

# Model-based clustering for multivariate functional data

Julien Jacques, Cristian Preda

▶ **To cite this version:**

Julien Jacques, Cristian Preda. Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis, 2014, 71, pp.92-106. 10.1016/j.csda.2012.12.004 . hal-00713334v2

**HAL Id: hal-00713334**
**https://hal.science/hal-00713334v2**

Submitted on 13 Oct 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Model-based clustering for multivariate functional data

Julien Jacques*, Cristian Preda

*Laboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France*

*MODAL team, INRIA Lille-Nord Europe, & Polytech'Lille*

**Abstract**

This paper proposes the first model-based clustering algorithm for multivariate functional data. After introducing multivariate functional principal components analysis (MFPCA), a parametric mixture model, based on the assumption of normality of the principal components, is defined and estimated by an EM-like algorithm. The main advantage of the proposed model is its ability to take into account the dependence among curves. Results on simulated and real datasets show the efficiency of the proposed method.

*Keywords:* Multivariate functional data, density approximation, model-based clustering, multivariate functional principal component analysis, EM algorithm.

## 1. Introduction

Functional data analysis or "data analysis with curves" is an active topic in statistics with a wide range of applications. New technologies allow to record data with accuracy and at high frequency (in time or other dimension), generating large volume of data. In medicine one has growth curves of children and patient's state evolution, in climatology one records weather parameters over decades, chemometric curves are analysed in chemistry and physics (spectroscopy) and special attention is paid to the evolution of indicators coming from economy and finance. See Ramsay and Silverman (2005) for more details.

The statistical model underlying data represented by curves is a stochastic process with continuous time, $X = \{X(t)\}_{t\in[0,T]}$. Most of the approaches dealing with functional data consider the univariate case, i.e. $X(t) \in \mathbb{R}$, $\forall t \in [0,T]$, a path of $X$ being represented by a single curve. Despite its evident interest, the multidimensional case,

$$\boldsymbol{X} = \{\boldsymbol{X}(t)\}_{t\in[0,T]} \quad \text{with} \quad \boldsymbol{X}(t) = (X^1(t), \ldots, X^p(t))' \in \mathbb{R}^p, \quad p \geq 2$$

*Corresponding author. Tel.: +33 320 436 760, Fax: +33 320 434 302

*Email addresses:* julien.jacques@polytech-lille.fr (Julien Jacques), cristian.preda@polytech-lille.fr (Cristian Preda)

is rarely considered in literature. In this case a path of $\boldsymbol{X}$ is represented by a set of $p$ curves. The dependency between these $p$ measures provides the structure of $\boldsymbol{X}$. One finds in Ramsay and Silverman (2005) a brief example of bivariate functional data, $\boldsymbol{X}(t) = (X^1(t), X^2(t))' \in \mathbb{R}^2$, as a model for gait data (knee and hip measures) used in the context of functional principal component analysis (FPCA) as an extension of the univariate case. For a more theoretical framework, we must go back to the pioneer works of Besse (1979) on random variables in a general Hilbert space. Saporta (1981) provides a complete analysis of multivariate functional data from the point of view of factorial methods (principal components and canonical analysis).

In this paper we consider the problem of clustering multivariate functional data. Cluster analysis aims to identify homogeneous groups of data without using any prior knowledge on the group labels of data. When clustering functional data, the main difficulty is due to infinite dimensional space that data belong to. Consequently, most of clustering algorithms for functional data consists of a first step of transforming the infinite dimensional problem into a finite dimensional one and a second step using a clustering method designed for finite dimensional data. Recently, several new approaches considers the k-means algorithm applied to a $B$-spline fitting (Abraham et al., 2003), to defined principal points of curves (Tarpey and Kinateder, 2003) or to a truncation of the Karhunen-Loeve expansion (Chiou and Li, 2007). Sangalli et al. (2010) also use a k-means algorithm to cluster misaligned curves. 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: James and Sugar (2003) define an approach particularly effective for sparsely sampled functional data, Ray and Mallick (2006) propose a nonparametric Bayes wavelet model for curves clustering based on a mixture of Dirichlet processes, Frühwirth-Schnatter and Kaufmann (2008) build a specific clustering algorithm based on parametric time series models, Bouveyron and Jacques (2011) extend the high-dimensional data clustering algorithm (HDDC, Bouveyron et al. (2007)) to the functional case and Jacques and Preda (2012) build a model-based clustering based on an approximation of the notion of functional variable density.

The case of multivariate functional data is more rarely considered in literature: Singhal and Seborg (2005) and Ieva et al. (2011) use a k-means algorithm based on specific distances between multivariate functional data, whereas Kayano et al. (2010) consider Self-Organizing Maps based on the coefficients of multivariate curves into orthonormalized Gaussian basis expansions. Tokushige et al. (2007) extend crisp and fuzzy k-means algorithms for multivariate functional data by considering specific distance between functions, but applied their algorithms only on univariate functional data.

In the finite dimensional setting, model-based clustering algorithms consider that 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 (Delaigle and Hall, 2010). Consequently, model-based clustering

algorithms assume a parametric distribution on some finite set of coefficients characterizing the curves. In Jacques and Preda (2012), the authors use the density surrogate defined in Delaigle and Hall (2010) to build a model-based clustering for univariate functional data. This density surrogate, based on the truncation of the Karhunen-Loeve expansion, relies on the probability density of the principal components of the curves (Ramsay and Silverman, 2005), which is assumed to be Gaussian.
In this paper we propose an extension of Jacques and Preda (2012) approach to multivariate functional data. For this, we firstly introduce principal component analysis for multivariate functional data and assume a cluster-specific Gaussian distribution for the principal component scores. The elements derived from FPCA are estimated using approximation of the multivariate curves into a finite dimensional functional space. The number of principal components used in the density surrogate as well as the computation of the principal component scores are cluster specific.

The main advantage of our model is its ability to take into account dependency between the $p$ curves of the multidimensional data, thanks to the principal component analysis for multivariate functional data.

The paper is organized as follows. Section 2 introduces principal components analysis for multivariate functional data. Estimation and approximation details are provided and the task of normalizing the curves is discussed. Section 3 defines an approximation of the probability density for multivariate functional random variable. The model-based clustering approach and parameter estimation *via* an EM-like algorithm are presented in Section 4. Comparisons with existing methods on simulated and real datasets are presented in Section 5, and a discussion concludes the paper in Section 6.

## 2. Principal component analysis for multivariate functional data (MFPCA)

Principal component analysis for multivariate functional data has already been suggested in Ramsay and Silverman (2005) and Berrendero et al. (2011). In Ramsay and Silverman (2005) the authors propose to concatenate the observations of the functions on a fine grid of points (or the coefficients in a suitable basis expansion) into a single vector and then to perform a standard principal component analysis (PCA) on these concatenated vectors. When a basis expansion is used, this method forces to consider only orthonormal basis since the metric induced by the scalar product between the basis functions is not taken into account. In Berrendero et al. (2011), the authors propose to summarize the curves with functional principal components instead of scalar ones as in usual FPCA. For this purpose, they carry out classical PCA for each value of the domain on which the functions are observed and suggest an interpolation method to build the functional principal components.
Our approach is closely related to Ramsay and Silverman (2005) but in addition, we take into account the possible use of non orthonormal basis. In particular, our method allows to use different basis for each dimension of the multivariate curves.
Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be an i.i.d. sample of $\boldsymbol{X}$. The observation of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ provides a

set of $n$ $p$-variate curves, called *multivariate functional data*.
From this set of multivariate curves, one can be interested in optimal representation of curves in a function space of reduced dimension (principal component analysis), or in clustering, by determining an optimal partition of the observed curves with respect to some distance or homogeneity criterion. In order to address these two questions in a formal way, we need the hypothesis that considers $\boldsymbol{X} = (X^1, \ldots, X^p)'$ as a $L_2$ continuous stochastic process:

$$\forall t \in [0,T], \quad \lim_{h\to 0} \mathbb{E}\left[\|\boldsymbol{X}(t+h) - \boldsymbol{X}(t)\|^2\right] = \lim_{h\to 0} \int_0^T \sum_{\ell=1}^{p} \mathbb{E}\left[(X^\ell(t+h) - X^\ell(t))^2\right] dt = 0.$$

Notice that $L_2$-continuity of $\boldsymbol{X}$ implies $L_2$-continuity of each component of $\boldsymbol{X}$, *i.e.* $X^\ell$ is a $L_2$-continuous stochastic process for all $\ell = 1, \ldots, p$. The $L_2$-continuity is a quite general hypothesis, as most of the real data applications satisfies this one.
Let $\mu_\ell = \{\mu_\ell(t) = \mathbb{E}[X_\ell(t)]\}_{t\in[0,T]}$ denotes the mean function of $X^\ell$ $(1 \leq \ell \leq p)$ and

$$\mu = (\mu^1, \ldots \mu^p)' = \mathbb{E}[\boldsymbol{X}],$$

denotes the mean function of $\boldsymbol{X}$.
The covariance operator $\mathcal{V}$ of $\boldsymbol{X}$:

$$\begin{array}{rcl} \mathcal{V} : L_2([0,T])^p & \to & L_2([0,T])^p \\ \boldsymbol{f} & \overset{\mathcal{V}}{\longmapsto} & \mathcal{V}\boldsymbol{f} = \displaystyle\int_0^T V(\cdot,t)\boldsymbol{f}(t)dt, \end{array}$$

is an integral operator with kernel $V$ defined by:

$$V(s,t) = \mathbb{E}\left[(\boldsymbol{X}(s) - \mu(s)) \otimes (\boldsymbol{X}(t) - \mu(t))\right], \quad s,t \in [0,T]$$

where $\otimes$ is the tensor product on $\mathbb{R}^p$. Thus, for any $s,t \in [0,T]$, $V(s,t)$ is a $p \times p$ matrix with elements

$$V(s,t)[j,\ell] = Cov(X^j(s), X^\ell(t)), \quad j,\ell = 1, \ldots p.$$

*2.1. Principal components analysis of $\boldsymbol{X}$*

Under the hypothesis of $L_2$-continuity, $\mathcal{V}$ is an Hilbert-Schmidt operator, i.e compact, self-adjoint and such that $\sum_{j\geq 1} \lambda_j^2 < +\infty$. The spectral analysis of $\mathcal{V}$ provides a countable set of positive eigenvalues $\{\lambda_j\}_{j\geq 1}$ associated to an orthonormal basis of eigenfunctions $\{\boldsymbol{f}_j\}_{j\geq 1}$, $\boldsymbol{f}_j = (f_j^1, \ldots, f_j^p)$, called *principal factors* and are solutions of:

$$\mathcal{V}\boldsymbol{f}_j = \lambda_j \boldsymbol{f}_j, \tag{1}$$

with $\lambda_1 \geq \lambda_2 \geq \ldots$ and $\int_0^T \sum_{\ell=1}^{p} f_j^\ell(t) f_{j'}^\ell(t) dt = 1$ if $j = j'$ and 0 otherwise.
The *principal components* $\mathrm{C}_j$ of $\boldsymbol{X}$ are zero-mean random variables defined as the projections of $\boldsymbol{X}$ on the eigenfunctions of $\mathcal{V}$:

$$\mathrm{C}_j = \int_0^T \langle \boldsymbol{X}(t) - \mu(t), \boldsymbol{f}_j(t) \rangle_{\mathbb{R}^p} dt = \int_0^T \sum_{\ell=1}^{p} (X^\ell(t) - \mu^\ell(t)) f_j^\ell(t) dt.$$

Similar to the univariate setting, the principal components $\{\mathrm{C}_j\}_{j\geq 1}$ are zero-mean uncorrelated random variables with variance $\lambda_j$, $j \geq 1$.
Saporta (1981) shows that the following Karhunen-Loeve expansion holds in multidimensional context:

$$\boldsymbol{X}(t) = \mu(t) + \sum_{j\geq 1} \mathrm{C}_j \boldsymbol{f}_j(t), \qquad t \in [0, T]. \tag{2}$$

Principal components and principal factors of MFPCA have the same interpretation as in the functional univariate case. The truncation of (2) at the first $q$ terms provides a reduced dimensional space where classical tools (clustering, regression, ...) from multivariate analysis can be used to describe $\boldsymbol{X}$.

*2.2. Computational methods for MFPCA*

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^p)$, be the observation of the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The estimators for $\mu(t)$ and $V(s,t)$, for $s, t \in [0, T]$, are:

$$\hat{\mu}(t) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i(t) \quad \text{and} \quad \hat{V}(s,t) = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i(s) - \hat{\mu}(s)) \otimes (\boldsymbol{x}_i(t) - \hat{\mu}(t)).$$

In Deville (1974) it has been shown that $\hat{\mu}$ and $\hat{V}$ converges to $\mu$ and $V$ in $L_2$-norm with convergences rate of $O(n^{-1/2})$.
Often in practice, data are observed at discrete time points and with some noise. In order to get the functional feature of data, smoothing and interpolation methods are used considering the true curve belongs to a finite dimensional space spanned by some basis of functions. This approximation also reduces the eigen-analysis problem (1) to the one in finite dimensional setting.
Let assume that each curve $x_i^\ell$ ($1 \leq i \leq n$, $1 \leq \ell \leq p$) can be expressed as a linear combination of basis functions $\{\phi_\ell^j\}_{j=1,m_\ell}$:

$$x_i^\ell(t) = \sum_{j=1}^{m_\ell} \mathrm{a}_{i\ell j} \phi_\ell^j(t), \quad t \in [0, T]. \tag{3}$$

This can be written with the matrix formulation

$$\boldsymbol{x}_i(t) = \Phi(t) \mathrm{a}_i'$$

with $\mathrm{a}_i = (\mathrm{a}_{i11}, \ldots, \mathrm{a}_{i1m_1}, \mathrm{a}_{i21}, \ldots, \mathrm{a}_{i2m_2}, \ldots, \mathrm{a}_{ip1}, \ldots, \mathrm{a}_{ipm_p})$ being the vector of the basis expansion coefficients, and

$$\Phi(t) = \begin{pmatrix} \phi_1^1(t) & \ldots & \phi_1^{m_1}(t) & 0 & & & \ldots & & & 0 \\ 0 & \ldots & 0 & \phi_2^1(t) & \ldots & \phi_2^{m_2}(t) & 0 & \ldots & & 0 \\ & & & & & \ldots & & & & \\ 0 & & & \ldots & & & 0 & \phi_p^1(t) & \ldots & \phi_p^{m_p}(t) \end{pmatrix}.$$

Let $\tilde{\mathrm{A}}$ be the $n \times \sum_{\ell=1}^{p} m_\ell$-matrix, whose rows are the vectors $\mathrm{a}_i$, and $\mathrm{M}(t)$ the $n \times p$ matrix with values $x_i^\ell(t)$ of functions $x_i^\ell$ at times $t \in [0, T]$ ($1 \leq i \leq n$, $1 \leq \ell \leq p$). With these notations, we have

$$\mathrm{M}(t) = \tilde{\mathrm{A}}\Phi'(t). \tag{4}$$

Under the basis expansion assumption (3), the estimator $\hat{V}$ of $V$, for all $s, t \in [0, T]$, is given by:

$$\hat{V}(s,t) = \frac{1}{n-1}(\mathrm{M}(s) - \hat{\mu}'(s))'(\mathrm{M}(t) - \hat{\mu}'(t)) = \frac{1}{n-1}\Phi(s)\mathrm{A}'\mathrm{A}\Phi'(t), \tag{5}$$

where $\mathrm{M}(s) - \hat{\mu}'(s)$ means that the row vector $\hat{\mu}'(s)$ is subtracted to each row of $\mathrm{M}(s)$, and $\mathrm{A} = (I_n - \mathbb{1}_n(1/n, \ldots, 1/n))\tilde{\mathrm{A}}$ where $I_n$ and $\mathbb{1}_n$ are respectively the identity $n \times n$-matrix and the unit column vector of size $n$.
From (1) and (5), each principal factor $\boldsymbol{f}_j$ belongs to the linear space spanned by the basis $\Phi$:

$$\boldsymbol{f}_j(t) = \Phi(t)\mathrm{b}_j' \tag{6}$$

with $\mathrm{b}_j = (\mathrm{b}_{j11}, \ldots, \mathrm{b}_{j1m_1}, \mathrm{b}_{j21}, \ldots, \mathrm{b}_{j2m_2}, \ldots, \mathrm{b}_{jp1}, \ldots, \mathrm{b}_{jpm_p})$.
Using the estimation $\hat{V}$ of $V$, the eigen problem (1) becomes

$$\int_0^T \hat{V}(s,t)\boldsymbol{f}_j(t)dt = \lambda_j \boldsymbol{f}_j(s),$$

which, by replacing $\hat{V}(s,t)$ and $\boldsymbol{f}_j(s)$ by their expressions given in (5) and (6), is equivalent to

$$\int_0^T \frac{1}{n-1}\Phi(t)\mathrm{A}'\mathrm{A}\Phi'(t)\boldsymbol{f}_j'(t)ds = \lambda_j\Phi(s)\mathrm{b}_j', \tag{7}$$

$$\Leftrightarrow \frac{1}{n-1}\Phi(s)\mathrm{A}'\mathrm{A}\underbrace{\int_0^T \Phi'(t)\Phi(t)dt}_{\mathrm{W}}\mathrm{b}_j' = \lambda_j\Phi(s)\mathrm{b}_j', \tag{8}$$

where $\mathrm{W} = \int_0^T \Phi'(t)\Phi(t)dt$ is defined as the symmetric block-diagonal $\sum_{\ell=1}^p m_\ell \times \sum_{\ell=1}^p m_\ell$-matrix of the inner products between the basis functions.
Since (8) is true for all $s$, we have:

$$\frac{1}{n-1}\mathrm{A'AWb}'_j = \lambda_j \mathrm{b}'_j.$$

By defining $\mathrm{u}_j = \mathrm{b}_j \mathrm{W}^{1/2}$, the multivariate functional principal component analysis is reduced to the usual PCA of the matrix $\frac{1}{\sqrt{n-1}}\mathrm{AW}^{1/2}$:

$$\frac{1}{n-1}\mathrm{W}^{1/2'}\mathrm{A'AW}^{1/2}\mathrm{u}'_j = \lambda_j \mathrm{u}'_j.$$

The coefficient $\mathrm{b}_j$, $j \geq 1$, of the principal factors $\boldsymbol{f}_j$ are obtained by $\mathrm{b}_j = (\mathrm{W}^{1/2'})^{-1}\mathrm{u}'_j$, and the principal component scores, are given by

$$\mathrm{C}_j = \mathrm{AWb}'_j \quad j \geq 1.$$

Note that the principal components scores $\mathrm{C}_j$ are also the solutions of the eigenvalues problem:

$$\frac{1}{n-1}\mathrm{AWA'C}_j = \lambda_j \mathrm{C}_j.$$

*2.3. Normed principal component analysis*

When the $X^\ell$'s components of $\boldsymbol{X}$ $(1 \leq \ell \leq p)$ are of different natures (different measure units for example), the question of normalizing data occurs naturally. It is well known that the principal components are defined as maximizing the variance with respect to the total variance $Trace(\mathcal{V}) = \sum_{j\geq 1} \lambda_j$. Since,

$$Trace(\mathcal{V}) = \int_0^T \sum_{j=1}^p \mathbb{V}ar(X^j(t))dt,$$

it is clear that components $X^j$ with large variances plays an important role in defining the principal components. This source of variability is in general not interesting and hides more interesting features of the data structure. For this reason, except in particular situations (same scale for all $X^j$'s, for example), normalization is suitable. As in the classical framework of PCA, this is done by introducing some metrics.
One way to introduce a metric in MFPCA is to consider it as a canonical analysis (Saporta, 1981), in which the principal components are defined as solutions of the following eigen problem:

$$\int_0^T \mathcal{P}_t(\mathrm{C}_j)dt = \lambda_j \mathrm{C}_j, \quad j \geq 1 \tag{9}$$

where $\mathcal{P}_t$ is the orthogonal projection operator associated with $\boldsymbol{X}$ defined as

$$\mathcal{P}_t(\mathrm{C}_j) =< \boldsymbol{X}(t), [V(t,t)]^{-1}\mathbb{E}[\boldsymbol{X}(t)\mathrm{C}_j] >_{\mathbb{R}^p} . \tag{10}$$

Combining (10) and (9) one obtains

$$\mathrm{C}_j = \int_0^T < \boldsymbol{X}(t) - \mu(t), \boldsymbol{f}_j(t) >_{\mathbb{R}^p} dt \tag{11}$$

where $\boldsymbol{f}_j$ is the solution of the eigenvector problem

$$\int_0^T [V(s,s)]^{-1}V(s,t)\boldsymbol{f}_j(t)dt = \lambda \boldsymbol{f}(s). \tag{12}$$

Clearly, $[V(s,s)]^{-1}$ must exist for each $s \in [0,T]$. Under this hypothesis, the principal factors of the normalized MFPCA are the eigenfunctions of the integral operator with kernel $[V(s,s)]^{-1}V(s,t)$ as in (12). The Karhunen-Loeve expansion of $\boldsymbol{X}$ becomes

$$\boldsymbol{X}(t) = \mu(t) + \sum_{j=1}^{\infty} \mathrm{C}_j[V(t,t)]\boldsymbol{f}_j(t),$$

where the principal components $\mathrm{C}_j$, defined by (11), have zero mean and variance $\lambda_j$.

*Normalization in practice.* Observe that if $R(t,t)$ is the square root of the matrix $V(t,t)$, *i.e.* $V(t,t) = R(t,t)R(t,t)'$, then the MFPCA of $\boldsymbol{X}$ with metric $V(t,t)$ is equivalent to the MFPCA of $\boldsymbol{Y}$ defined by

$$\boldsymbol{Y}(t) = R(t,t)^{-1}\boldsymbol{X}(t),$$

with identity metric as in Section 2.1. In practice, if $\boldsymbol{X}$ is observed at times $t_1, \ldots, t_r$, $r > 1$, then $\boldsymbol{Y}$ is defined from $\boldsymbol{X}$ as

$$\boldsymbol{Y}(t_i) = R(t_i,t_i)^{-1}\boldsymbol{X}(t_i), \quad i = 1, \ldots, r$$

and approximation and estimation methodology in Section 2.2 is applied to $\boldsymbol{Y}$.

## 3. Approximation of the density for multivariate functional data

As the notion of probability density is not well defined for functional data (Delaigle and Hall, 2010), we can use an approximation of the density based on the Karhunen-Loeve expansion (2). Considering the principal components indexed upon the descending order of the eigenvalues ($\lambda_1 \geq \lambda_2 \geq \ldots$), and denoting $\boldsymbol{X}^{(q)}$ as the approximation of $\boldsymbol{X}$ by truncating (2) at the $q$ first terms, $q \geq 1$, we have

$$\boldsymbol{X}^{(q)}(t) = \mu(t) + \sum_{j=1}^{q} \mathrm{C}_j \boldsymbol{f}_j(t). \tag{13}$$

Then, $\boldsymbol{X}^{(q)}$ is the best approximation of $\boldsymbol{X}$, under the mean square criterion, among all the approximations of the same type (linear combination of $q$ deterministic functions of $t$ with random coefficients, Saporta (1981)).

Based on the approximation $\boldsymbol{X}^{(q)}$ of $\boldsymbol{X}$, Delaigle and Hall (2010) shows that, in the case $p = 1$, the probability of $\boldsymbol{X}$ belonging to a ball of radius $h$ centred at $\boldsymbol{x} \in L_2([0,T])^p$ can be written as

$$\log P(\|\boldsymbol{X} - \boldsymbol{x}\| \leq h) = \sum_{j=1}^{q} \log f_{\mathrm{C}_j}(\mathrm{c}_j(\boldsymbol{x})) + \xi(h, q(h)) + o(q(h)), \tag{14}$$

where $f_{C_j}$ is the probability density of $\mathrm{C}_j$ and $\mathrm{c}_j(\boldsymbol{x})$ is the $j$th principal component score of $\boldsymbol{x}$, $\mathrm{c}_j(\boldsymbol{x}) =< \boldsymbol{f}_j, \boldsymbol{x} >_{L_2([O,T])^p}$. The functions $q$ and $\xi$ are such that $q$ grows to infinity when $h$ tends to zero and $\xi$ depends only on $h$ and $q(h)$.

The equality (14) suggests the use of the multivariate probability density of the vector $\mathrm{C}^{(q)} = (\mathrm{C}_1, \ldots, \mathrm{C}_q)$ as a surrogate $f_{\boldsymbol{X}}^{(q)}$ for the "density" of $\boldsymbol{X}$:

$$f_{\boldsymbol{X}}^{(q)}(\boldsymbol{x}) = \prod_{j=1}^{q} f_{\mathrm{C}_j}(\mathrm{c}_j(\boldsymbol{x})). \tag{15}$$

Jacques and Preda (2012) use successfully a similar density surrogate for the clustering of univariate functional data.

## 4. A model based-clustering for multivariate functional data

The aim of model-based clustering is to identify homogeneous groups of data sampled from a mixture density model. In this section, we build a mixture model based on the surrogate (15) for the density of $\boldsymbol{X}$.

Let us consider that there exists a latent group variable Z, of $K$ categories ($K$ groups), $\mathrm{Z} = (\mathrm{Z}^1, \ldots, \mathrm{Z}^K) \in \{0,1\}^K$ such that $\mathrm{Z}^g = 1$ indicates that the multivariate curve $\boldsymbol{X}$ belongs to the cluster $g$, $1 \leq g \leq K$, and $\mathrm{Z}^g = 0$ otherwise. For each $i = 1, \ldots, n$, let $\mathrm{Z}_i$ be the group indicator corresponding to $\boldsymbol{X}_i$.

In the following we suppose that $\boldsymbol{X}_{|\mathrm{Z}^g=1}$ is such that each $X^{\ell}{}_{|\mathrm{Z}^g=1}$ is a zero-mean stochastic process ($1 \leq \ell \leq p$). The number $K$ of groups is assumed to be known. In the contrary case, an approximation of the BIC criterion (Schwarz, 1978), based on the approximated likelihood (17), could be used to select $K$.

### *4.1. The mixture model*

Let assume that each couple $(\boldsymbol{X}_i, \mathrm{Z}_i)$, $i = 1, \ldots, n$, is an independent realization of the random vector $(\boldsymbol{X}, \mathrm{Z})$ where $\boldsymbol{X}$ has a density surrogate depending on its group belonging:

$$f_{\boldsymbol{X}_{|\mathrm{Z}^g=1}}^{(q_g)}(\boldsymbol{x}; \Sigma_g) = \prod_{j=1}^{q_g} f_{\mathrm{C}_{j\,|\mathrm{Z}^g=1}}(\mathrm{c}_{j,g}(\boldsymbol{x}); \lambda_{j,g})$$

where $q_g$ is the number of principal components retained in the approximation (15) for the group $g$, $\mathrm{c}_{j,g}(\boldsymbol{x})$ is the $j$th principal component score of $\boldsymbol{X}_{|\mathrm{Z}^g=1}$ for $\boldsymbol{X}=\boldsymbol{x}$, $f_{\mathrm{C}_{j\,|\mathrm{Z}^g=1}}$ its probability density and $\Sigma_g$ the diagonal matrix of the principal components variances $(\lambda_{1,g},\ldots,\lambda_{q_g,g})$. Conditionally on the group $g$, the probability density $f_{\mathrm{C}_{j\,|\mathrm{Z}^g=1}}$ of the $j$th principal component of $\boldsymbol{X}$ is assumed to be the univariate Gaussian density with zero mean (the principal component are centred) and variance $\lambda_{j,g}$. This assumption is in particular satisfied when $\boldsymbol{X}_{|\mathrm{Z}^g=1}$ is a Gaussian process.

**Remark** (Data generation)**.** *For a given cluster $g$, $1\leq g\leq K$, provided that the $q_g$ eigenfunctions $\boldsymbol{f}_j$ and eigenvalues $\lambda_j$ of the covariance operator of $\boldsymbol{X}_{|Z^g=1}$ are known, then, generating an approximation $\boldsymbol{X}^{(q_g)}_{|Z^g=1}$ of $\boldsymbol{X}_{|Z^g=1}$ reduces to generating a real random variables $C_j$ according to centred Gaussian distributions with variance $\lambda_j$ $(1\leq j\leq q_g)$. Of course, that does not generate the true $\boldsymbol{X}_{|Z^g=1}$. However, the main structure of clusters is assumed to be characterized by this type of approximations.*

The vector $\mathrm{Z}=(\mathrm{Z}^1,\ldots,\mathrm{Z}^K)$ is assumed to have one-order multinomial distribution $\mathcal{M}_1(\pi_1,\ldots,\pi_K)$, with $\pi_1,\ldots,\pi_K$ the mixing proportions $(\sum_{g=1}^K\pi_g=1)$. Under this model we can deduce a surrogate for the unconditional density of $\boldsymbol{X}$:

$$f^{(q)}_{\boldsymbol{X}}(\boldsymbol{x};\theta)=\sum_{g=1}^{K}\pi_g\prod_{j=1}^{q_g}f_{\mathrm{C}_{j\,|\mathrm{Z}^g=1}}(\mathrm{c}_{j,g}(\boldsymbol{x});\lambda_{j,g}) \tag{16}$$

where $\theta=\{(\pi_g,\lambda_{1,g},\ldots,\lambda_{q_g,g})_{1\leq g\leq K}\}$ and $q=(q_1,\ldots,q_K)$. From this density surrogate, we deduce the pseudo likelihood:

$$l^{(q)}(\theta;\underline{\boldsymbol{X}})=\prod_{i=1}^{n}\sum_{g=1}^{K}\pi_g\prod_{j=1}^{q_g}\frac{1}{\sqrt{2\pi\lambda_{j,g}}}\exp\left(-\frac{1}{2}\frac{\mathrm{C}^2_{i,j,g}(\boldsymbol{X}_i)}{\lambda_{j,g}}\right) \tag{17}$$

where $\mathrm{C}_{i,j,g}(\boldsymbol{X}_i)$ is the $j$th principal score of the curve $\boldsymbol{X}_i$ for the group $g$ and $\underline{\boldsymbol{X}}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$.

**Remark** (Identifiability)**.** *When the approximation orders are different, identifiability of the mixture model (16) is straightforward. When all the approximation orders are equal, the identifiability of model (16) deduces directly from the identifiability of mixture of multivariate Gaussian (Titterington et al., 1985).*

### *4.2. Parameter estimation*

In the unsupervised context the estimation of the mixture model parameters is not so straightforward as in the supervised context since the groups indicators $\mathrm{Z}_i$ are unknown. On one hand, we need to use an iterative algorithm which alternates the estimation of the group indicators, the estimation of the PCA scores for each group and then the estimation of the mixture model parameters. On the other hand, the parameter

$q = (q_1, \ldots, q_g)$ will be estimated by an empirical method, similar to those used to select the number of components in usual PCA.

A classical way to maximise a mixture model likelihood when data are missing (here the clusters indicators $\mathrm{Z}_i$) is to use the iterative EM algorithm (Dempster et al., 1977; McLachlan and Peel, 2000). In this work we use an EM-like algorithm including, between the standard E and M steps, a first step in which the principal components scores of each group are updated and a second one in which the approximation order $q$ are selected. Our EM-like algorithm consists in maximizing the pseudo completed log-likelihood

$$L_c^{(q)}(\theta; \underline{\boldsymbol{X}}, \underline{\mathrm{Z}}) = \sum_{i=1}^{n} \sum_{g=1}^{K} \mathrm{Z}_i^g \left( \log \pi_g + \sum_{j=1}^{q_g} \log f_{\mathrm{C}_{j\,|\mathrm{Z}_i^g=1}}(\mathrm{C}_{i,j,g}(\boldsymbol{X}_i); \lambda_{j,g}) \right),$$

where $\underline{\mathrm{Z}} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_n)$. Let $\theta^{(h)}$ contains the current values of the estimated parameters at step $h$, $h \geq 1$.

*E step.* As the group indicators $\mathrm{Z}_i^g$'s are unknown, the E step consists in computing the conditional expectation of the pseudo completed log-likelihood:

$$\mathcal{Q}(\theta; \theta^{(h)}) = E_{\theta^{(h)}}[L_c^{(q)}(\theta; \underline{\boldsymbol{X}}, \underline{\mathrm{Z}}) | \underline{\boldsymbol{X}} = \underline{\boldsymbol{x}}] = \sum_{i=1}^{n} \sum_{g=1}^{K} t_{i,g} \left( \log \pi_g + \sum_{j=1}^{q_g} \log f_{\mathrm{C}_{j\,|\mathrm{Z}_i^g=1}}(\mathrm{c}_{i,j,g}(\boldsymbol{x}_i); \lambda_{j,g}) \right)$$

where $\mathrm{c}_{i,j,g}(\boldsymbol{x}_i)$ is the value of the random variable $\mathrm{C}_{i,j,g}(\boldsymbol{X}_i)$ for $\boldsymbol{X}_i = \boldsymbol{x}_i$, $t_{i,g}$ is the probability for the multidimensional curve $\boldsymbol{X}_i$ to belong to the group $g$ conditionally to $\mathrm{C}_{i,j,g}(\boldsymbol{X}_i) = \mathrm{c}_{i,j,g}(\boldsymbol{x}_i)$:

$$t_{i,g} = E_{\theta^{(h)}}[\mathrm{Z}_i^g | \underline{\boldsymbol{X}} = \underline{\boldsymbol{x}}] \simeq \frac{\pi_g \prod_{j=1}^{q_g} f_{\mathrm{C}_{j\,|\mathrm{Z}_i^g=1}}(\mathrm{c}_{i,j,g}(\boldsymbol{x}_i); \lambda_{j,g})}{\sum_{l=1}^{K} \pi_l \prod_{j=1}^{q_l} f_{\mathrm{C}_{j\,|\mathrm{Z}_i^l=1}}(\mathrm{c}_{i,j,l}(\boldsymbol{x}_i); \lambda_{j,l})}. \tag{18}$$

The approximation in (18) is due to the use of the surrogate for the density of $\boldsymbol{X}$ given by (15).

*Principal score updating step.* The computation of the principal component scores has been described in Section 2.2. Here, the principal component scores $\mathrm{C}_{i,j,g}$ of the multidimensional curve $\boldsymbol{X}_i$ in the group $g$ is updated according to the current conditional probability $t_{i,g}$ computed in the previous E step. This computation is carried out by weighting the importance of each curve in the construction of the principal components with the $t_{i,g}$'s. Consequently, the first step consists in centring the curve $\boldsymbol{X}_i$ within the group $g$ by subtracting the mean curve computed using the $t_{i,g}$'s: the basis expansion coefficients matrix A becomes $\mathrm{A}_g = (I_n - \mathbb{1}_n(t_{1,g}, \ldots, t_{n,g}))\tilde{\mathrm{A}}$. The $j$th principal component scores $\mathrm{C}_{j,g}$ is then the $j$th eigenvector of the matrix $\mathrm{A}_g \mathrm{W} \mathrm{A}_g' T_g$ associated with the $j$th eigenvalue $\lambda_{j,g}$, with $T_g = \mathrm{diag}(t_{1,g}, \ldots, t_{n,g})$.

*Group specific dimension $q_g$ estimation step.* 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 at the previous step, the scree-test of Cattell (1966) 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.

*M step.* The M step consists of computing the mixture model parameters $\theta^{(h+1)}$ which maximizes $\mathcal{Q}(\theta; \theta^{(h)})$ according to $\theta$. It leads simply to the following estimators

$$\pi_g^{(h+1)} = \frac{1}{n}\sum_{i=1}^{n} t_{i,g}, \quad \text{and} \quad {\lambda_{j,g}}^{(h+1)} = \lambda_{j,g}, \quad 1 \leq j \leq q_g,$$

where $\lambda_{j,g}$ is the variance of the $j$th principal component of the cluster $g$ computed in the principal score updating step.

*Convergence and numerical considerations.* The proposed estimation algorithm is not a proper EM algorithm, since the growth of the pseudo likelihood is not guaranteed between two steps. The main reason is the use of a pseudo likelihood built on an approximation of the notion of density. Indeed, since only a finite number of principal components are used, and since these principal components are computed with different weights at each iteration of the algorithm (the $t_{i,g}$'s), the 'data' on which the pseudo likelihood is computed, *i.e.* the principal components scores, are not the same at each step. To avoid this phenomenon, all principal components should be used, which is generally not applicable for functional data since they form an infinite set. Thus, the values of the likelihood can not be directly compared between two iterations. The properties of the EM algorithm, which guarantee the convergence to a local maxima of the likelihood in the classical framework does not work any more. Moreover, the approximation orders $q_g$, $1 \leq g \leq K$, are updated at each iteration, and this can also induce an artificial increase or decrease of the pseudo likelihood.
In practice, we adopt the following empirical strategy which allows us to perform numerical applications:

- the algorithm is launched several times with random initializations for a small number of iterations, and the best reached solution is used as the initialization point for a longer algorithm (Biernacki, 2004). Typically, 10 small runs with 10 iterations are used in the following experiments.
- the values of $q_g$ are initialized to 1, and they are only allowed to increase.
- the number $S$ of iterations is set as follows: for a given $S$ (200 for instance), the algorithm is executed 20 times, and $S$ is considered to be large enough if the maximum of the pseudo likelihood has been achieved before $3S/4$ iterations for the 20 executions.

Anticipating the application results, Figure 4 and Figure 6 illustrate the trajectories of the pseudo likelihood and the approximation orders on simulated and real datasets.

## 5. Numerical experiments

This section is devoted to compare our approach, which we will call *Funclust* – as in univariate case (Jacques and Preda, 2012)– with other existing methods. The evaluation of a clustering algorithm is always a difficult and subjective task. Following Guyon et al. (2009), three evaluations strategies are considered in this paper. First, Funclust is compared to other clustering methods for univariate functional data using three classification benchmark datasets. Second, a simulation study allows to compare Funclust with another clustering method for multivariate functional data based on k-means. Third, a real clustering application on a climatology dataset is carried out. The clusters obtained by Funclust and the k-means based method are then compared from the interpretation point of view.

**Remark** (Data registration)**.** *When working with functional data, a curve registration step is often needed to remove the amplitude and phase variation of curves (Ramsay and Silverman, 2005, Chap. 7). In our opinion, in the clustering context, the amplitude and phase variability of curves are interesting elements to define clusters. For instance, in the Canadian weather example which will be analysed in the sequel, the geographical interpretation of the clusters is mainly due to amplitude variability. Similarly, for the Growth dataset, it is shown in Liu and Yang (2009) that performing registration before or simultaneously with clustering failed in retrieving the gender of subjects, probably because the main gender difference is due to a time wrapping effect. For this reason, we do not perform data registration in this work before our clustering study.*

The **R** code for Funclust is available on request from the authors.

### *5.1. Benchmark study in the case of univariate functional data*

*The data.* Three real datasets are considered: the *Kneading*, *Growth*, and *ECG* datasets. These three datasets, already studied in Jacques and Preda (2012), are plotted on Figure 1. The first dataset (Kneading) 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 Leveder et al. (2004). There are 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. One obtains 115 kneading curves observed at 241 equispaced instants of time in the interval $[0, 480]$. The 115 flours produce cookies of different quality: 50 of them have produced cookies of *good* quality, 25 produced *medium* quality and 40 *low* quality. This data has been already studied in a supervised classification context (Leveder et al., 2004; Preda et al., 2007). This data is known to be hard to discriminate, even for supervised classifiers, partly because of the medium class. The second dataset (Growth) comes from the Berkeley growth study (Tuddenham and Snyder, 1954) and is

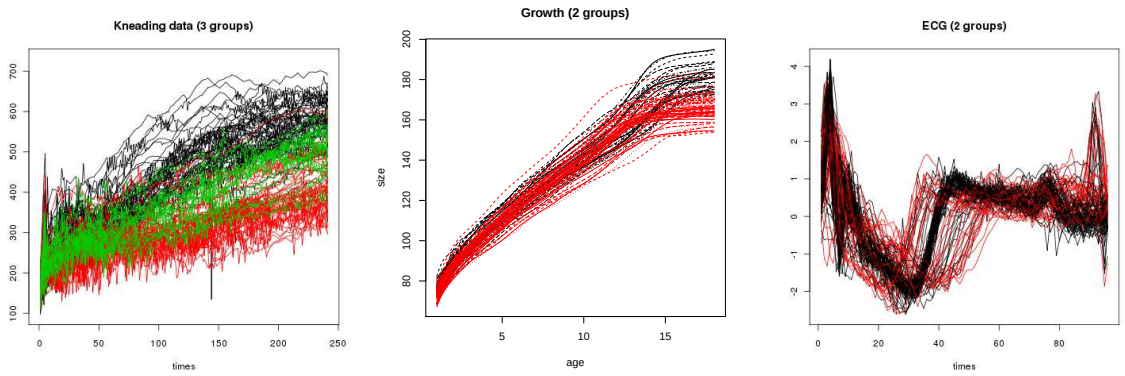



Figure 1: *Kneading*, *Growth* and *ECG* datasets.

available in the *fda* package of the software **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 last dataset (ECG) is taken from the *UCR Time Series Classification and Clustering* website[1]. This dataset consists of 200 curves from 2 groups sampled at 96 time instants, and has already been studied in Olszewski (2001).

*Experimental set-up.* For each dataset, the labels indicating the group membership of observations are available. These labels have been provided by human experts (Kneading and ECG datasets) or simply by the nature of the individuals (gender for the Growth dataset). In order to compare the ability of the clustering methods to retrieve the class labels, we choose to use the correct classification rate (CCR) which measures the adequacy of the resulting clusters with the known partition. This measure varies between 0 and 1, and larger the CCR, better the correspondence between the clustering and the known partition. In order to deal with the labelling problem, all the possible permutations are tested to label the $K$ groups, and the best CCR is retained.
In this benchmark study, Funclust is compared to three challengers dedicated to the clustering of univariate functional data: FunHDDC (Bouveyron and Jacques, 2011) and $k$-centres (kCFC, Chiou and Li (2007)) which are the closest methods and *fclust* (James and Sugar, 2003) which is known to be a good challenger. Note that since no code is available for kCFC (to the best of our knowledge), only the comparison on the Growth dataset is possible, thanks to the classification results presented in Chiou and Li (2007). The number of iterations and the initialization are set as explained in Section 4.2. The threshold of the Cattell scree test allowing to select the approximation order $q_k$ is fixed to 0.05. With this experimental set-up, Funclust estimation is obtained in about 30 seconds for each dataset, on a laptop (2.80GHz CPU) and with a code in **R** software.

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

*Results.* The estimated approximation orders $q_g$ for Funclust are the following: Kneading ($q_1 = 2,\ q_2 = 1,\ q_3 = 3$), Growth ($q_1 = 2,\ q_2 = 3$), ECG ($q_1 = 9,\ q_2 = 4$). These $q_g$'s are relatively close (or slightly greater) to the numbers $q_g^*$'s of principal components retained by the Cattell scree test (with the same threshold of 0.05) when carrying out FPCA on the true classes: Kneading ($q_1^* = 1,\ q_2^* = 1,\ q_3^* = 2$), Growth ($q_1^* = 1,\ q_2^* = 1$), ECG ($q_1^* = 4,\ q_2^* = 5$).
The correct classification rates (CCR) according to the known partitions are given in Table 1. Funclust performs better than the other methods on two datasets among three (Kneading and ECG). On the last dataset, the results are relatively poor (69.89% whereas some method are about 97%), but they can be greatly increased (95.70%) if the dimensions $q_g$ are fixed to 2. 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.

| method | Kneading | Growth | ECG |
|---|---|---|---|
| Funclust | **66.96** | 69.89 | **84** |
| FunHDDC | 62.61 | 96.77 | 75 |
| fclust | 64 | 69.89 | 74.5 |
| kCFC | - | 93.55 | - |

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

*5.2. Simulation study in the case of multivariate functional data*

*The data.* In this simulation study, the number of clusters is assumed to be known: K=2. A sample of $n = 50$ curves are simulated according to the following model inspired by Ferraty and Vieu (2003) and Preda (2007): for $t \in [1, 21]$,

$$\begin{aligned}
\text{Class 1} \ : \quad & X_1(t) = -5 + t/2 + U_2 h_3(t) + U_3 h_2(t) + \sqrt{0.1}\epsilon(t), \\
& X_2(t) = -5 + t/2 + U_1 h_1(t) + U_2 h_2(t) + U_3 h_3(t) + \sqrt{0.5}\epsilon(t), \\
\text{Class 2} \ : \quad & X_1(t) = U_3 h_2(t) + \sqrt{10}\epsilon(t), \\
& X_2(t) = U_1 h_1(t) + U_3 h_3(t) + \sqrt{0.5}\epsilon(t),
\end{aligned}$$

where $U_1 \sim \mathcal{N}(0.5, 1/12)$, $U_2 \sim \mathcal{N}(0, 1/12)$ and $U_3 \sim \mathcal{N}(0, 2/3)$ are independent Gaussian variables and $\epsilon(t)$ is a white noise, independent of $U_i$'s and of unit variance. The function $h_1$, $h_2$ and $h_3$ (plotted on Figure 2) are defined, for $t \in [1, 21]$, by $h_1(t) = (6 - |t - 11|)_+$ where $(\cdot)_+$ mean the positive part, $h_2(t) = (6 - |t - 7|)_+$ and $h_3(t) = (6 - |t - 15|)_+$. The mixing proportions $\pi_i$'s are chosen to be equal, and the curves are observed in 1001 equidistant points ($t = 1, 1.02, \ldots, 21$). Figure 3 plots the simulated curves.

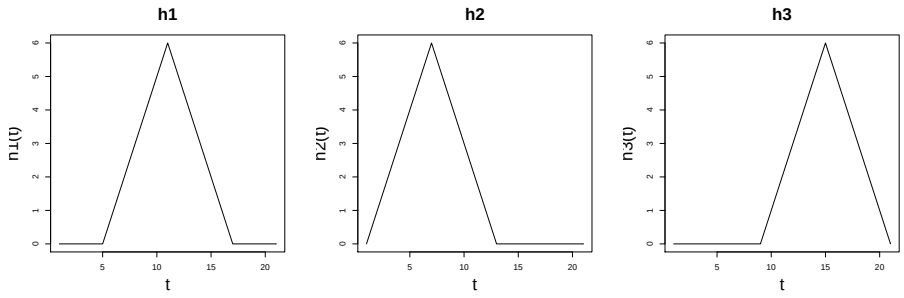



Figure 2: Function $h_1(t) = (6-|t-11|)_+$, $h_2(t) = (6-|t-7|)_+$ and $h_3(t) = (6-|t-15|)_+$ for $t \in [1, 21]$.

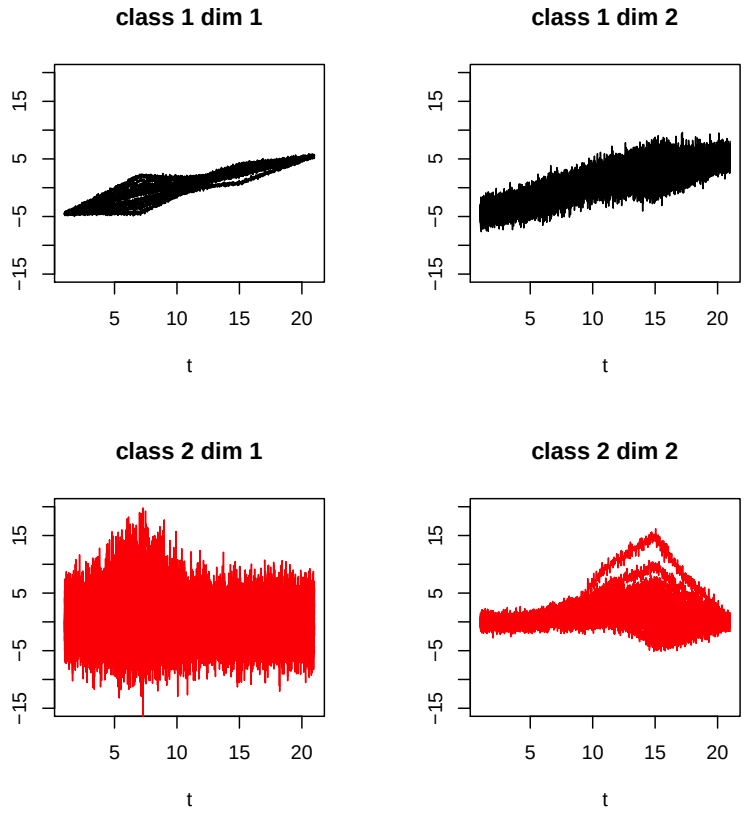



Figure 3: Bi-dimensional simulated curves for class 1 (top) and class 2 (bottom).

*Experimental set-up.* Funclust is compared to the multivariate functional data clustering methods described in Ieva et al. (2011), based on k-means method with the following

distances:

$$d_1(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{j=1}^{2}\int_0^T (X_j(t)-Y_j(t))^2 dt} \quad \text{and} \quad d_2(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{j=1}^{2}\int_0^T (X_j'(t)-Y_j'(t))^2 dt},$$

where $X_j'(t)$ is the derivative of $X_j(t)$. A discussion on these distances in the case of univariate functional data can be found in Ferraty and Vieu (2003). Since no public implementation is available for this method, we built our own implementation in the **R** software. In particular, the distance $d_2$ was computed using the package *fda*.
In addition to these methods, quoted with *kmeans-*$d_1$ and *kmeans-*$d_2$ in the sequel, we consider usual k-means applied on the values of the functions at each observation points $t = 1, 1.02, \ldots, 21$ (quoted as *kmeans-points*) and on the coefficients resulting from a linear spline smoothing with 30 equidistant knots (*kmeans-spline*). Linear spline smoothing has also been used by Funclust, with initialization and iterations number fixed following Section 4.2, and with a Cattell scree test threshold fixed to 0.05. Since both components $X_1$ and $X_2$ have similar covariance structure, the curves have not been normalized.

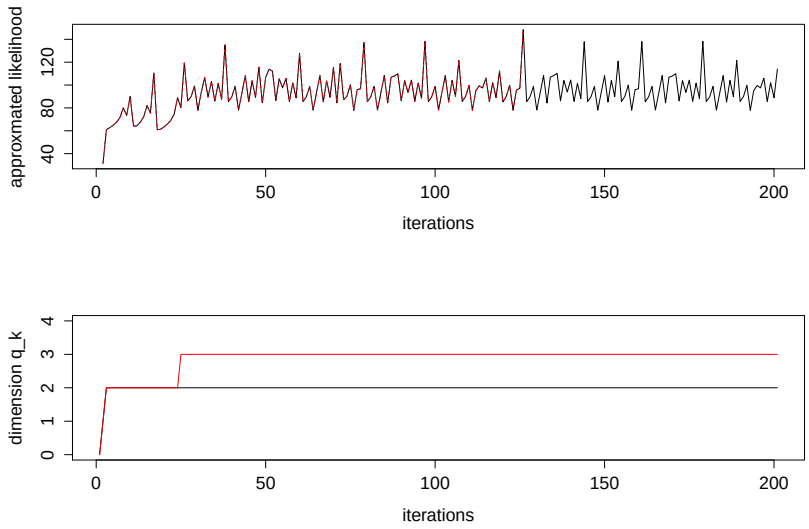



Figure 4: Convergence of the pseudo EM algorithm (top: pseudo likelihood evolution, bottom: approximation orders evolutions). The red part in the pseudo likelihood stops when the maximum is achieved.

*Results.* The convergence of Funclust is illustrated by Figure 4. Table 2 presents the mean and standard deviation of the correct classification rates (CCR), for 100 simulations. The results confirm the good behaviour of Funclust that we have already noticed for univariate functional data.

| method | Funclust | kmeans-$d_1$ | kmeans-$d_2$ | kmeans-points | kmeans-spline |
|---|---|---|---|---|---|
| Mean CCR | **86.80** | 86.32 | 85.76 | 80.60 | 86.14 |
| Std CCR | 14.51 | 16.30 | 10.28 | 14.94 | 15.80 |

Table 2: Mean and standard deviation (for 100 simulations) of correct classification rates (CCR) in percentage for Funclust, kmeans-$d_1$, kmeans-$d_2$ and k-means applied on observations points and spline coefficients.

*5.3. Canadian temperature and precipitation*

In this last application, the Canadian temperature and precipitation data (available in the **R** package *fda* and presented in detail in Ramsay and Silverman (2005)) are used to compared Funclust with Ieva et al. (2011)'s method (kmeans-$d_1$ and kmeans-$d_2$). The dataset consists in the daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994. The goal is to provide a clustering into 4 groups, and to give a geographical interpretation of the resulting clusters.
Since the units of both curves are different (Celsius degrees and millimetres), the data are normalized following methodology presented at the end of Section 2.3. Figure 5 plots original and reduced curves. The curve has been smoothed using Fourier basis with 65 knots, as in Ramsay and Silverman (2005).
Funclust, *kmeans-$d_1$* and *kmeans-$d_2$* are applied on this dataset. For Funclust, the initialization and the iterations number has been chosen following Section 4.2 and the threshold of the Cattell scree test has been fixed to 0.2. The convergence of Funclust is illustrated by Figure 6. Figure 7 presents the clustering into 4 groups of the Canadian weather stations using Funclust. We can observe four distinct groups of stations. The green group is mostly made of northern continental stations, whereas Atlantic stations and southern continental stations are mostly gathered in the red group. The blue group mostly contains Pacific stations and the last group (black) contains only the northernmost station Resolute (N.W.T.). We recall that all these results have been obtained without using the geographical positions of the stations.

In comparison, Figure 8 shows the clustering with *kmeans-$d_1$* and *kmeans-$d_2$* methods, using the same normalized curves and the same basis approximation. This clustering seems less pertinent than Funclust clustering since the blue and red group contains both Atlantic and Pacific stations. Nevertheless, the black group mainly contains continental stations whereas in Funclust clustering, continental and Atlantic stations are gathered together. This last fact is probably due to the Resolute station which is so different from the others (the temperature and precipitation are the lowest) that Funclust clusters this station alone apart in a group.

## 6. Discussion

In this paper we propose a clustering procedure for multivariate functional data based on an approximation of the notion of density for multivariate random function. We

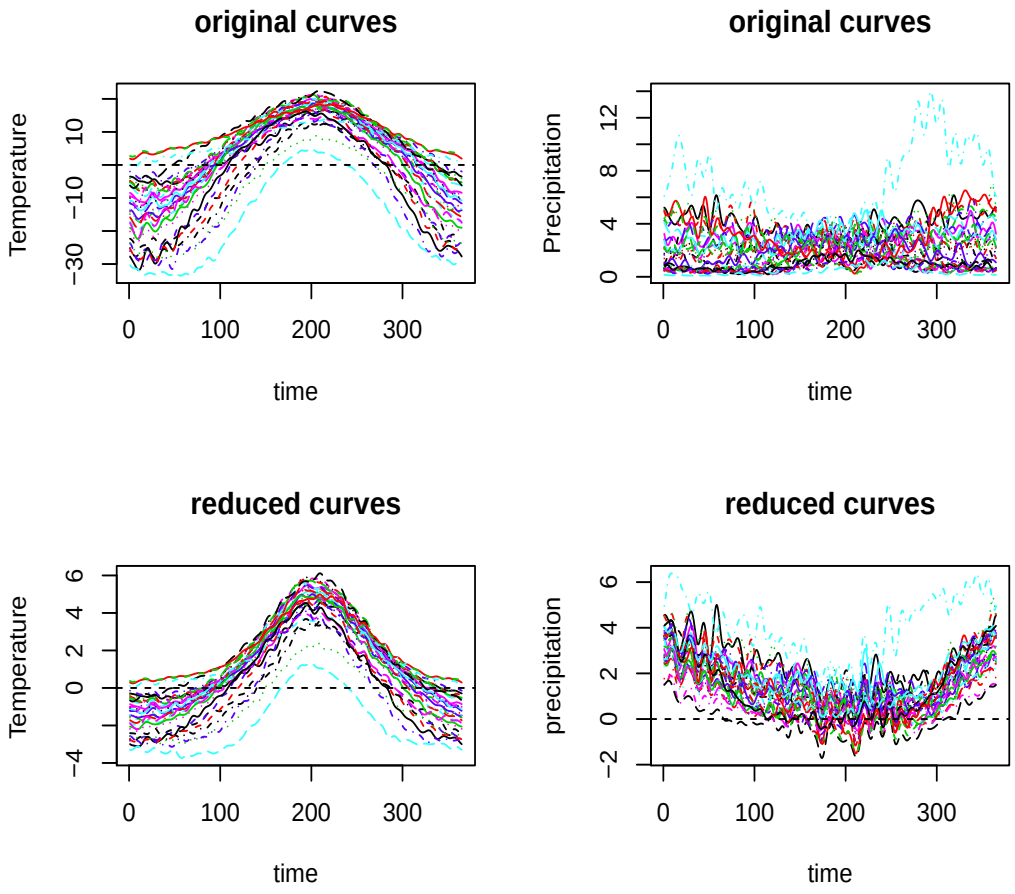



Figure 5: Temperature and precipitation curves for 35 locations in Canada, averaged over 1960 to 1994. The top figures are the original curves and the bottom figures are the reduced ones.

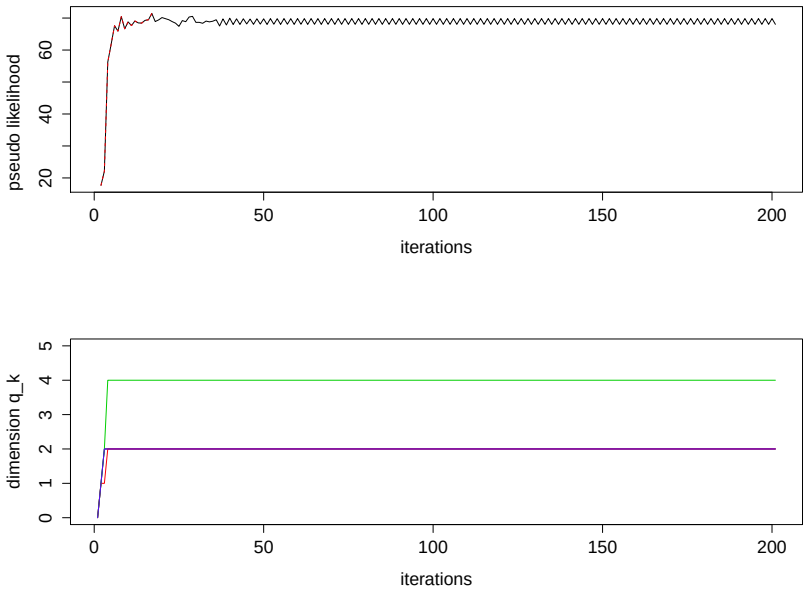



Figure 6: Convergence of the pseudo EM algorithm (top: pseudo likelihood evolution, bottom: approximation orders evolutions). The red part in the pseudo likelihood stops when the maximum is achieved.




Figure 7: Funclust clustering using the reduced curves into 4 groups.

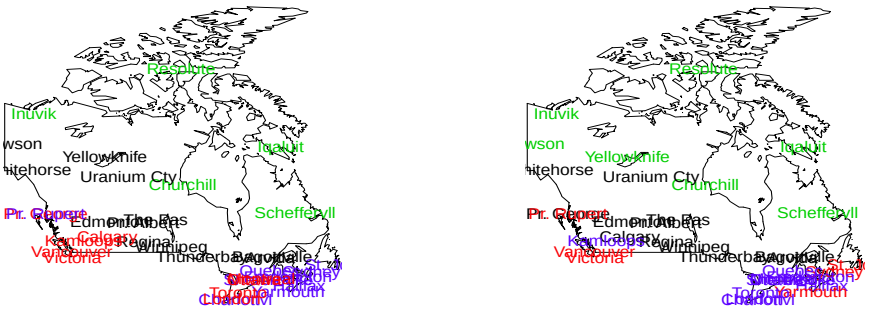



Figure 8: K-means clustering using the reduced curves into 4 groups, with distance $d_1$ (left) and $d_2$ right.

introduce the principal component analysis of multivariate functional data. Assuming normality for the principal components within clusters, we define a mixture model for multivariate functional data clustering. The obtained model is an extrapolation of the finite dimensional Gaussian mixture model to the infinite dimensional setting. With respect to other clustering techniques for multivariate functional data, our methodology has the advantage to take into account the dependency between the curves defining the multivariate data. An EM-like algorithm is proposed for parameter estimation. The results obtained on real and simulated data illustrate the efficiency of our methodology.

Some questions still remain open and further research are to be undertaken to provide answers. First of all, as previously discussed, the selection of the approximation orders is a great challenge for which we actually use an empirical method. Moreover, since only an approximation of the likelihood is available, the question of the convergence of the estimation algorithm is currently without response. However, the heuristic strategy used in this paper provide interesting clusters.

## References

Abraham, C., Cornillon, P. A., Matzner-Løber, E., Molinari, N., 2003. Unsupervised curve clustering using B-splines. Scandinavian Journal of Statistics. Theory and Applications 30 (3), 581–595.

Berrendero, J., Justel, A., Svarc, M., 2011. Principal components for multivariate functional data. Computational Statistics and Data Analysis 55, 2619–263.

Besse, P., 1979. Etude descriptive d'un processus. Ph.D. thesis, Université Paul Sabatier, Toulouse.

Biernacki, C., 2004. Initializing EM using the properties of its trajectories in Gaussian mixtures. Statistics and Computing 14 (3), 267–279.

Bouveyron, C., Girard, S., Schmid, C., 2007. High Dimensional Data Clustering. Computational Statistics and Data Analysis 52, 502–519.

Bouveyron, C., Jacques, J., 2011. Model-based clustering of time series in group-specific functional subspaces. Advances in Data Analysis and Classification 5 (4), 281–300.

Cattell, R., 1966. The scree test for the number of factors. Multivariate Behaviour Research 1 (2), 245–276.

Chiou, J.-M., Li, P.-L., 2007. Functional clustering and identifying substructures of longitudinal data. Journal of the Royal Statistical Society. Series B. Statistical Methodology 69 (4), 679–699.

Delaigle, A., Hall, P., 2010. Defining probability density for a distribution of random functions. The Annals of Statistics 38, 1171–1193.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B. Methodological 39 (1), 1–38.

Deville, J., 1974. Méthodes statistiques et numériques de l'analyse harmonique. Annales de l'INSEE 15, 3–101.

Ferraty, F., Vieu, P., 2003. Curves discrimination: a nonparametric approach. Computational Statistics and Data Analysis 44, 161–173.

Frühwirth-Schnatter, S., Kaufmann, S., 2008. Model-based clustering of multiple time series. Journal of Business and Economic Statistics 26, 78–89.

Guyon, I., Von Luxburg, U., Williamson, R., 2009. Clustering: Science or art. In: NIPS 2009 Workshop on Clustering Theory.

Ieva, F., Paganoni, A., Pigoli, D., Vitelli, V., 2011. ECG signal reconstruction, landmark registration and functional classification. In: 7th Conference on Statistical Computation and Complex System. Padova.

Jacques, J., Preda, C., 2012. Model-based clustering of functional data. In: 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, pp. 459–464.

James, G., Sugar, C., 2003. Clustering for sparsely sampled functional data. Journal of the American Statistical Association 98 (462), 397–408.

Kayano, M., Dozono, K., Konishi, S., 2010. Functional Cluster Analysis via Orthonormalized Gaussian Basis Expansions and Its Application. Journal of Classification 27, 211–230.

Leveder, C., Abraham, C., Cornillon, P. A., Matzner-Løber, E., Molinari, N., 2004. Discrimination de courbes de pétrissage. In: Chimiométrie 2004. Paris, pp. 37–43.

Liu, X., Yang, M., 2009. Simultaneous curve registration and clustering for functional data. Computational Statistics and Data Analysis 53, 1361–1376.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley Interscience, New York.

Olszewski, R., 2001. Generalized feature extraction for structural pattern recognition in time-series data. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Preda, C., 2007. Regression models for functional data by reproducing kernel hilbert spaces methods. Journal of Statistical Planning and Inference 137, 829–840.

Preda, C., Saporta, G., Lévéder, C., 2007. PLS classification of functional data. Computational Statistics 22 (2), 223–235.

Ramsay, J. O., Silverman, B. W., 2005. Functional data analysis, 2nd Edition. Springer Series in Statistics. Springer, New York.

Ray, S., Mallick, B., 2006. Functional clustering by Bayesian wavelet methods. Journal of the Royal Statistical Society. Series B. Statistical Methodology 68 (2), 305–332.

Sangalli, L., Secchi, P., Vantini, S., Vitelli, V., 2010. K-means alignment for curve clustering. Computational Statistics and Data Analysis 54 (5), 1219–1233.

Saporta, G., 1981. Méthodes exploratoires d'analyse de données temporelles. Cahiers du Buro 37–38.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Singhal, A., Seborg, D., 2005. Clustering multivariate time-series data. Journal of Chemometrics 19, 427–438.

Tarpey, T., Kinateder, K., 2003. Clustering functional data. Journal of Classification 20 (1), 93–114.

Titterington, D. M., Smith, A. F. M., Makov, U. E., 1985. Statistical analysis of finite mixture distributions. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester.

Tokushige, S., Yadohisa, H., Inada, K., 2007. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. Computational Statistics 22, 1–16.

Tuddenham, R., Snyder, M., 1954. Physical growth of california boys and girls from birth to eighteen years. Universities of Califiifornia Public Child Development 1, 188–364.