Clustering multivariate functional data

Julien Jacques, Cristian Preda

Clustering multivariate functional data

Julien JACQUES, *University Lille 1 & CNRS & Inria*, julien.jacques@polytech-lille.fr
Cristian PREDA, *University Lille 1 & CNRS & Inria*, cristian.preda@polytech-lille.fr

Abstract. Model-based clustering is considered for Gaussian multivariate functional data as an extension of the univariate functional setting. Principal components analysis is introduced and used to define an approximation of the notion of density for multivariate functional data. An EM like algorithm is proposed to estimate the parameters of the reduced model. Application on climatology data illustrates the method.

Keywords. Multivariate functional data, density approximation, model-based clustering, 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 monograph [14] for more details.

The statistical model underlying data represented by curves is a real-valued 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,

$$\mathbf{X}_t = (X_1(t), \ldots, X_p(t)) \in \mathbb{R}^p, \quad p \geq 2$$

is, curiously, rarely considered in literature. In this case a path of $\mathbf{X}$ is represented by a set of p curves. The dependency between these p measures provides the structure of $\mathbf{X}$. One finds in Ramsay and Silverman [14] a brief example of bivariate functional data, $\mathbf{X}_t = (X_1(t), X_2(t)) \in \mathbb{R}^2$, 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 [3] on random variables with values into a general Hilbert space. Saporta [17] 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. The main difficulty in clustering functional data arises because of the infinite dimensional space data belong. Consequently, most of clustering algorithms for functional data consists in a first step of transforming the infinite dimensional problem into a finite dimensional one and in a second step of using a model-based clustering method designed for finite dimensional data. Examples of theses works in the case of univariate functional data are numerous. A lot of them consider the k-means algorithm, applied on a B-spline fitting [1], on defined principal points of curves [19] or on the truncated Karhunen-Loeve expansion [6]. [16] use also a k-means algorithm for clustering 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: [11] define an approach particularly effective for sparsely sampled functional data, [15] propose a nonparametric Bayes wavelet model for clustering of functional data based on a mixture of Dirichlet processes, [8] build a specific clustering algorithm based on parametric time series models, [5] extend the high-dimensional data clustering algorithm (HDDC, [4]) to the functional case. The case of multivariate functional data is more rarely considered in literature: [18] and [9] use a k-means algorithm based on specific distances between multivariate functional data, whereas [12] consider Self-Organizing Maps based on coefficients of multivariate curves into orthonormalized Gaussian basis expansions.

In the finite dimensional setting, model-based clustering algorithms consider that the data arise from a mixture of density probability. This is not directly applicable to functional data since the notion of density probability 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 [10], the authors use the density surrogate defined in [7] 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 density probability of the first principal components [14] of the curves. This paper proposes an extension of [10] to multivariate functional data. For this, we firstly propose principal component analysis for multivariate functional data. Our model then assumes a cluster-specific Gaussian distribution for the principal component scores. The number of first principal components as well as the computation of the principal component scores are cluster specific.

The paper is organized as follows. In the second section we introduce the model for multivariate functional data and present the principal components analysis of $\mathbf{X}$. Section 3 defines an approximation of the probability density for multivariate functional random variable. The model-based clustering approach and parameters estimation by the mean of an EM-like algorithm are presented in Section 4. Numerical examples on weather data illustrating our approach are presented in Section 5.

2 Principal component analysis for multivariate functional data

Principal component analysis for multivariate functional data has already been suggested in [14] and [2]. In [14] the authors propose to concatenate the functions into a single long function for each observation and then perform FPCA for the concatenated functions. In [2], they propose to not summarize the curves with real principal components as in FPCA, but with functional ones. For this, they carry out a classical multivariate PCA for each value of the domain on which the functions are observed and suggest some interpolation method to build functional principal components. Our approach is similar to [14] but by allowing the use of different basis for the different dimensions of the multivariate curves.

Let Ω be a population space of statistical units (subjects, regions, etc.) and $s_n = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a random sample of size n drawn from Ω. Let $\mathbf{X}$ be a random variable defined on Ω associating to $\omega \in \Omega$ a set of p curves, $p \geq 2$, each one defined on the finite interval $[0, T]$, $0 < T < \infty$, i.e

$$\mathbf{X}(\omega) = \{(X_{\omega,1}(t), \ldots, X_{\omega,p}(t)), t \in [0, T]\}.$$

The observation of $\mathbf{X}$ on the sample s_n provides the set $\{\mathbf{X}(\omega_1), \ldots, \mathbf{X}(\omega_n)\}$ of multivariate curves called *multivariate functional data*. From the s_n curves, one can be interested in optimal representation of curves in a reduced dimensional function space (principal component analysis), or in clustering, by determining an optimal partition of $\mathbf{X}$ 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 $\mathbf{X} = (X_1, \ldots, X_p)$ such that X_ℓ, $\ell = 1, \ldots, p$ are $L_2([0, T])$-valued random variables and $\mathbf{X}$ is a L_2 continuous stochastic process,

$$\lim_{h \to 0} \mathbb{E}\left[\|\mathbf{X}(t+h) - \mathbf{X}(t)\|^2\right] = \lim_{h \to 0} \int_0^T \sum_{\ell=1}^{p} \mathbb{E}\left[(X_\ell(t+h) - X_\ell(t))^2\right] = 0.$$

Let denote by $\mu_\ell = \{\mu_\ell(t) = \mathbb{E}[X_\ell(t)], t \in [0, T]\}$ the mean function of X_ℓ and by

$$\mu = (\mu_1, \ldots \mu_p) = \mathbb{E}[\mathbf{X}],$$

the mean function of $\mathbf{X}$. The covariance operator of $\mathbf{X}$ is defined as an integral operator C with kernel

$$C(t, s) = \mathbb{E}\left[(\mathbf{X}(t) - \mu(t)) \otimes (\mathbf{X}(s) - \mu(s))\right],$$

where $\otimes$ is the tensor product on $\mathbb{R}^p$. Thus, $C(t, s)$ is a $p \times p$ matrix with elements

$$C(t, s)[i, j] = Cov(X_i(t), X_j(s)), \quad i, j = 1, \ldots p.$$

The covariance operator of $\mathbf{X}$, $C : \{L_2([0, T])\}^p \to \{L_2([0, T])\}^p$ is defined by

$$\mathbf{f} \stackrel{C}{\longmapsto} \mathbf{g}, \quad \mathbf{g}(t) = \int_0^T C(t, s)\mathbf{f}(s)ds, \quad t \in [0, T],$$

where $\mathbf{f} = (f_1, \ldots, f_p)$ and $\mathbf{g} = (g_1, \ldots, g_p)$ are elements of $\{L_2([0, T])\}^p$.

Principal components analysis of X

Under the hypothesis of L_2-continuity, C 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 C provides a countable set of positive eigenvalues $\{\lambda_j\}_{j\geq 1}$ associated to an orthonormal basis of eigen-functions $\{\mathbf{f}_j\}_{j\geq 1}$, $\mathbf{f}_j = (f_{j,1}, \ldots, f_{j,p})$, called principal factors:

$$C\mathbf{f}_j = \lambda_j \mathbf{f}_j, \tag{1}$$

with $\lambda_1 \geq \lambda_2 \geq \ldots$ and $\langle \mathbf{f}_i, \mathbf{f}_j \rangle_{\{L_2([0,T])\}^p} = \int_0^T \sum_{\ell=1}^p f_{i,\ell}(t) f_{j,\ell}(t) = \delta_{i,j}$ with $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

The principal components C_j of $\mathbf{X}$ are zero-mean random variables defined as the projections of $\mathbf{X}$ on the eigenfunctions of C,

$$C_j = \int_0^T \langle \mathbf{X}(t) - \mu(t), \mathbf{f}_j(t) \rangle_{\mathbb{R}^p} dt = \int_0^T \sum_{\ell=1}^p (X_\ell(t) - \mu_\ell(t)) f_{j,\ell}(t) dt.$$

Let recall that, as in the univariate setting, the principal components $\{C_j\}_{j\geq 1}$ are zero-mean uncorrelated random variables with variance $\mathbb{V}(C_j) = \lambda_j$, $j \geq 1$.

The following Karhunen-Loeve expansion holds [17],

$$\mathbf{X}(t) = \mu(t) + \sum_{j\geq 1} C_j \mathbf{f}_j(t),$$

and the approximation of order q of $\mathbf{X}$, $q \in \mathbb{N}^*$,

$$\mathbf{X}^{(q)}(t) = \mu(t) + \sum_{j=1}^{q} C_j \mathbf{f}_j(t),$$

is the best approximation of this form under the mean square criterion.

Estimation and computational methods

Let consider the random sample of size n, $s_n = \{\omega_1, \omega_2, \ldots, \omega_n\}$, and denote by $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' = \mathbf{X}(\omega_i)'$. The estimators for μ and C are

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i,$$

and

$$\hat{C}(t,s) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i(t) - \hat{\mu}(t)) \otimes (\mathbf{x}_i(s) - \hat{\mu}(s)).$$

Approximation into a finite basis of functions

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 reduces also the eigen-analysis problem (1) to one in finite dimensional setting.

Let assume that each curve $x_{i\ell}$ $(1 \leq \ell \leq p)$ can be expressed as a linear combination of basis functions $\Phi_\ell = (\phi_{\ell 1}, \ldots, \phi_{\ell q_\ell})$:

$$x_{i\ell}(t) = \sum_{j=1}^{q_\ell} \xi_{i\ell j} \phi_{\ell j}(t).$$

This can be written with the matrix formulation

$$\mathbf{x}_i = \mathbf{\Phi} \mathbf{a}_i'$$

with

$$\mathbf{\Phi} = \begin{pmatrix} \phi_{11} & \ldots & \phi_{1q_1} & 0 & & & \ldots & & & 0 \\ 0 & \ldots & 0 & \phi_{21} & \ldots & \phi_{2q_2} & 0 & \ldots & & 0 \\ & & & & & \ldots & & & & \\ 0 & & & \ldots & & & 0 & \phi_{p1} & \ldots & \phi_{pq_p} \end{pmatrix}$$

and

$$\mathbf{a}_i = (\xi_{i11}, \ldots, \xi_{i1q_1}, \xi_{i21}, \ldots, \xi_{i2q_2}, \ldots, \xi_{ip1}, \ldots, \xi_{ipq_p}).$$

For a set of n sample paths $\underline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ of $\mathbf{X}$, we have

$$\underline{\mathbf{x}} = \tilde{\mathbf{A}} \Phi' \tag{2}$$

with $\tilde{\mathbf{A}}$ the $n \times \sum_{\ell=1}^{p} q_\ell$-matrix, whose rows are the $\mathbf{a}_i$ which contain the basis expansion coefficients of the $\mathbf{x}_i$.

Under the previous basis expansion assumption, the covariance matrix estimator $\hat{C}(t,s)$ is

$$\hat{C}(t,s) = \frac{1}{n}(\underline{\mathbf{x}}(t) - \hat{\mu}(t))'(\underline{\mathbf{x}}(s) - \hat{\mu}(s)) = \frac{1}{n}\Phi(t)\mathbf{A}'\mathbf{A}\Phi'(s),$$

with $\mathbf{A} = (I_n - \mathbb{1}_n(1/n, \ldots, 1/n))\tilde{\mathbf{A}}$.
From Equation (1), each principal factor $\mathbf{f}_j$ admits the same basis expansion as the observed curve (see Equation (3)):

$$\mathbf{f}_j = \Phi \mathbf{b}_j'$$

with $\mathbf{b} = (b_{j11}, \ldots, b_{j1q_1}, b_{j21}, \ldots, b_{j2q_2}, \ldots, b_{jp1}, \ldots, b_{jpq_p})$.

With these assumptions and notations

$$C\mathbf{f}_j = \int_0^T C(t,s)\mathbf{f}_j(s)ds = \int_0^T \frac{1}{n}\Phi(t)\mathbf{A}'\mathbf{A}\Phi'(s)\mathbf{f}_j(s)ds \tag{3}$$

$$= \int_0^T \frac{1}{n}\Phi(t)\mathbf{A}'\mathbf{A}\Phi'(s)\Phi(s)\mathbf{b}_j'ds = \frac{1}{n}\Phi(t)\mathbf{A}'\mathbf{A}\underbrace{\int_0^T \Phi'(s)\Phi(s)ds}_{\mathbf{W}}\mathbf{b}_j' \tag{4}$$

where $\mathbf{W} = \int_0^T \Phi'(s)\Phi(s)ds$ is the symmetric block-diagonal $\sum_{\ell=1}^p q_\ell \times \sum_{\ell=1}^p q_\ell$-matrix of the inner products between the basis functions.

The eigen-analysis problem (1) is then

$$\frac{1}{n}\Phi(t)\mathbf{A}'\mathbf{A}\mathbf{W}\mathbf{b}_j' = \lambda_j \Phi(t)\mathbf{b}_j'$$

which becomes, since it should be true for all t

$$\frac{1}{n}\mathbf{A}'\mathbf{A}\mathbf{W}\mathbf{b}_j' = \lambda_j \mathbf{b}_j'.$$

By defining $\mathbf{u}_j = \mathbf{b}_j\mathbf{W}^{1/2}$, the eigen-analysis problem (1) can be approximated by the usual PCA of the matrix $\frac{1}{\sqrt{n}}\mathbf{A}\mathbf{W}^{1/2}$:

$$\frac{1}{n}\mathbf{W}^{1/2\prime}\mathbf{A}'\mathbf{A}\mathbf{W}^{1/2}\mathbf{u}_j' = \lambda_j \mathbf{u}_j'.$$

The principal factors can finally be obtained by $\mathbf{b}_j = (\mathbf{W}^{1/2\prime})^{-1}\mathbf{u}_j'$, and the principal component scores, $c_j = \mathbf{A}\mathbf{W}\mathbf{b}_j'$. The principal components scores c_j are also solutions of the eigenvalues problem:

$$\frac{1}{n}\mathbf{A}W\mathbf{A}'c_j = \lambda_j c_j.$$

3 Approximation of the density for multivariate functional data

As the notion of probability density is not well defined for functional data, we can use an approximation of the density based on the Karhunen-Loeve expansion, adapted here to the multidimensional nature of the data:

$$\mathbf{X}(t) - \mu(t) = \sum_{j=1}^{\infty} C_j \mathbf{f}_j(t).$$

From this expansion, we propose the following approximation of the density of $\mathbf{X}$:

$$f_{\mathbf{X}}^{(q)}(\mathbf{x}) = \prod_{j=1}^{q} f_{C_j}(c_j(\mathbf{x})), \tag{5}$$

where f_{C_j} is the probability density function of the jth principal components C_j. Delaigle *et al.* [7] show, for univariate functional data, that the approximation error, which decreases when q grows, is under control.

4 A model based-clustering for multivariate functional data

The aim of model-based clustering is to identify homogeneous groups of data from a mixture densities model. In this section, we build a mixture model based on the approximation (5) of the density of $\mathbf{X}$. In the following we suppose that $\mathbf{X}$ is such that each X_ℓ is a zero-mean Gaussian stochastic process ($1 \leq \ell \leq p$). For each $i = 1, \ldots, n$, let associate to the ith observation $\mathbf{X}_i$ of $\mathbf{X}$ the categorical variable $\mathbf{Z}_i$ indicating the group $\mathbf{X}_i$ belongs: $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,K}) \in \{0,1\}^K$ is such that $Z_{i,g} = 1$ if $\mathbf{X}_i$ belongs to the cluster g, $1 \leq g \leq K$, and 0 otherwise. The number K of groups is assumed to be known.

The mixture model

Let assume that each couple $(\mathbf{X}_i, \mathbf{Z}_i)$ is an independent realization of the random vector $(\mathbf{X}, \mathbf{Z})$ where $\mathbf{X}$ has an approximated density depending on its group belonging:

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

where q_g is the number of the first principal components retained in the approximation (5) for the group g, $c_{j,g}(\mathbf{x})$ is the jth principal component score of $\mathbf{X}_{|Z_g=1}$ for $\mathbf{X} = \mathbf{x}$, $f_{C_{j|Z_g=1}}$ its probability density and Σ_g the diagonal matrix of the principal components variances $\mathrm{diag}(\lambda_{1,g}, \ldots, \lambda_{q_g,g})$. Conditionally on the group, the probability density $f_{C_{j|Z_g=1}}$ of the jth principal component of $\mathbf{X}$ 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 $\mathbf{X}_{|Z_g=1}$ is a Gaussian process.

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

$$f^{(q)}_{\mathbf{X}}(\mathbf{x}; \theta) = \sum_{g=1}^{K} \pi_g \prod_{j=1}^{q_g} f_{C_{j|Z_g=1}}(c_{j,g}(\mathbf{x}); \lambda_{j,g}) \tag{6}$$

where $\theta = (\pi_g, \lambda_{1,g}, \ldots, \lambda_{q_g,g})_{1 \leq g \leq K}$ have to be estimated and $q = (q_1, \ldots, q_K)$. From this approximated density, we deduce an approximated likelihood:

$$l^{(q)}(\theta; \underline{\mathbf{X}}) = \prod_{i=1}^{n} \sum_{g=1}^{K} \pi_g \prod_{j=1}^{q_g} \frac{1}{\sqrt{2\pi\lambda_{j,g}}} \exp\left(-\frac{1}{2}\frac{C^2_{i,j,g}}{\lambda_{j,g}}\right) \tag{7}$$

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

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 Z_i are unknown. On the one hand, we need to use an iterative algorithm which alternate 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 $q = (q_1, \ldots, q_g)$ must 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 $\mathbf{Z}_i$) is to use the iterative EM algorithm [13]. In this work we use an EM like algorithm including, between the standard E and M steps, a step in which the principal components scores of each group are updated. Our EM like algorithm consists in maximizing the approximated completed log-likelihood

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

Let $\theta^{(h)}$ be the current value of the estimated parameter at step h, $h \geq 1$.

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

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

where $t_{i,g}$ is the probability for the multidimensional curve $\mathbf{X}_i$ to belong to the group g conditionally to $C_{i,j,g} = c_{i,j,g}$:

$$t_{i,g} = E_{\theta^{(h)}}[Z_{i,g}|\underline{\mathbf{X}}=\underline{\mathbf{x}}] \simeq \frac{\pi_g\prod_{j=1}^{q_g} f_{C_{j|Z_{i,g}=1}}(c_{i,j,g})}{\sum_{l=1}^{K}\pi_l\prod_{j=1}^{q_l} f_{C_{j|Z_{i,l}=1}}(c_{i,j,l})}. \tag{8}$$

The approximation in (8) is due to the use of the approximation of the density of $\mathbf{X}$ by (5).

Principal score updating step. The computation of the principal component scores has been described in Section 2. Here, the principal component scores $C_{i,j,g}$ of the multidimensional curve $\mathbf{X}_i$ in the group g, is updated depending on 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 conditional probabilities $T_g = \mathrm{diag}(t_{1,g},\ldots,t_{n,g})$. Consequently, the first step consists in centring the curve $\mathbf{X}_i$ within the group g by subtraction of the mean curve computed using the $t_{i,g}$'s: the basis expansion coefficients matrix $\mathbf{A}$ in Equation (2) becomes $\mathbf{A}_g = (I_n - \mathbb{I}_n(t_{1,g},\ldots,t_{n,g}))\tilde{\mathbf{A}}$, where I_n and $\mathbb{I}_n$ 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 $\mathbf{A}_g W \mathbf{A}_g' T_g$ associated to the jth eigenvalue $\lambda_{j,g}$.

Group specific dimension q_g estimation step. The estimation of the group specific dimension q_g is an open problem. Indeed, no likelihood criterion can be directly used since the likelihood is directly function of the dimensions q_g. In particular, growing q_g leads to add in the density approximation, above a given order, principal components with small variances, which could artificially increase the approximated likelihood. The estimation of q_g and the investigation of the behaviour of our algorithm when q_g changed is a complete research subject, out of the topic of this paper. In this work we simply propose to use, once the group specific principal component scores have been computed, classical empirical criteria as the proportion of the explained variance in order to select each group specific dimension q_g.

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

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

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

Since only an approximation of the density is available, we stop the EM like algorithm at the convergence of the classification (same classification for a given number of consecutive steps).

5 Illustration: Canadian temperature and precipitation

In this illustrative application, the Canadian temperature and precipitation data (available in the **R** package *fda* and presented in detail in [14]) are used to illustrate the main features of the proposed multivariate functional clustering method. The dataset consists in the daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994. In this study, we compare the classification obtained using separately the temperature data and the precipitation data with the classification obtained using both curves. Comparison with other multivariate functional data custering methods will be carried out in order to be presented during the COMPSTAT conference.

The results presented hereafter have been obtained with the following experimental setup: the percentage of explained variance is fixed at 95%, the number of clusters at 4, and the convergence of the algorithm at 3 identical consecutive classifications (with a maximum number of iterations equal to 100).

Figure 1 presents the result for clustering of the Canadian weather stations using respectively the temperature curves and the precipitation curves. For each clustering, four graphs are plotted: the approximated likelihood and the approximation orders q_k evolutions during the algorithm iterations, the curves clustering and the geographical positions of the Canadian weather stations according to their estimated group belonging. Figure 2 plots the same informations for the clustering of multivariate (precipitation and temperature) curves.

On the one hand, the classification of the Canadian weather stations using the temperature curves exhibits a distinction between the stations according to their latitude: the red group is composed of the stations having the highest temperatures, located in the South of the Canada, the black group is composed of stations of the North of Canada, with colder temperatures than the red group, and the green group contains only one station, Resolute (N.W.T.), which is the coldest and northernmost station. Let notice that even if we ask for four groups, the convergence of the algorithm lead to three classes (the forth being empty).

On the other hand, the classification using the precipitation curves seems to be related to the proximity of one of the Atlantic and Pacific oceans: if the larger group (the red one) is mainly composed of continental stations, the blue group is composed of Atlantic stations and the green and black groups contain essentially Pacific stations. Let notice that the black group, which contains only one station, Prince Rupert (B.C.), is separated from the other Pacific stations, because its precipitation curve is very atypical: the precipitation are by far the most important among the precipitation of all the weather stations, mainly in autumn and winter.

Using both precipitation and temperature curves provides a finer description of the Canadian weather stations. Indeed, we can show on Figure 2 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 black group. The red group mostly contains Pacific stations and the last group (blue) 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.

From an algorithmic point of view, we find that the approximated likelihood is globally increasing during the iterations. Let remark also that this approximated likelihood can be particularly affected when the approximation orders move: on Figure 1, the approximated likelihood decreases when one or several approximation orders decrease. Our last remark concerns these approximations orders: it seems coherent and efficient, on this application, to allow different

orders for each cluster.

6 Discussion

In this paper we propose a clustering procedure for multivariate functional data based on an approximation of the notion of density of a multivariate random function. The main tool is the principal component analysis of multivariate functional data, and the use of the probability densities of the principal components scores. Assuming that the multivariate functional data are sample of a multivariate Gaussian process, the resulting mixture model is an extrapolation of the finite dimensional Gaussian mixture model to the infinite dimensional setting. In comparison of usual clustering techniques, which mostly consist in approximating the functional data into a finite basis and then using a clustering algorithm for finite dimensional data, our multivariate functional clustering procedure has the advantage to take into account the dependency between each univariate functional data composing the multivariate data. An EM like algorithm is proposed for the parameter estimation, with a stopping criterion based on the convergence of the classification. The interest of our model is illustrated by the clustering of Canadian weather stations using multivariate functional data: the annual precipitation and temperature curves. It appears that using both precipitation and temperature leads to a more precise classification of the stations than using separately the precipitation curves or the temperature curves.

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, usual questions which are the selection of the number of clusters or the proofs of the convergence of the estimation algorithm are currently without response.

Bibliography

[1] C. Abraham, P. A. Cornillon, E. Matzner-Løber, and N. Molinari. Unsupervised curve clustering using B-splines. *Scand. J. Statist.*, 30(3):581–595, 2003.

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

[3] P. Besse. *Etude descriptive d'un processus*. PhD thesis, Université Paul Sabatier, Toulouse, 1979.

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

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

[6] Jeng-Min Chiou and Pai-Ling Li. Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(4):679–699, 2007.

[7] A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38:1171–1193, 2010.

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

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

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

[11] G.M. James and C.A. Sugar. Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.*, 98(462):397–408, 2003.

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

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

[14] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[15] Shubhankar Ray and Bani Mallick. Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(2):305–332, 2006.

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

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

[18] A. Singhal and D.E. Seborg. Clustering multivariate time-series data. *Journal of Chemometrics*, 19:427–438, 2005.

[19] T. Tarpey and K.J. Kinateder. Clustering functional data. *J. Classification*, 20(1):93–114, 2003.

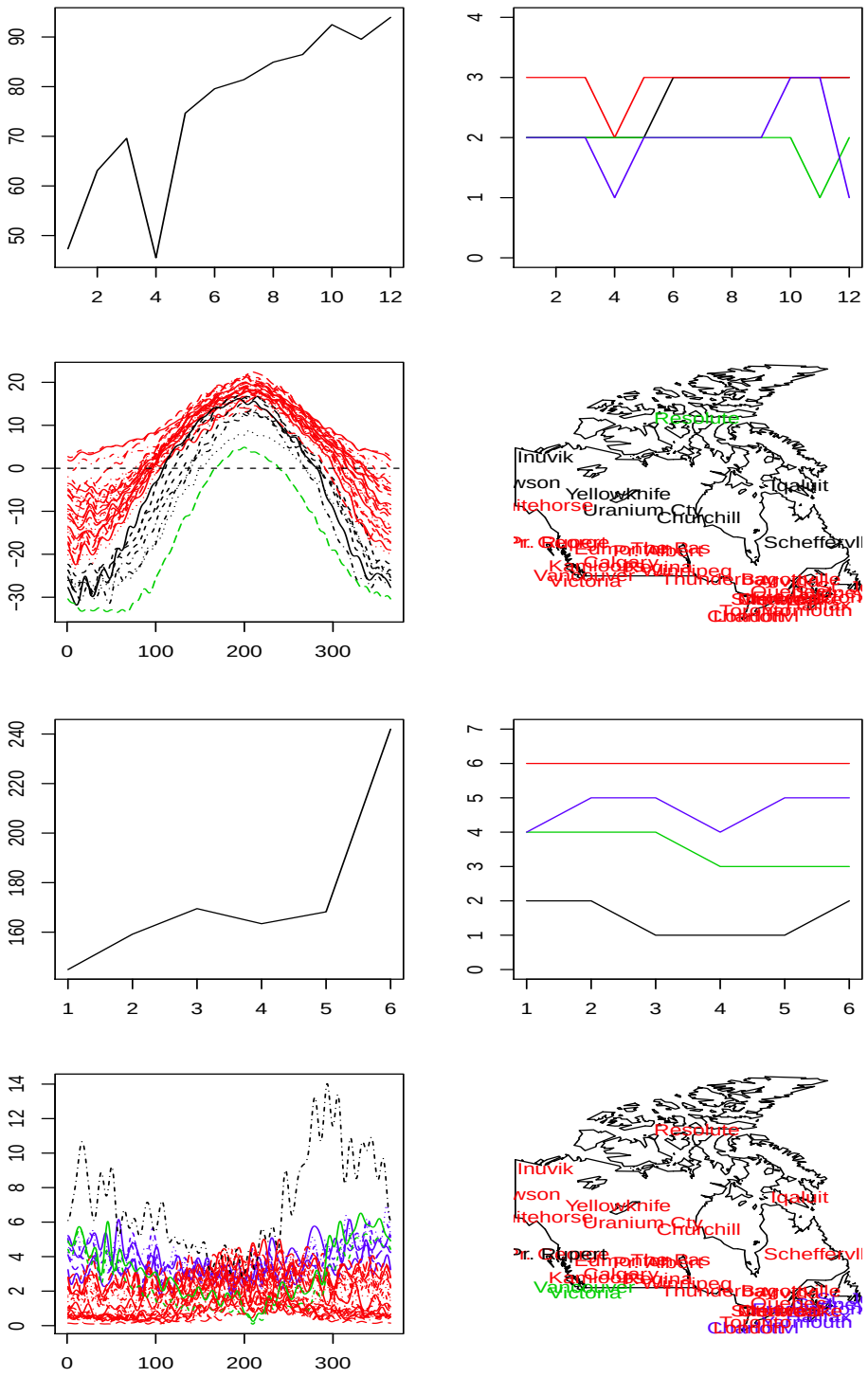

Figure 1. Separate clustering of the Canadian weather stations using respectively the temperature curves (four top graphs) and the precipitation curves (four bottom graphs). For each group of 4 graphs, we have from the top left to the bottom right: the likelihood evolution, the approximation order evolution, the curves classification and the corresponding geographical positions of the weather stations.

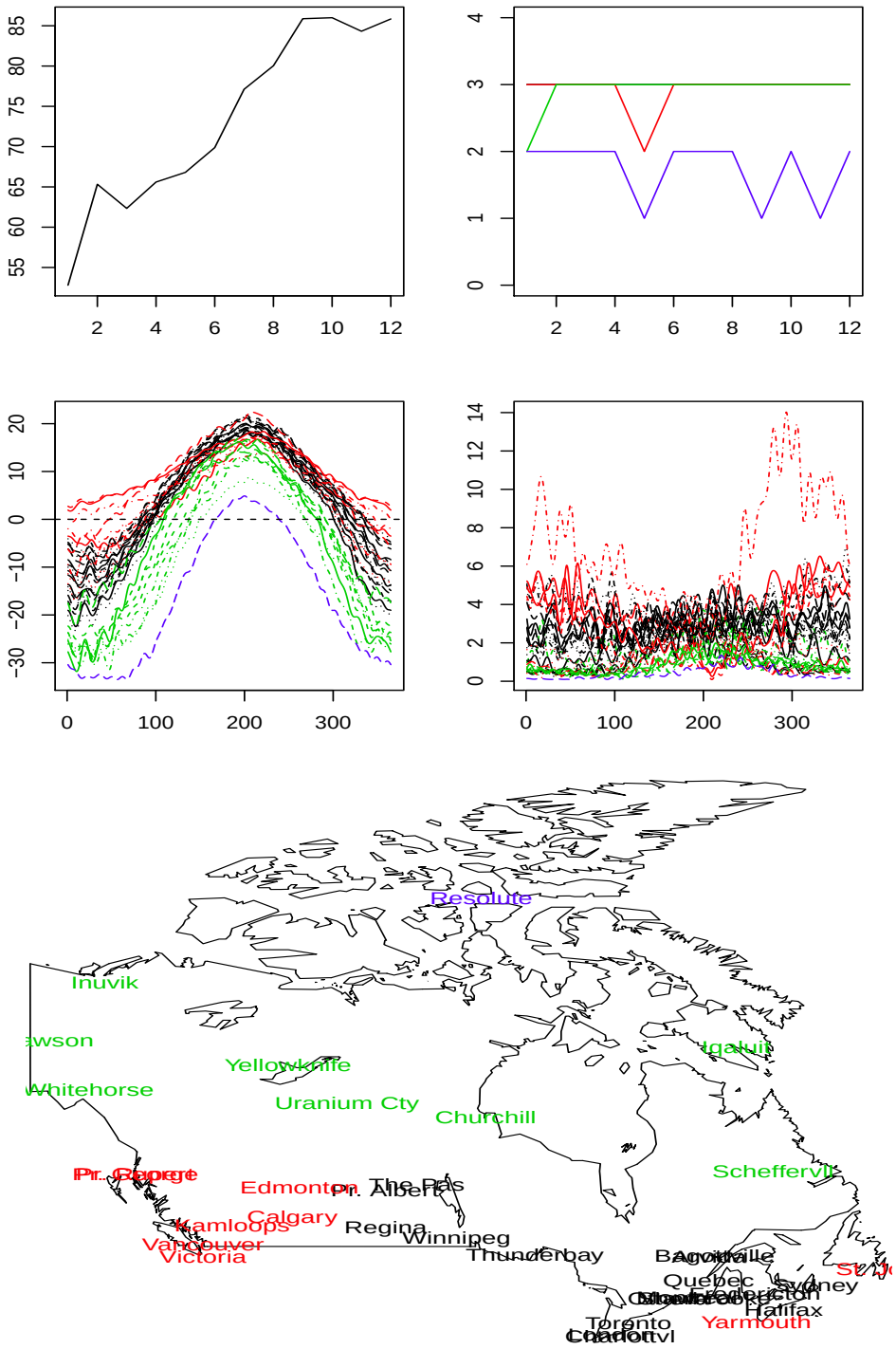

Figure 2. Clustering of the Canadian weather stations using both the temperature and the precipitation curves. The top four graphs represents the likelihood evolution (top left), approximation order evolution (top right) and the temperature curves and precipitation curves classifications. The last graph represents the geographical positions of the Canadian weather stations according to their estimated group belonging.