^{(l+1)}$ using Eq. (14)
      $l\leftarrow l+1$
    end while
    $\mathbf{w}_{gk}^{(q+1)}\leftarrow\mathbf{w}_{gk}^{(l)}$
    $q\leftarrow q+1$
  end for
end while
$\hat{\boldsymbol{\Psi}}=(\alpha_{g1}^{(q)},\ldots,\alpha_{gK_g}^{(q)},\boldsymbol{\Psi}_{g1}^{(q)},\ldots\boldsymbol{\Psi}_{gK_g}^{(q)})$

Output: $\hat{\boldsymbol{\Psi}}_g$ the maximum likelihood estimate of $\boldsymbol{\Psi}_g$

3.4. Curve classification and approximation with the FMDA-MixRHLP approach

This section relates to the approximation of each class of curves by a single or several curve models in the case of a dispersed class, and the class prediction for new observed curves based on the learned class parameters. Once we have an estimate $\hat{\boldsymbol{\Psi}}_g$ of the parameters of the functional mixture density MixRHLP (provided by the EM algorithm) for each class, a new curve $\mathbf{x}_i$ is then assigned to the class maximizing the posterior probability (MAP principle) using Eq. (1). This therefore leads us to

the functional mixture discriminant analysis classification rule (FMDA-MixRHLP) which is particularly adapted to deal with the problem of classes composed of several sub-classes and to further handle the problem of regime changes within each sub-class.

Concerning the curves approximation, each sub-class k of class g is summarized by approximating it by a single "mean" curve, which we denote by $\hat{\mathbf{x}}_{gk}$. Each point $\hat{x}_{gkj}$ $(j=1,\ldots,m)$ of this mean curve is defined by the conditional expectation $\hat{x}_{gkj}=\mathbb{E}[x_{ij}|y_i=g,\ z_{gi}=k,t_j;\hat{\boldsymbol{\Psi}}_{gk}]$ given by

$$\begin{aligned}\hat{x}_{gkj} &= \int_{\mathbb{R}} x_{ij}\, p(x_{ij}|y_i=g,z_{gi}=k,t_j;\hat{\boldsymbol{\Psi}}_{gk})\, dx_{ij}\\ &= \int_{\mathbb{R}} x_{ij} \sum_{\ell=1}^{L_{gk}} \pi_{gk\ell}(t_j;\hat{\mathbf{w}}_{gk})\mathcal{N}(x_{ij};\hat{\boldsymbol{\beta}}_{gk\ell}^T\mathbf{t}_j,\hat{\sigma}^2_{gk\ell})\, dx_{ij}\\ &= \sum_{\ell=1}^{L_{gk}} \pi_{gk\ell}(t_j;\hat{\mathbf{w}}_{gk})\hat{\boldsymbol{\beta}}_{gk\ell}^T\mathbf{t}_j \end{aligned} \tag{15}$$

which is a sum of polynomials weighted by the logistic probabilities $\pi_{gk\ell}$ that model the regime variability over time.

3.5. Model selection

The number of sub-classes (clusters) K_g for each class g $(g=1,\ldots,G)$ and the number regimes L_{gk} for each sub-class can be computed by maximizing some information criteria e.g., the Bayesian Information Criterion (BIC) [28]:

$$\mathrm{BIC}(K,R,p)=\mathcal{L}(\hat{\boldsymbol{\Psi}}_g)-\frac{\nu_{\boldsymbol{\Psi}_g}}{2}\log(n), \tag{16}$$

where $\hat{\boldsymbol{\Psi}}_g$ is the maximum likelihood estimate of the parameter vector $\boldsymbol{\Psi}_g$ provided by the EM algorithm, $\nu_{\boldsymbol{\Psi}_g}=K_g-1+\sum_{k=1}^{K_g}\nu_{\boldsymbol{\Psi}_{gk}}$ is the number of free parameters of the MixRHLP model, K_g-1 being the number of mixing proportions and $\nu_{\boldsymbol{\Psi}_{gk}}=(p+4)L_{gk}-2$ represents the number of free parameters of each RHLP model associated with sub-class k, and n is the number of curves belonging to the training set of the considered class. Note that in Gui and Li [17] the number of sub-classes are fixed by the user.

In practice, the model selection procedure consists in specifying a maximum values of $(K_{max},R_{max},p_{max})$ and then running the EM algorithm on each class of curves for $k=1,\ldots,K_{max}$, $R=1,\ldots,R_{max}$ and $p=0,\ldots,K_{max}$ and the corresponding BIC value is stored. The triplet corresponding to the highest value of BIC is then selected. These computations for selecting three values can be computationally more expensive compared to the ones in classical model selection namely for standard mixture where only the number of cluster has to be selected. However, we notice that for small values of (K,R,p), the computational cost is around few minutes and is not dramatically high, compared to approaches involving dynamic programming namely when using piecewise regression or when training approaches requiring MCMC sampling. Furthermore, it can be noticed that in some real situations, such as the one we present later, one or more values can be known (i.e., fixed by the experts namely the number of regimes and the structure of regimes for which a three-polynomial degree is well adapted) and the corresponding model selection procedure is quite fast.

The next section, we evaluate the proposed approach and perform comparisons with alternative ones.

4. Experimental study

This section is dedicated to the evaluation of the proposed approach. We tested it on simulated data, the waveform benchmark curves of Breiman et al. [2] and real data from a railway diagnosis application [7,8,27].

We performed comparisons with alternative functional discriminant analysis approaches using a polynomial regression (PR) or a spline regression (SR) model [21], and the one that uses a single RHLP model per class as in Chamroukhi et al. [8]. These alternatives will be abbreviated as FLDA-PR, FLDA-SR and FLDA-RHLP, respectively. We also considered alternative functional mixture discriminant analysis approaches that use polynomial regression mixtures (PRM), and spline regression mixtures (SRM) as in Gui and Li [17] which will be abbreviated as FMDA-PRM and FMDA-SRM respectively.

We used two evaluation criteria. The first one is the misclassification error rate computed by a 5-fold cross-validation procedure and concerns the performance of the approaches in terms of curve classification. The second one is the square error between the observed curves and the estimated mean curves, which is equivalent to the intra-class inertia, and regards the performance of the approaches with respect to the curves modeling and approximation. For FLDA approaches, as each class g is approximated by a single mean curve $\hat{\mathbf{x}}_g$, this error criterion is therefore given by $\sum_g\sum_{i|y_i=g}\|\mathbf{x}_i-\hat{\mathbf{x}}_g\|^2$. While, for FMDA approaches, each class g is summarized by several (K_g) mean curves $\{\hat{\mathbf{x}}_{gk}\}$, each of them summarizes a sub-class k, and the intra-class inertia in this case is therefore given by $\sum_g\sum_{k=1}^{K_g}\sum_{i|y_i=g,z_{gi}=k}\|\mathbf{x}_i-\hat{\mathbf{x}}_{gk}\|^2$. Notice that each point of the estimated mean curve for each sub-class is given by a polynomial function or a spline function for the case of polynomial regression mixture classification (FMDA-PRM) or spline regression mixture classification (FMDA-SRM) respectively, or by Eq. (15) for the case of the proposed FMDA-MixRHLP approach.

4.1. Experiments on simulated curves

In this section, we consider simulated curves issued from two classes of piecewise noisy functions. The first class has a complex shape and is composed of three sub-classes (see Fig. 1), while the second one is a homogeneous class. Each sub-class is composed of 50 curves and each curve consists of three regimes and is composed of 200 points.

Fig. 2 shows the obtained modeling results for the complex-shaped class shown in Fig. 1. It can be observed that the proposed approach accurately decomposes the class into homogeneous sub-classes of curves by automatically determining the sub-classes and the underlying hidden regimes for each sub-class. Furthermore, the flexibility of the logistic process used to model the hidden regimes allows for accurately approximating both abrupt and/or smooth regime changes within each sub-class. This can be clearly seen on the logistic probabilities which vary over

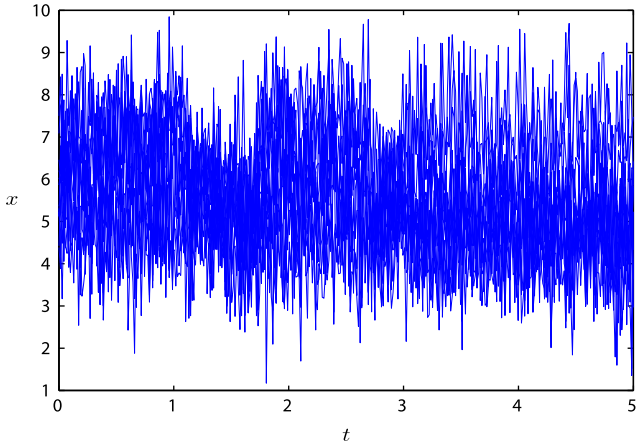

Fig. 1. Simulated curves from a complex-shaped class composed of three sub-classes, each of them is composed of three constant regimes.

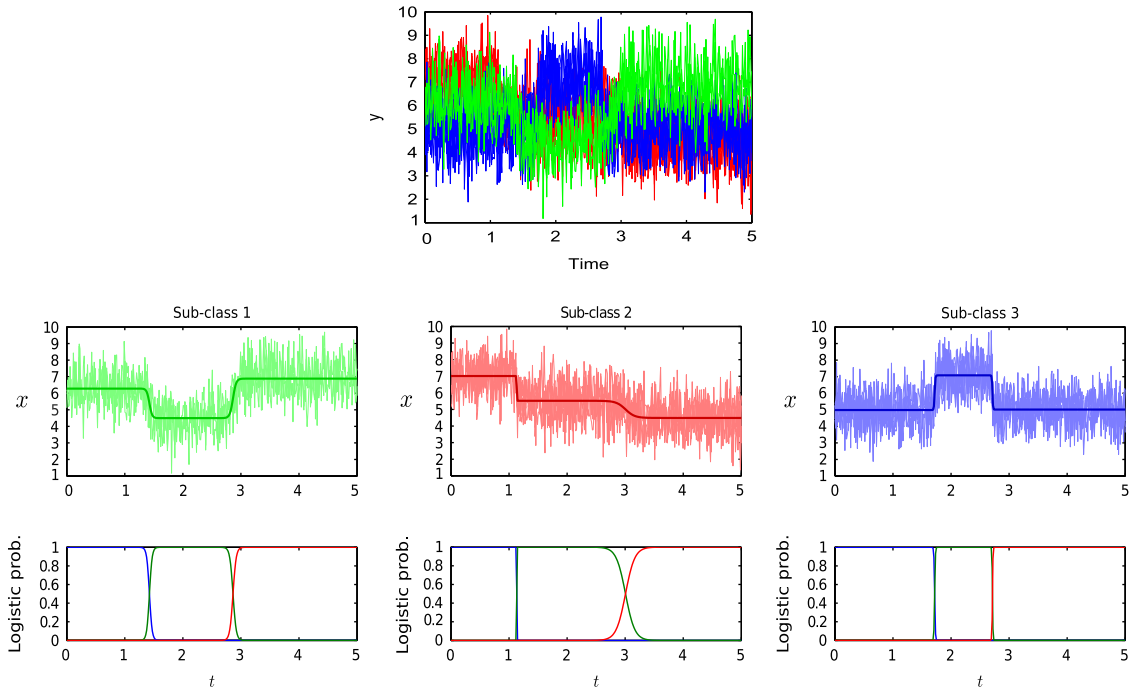

Fig. 2. The estimated sub-classes colored according to the partition given by the EM algorithm for the proposed approach (top); then are presented separately each sub-class of curves with the estimated mean curve in bold line (top sub-plot) and the corresponding logistic probabilities that govern the hidden regimes (bottom sub-plot). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Table 1
Obtained results for the simulated curves.

Approach	Classification error rate (%)	Intra-class inertia
FLDA-PR	21	7.1364×10^3
FLDA-SR	19.3	6.9640×10^3
FLDA-RHLP	18.5	6.4485×10^3
FMDA-PRM	11	6.1735×10^3
FMDA-SRM	9.5	5.3570×10^3
FMDA-MixRHLP	**5.3**	$\mathbf{3.8095} \times 10^3$

time according to both which regime is active or not and how is the transition from one regime to another over time (i.e., abrupt or smooth transition from one regime to another). We also notice that, approximating this class with a single mean curve, which is the case when using FLDA approaches (i.e., FMDA-PR or FMDA-SR), fails; the class is clearly heterogeneous. Using FMDA approaches based on polynomial or spline regression mixture (i.e., FMDA-PRM or FMDA-SRM) does not provide significant modeling improvements. This is due the fact that, as we can clearly see it on the data, the sub-classes present abrupt and smooth regime changes for which these two approaches are not well adapted. This is confirmed on the obtained intra-class inertia results given in Table 1. Table 1 indeed shows that the smallest value of intra-class inertia is obtained for the proposed FMDA-MixRHLP approach. The proposed functional mixture discriminant analysis approach based on hidden logistic process regression (FMDA-MixRHLP) outperforms the alternative FMDA based on polynomial regression mixtures (FMDA-PRM) or spline regression mixtures (FMDA-SRM). This performance is attributed to the flexibility of the MixRHLP model thanks to the logistic process which is well adapted for modeling the regime changes. We can also observe in Table 1 that, as expected, the FMDA approaches outperforms the FLDA approaches. This is attributed to the fact that, in this case of heterogeneous class, FLDA provides a rough class approximation

Table 1 also shows the misclassification error rates obtained with the proposed FMDA-MixRHLP approach and the alternative approaches. It can be seen that, also in terms of curve classification the FMDA approaches provide better results compared to FLDA approaches. This is due to the fact that using a single model for complex-shaped classes (i.e., when using FLDA approaches) is not adapted as it does not take into account the class dispersion when modeling the class conditional density. On the other hand, the proposed FMDA-MixRHLP approach provides a better modeling and therefore a more accurate class prediction.

We also performed experiments to select the best model for this data set. The true values for the dispersed class are ($K=3$, $R=3, p=0$) and for the other class which is homogeneous are ($K=1, R=3, p=0$). For the procedure of model selection as described in Section 3.5, the values of ($K_{max}, R_{max}, p_{max}$) were set to ($K_{max}=4$, $R_{max}=4, p_{max}=4$) and the proposed EM algorithm was applied to select the best model according to the highest BIC values. The process of model selection was repeated for 100 random samples.

The percentage of choosing the best model for the first class composed of three sub-classes is 91% while only 9% were obtained for the model ($K=3, R=3, p=1$). This is attributed to the fact that the constant regimes may be approximated as well by a polynomial of order 1 (linear function). The percentage of choosing the best model for the second homogeneous class is equal to 94%, the model corresponding to ($K=1, R=3, p=1$) were selected in only 6% of cases.

In the next section, the proposed approach is applied on the waveform curves of Breiman et al. [2].

4.2. Waveform curves of Breiman

In this section, we also illustrate our proposed approach on the waveform curves. The waveform data introduced by Breiman et al. [2] consist of a three-class problem where each curve is generated as follows:

- $\mathbf{x}_i(t) = uf_1(t) + (1-u)f_2(t) + \epsilon_t$ for the class 1;
- $\mathbf{x}_i(t) = uf_2(t) + (1-u)f_3(t) + \epsilon_t$ for the class 2;
- $\mathbf{x}_i(t) = uf_1(t) + (1-u)f_3(t) + \epsilon_t$ for the class 3,

where u is a uniform random variable on (0,1), $f_1(t) = \max(6-|t-11|,0)$; $f_2(t) = f_1(t-4)$; $f_3(t) = f_1(t+4)$ and ϵ_t is a zero-mean Gaussian noise with unit standard deviation. The temporal interval considered for each curve is (0,20) with a constant period of sampling of 1 s. For the experiments considered here, in order to have a heterogeneous class, we combine both class 1 and class 2 to form a single class called class 1. Class 2 will therefore used to refer to class 3 in the previous description of the waveform data. Fig. 3 (top) shows curves from each of the two classes.

Fig. 3 (middle) shows the obtained modeling results for each of the two classes by applying the proposed approach. We can see that the two sub-classes for the first class are well identified. These two sub-classes (clusters) are shown separately in Fig. 3 (bottom) with their corresponding mean curves. We notice that for this data set, all FMDA approaches provide very similar results regarding both the classification and the approximation since, as it can be seen, the complexity for this example is only related to the dispersion of the first class into sub-classes, and there are no explicit regime changes; each sub-class can therefore also be accurately approximated by a polynomial or a spline function.

4.3. Experiments on real data

In this section, we used a database issued from a railway diagnosis application as studied in Chamroukhi et al. [8,7] and Samé et al. [27]. The application context is the remote monitoring of the railway infrastructure components and more particularly the switch mechanism (also called points mechanism). The railway switch allows for guiding trains from one track to another at a railway junction, and is driven by an electric motor. The problem consists in accurately detecting possible defect on the system in order to alert the maintenance services. The used data are the curves of the instantaneous electrical power consumed during the switch actuation period. Each curve consists of 564 points sampled at 100 Hz in the range of (0;5.64) s (e.g., see Fig. 4). The switch actuation consists of five successive mechanical motions of different parts of the mechanism.

1. starting phase,
2. points unlocking phase,
3. points translation phase,
4. points locking phase,
5. friction phase.

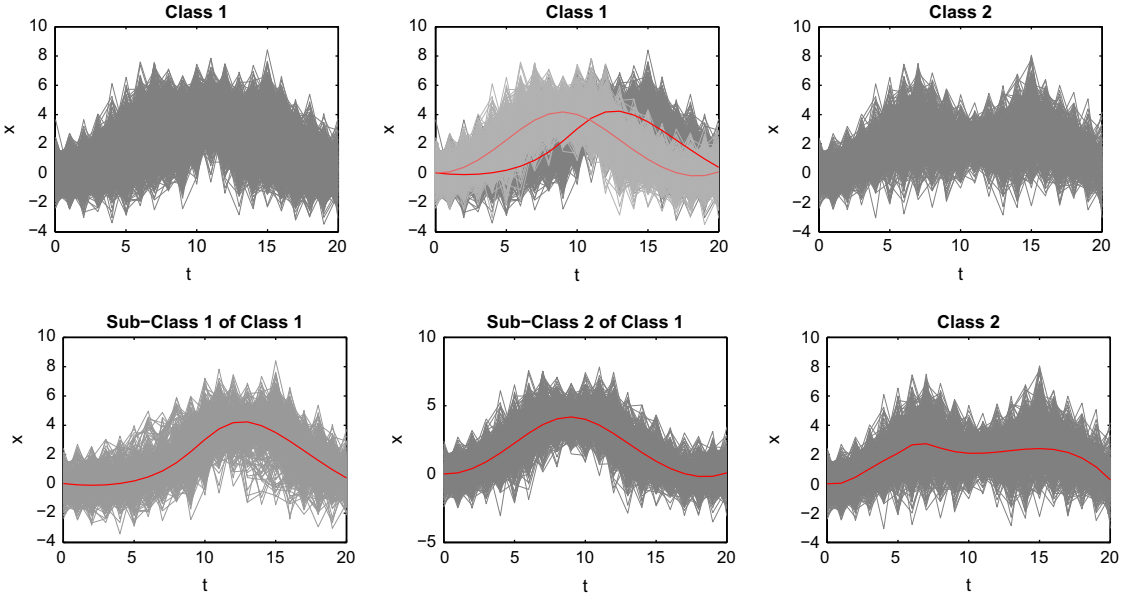

Fig. 3. Modeling results for the waveform curves: (top) the waveforms (500 curves per class) where the first class is composed of two sub-classes, (middle) the waveforms and the estimated subclasses for class 1 and the corresponding mean curves for each class, and (bottom) the two subclasses of class 1 shown separately with their corresponding mean curves.

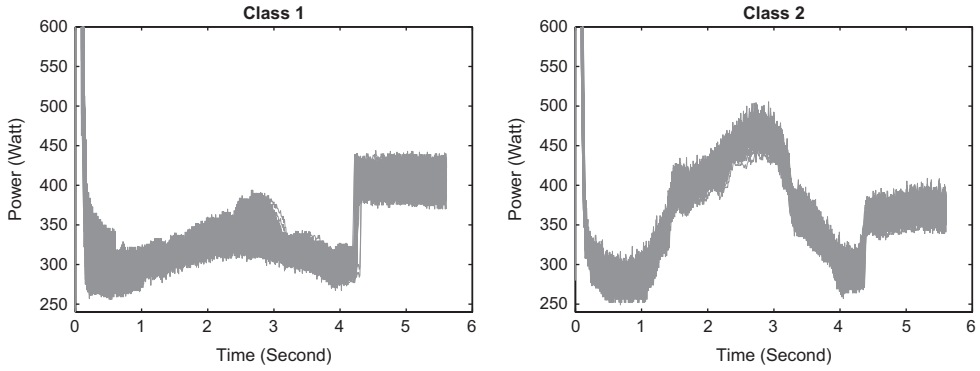

Fig. 4. Seventy five switch operation curves from the first class (left) and 45 curves from the second class (right).

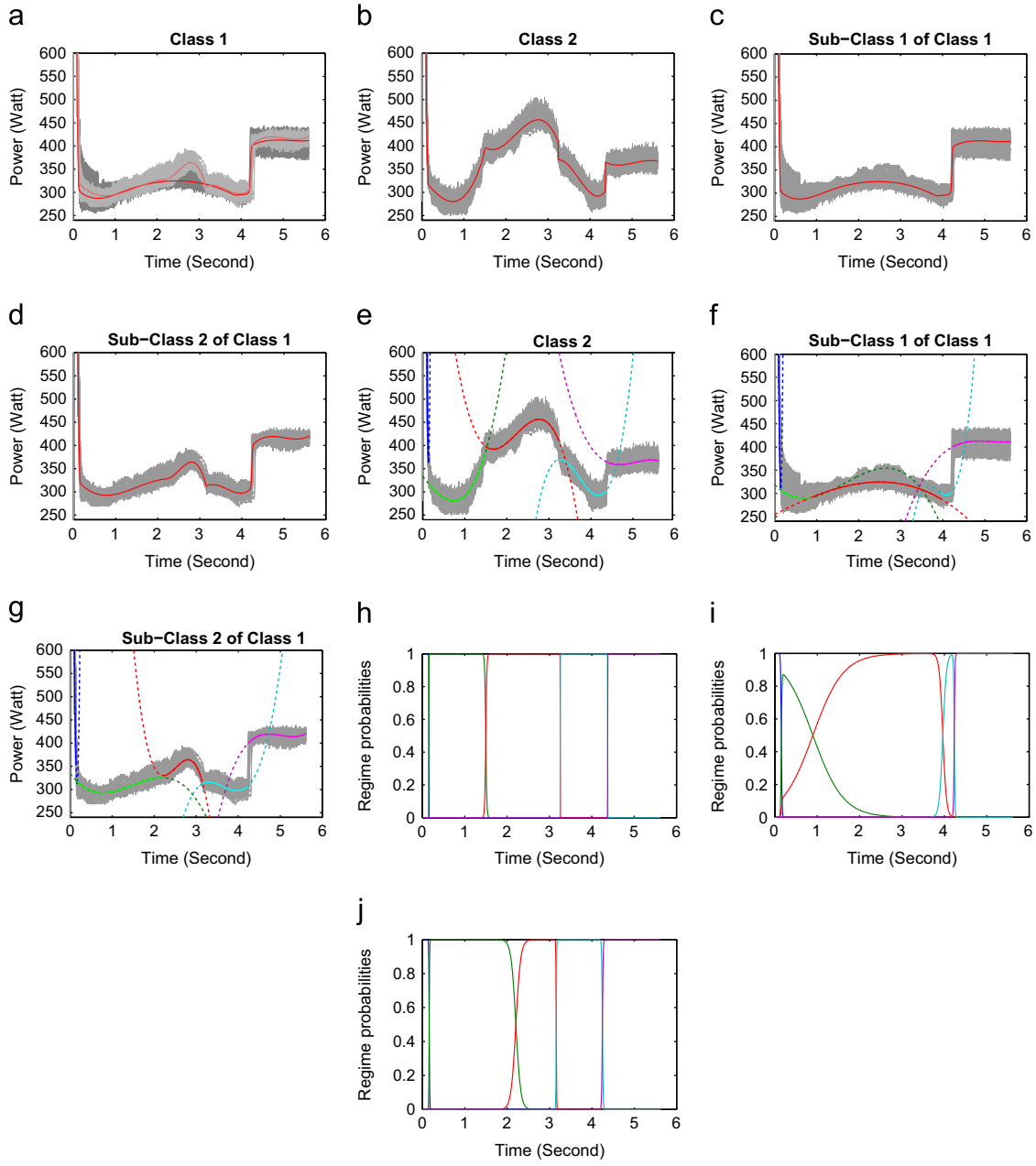

Fig. 5. Results obtained with the proposed model for the real curves. The estimated sub-classes for class 1 and the corresponding mean curves provided by the proposed approach (a); then, we show separately each sub-class of class 1 with the estimated mean curve presented in a bold line (c,d), the polynomial regressors (degree $p=3$) (f,g) and the corresponding logistic proportions that govern the hidden processes (i,j). Similarly, for class 2, we show the estimated mean curve in bold line (b), the polynomial regressors (e) and the corresponding logistic proportions.

Let us notice that the shape and the duration of each phase can vary from one situation of curves (class) to another according to the state of the system (e.g., with a defect, without defect). The used database is composed of 120 labeled real switch operation curves. In Chamroukhi et al. [8,7] and Samé et al. [27], the data were used to perform classification of three classes: no defect, with a minor defect and with a critical defect. In this study, we rather consider two classes where the first one is composed of the curves with no defect and with a minor defect so that the decision will be either with or without defect. The goal is therefore to provide an accurate automatic modeling especially for class 1 which is henceforth dispersed into two sub-classes. The cardinal numbers of the classes are $n_1 = 75$ and $n_2 = 45$ respectively. Fig. 4 shows each class of curves, where the first class is composed of two sub-classes.

Fig. 5 shows the modeling results provided by the proposed approach for each of the two classes. It shows the two sub-classes estimated for class 1 and the corresponding mean curves for the two classes. We also present the estimated polynomial regressors for each set of curves and the corresponding probabilities of the logistic process that govern the regime changes over time. We see that the proposed method ensures both an accurate decomposition of the complex shaped class into sub-classes and at the same time, a good approximation of the underlying regimes within each homogeneous set of curves. Indeed, it can be seen that the logistic process probabilities are close to 1 when the ℓth regression model seems to be the best fit for the curves and vary over time according to the smoothness degree of regime transition.

Then, the obtained classification results, by considering the FLDA approaches and the FMDA approaches (which are more

Table 2
Obtained results for the real curves of switch operations.

Approach	Classification error rate (%)	Intra-class inertia
FLDA-PR	11.5	10.7350×10^9
FLDA-SR	9.53	9.4503×10^9
FLDA-RHLP	8.62	8.7633×10^9
FMDA-PRM	9.02	7.9450×10^9
FMDA-SRM	8.50	5.8312×10^9
FMDA-MixRHLP	**6.25**	$\mathbf{3.2012} \times 10^9$

Table 3
Number of free parameters for each of the used models for class 1.

Model	Number of parameters for class 1
FLDA-PR	5
FLDA-SR	15
FLDA-RHLP	33
FMDA-PRM	11
FMDA-SRM	31
FMDA-MixRHLP	67

competitive) and gave the best results for simulations, are given in Table 2.

We can see that, although the classification results are similar for the FMDA approaches, the difference in terms of curves modeling (approximation) is significant, for which the proposed FMDA-MixRHLP approach clearly outperforms the alternative ones. This is attributed to the fact that the use of polynomial regression mixtures for FMDA-PRM or spline regression mixtures (FMDA-SRM) does not fit at best the regime changes compared to the proposed model. However, even the proposed approach provides the better results, we note that we have many parameters to estimate as summarized by Table 1 for the complex class (class 1). On the other hand, we note that for these values of (K,L,p) provided by the experts (Table 3), there is no over-fitting.

We also note that, for this real data, in terms of required computational effort to train each of the compared methods, the FLDA approaches are faster than the FMDA ones. In FLDA, both the polynomial regression and the spline regression approaches are analytic and does not require a numerical optimization scheme. The FLDA-RHLP is based on an EM algorithm which, while therefore performs in an iterative way, the learning scheme is quite fast and is in mean around 1 min for the described real data, and outperforms the alternative piecewise regression using dynamic programming and significantly improves the results. Detailed comparisons have been given in Chamroukhi et al. [8], namely in terms of computational time. On the other hand, the alternative FMDA approaches, that is the regression mixture and the spline regression mixture-based approaches still faster and their EM algorithm requires only few seconds to converge. However, these approaches are clearly not adapted for the regime changes problem; to do that, one needs to built a piecewise regression-based model which requires dynamic programming and therefore a dramatical computational cost especially for large curves, and still only adapted to abrupt regime changes. As stated in Section 3.5, the training procedure for the proposed approach is not dramatically time consuming, the training for the data of class 1 (which is the more complex class) requires a mean computational time of 2.98 min on a Matlab software using a standard laptop CPU of 2 GHz.

For model selection for this real data set, we notice that the number of regimes is fixed by the experts ($L=5$) and equals the number of electromechanical phases of the switch operation [7,8]. The number of sub-classes for class 1 is $K=2$ as we have no-defect sub-class and a sub-class corresponding to curves with a minor defect. The polynomial degree which is well adapted to the regime shape for the curves is $p=3$ (this was a preliminary choice made in conjunction with the expert [7,8] and a model selection procedure in Samé et al. [27] have confirmed this choice).

5. Conclusion

In this paper, we presented a new mixture model-based approach for functional data classification. The discrimination approach includes an unsupervised task that consists in clustering dispersed classes into sub-classes and determining the underlying unknown regimes for each sub-class. The proposed functional discriminant analysis approach uses a specific mixture of hidden process regression model for each class, which is particularly adapted for modeling complex-shaped classes of curves presenting regime changes. The parameters of each class are estimated in an unsupervised way by a dedicated EM algorithm and a model selection procedure is presented. The experimental results on simulated data and real data and comparisons to alternative approaches demonstrate the effectiveness of the proposed approach. In a first time, a future work will mainly concern learning the MixRHLP model of each class by maximizing a classification likelihood criterion, in which we will mainly be interested in classification, rather than maximizing a likelihood criterion as in this approach where we mainly focus on model estimation. This will rely on the Classification EM (CEM) algorithm [3]. Another future perspective is to build a fully Bayesian model for functional data to explicitly incorporate some prior knowledge on the data structure and to better control the model complexity.

References

[1] R. Bellman, On the approximation of curves by line segments using dynamic programming, Commun. Assoc. Comput. Mach. (CACM) 4 (6) (1961) 284.
[2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, New York, 1984.
[3] G. Celeux, G. Govaert, A classification EM algorithm for clustering and two stochastic versions, Comput. Stat. Data Anal. 14 (1992) 315–332.
[4] F. Chamroukhi, Hidden Process Regression for Curve Modeling, Classification and Tracking, Ph.D. Thesis, Université de Technologie de Compiègne, Compiègne, France, 2010.
[5] F. Chamroukhi, H. Glotin, C. Rabouy, Functional mixture discriminant analysis with hidden process regression for curve classification, in: Proceedings of XXth European Symposium on Artificial Neural Networks (ESANN), April 2012.
[6] F. Chamroukhi, A. Samé, G. Govaert, P. Aknin, A regression model with a hidden logistic process for feature extraction from time series, in: International Joint Conference on Neural Networks (IJCNN), 2009.
[7] F. Chamroukhi, A. Samé, G. Govaert, P. Aknin, Time series modeling by a regression approach based on a latent process, Neural Networks 22 (5–6) (2009) 593–602.
[8] F. Chamroukhi, A. Samé, G. Govaert, P. Aknin, A hidden process regression model for functional data description. Application to curve discrimination, Neurocomputing 73 (March (7–9)) (2010) 1210–1221.
[9] S. Dabo-Niang, F. Ferraty, P. Vieu, On the using of modal curves for radar waveforms classification, Comput. Stat. Data Anal. 51 (10) (2007) 4878–4890.
[10] C. Deboor, A Practical Guide to Splines, Springer-Verlag, 1978.
[11] A. Delaigle, P. Hall, N. Bathia, Component wise classification and clustering of functional data, Biometrika 99 (2) (2012) 299–313.
[12] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1) (1977) 1–38.
[13] F. Ferraty, P. Vieu, Curves discrimination: a nonparametric functional approach, Comput. Stat. Data Anal. 44 (1–2) (2003) 161–173.
[14] S.J. Gaffney, Probabilistic Curve-aligned Clustering and Prediction with Regression Mixture Models, Ph.D. Thesis, Department of Computer Science, University of California, Irvine, 2004.
[15] S.J. Gaffney, P. Smyth, Joint probabilistic curve clustering and alignment, in: Advances in Neural Information Processing Systems (NIPS), 2004.

[16] P. Green, Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, J. R. Stat. Soc. B 46 (2) (1984) 149–192.
[17] J. Gui, H. Li, Mixture functional discriminant analysis for gene function classification based on time course gene expression data, in: Proceedings of the Joint Statistical Meeting (Biometric Section), 2003.
[18] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. R. Stat. Soc. B 58 (1996) 155–176.
[19] G. Hébrail, B. Hugueney, Y. Lechevallier, F. Rossi, Exploratory analysis of functional data via clustering and optimal segmentation, Neurocomputing 73 (March (7–9)) (2010) 1125–1141.
[20] J.Q. Shi, R. Murray-Smith, D.M. Titterington, Hierarchical Gaussian process mixtures for regression, Stat. Comput. 15 (1) (2005) 31–41.
[21] G.M. James, T.J. Hastie, Functional linear discriminant analysis for irregularly sampled curves, J. R. Stat. Soc. Ser. B 63 (2001) 533–550.
[22] V.E. McGee, W.T. Carleton, Piecewise regression, J. Am. Stat. Assoc. 65 (1970) 1109–1124.
[23] G.J. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, 1997.
[24] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[25] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer Series in Statistics, Springer, 2005.
[26] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006.
[27] A. Samé, F. Chamroukhi, G. Govaert, P. Aknin, Model-based clustering and segmentation of time series with changes in regime, Adv. Data Anal. Classification 5 (4) (2011) 1–21.
[28] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[29] J. Shi, B. Wang, R. Murray-Smith, D. Titterington, Gaussian process functional regression modeling for batch data, Biometrics 63 (2007) 714–723.
[30] J.Q. Shi, T. Choi, Gaussian Process Regression Analysis for Functional Data, Chapman & Hall/CRC Press, 2011.
[31] H. Stone, Approximation of curves by line segments, Math. Comput. 15 (73) (1961) 40–47.
[32] D. Titterington, A. Smith, U. Makov, Statistical Analysis of Finite Mixture Distributions, John Wiley & Sons, 1985.

Faicel Chamroukhi received his Ph.D. in statistical machine learning from Compiègne University of Technology in 2010 and his Master's degree from PARIS 6 University in 2007 in Engineering entitled "Signals Systems Images and Robotics". Since 2011, he is an Associate Professor in Computer Science at Southern University of Toulon-Var Computer Science Department and the Information Sciences and Systems Lab (LSIS) UMR CNRS 7296. His research interests include statistical learning, functional data analysis pattern recognition signal and image processing, and their applications.

Hervé Glotin is a professor at the Insitut Universitaire de France (since 2011) and University of Toulon (since 2010), in the Systems & Information Sciences CNRS lab. He is leading the DYNI team on stochastic multimodal information retrieval. He received his master 1 in computer science from University Pierre et Marie Curie-Paris. In his master 2 thesis in artificial intelligence in Grenoble he proposed the first modelisation of vocalic system evolution, addressing the emergence of a common phonetic code in a society of communicating speech agents using evolutionary learning, which has been extended in many other works. He carried out his Ph.D. at the Institute of Perceptual Artificial Intelligence (IDIAP) CH and Institute of Spoken Communication - Perception Team Grenoble on "Robust adaptive multi-stream automatic speech recognition using voicing and localization cues". In 2000 he has been involved as expert at Johns Hopkins CSLP lab with IBM human language team in audiovisual Large Vocabulary Speech Recognition. After two years as research engineer at CNRS lab on phonology and Semantic analyses, he has been an assistant professor position at University of Toulon in 2003. He then conducted researches on multimodal pattern analysis and retrieval systems, audiovisual indexing, cognitive models and bioacoustics. He is co-author of one hundred of international refereed articles, and of an international (US, CANADA, etc.) patent on a real-time bioacoustic indexing algorithm. He is chairing since eight years the summer school in Multimodal Information Retrieval ERMITES. His research interests include signal processing, scene Understanding (vision, audition, bioacoustics), cognitive Systems and machine learning.

Allou Samé is holding a Master's Degree in Computer Science and Data Analysis from Compiègne University of Technology in 2001. He received his Ph.D. in the field of "real time clustering" from Compiègne University of Technology in 2004. Since 2006, he is working as researcher at the French National Institute for Transport and Safety Research (IFSTTAR).