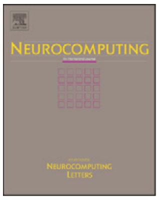
Neurocomputing 112 (2013) 164–171

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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



Funclust: A curves clustering method using functional random variables density approximation

Julien Jacques [a,b,c,*], Cristian Preda [a,b,c]

[a] Laboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France
[b] MODAL team, INRIA Lille-Nord Europe, France
[c] Polytech'Lille, France

ARTICLE INFO

Available online 7 March 2013

Keywords:
Functional data
Model-based clustering
Random variable density
Functional principal component analysis

ABSTRACT

A new method for clustering functional data is proposed under the name Funclust. This method relies on the approximation of the notion of probability density for functional random variables, which generally does not exist. Using the Karhunen–Loeve expansion of a stochastic process, this approximation leads to define an approximation for the density of functional variables. Based on this density approximation, a parametric mixture model is proposed. The parameter estimation is carried out by an EM-like algorithm, and the maximum a posteriori rule provides the clusters. The efficiency of Funclust is illustrated on several real datasets, as well as for the characterization of the Mars surface.

1. Introduction

Cluster analysis aims to identify homogeneous groups of data without using any prior knowledge on the group labels of data. Several methods, from hierarchical clustering [1] or k-means [2] to more recent probabilistic model-based clustering algorithms [3,4], have been proposed. A particular type of data for which clustering is a difficult task is the functional data (curves or trajectories [5]). The main difficulty in clustering such data arises because of the infinite dimensional space that the data belong to.

Consequently, most of the clustering algorithms for functional data consist in a first step of transforming the infinite dimensional problem into a finite dimensional one and in a second step, using a model-based clustering method designed for finite dimensional data. The representation of functions in a finite dimensional space can be carried out in several ways: discretizing the time interval, approximating data into a finite basis of functions or using some dimension reduction techniques such as functional principal component analysis (FPCA [5]). Note that using time interval discretization, we need to observe all curves at the same time stamps. The size of discretization being generally large, regularized clustering algorithm should be used [3,4,6–8]. The approximation of data (curves) into a finite dimensional space of functions – using a basis of functions such as spline or Fourier – has the advantage to take into account the possible measurement errors. Indeed, in the presence of such errors, a least square approximation approach can be used to estimate the coefficients of the basis approximation, whereas an interpolation method can be used if the data are observed without noise. More about smoothing functional data is presented in [5].

In the framework of clustering, the main contributions use the k-means algorithm, applied on a B-spline fitting [9], on defined principal points of curves [10], on the truncated Karhunen–Loeve expansion [11] or more recently on wavelets [12]. As in the finite dimensional setting, where Gaussian model-based clustering generalizes the k-means algorithm, some other works introduce more sophisticated model-based techniques: [13] define an approach particularly effective for sparsely sampled functional data, [14] propose a non-parametric Bayes wavelet model for clustering of functional data based on a mixture of Dirichlet processes, [15] build a specific clustering algorithm based on parametric time series models, [16] extend the high-dimensional data clustering (HDDC [7]) algorithm to the functional case.

Although we are mainly interested in model-based clustering algorithms, we mention several other approaches which contributed to the field of functional data clustering: [10] use k-means with distance L_2 on Gaussian process, [17] apply Self-Organized Map onto curves coefficients into an orthogonal basis expansions, [18] consider hierarchical clustering using specific semi-metric between curves, [19] define crisp and fuzzy k-means for functional data with time-dependent partition, [20,21] cluster piecewise estimation of the curves using dynamic programming algorithms, [22] use k-means algorithm using dissimilarity between curve based on a tail-dependence indices, and more recently [23] propose the utilization of divergences as dissimilarity measure in the Fuzzy c-Means algorithm

[*] Corresponding author at: Laboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France.
E-mail address: julien.jacques@polytech-lille.fr (J. Jacques).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.042

and [24] extend the Conn-Index for fuzzy prototype vector quantization clustering method.

In the finite dimensional setting, model-based clustering algorithms assume that the data is sampled from a mixture of probability densities. This is not directly applicable to functional data since the notion of probability density generally does not exist for functional random variable. Consequently, model-based clustering algorithms previously cited assume a parametric distribution on a finite series of coefficients characterizing the curves.

In the present paper, we use the density approximation defined in [22] to build our model-based clustering. This density approximation, based on the truncation of the Karhunen–Loeve expansion, relies on the probability density of the first principal components [5] of the curves. Our model assumes a cluster-specific Gaussian distribution for the principal component scores. The number of principal components as well as the computation of the principal component scores is cluster specific.

The most related methods are the k-centres algorithm (*kCFC*, [11]) and the *FunHDDC* algorithm [16]. In [11], the k-means algorithm is based on the distance between the truncated Karhunen–Loeve expansion of the curves. As for our model, different truncation orders are allowed for each cluster. But, contrary to our model, the k-means algorithm assumed equal within-cluster variations. Moreover, the estimation algorithm used in k-means performed classification at each iteration, whereas only a fuzzy partition is used in our algorithm. These differences are similar to the differences between k-means and more general Gaussian mixture models: k-means assumes equal diagonal covariance matrices for each cluster, whereas Gaussian mixture models allow more general covariance structures; k-means uses a CEM (Classification Expectation Maximisation [25]) algorithm whereas Gaussian mixture models are generally estimated more efficiently by the EM (Expectation Maximisation [26]) algorithm. In [16], the authors assume a parsimonious Gaussian model on the principal component scores issued from cluster-specific functional principal components analysis (FPCA). Real-data applications (Section 4) will illustrate numerically these differences between our methods, kCFC and FunHDDC.

The paper is organized as follows. Section 2 presents the approximation for the probability density of a functional random variable introduced in [22]. Model-based clustering using this approximation as well as the model estimation procedure, based on the EM algorithm, is presented in Section 3. Finally, Section 4 compares our method with other clustering algorithms on real datasets. An application to the characterization of the surface of Mars using clustering of spectrum concludes the paper.

2. Density approximation for functional data

Let X be a functional random variable with values in $L_2([0,T])$, $T>0$, and assume that X is a L_2-continuous stochastic process, $X=\{X(t),\ t\in[0,T]\}$. Let $\underline{X}=(X_1,\ldots,X_n)$ be an i.i.d sample of size n from the same probability distribution as X. $\underline{X}$ is generally called a sample of *functional data* for which the underlying model is X.

It is well known that the notion of probability density for this type of random variables is not well defined. In [18] a non-parametric approach for the estimation of probability density is presented as an extension of the multivariate finite case. This non-parametric approximation is not helpful in the context of model-based approaches.

Our work is based on the idea developed in [22] where an “approximation density” for X is proposed using the Karhunen–Loeve expansion (or principal component analysis (PCA))

$$X(t)=\mu(t)+\sum_{j=1}^{\infty}C_j\psi_j(t), \tag{1}$$

where μ is the mean function of X, $C_j=\int_0^T(X(t)-\mu(t))\psi_j(t)\,dt$, $j\geq 1$, are zero-mean random variables (called principal components) and ψ_j’s form an orthonormal system of eigen-functions of the covariance operator of X

$$\int_0^T Cov(X(t),X(s))\psi_j(s)\,ds=\lambda_j\psi_j(t),\quad \forall t\in[0,T].$$

Notice that the principal components C_j’s are uncorrelated random variables of variance λ_j. Considering the principal components indexed upon the descending order of the eigenvalues ($\lambda_1\geq\lambda_2\geq\cdots$), let $X^{(q)}$ denote the approximation of X by truncating (1) at the q first terms, $q\geq 1$

$$X^{(q)}(t)=\mu(t)+\sum_{j=1}^{q}C_j\psi_j(t). \tag{2}$$

Then, $X^{(q)}$ is the best approximation of X, under the mean square criterion, among all the approximations of the same type (linear combination of q deterministic functions of t with random coefficients, [27]). Denoting by $\|\cdot\|$ the usual norm on $L_2([0,T])$, we have

$$\mathbb{E}(\|X-X^{(q)}\|^2)=\sum_{j\geq q+1}\lambda_j\quad\text{and}\quad \|X-X^{(q)}\|\overset{\text{m.s.}}{\underset{q\to\infty}{\to}}0. \tag{3}$$

Without loss of generality, we will suppose in the following that X is a zero-mean stochastic process, i.e. $\mu(t)=0$, $\forall t\in[0,T]$.

Based on the approximation of X by $X^{(q)}$, in [22] it is shown that the probability of X to belong to a ball of radius h centred in $x\in L_2[0,T]$ can be written as

$$\log P(\|X-x\|\leq h)=\sum_{j=1}^{q}\log f_{C_j}(c_j(x))+\xi(h,q(h))+o(q(h)), \tag{4}$$

where f_{C_j} is the probability density of C_j and $c_j(x)$ is the jth principal component score of x, $c_j(x)=\langle x,\psi_j\rangle_{L_2}$. The functions q and ξ are such that q grows to infinity when h decreases to zero and ξ is depending only on h. Thus, the dependency of $\log P(\|X-x\|\leq h)$ with x is contained in the term $\sum_{j=1}^{q}\log f_{C_j}(c_j(x))$. Since the notion of probability density can be seen in the finite dimensional case as the limit of $P(\|X-x\|\leq h)/h$ when h tends to 0, [22] suggests the use of $\prod_{j=1}^{q}f_{C_j}(c_j(x))$ as an approximation for the density of X. In the sequel we give some additional justifications to this approximation.

Moreover, observe that we have, $\forall h>0$, $x\in L_2[0,T]$,

$$P(\|X^{(q)}-x\|\leq h-\|X-X^{(q)}\|)\leq P(\|X-x\|\leq h)\leq P(\|X^{(q)}-x\|\leq h+\|X-X^{(q)}\|). \tag{5}$$

The relations (3) and (5) also suggest that the probability $P(\|X-x\|\leq h)$ could be approximated by $P(\|X^{(q)}-x\|\leq h)$.

Let $f_X^{(q)}$ denote the joint probability density of $C^{(q)}=(C_1,\ldots,C_q)$. If $x=\sum_{j\geq 1}c_j(x)\psi_j$ and $x^{(q)}=\sum_{j=1}^{q}c_j(x)\psi_j$ then

$$P(\|X^{(q)}-x\|\leq h)=\int_{\mathcal{D}_x^{(q)}}f_X^{(q)}(y)\,dy, \tag{6}$$

where $\mathcal{D}_x^{(q)}=\{y\in\mathbb{R}^q:\|y-x^{(q)}\|_{\mathbb{R}^q}\leq\sqrt{h^2-\sum_{j\geq q+1}c_j^2(x)}\}$. Eqs. (5) and (6) suggest that the density $f_X^{(q)}$ can then be used as an approximation of the density of X. Moreover, when X is a Gaussian process, the principal components C_j are Gaussian and independent. The density

$f_X^{(q)}$ is then

$$f_X^{(q)}(x) = \prod_{j=1}^{q} f_{C_j}(c_j(x)), \qquad (7)$$

with f_{C_j} the Gaussian centred density of variance λ_j.

These results justify at least theoretically, the use of the principal component densities f_{C_j} to approximate the notion of probability density of X. In particular, it gives a theoretical justification to the method *kCFC* [11] which applies k-means on the principal components.

3. Model-based clustering for functional data

Several clustering algorithms for functional data used a truncation of the Karhunen–Loeve expansion [11,28]. In these works, the truncation is used in order to define a distance between function, which relies on the difference between the first Karhunen–Loeve expansion coefficients. The approximation provided in (7) allows to define more general model-based clustering by considering that the observed curves are sampled from a mixture of such densities.

Let us consider that there exists a latent group variable Z, of K modalities (K groups), such that $Z = Z_1, \ldots, Z_K$ with $Z_g = 1$ if X belongs to the cluster g, $1 \le g \le K$, and 0 otherwise. Conditionally on Z_g=1, let us assume that X is a Gaussian random variable of density $f_{X_{|Z_g=1}}^{(q_g)}(x)$. Here, q_g is the number of principal components used to approximate the density of X conditionally on the group g (Z_g=1). For each $i = 1, \ldots, n$, let us associate to X_i the corresponding categorical variable Z_i indicating the group X_i belongs.

3.1. The mixture model

Let us assume that each couple (X_i, Z_i) is an independent realization of the random vector (X,Z). Given a group Z_g=1, we consider the approximation (7) of the density of $X_{|Z_g=1}$ being

$$f_{X_{|Z_g=1}}^{(q_g)}(x;\Sigma_g) = \prod_{j=1}^{q_g} f_{C_{j|Z_g=1}}(c_{j,g}(x);\lambda_{j,g})$$

where q_g is the number of the first principal components retained in the approximation (7) for the group g, $c_{j,g}(x)$ is the jth principal component score of $X_{|Z_g=1}$ for X=x, $f_{C_{j,g}}$ is its probability density and Σ_g is the diagonal matrix diag$(\lambda_{1,g}, \ldots, \lambda_{q_g,g})$. Conditionally on the group, the probability density $f_{C_{j,g}}$ of the jth principal component of X is assumed to be the univariate Gaussian density with zero mean (the principal component are centred) and variance $\lambda_{j,g}$. This assumption is satisfied when $X_{|Z_g=1}$ is a Gaussian process.

The vector $Z = (Z_1, \ldots, Z_K)$ is assumed to have one order multinomial distribution $\mathcal{M}_1(\pi_1, \ldots, \pi_K)$, where $\pi_1, \ldots, \pi_K$ are the mixing probabilities ($\sum_{g=1}^{K} \pi_g = 1$). Under this model it follows that the unconditional approximated density of X is given by

$$f_X^{(q)}(x;\theta) = \sum_{g=1}^{K} \pi_g \prod_{j=1}^{q_g} f_{C_{j,g}}(c_{j,g}(x);\lambda_{j,g}) \qquad (8)$$

where $\theta = (\pi_g, \lambda_{1,g}, \ldots, \lambda_{q_g,g})_{1 \le g \le K}$ have to be estimated and $q = (q_1, \ldots, q_K)$. By extrapolation of the finite dimensional setting, we define a *pseudo-likelihood* by

$$l^{(q)}(\theta;\underline{X}) = \prod_{i=1}^{n} \sum_{g=1}^{K} \pi_g \prod_{j=1}^{q_g} \frac{1}{\sqrt{2\pi\lambda_{j,g}}} \exp\left(-\frac{1}{2}\frac{C_{i,j,g}^2}{\lambda_{j,g}}\right) \qquad (9)$$

where $C_{i,j,g} = C_{j,g}(X_i)$ is the jth principal score of the curve X_i belonging to the group g.

3.2. Parameter estimation

In the unsupervised context the estimation of the mixture model parameters is not as straightforward as in the supervised context since the group's labels Z_i are unknown. A classical way to maximize a mixture model likelihood when data are missing (here the clusters indicators Z_i) is to use the iterative EM algorithm [29]. In this work we use an EM-like algorithm including in the M step the computation of the principal components scores of each group and the selection of the group specific dimension q_g. Our EM-like algorithm consists in maximizing the pseudo completed log-likelihood

$$L_c^{(q)}(\theta;\underline{X},\underline{Z}) = \sum_{i=1}^{n} \sum_{g=1}^{K} Z_{i,g}\left(\log \pi_g + \sum_{j=1}^{q_g} \log f_{C_{j,g}}(C_{i,j,g})\right),$$

which is easier to maximize than its incomplete version (9), and leads to the same estimate. Let $\theta^{(h)}$ be the current value of the estimated parameter at step h, $h \ge 1$.

E step: As the group indicators $Z_{i,g}$'s are unknown, the E step consists in computing the conditional expectation of the pseudo completed log-likelihood

$$\begin{aligned}\mathcal{Q}(\theta;\theta^{(h)}) &= E_{\theta^{(h)}}[L_c^{(q)}(\theta;\underline{X},\underline{Z})|\underline{X} = \underline{x}] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{K} \tau_{i,g}\left(\log \pi_g + \sum_{j=1}^{q_g} \log f_{C_{j,g}}(c_{i,j,g})\right)\end{aligned}$$

where $\tau_{i,g}$ is the probability for the curve X_i to belong to the group g conditionally to $C_{i,j,g} = c_{i,j,g}$:

$$\tau_{i,g} = E_{\theta^{(h)}}[Z_{i,g}|\underline{X} = \underline{x}] \simeq \frac{\pi_g \prod_{j=1}^{q_g} f_{C_{j,g}}(c_{i,j,g})}{\sum_{l=1}^{K} \pi_l \prod_{j=1}^{q_l} f_{C_{j,l}}(c_{i,j,l})}. \qquad (10)$$

The approximation (10) is due to the use of the approximation of the density of X given by (7).

M step: The M step is composed of three stages:

1. *Principal score update*: The principal components $C_{j,g}$ of group g are computed by weighting the curves according to the conditional probabilities $\tau_{i,g}$ $(1 \le i \le n)$ computed in the E step. The estimation of the principal components is described in Section 3.3.
2. *Group specific dimension q_g selection*: The estimation of the group specific dimension q_g is an open problem. In this work we propose to use, once the group specific FPCA have been computed, the scree-test of Cattell [30] in order to select each group specific dimension q_g. The advantage of using this test is that one hyperparameter (the threshold of the Cattell scree-test) allows to estimate K approximation orders.
3. *Parameters update*: The M step consists in computing the mixture model parameters $\theta^{(h+1)}$ which maximizes $\mathcal{Q}(\theta;\theta^{(h)})$. The variance $\lambda_{j,g}$ of the jth principal component for cluster g has already been computed in the principal score update step. For the mixing proportions, the usual estimator is obtained:

   $$\pi_g^{(h+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{i,g}.$$

   Let us recall that the mean of the principal component $C_{j,g}$ is not considered since it is 0. The average shape of the curves of a cluster is taken into account in the computing of the principal components $C_{j,g}$ of the cluster.

Stopping criterion: When using an EM algorithm, usual stopping criterion is based on the growth of the likelihood. In our work, since the group specific approximation orders can change between two steps of the algorithm, the likelihood can artificially change (increase or decrease). In practice, we notice quite often that the estimation algorithm is hesitating between approximation orders, which prevents convergence of the pseudo-likelihood. For this reason, the algorithm often stops on the maximum number of iterations allowed. In this case, the retained solution is the solution maximizing the pseudo-likelihood.

The proposed mixture model and the corresponding estimation algorithm will be called *Funclust* in what follows.

3.3. Estimation and approximation for functional principal component analysis (FPCA)

Except some theoretical models (e.g. Brownian motion, Poisson process), the mean and the covariance functions of the stochastic process X are unknown. They are estimated from an i.i.d. sample of X, $\{X_1,\dots,X_n\}$, $n>1$, by

$$\hat{\mu}(t)=\frac{1}{n}\sum_{i=1}^{n}X_i(t),\quad t\in[0,T],$$

and

$$\widehat{Cov}(t,s)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i(t)-\hat{\mu}(t))(X_i(s)-\hat{\mu}(s)).$$

In the context of functional principal components, the asymptotic properties of these estimators are studied in [3,31]. Under the existence condition of fourth moment of X, in [32] convergences rates for the estimators of the eigenvalues and of the eigenfunction of the integral operator with kernel $\widehat{Cov}(t,s)$ are provided. See also [33] for more details.

3.3.1. Smoothing and interpolating curves

In practice, a new problem appears because of the continuous-time feature of the X_i's. In practice, a curve $X_i=\{X_i(t),t\in[0,T]\}$ is usually observed only in a discrete set of time-points, $\{X(t_{i,s}),0\le s\le m_i,t_{i,s}\in[0,T]\}$, that is, we have only discrete observations of each sample path X_i at a discrete set of knots $\{t_{i,s}:s=1,\dots,m_i\}$. Because of this, the first step in functional data analysis is often the reconstruction of the functional form of data from discrete observations. In [34] it is shown that this is equivalent to the choice of a metric in the space of discrete observations. The most common solution to this problem is to consider that sample paths belong to a finite dimensional space of functions spanned by a basis of functions $\{\phi_j\}_{j=1,\dots p}$ (see, for example, [5]).

$$X_i(t)=\sum_{j=1}^{p}\gamma_{i,j}\phi(t),\quad p\ge 1.$$

An alternative way of solving this problem is based on non-parametric smoothing of functions (see [18]).

Sample paths basis coefficients $\gamma_{i,j}$'s are estimated from discrete-time observations by using an appropriate numerical method. If the functional predictor is observed with error

$$X_i^{obs}(t_{i,s})=X_i(t_{i,s})+\varepsilon_{is},\quad s=0,\dots,m_i,$$

least square smoothing is used after choosing a suitable basis, for example, trigonometric functions, B-splines or wavelets (see [5] for a detailed study). In this case, the basis coefficients of each sample path X_i are approximated by

$$\widehat{\gamma}_i=(\Theta_i'\Theta_i)^{-1}\Theta_i'X_i^{obs},$$

with $\Theta_i=(\phi_j(t_{is}))_{1\le i\le n,1\le s\le m_i}$ and $X_i^{obs}=(X_i^{obs}(t_{i,0}),\dots,X_i^{obs}(t_{i,m_i}))'$.

The choice of the basis functions as well as the dimension of this basis are quite subjective. If the sample paths of X are smooth and periodic then Fourier basis could be a good choice. However, the optimal properties of cubic B-spline functions make them the first choice for smoothing noisy data. See for example the monograph [35] and, in the context of functional data, see [5].

If the sample curves are observed without error, an interpolation procedure can be used. For example, in [36] quasi-cubic spline interpolation for reconstructing annual temperatures curves from monthly values is proposed. More about interpolation of functional data is provided in [27].

3.3.2. FPCA computation

Let Γ be the $n\times p$ expansion coefficients γ_{ij} matrix and W be the matrix of the inner products between the basis functions $w_{j\ell}=\int_0^T\phi_j(t)\phi_\ell(t)\,dt$ $(1\le j,\ell\le p)$. We explain here the computation of the principal component $C_{j,g}$ of group g appearing in the M step as previously described. This computation is carried out by weighting the importance of each curve in the construction of the principal components with the conditional probabilities $T_g=\mathrm{diag}(\tau_{1,g},\dots,\tau_{n,g})$. Consequently, the first step consists in centring the curve X^i within the group g by subtraction of the mean curve computed using the $\tau_{i,g}$'s. The expansion coefficients of the centred curves are given by

$$\Gamma_g=(I_n-\mathbb{1}_n(\tau_{1,g},\dots,\tau_{n,g}))\Gamma,$$

where I_n and $\mathbb{1}_n$ are respectively the identity $n\times n$-matrix and the unit n-vector. The jth principal component scores $C_{j,g}$ is then the jth eigenvector of the matrix $\Gamma_gW\Gamma_g'T_g$ associated with the jth eigenvalue $\lambda_{j,g}$

$$\Gamma_gW\Gamma_g'T_gC_{j,g}=\lambda_{j,g}C_{j,g}.$$

Note that usual FPCA computation occurs if $T_g=1/n(I_n)$.

4. Applications

4.1. Clustering evaluation

Before validating the proposed clustering method on numerical applications, we have to choose an evaluation strategy, which remains an open questions in clustering. In lot of works, classification benchmark datasets are commonly used to validate and compare clustering models (see for instance [3,37,38]). As mentioned in several works [39,40], this strategy can be sometimes dangerous and misleading. Indeed, this evaluation strategy relies on the assumption that class labels coincide with cluster structure, which can be true for some datasets and not for others. Another strategy can be the use of artificial datasets. But this strategy can also be criticized, since it evaluates the clustering only under particular assumption on the data generating process. [41] argues that the best way to evaluate clustering is probably to work on real world datasets, and to explain how the obtained clusters make sense.

In this section, each of these three strategies will be used. First, a simulation study will be carried out to compare Funclust with two challengers for functional data clustering as well as usual clustering methods for finite dimensional data applied on FPCA scores. In a second part, the comparison is based on the three classification datasets. Finally, a clustering of the surface of the soil of Mars will be estimated with Funclust, and a physical interpretation of the clusters will be used to validate the usefulness of the obtained clustering.

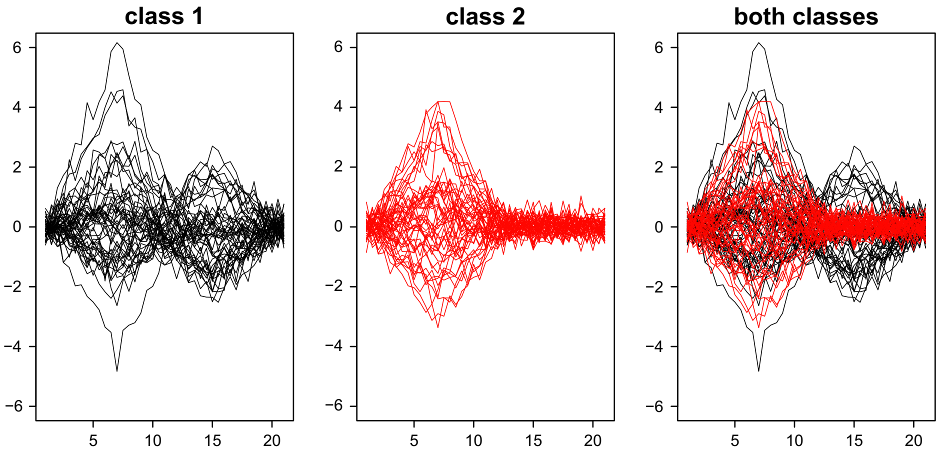


Fig. 1. Class 1 (left), Class 2 (center) and both classes (right).

4.2. Simulation study

In this simulation, the number of clusters is assumed to be known: $K=2$. A sample of $n=100$ curves are simulated according to the following model inspired by [42,43]

$$\text{Class 1}: \; X(t)=U_1h_1(t)+U_2h_2(t)+\epsilon(t), \quad t\in[1,21],$$

$$\text{Class 2}: \; X(t)=U_1h_1(t)+\epsilon(t), \quad t\in[1,21],$$

where U_1 and U_2 are independent Gaussian variables such that $\mathbb{E}[U_1]=\mathbb{E}[U_2]=0$, $\mathbb{V}ar(U_1)=\mathbb{V}ar(U_2)=1/12$ and $\epsilon(t)$ is a white noise, independent of U_i's and such that $\mathbb{V}ar(\epsilon_t)=1/12$. The functions h_1 and h_2 are defined, for $t\in[1,21]$, by $h_1(t)=6-|t-7|$ and $h_2(t)=6-|t-15|$. The mixing proportions π_i's are chosen to be equal, and the curves are observed in 41 equidistant points ($t=1, 1.5,\ldots,21$). Fig. 1 plots the simulated curves. The functional form of the data is reconstructed using linear spline smoothing (with 30 equidistant knots).

Funclust is compared with the two challengers for functional data clustering, FunHDDC [16] and *fclust* [13], and the three clustering methods traditionally devoted to clustering finite-dimensional data applied on the FPCA scores: Gaussian mixture models on the FPCA scores (GMM, [4]) *via* the *Rmixmod* package for **R**, *k*-means [2] and hierarchical clustering (packages *kmeans* and *hclust*). The selection of the number of FPCA components is carried out by the Cattell scree test. For FunHDDC and GMM, which proposes several models, the best model according to BIC has been retained. Fig. 2 shows the correct classification rates over 100 simulations, which exhibited better results for Funclust on this simulation set-up.

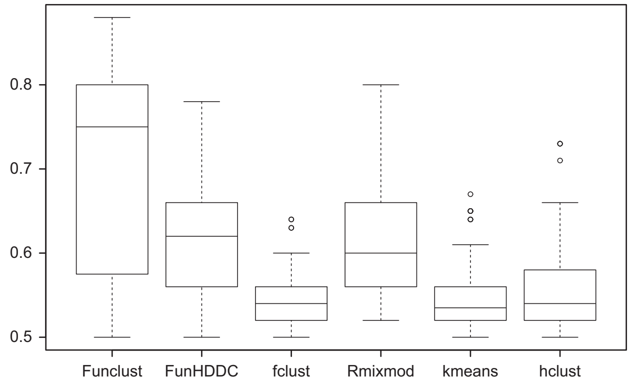


Fig. 2. Correct classification rates over 100 simulations.

4.3. Benchmark study

Funclust is now compared with other clustering methods on the basis of the capacity to find the class labels of the three classification datasets.

4.3.1. The data

The three real datasets are considered: the *Kneading*, *Growth*, and *ECG* datasets. These three datasets are plotted in Fig. 3. The Kneading dataset comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process. The kneading dataset is described in detail in [44]. There are 115 different flours for which the dough resistance is measured during the kneading process for 480 s. One obtains 115 kneading curves observed at 241 equispaced instants of time in the interval [0, 480]. The 115 flours produce cookies of different qualities: 50 of them have produced cookies of *good* quality, 25 produced *medium* quality and 40 *low* quality. This data have been already studied in a supervised classification context [44,45]. They are known to be hard to discriminate, even for supervised classifiers, partly because of the medium quality class. Taking into account that the resistance of dough is a smooth curve measured with error, and following previous works on this data [44,45], least squares approximation on a basis of cubic B-spline functions (with 18 knots) is used to reconstruct the true functional form of each sample curve. The Growth dataset comes from the Berkeley growth study [46] and is available in the *fda* package of **R**. In this dataset, the heights of 54 girls and 39 boys were measured at 31 stages, from 1 to 18 years. The goal is to cluster the growth curves and to determine whether the resulting clusters reflect gender differences. The ECG dataset is taken from the *UCR Time Series Classification and Clustering* website.[1] This dataset consists of 200 electrocardiogram from 2 groups of patients sampled at 96 time instants, and has already been studied in [47]. For these two datasets, the same basis functions as for the Kneading dataset has been arbitrarily chosen (20 cubic B-splines).

4.3.2. Experimental set-up

In this benchmark study, Funclust is compared with FunHDDC and *fclust*, as in the simulation study. The Growth dataset allows an additional comparison with *k*-centres (kCFC, [11]), since they

[1] http://www.cs.ucr.edu/~eamonn/time_series_data/.

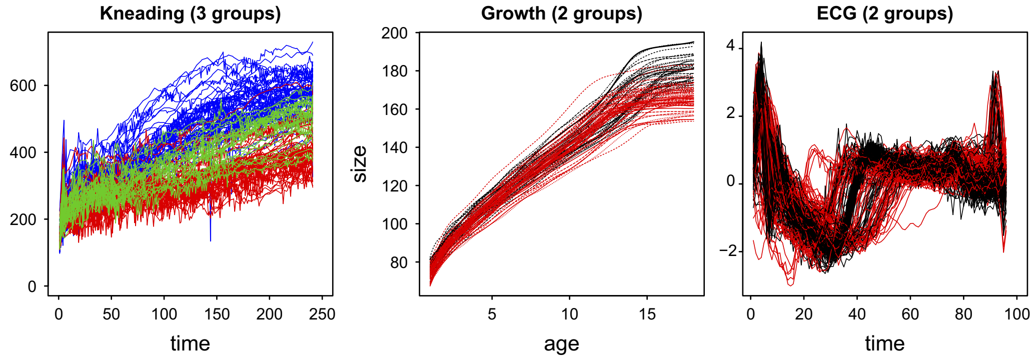


Fig. 3. *Kneading*, *Growth* and *ECG* datasets.

Table 1
Correct classification rates (CCR) in percentage for Funclust, FunHDDC (best model according BIC), fclust, kCFC and usual non-functional methods on the Kneading, Growth and ECG datasets.

Functional methods	Kneading	2-step methods	Kneading		
			Discretized (241 instants)	Spline coeff. (20 splines)	FPCA scores (4 components)
Funclust	**66.96**	HDDC	66.09	53.91	44.35
FunHDDC	62.61	MixtPPCA	65.22	64.35	62.61
fclust	64	GMM	63.48	50.43	60
kCFC	-	*k*-means	62.61	62.61	62.61
		hclust	63.48	63.48	63.48
	Growth	2-step methods	Growth		
			Discretized (350 instants)	Spline coeff. (20 splines)	FPCA scores (2 components)
Funclust	69.89	HDDC	56.99	50.51	**97.85**
FunHDDC	96.77	MixtPPCA	62.36	50.53	**97.85**
fclust	69.89	GMM	65.59	63.44	95.70
kCFC	93.55	*k*-means	65.59	66.67	64.52
		hclust	51.61	75.27	68.81
	ECG	2-step methods	ECG		
			Discretized (96 instants)	Spline coeff. (20 splines)	FPCA scores (19 components)
Funclust	**84**	HDDC	74.5	73.5	74.5
FunHDDC	75	MixtPPCA	74.5	73.5	74.5
fclust	74.5	GMM	81	80.5	81.5
kCFC	–	*k*-means	74.5	72.5	74.5
		hclust	73	76.5	64

present in [11] the correct classification rate they obtained on the Growth dataset (their code are not available to the best of our knowledge). Concerning the finite-dimensional methods to which Funclust is compared, we added to GMM, *k*-means and hierarchical clustering, the two methods dedicated to the clustering of high-dimensional data: *HDDC* [7] and *MixtPPCA* [6] (*HDclassif* package). These methods for finite-dimensional data have been applied on the FPCA scores with choice of the number of components with the Cattell scree test, but also directly on the discrete observations of the curves and on the coefficients in the cubic *B*-spline basis approximation.

Details for Funclust: The maximum number of iterations is fixed to 200. Note that for these three applications, the maximum number of iterations has always been reached. Nevertheless, since the iterations corresponding to the retained solutions (according to the best pseudo-likelihood) were always relatively far from the last one, we assume this maximum number of iterations as sufficient. The threshold of the Cattell scree test allowing to select the approximation order q_k is fixed to 0.05. In order to avoid convergence to a local maximum of the pseudo-likelihood, our EM-like algorithm has been initialized with the best solutions of 20 small EM-like algorithms with 20 iterations each [48]. With this experimental set-up, Funclust estimation is obtained in about 30 s for each dataset, on a laptop (2.80 GHz CPU) and with a code in **R** software.

4.3.3. Results

The estimated approximation order q_g for Funclust are the following: Kneading ($q_1=2$, $q_2=1$, $q_3=3$), Growth ($q_1=2$, $q_2=3$), ECG ($q_1=9$, $q_2=4$). The correct classification rates (CCR) according to the known partitions are given in Table 1. Funclust performs better to estimate the class label than all the other methods on two datasets among three (Kneading and ECG). On the last dataset, the results are relatively poor (69.89% accurate whereas some method are about 97% accurate), but the performance can be greatly increased (95.70%) if the dimensions q_g are fixed to 2 (as the number of FPCA scores used by the non-functional methods). This dataset illustrates that the choice of the approximation order is a very important question, and that further works have to be carried out in this direction. A last remark concerns the use of non-functional methods. These methods can sometimes perform very well to estimate the class label, but the main problem is that, in the present unsupervised

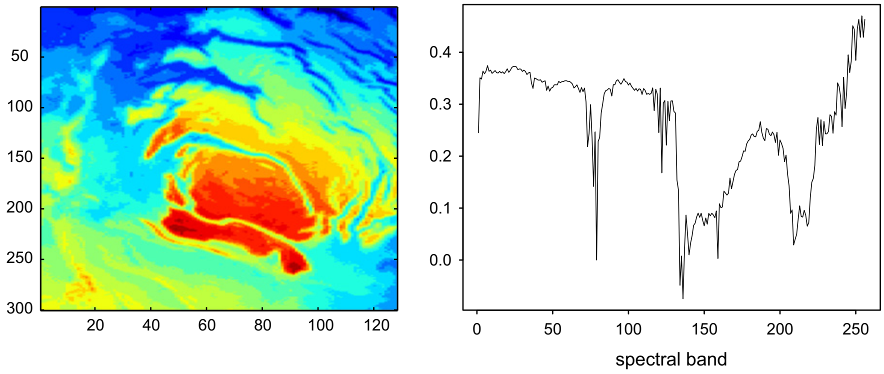


Fig. 4. Mars data: image of the studied zone.

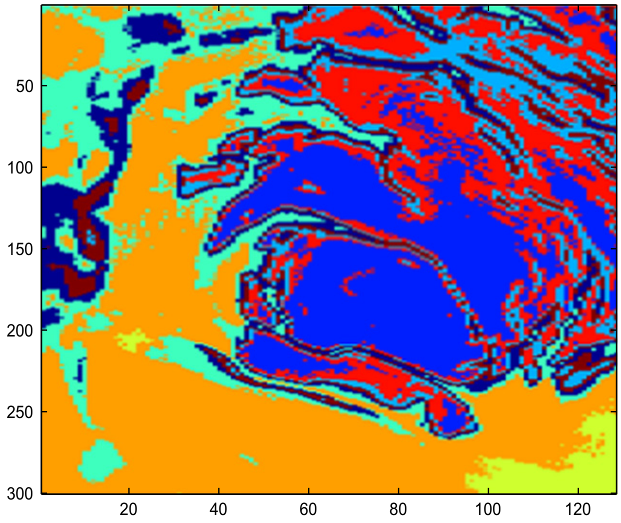


Fig. 5. Funclust clustering in 8 groups.

context, we have no way to choose between working with the discrete data, with the spline coefficients or with the FPCA scores. For instance, HDDC and MixtPPCA are very well performing on the Growth dataset using the FPCA scores, but they are very poor using the discrete data or the spline coefficients.

4.4. *Application to Mars surface characterization*

This data, provided by the Laboratory of Planetology of Grenoble [49,50], were acquired by the OMEGA imaging satellite. The soil of Mars has been observed with a resolution between 300 and 3000 m depending on the altitude of the satellite. It was acquired for each pixel a spectra whose wavelengths range from 0.36 to 5.2 μm and stored this information in a vector of 256 dimensions. The purpose of this preliminary study is to characterize the composition of the surface of Mars by determining zones composed of similar material. The number of groups has been fixed to 8, since the experts expect 8 main classes of mineralogical. The analysis of some spectra in each cluster will allow the expert to indicate to which mineralogical corresponds each cluster. A photography of size 300 × 128 pixels of the surface of Mars (left image of Fig. 4) is considered, each of 38,400 pixels being described by a spectrum (right image of Fig. 4).

The clusters resulting from Funclust is represented in Fig. 5. This clustering seems to be in accordance with the photography of Fig. 4 (we recall that no spatial information has been used for this clustering). Moreover, the experts of the Laboratory of Planetology of Grenoble particularly appreciated that our method is able to detect specific cluster, in the form of edging, at the border of the main areas: for instance the magenta and cyan clusters separate the main blue and orange area. Analyzing some spectra in each cluster has allowed to deduce that these clusters reflect the presence of particular materials (mixture of carbonate and ice) at the border of main materials (ice and dust).

5. Conclusion

In this paper we propose a new clustering procedure for functional data based on an approximation of the notion of density of a random function. The main tool is the use of the probability densities of the principal components scores. Assuming that the functional data are sampled from a Gaussian process, the resulting mixture model is an extrapolation of the finite dimensional Gaussian mixture model to the infinite dimensional setting. We defined an EM-like algorithm for the parameter estimation and performed several numerical applications, in order to show the performance of this approach with respect to usual clustering procedures.

Future work is devoted to investigate the choice of the approximation orders. We observed in our application study that a bad choice of these dimensions can drastically deteriorate the clustering results. However, allowing the approximation order to change in the estimation algorithm leads to lose the properties of the EM algorithm. In particular, the pseudo-likelihood is not necessarily increasing, and we have to stop the algorithm after a given number of iterations and to choose the best reached solutions.

References

[1] J.H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.
[2] J. Hartigan, M. Wong, Algorithm as 1326: a *k*-means clustering algorithm, Appl. Stat. 28 (1978) 100–108.
[3] J. Banfield, A. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (1993) 803–821.
[4] G. Celeux, G. Govaert, Gaussian parsimonious clustering models, J. Pattern Recognition Soc. 28 (1995) 781–793.
[5] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer Series in Statistics, second ed., Springer, New York, 2005.
[6] M.E. Tipping, C. Bishop, Mixtures of principal component analyzers, Neural Comput. 11 (1999) 443–482.
[7] C. Bouveyron, S. Girard, C. Schmid, High dimensional data clustering, Comput. Stat. Data Anal. 52 (2007) 502–519.
[8] J. Jacques, C. Bouveyron, S. Girard, O. Devos, L. Duponchel, C. Ruckebusch, Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data, J. Chemometrics 24 (2010) 719–727.
[9] C. Abraham, P.A. Cornillon, E. Matzner-Løber, N. Molinari, Unsupervised curve clustering using B-splines, Scand. J. Stat. 30 (2003) 581–595.
[10] T. Tarpey, K. Kinateder, Clustering functional data, J. Classification 20 (2003) 93–114.

[11] J.-M. Chiou, P.-L. Li, Functional clustering and identifying substructures of longitudinal data, J. R. Stat. Soc. Ser. B 69 (2007) 679–699.
[12] A. Antoniadis, X. Brossat, J. Cugliari, J.-M. Poggi, Clustering functional data using wavelets, Rapport de recherche RR-7515, INRIA, 2011.
[13] G. James, C. Sugar, Clustering for sparsely sampled functional data, J. Am. Stat. Assoc. 98 (2003) 397–408.
[14] S. Ray, B. Mallick, Functional clustering by Bayesian wavelet methods, J. R. Stat. Soc. Ser. B 68 (2006) 305–332.
[15] S. Frühwirth-Schnatter, S. Kaufmann, Model-based clustering of multiple time series, J. Bus. Econ. Stat. 26 (2008) 78–89.
[16] C. Bouveyron, J. Jacques, Model-based clustering of time series in group-specific functional subspaces, Adv. Data Anal. Classification 5 (2011) 281–300.
[17] F. Rossi, B. Conan-Guez, A. El Golli, Clustering functional data with the som algorithm, in: Proceedings of ESANN 2004, Bruges, Belgium, pp. 305–312.
[18] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer Series in Statistics, Springer, New York, 2006.
[19] S. Tokushige, H. Yadohisa, K. Inada, Crisp and fuzzy *k*-means clustering algorithms for multivariate functional data, Comput. Stat. 22 (2007) 1–16.
[20] B. Hugueney, G. Hébrail, Y. Lechevallier, F. Rossi, Simultaneous clustering and segmentation for functional data, in: Proceedings of ESANN 2009, Bruges, Belgium, pp. 281–286.
[21] G. Hébrail, B. Hugueney, Y. Lechevallier, F. Rossi, Exploratory analysis of functional data via clustering and optimal segmentation, Neurocomputing/EEG Neurocomputing 73 (2010) 1125–1141.
[22] A. Delaigle, P. Hall, Defining probability density for a distribution of random functions, Ann. Stat. 38 (2010) 1171–1193.
[23] T. Geweniger, M. Kästner, T. Villmann, Optimization of parametrized divergences in fuzzy c-means, in: Proceedings of ESANN 2011, Bruges, Belgium, pp. 11–16.
[24] T. Geweniger, M. Kästner, M. Lange, T. Villmann, Modified conn-index for the evaluation of fuzzy clusterings, in: Proceedings of ESANN 2012, Bruges, Belgium, pp. 465–470.
[25] G. Celeux, G. Govaert, A classification EM algorithm for clustering and two stochastic versions, Comput. Stat. Data Anal. 14 (1992) 315–332.
[26] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1977) 1–38.
[27] G. Saporta, Méthodes exploratoires d'analyse de données temporelles, Cahiers du Buro, 1981, pp. 37–38.
[28] J. Peng, H.-G. Müller, Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions, Ann. Appl. Stat. 2 (2008) 1056–1077.
[29] G. McLachlan, D. Peel, Finite Mixture Models, Wiley Interscience, New York, 2000.
[30] R. Cattell, The screen test for the number of factors, Multivar. Behav. Res. 1 (1966) 245–276.
[31] J. Dauxois, A. Pousse, Y. Romain, Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference, J. Multivar. Anal. 12 (1982) 136–154.
[32] J. Deville, Méthodes statistiques et numériques de l'analyse harmonique, Ann. l'INSEE (1974) 3–101.
[33] J. Rice, B. Silverman, Estimating the mean and covariance structure nonparametrically when the data are curves, J. R. Stat. Soc. (B) 53 (1991) 233–243.
[34] P. Besse, Approximation spline et optimalité en Analyse en Composantes Principales, Ph.D. Thesis, Université Toulouse III, 1989.
[35] C. De Boor, A Practical Guide to Splines, Springer, 2001.
[36] M. Escabias, A. Aguilera, M. Valderrama, Principal component estimation of functional logistic regression: discussion of two different approaches, J. Nonparametric Stat. 16 (2004) 365–384.
[37] J. Baek, G. McLachlan, L. Flack, Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010).
[38] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, 2nd ed., Springer-Verlag, New York, 2009.
[39] I. Guyon, U. Von Luxburg, R. Williamson, Clustering: Science or art, in: NIPS 2009 Workshop on Clustering Theory.
[40] I. Färber, S. Günnemann, H. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, A. Zimek, On using class-labels in evaluation of clusterings, in: MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD 2010, Washington, DC.
[41] J. Kogan, Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, 2007.
[42] F. Ferraty, P. Vieu, Curves discrimination: a nonparametric approach, Comput. Stat. Data Anal. 44 (2003) 161–173.
[43] C. Preda, Regression models for functional data by reproducing Kernel Hilbert spaces methods, J. Stat. Plann. Inference 137 (2007) 829–840.
[44] C. Leveder, C. Abraham, P. A. Cornillon, E. Matzner-Løber, N. Molinari, Discrimination de courbes de pétrissage, in: Chimiométrie 2004, Paris, pp. 37–43.
[45] C. Preda, G. Saporta, C. Lévéder, PLS classification of functional data, Comput. Stat. 22 (2007) 223–235.
[46] R. Tuddenham, M. Snyder, Physical Growth of California Boys and Girls from Birth to Eighteen Years. University of California Publications in Child Development, vol. 1, 1954, pp. 188–364.
[47] R. Olszewski, Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data, Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, 2001.
[48] C. Biernacki, Initializing EM using the properties of its trajectories in Gaussian mixtures, Stat. Comput. 14 (2004) 267–279.
[49] J.-P. Bibring, 42 co-authors, OMEGA: Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité, ESA SP-1240: Mars Express: the Scientific Payload, p. 37–49.
[50] C. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes, S. Girard, Retrieval of Mars surface physical properties frim OMEGA hyperspectral images using regularized sliced inverse regression, J. Geophys. Res. 114 (2009) E06005.

Julien Jacques is an assistant professor of statistics at University of Lille 1, France, since 2006. His current research in computational statistics concerns the design of probabilistic generative models for functional data, ranking or ordinal data, with applications in clustering. He is a member of the French Society of Statistics and of the International Association for Statistical Computing.

Cristian Preda is a professor of statistics at the University of Lille 1, France. He taught in the Faculty of Medicine before coming to the Department of Mathematics at the University of Lille 1 in 2008. His current research concerns regression models for functional data analysis and approximations for scan statistics distributions with applications in biostatistics. He is a member of the French Society of Statistics and of the Romanian Society of Statistics and Probability.