Clustering for functional data

E.G. Bongiorno[1,*], A. Goia[1]

[1] *Università del Piemonte Orientale; enea.bongiorno@unipmn.it, aldo.goia@unipmn.it*
[*] *Corresponding author*

Abstract. *This work proposes an unsupervised classification algorithm for curves. It extends the density based multivariate cluster approach to the functional framework. In particular, the modes of the small-ball probability are used as starting points to build the clusters. A simulation study is proposed.*

Keywords. *density based clustering; small-ball probability; Karhunen-Loève decomposition.*

Introduction

Cluster analysis, or unsupervised classification, is a set of techniques to segment a collection of data into subsets. When data are curves, or *functional data* (see e.g. [3] and [7] for monographs on this topic), the classical multivariate approaches can not be directly used, due to problems related to the dimensionality of the space to which the data belong. Hence, a variety of specific clustering methods have been introduced in such framework: see for instance [5] and references therein for a recent survey on this topic.

Among the multivariate clustering approaches, an important class is those of the so-called "density oriented" methods. The main idea dates back to Hartigan (see [4]), where clusters are identified as the connected components of the level set (at a given threshold c) of the (multivariate) distribution f of the data; i.e. the connected components of $\{f > c\}$. Differently from the multivariate case, working with functional data, a definition of the density distribution (in the classical sense) is not available. Thus, to implement an equivalent of the Hartigan's approach in the functional context, one can refer to surrogate densities, as the one defined in [2] and based on the Karhunen-Loève truncated expansion (namely, the so-called Functional Principal Component Analysis). Following this principle, in [6] a model based clustering approach has been introduced: in particular, assuming that the underlying distribution of the functional principal scores is a gaussian mixture, the authors use a maximum likelihood and expectation maximization approach to identify the distribution parameters and hence the mixture. Clearly, the distributional assumption in [6] can appear restrictive: in this work we propose a "distribution free" approach, based on the non-parametric estimation of the joint density of a fixed number of principal component

scores. The main idea rests in finding the local maxima of such density (i.e., the modes) and in defining each cluster as the set of observations included in the largest level set that contains only one maximum.

The paper is structured as follows. In Section 1, we introduce the theoretical framework and the clustering method, while, in Section 2, the proposed method abilities are illustrated through an application to simulated data.

1 The clustering approach

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{L}^2_{[0,1]}$ be the Hilbert space of square integrable real functions on $[0,1]$ endowed with the inner product $\langle g,h\rangle = \int_0^1 g(t)\,h(t)\,dt$ and the induced norm $\|g\|^2 = \langle g,g\rangle$. On Ω, define the $\mathcal{L}^2_{[0,1]}$ valued Random Curve (RC) X. Denote by $X^\mu = \{\mathbb{E}[X(t)], t \in [0,1]\}$ and $\Sigma[\cdot] = \mathbb{E}[\langle X - X^\mu, \cdot\rangle (X - X^\mu)]$ its mean function and covariance operator respectively. Consider a sample of n curves X_i, being i.i.d. as the RC X. Thus the empirical versions of X^μ and Σ are: $\overline{X}_n(t) = \frac{1}{n}\sum_i X_i(t)$, $\widehat{\Sigma}_n[\cdot] = \frac{1}{n}\sum_i \langle X_i - \overline{X}_n, \cdot\rangle (X_i - \overline{X}_n)$.

Suppose that Ω is partitioned in K (unknown) groups Ω_k ($k = 1,\ldots,K$) each one with a RC unimodal specific distribution $\mathbb{P}(X \in \cdot \mid \Omega_k)$. Our aim is to determine the groups and to classify each observed X_i by means of a local version of the Hartigan's clustering idea generalized in the functional statistical context. Since it is not possible to define a probability density (in the sense of the Radon-Nikodym derivative with respect to some underlying measure) for functional data, we follow a similar thinking as in [2], where an approximation of the small-ball probability $p(x_0,\varepsilon) = \mathbb{P}(\|X - x_0\| < \varepsilon)$ (for small values of ε) is provided.

In this view, a crucial tool is the Karhunen-Loève expansion (see e.g. [1]): denoting by $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ the decreasing to zero sequence of non–negative eigenvalues and their associated orthonormal eigenfunctions of the covariance operator Σ, the RC X may be represented by

$$X(t) = X^\mu(t) + \sum_{j=1}^{\infty} \theta_j \xi_j(t), \qquad 0 \le t \le 1, \tag{1}$$

where $\theta_j = \langle X - X^\mu, \xi_j\rangle$ are the so-called principal components (PCs) of X satisfying

$$\mathbb{E}[\theta_j] = 0, \qquad Var(\theta_j) = \lambda_j, \qquad \mathbb{E}[\theta_j \theta_{j'}] = 0, \;\; j \neq j'.$$

Proposition 1, whose proof is based on similar arguments of Lemma 13.6 in [3], provides an asymptotic representation of the small-ball probability in terms of the density of the first r PCs.

Proposition 1 *Let r be a finite positive integer. Define the r-dimensional random vector $\mathbf{W} = (W_1 \ldots, W_r)'$, with $W_j = \langle X - x_0, \xi_j\rangle^2$, and assume that*
(i) it has density f_r continuous and strictly positive at $\mathbf{w} = (w_1,\ldots,w_r)'$,
(ii) $\sup_{j\ge 1}\{\mathbb{E}[W_j]/\lambda_j\} = M < \infty$, with M a positive constant.
Then there exists an r_0 large enough such that for any $r > r_0$ whenever ε tends to zero, it holds: $p(l_0,\varepsilon) \sim f_r(\mathbf{w})\,\varepsilon^r \pi^{r/2}/\Gamma(1 + r/2)$.

Thanks to the previous result, we can define the following unsupervised classification algorithm:

1. Obtain an estimate of the covariance operator and of eigenelements;
2. Fix r, compute $\widehat{f}_{r,n}$ (an estimation of the joint distribution density f_r), and look for its local maxima $\widehat{m}_k$ ($k = 1,\ldots,\widehat{K}$);
3. *Finding Prototypes*: for each k in $\{1,\ldots,\widehat{K}\}$, the k-th "prototypes" group is formed by those X_i whose estimated PCs belong to the largest level set of $\widehat{f}_{r,n}$ that contains only the maximum m_k.
4. Connect the unclassified X_i with the $\widehat{K}$ prototypes groups by means of a k-NN.

Attention must be payed choosing r: it should be small enough to avoid the well-known "curse of dimensionality" in estimating $\widehat{f}_r$ but it should be large enough to guarantee a good Karhunen-Loève approximation. For the former task, a kernel density approach require a tuning procedure for the bandwidth since the estimated number of clusters depends on the chosen bandwidth. Note that under the assumption that PCs are independent, the approximation of $p(x_0,\varepsilon)$ depends on the product of the marginal densities of PCs (see e.g. [2]) and, even in this case, the bandwidth choice still needs attention.

2 A simulation example

Simulation setting **-** In order to generate the sample mixture process, we use the functional basis expansion:

$$X_i^{(k)}(t) = \sum_{l=0}^{L} \sqrt{\lambda_l}\tau_{i,l}^{(k)}\varphi_l(t), \qquad t \in [0,1],\ i = 1,\dots,N \ \text{ and } \ k = 1,\dots,K,$$

where $K = 3$ is the number of generated groups for each of which $N = 100$ curves are simulated. The orthonormal basis functions $\{\varphi_l(t)\}$ play the role of eigenfunctions with the corresponding eigenvalues $\{\lambda_l\}$, $l = 1,\dots,L$. Here, we set $L = 150$, $\lambda_l = 0.7^{-l}$ and we choose the Fourier basis

$$\varphi_l(t) = \begin{cases} \sqrt{2}\sin(2\pi m t - \pi), & l = 2m-1; \\ \sqrt{2}\cos(2\pi m t - \pi), & l = 2m. \end{cases}$$

For each $k \in \{1,2,3\}$ and for a fixed $i \in \{1,\dots,N\}$, in order to have uncorrelated but dependent random coefficients $(\tau_{i,l}^{(k)})_{l=1}^{L}$, they are generated as a multivariate shifted t-Student with 10 degrees of freedom, with location parameters

$$\mu^{(k)} = \begin{cases} (\frac{5}{2}, -\frac{5}{2}, \dots, \frac{5}{2}, -\frac{5}{2}) & k = 1, \\ -\mu^{(1)} & k = 2, \\ \mu^{(1)} + \mu^{(2)} & k = 3, \end{cases}$$

and identity covariance matrix.

Numerical Result **-** We estimate the empirical mean, the covariance operator and its eigenelements. Figure 1 shows 30 curves from the sample $\{X_i\}_{i=1}^{300}$ and the empirical mean curves of the three distributions of the used mixture. Since the first two PCs explain the 93,39% of the variability, we implement the algorithm above with $r = 2$. Figure 2 shows the contour plot of $\widehat{f}_2$, the estimated modes $\widehat{m}_k$ and the "prototypes regions" (dashed and bolded contour lines) associated to these modes. After the k-NN procedure, we obtain three groups containing 112, 99 and 89 processes respectively, with a missclassification error equal to 8,33%. Figure 3 depicts such obtained clusters of curves on the sample.

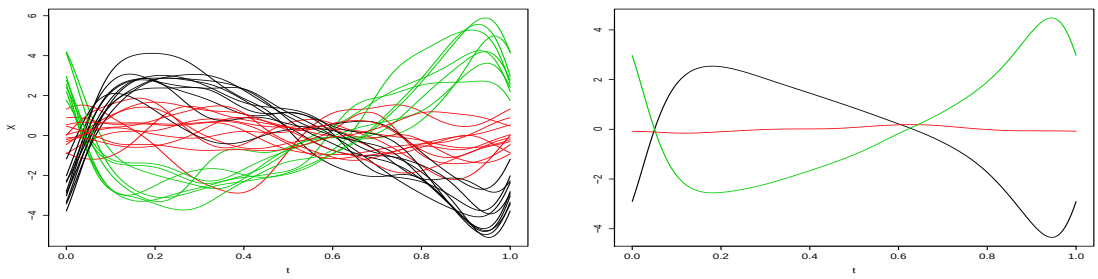

Figure 1: On the left, 30 curves from the sample $\{X_i\}_{i=1}^{300}$. On the right, the empirical mean curves of the three distributions of the used mixture.

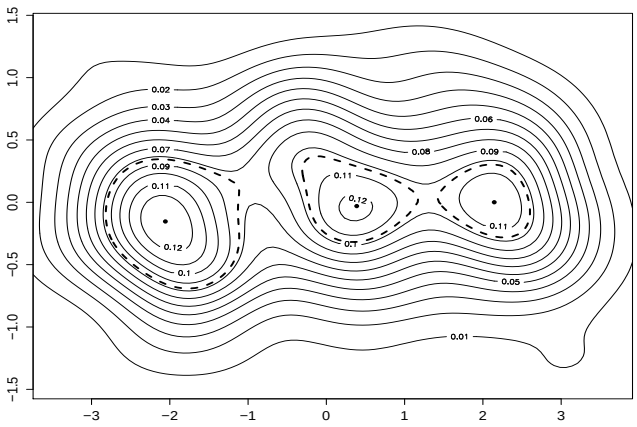

Figure 2: Contour lines of $\widehat{f}_2$, modes and level sets characterizing the prototypes.

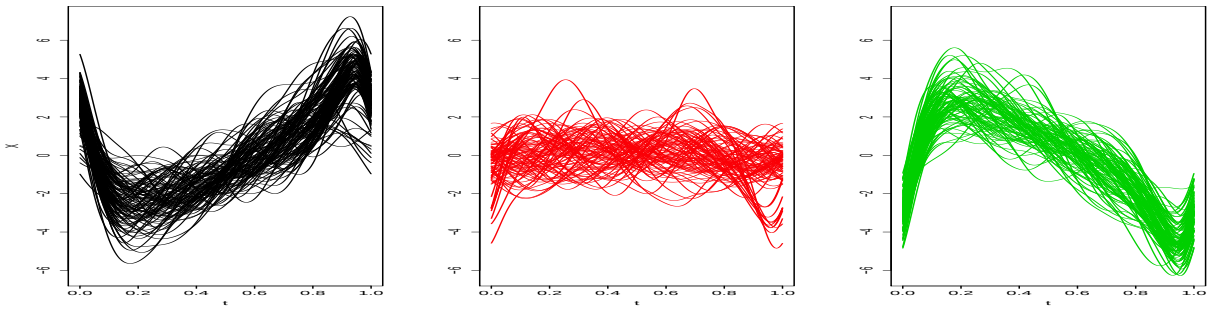

Figure 3: Computed clusters.

References

[1] D. Bosq: Linear Processes in Function Spaces: Theory and Applications. Lectures Notes in Statistics, 149, Springer–Verlag, Berlin (2000)

[2] A. Delaigle, P. Hall: Defining probability density for a distribution of random functions. Ann. Statist. **38**, no.2, 1171–1193 (2010)

[3] F. Ferraty, P. Vieu: Nonparametric functional data analysis. Theory and practice. Springer Series in Statistics (2006)

[4] J. A. Hartigan: Clustering Algorithm. Wiley, New York (1975)

[5] J. Jacques, C. Preda: Functional data clustering : a survey. Adv. Data Anal. Classif. (2014) To appear

[6] J. Jacques, C. Preda: Model-based clustering for multivariate functional data. Comput. Statist. Data Anal. **71**, 92–106 (2014)

[7] J. O. Ramsay, B. W. Silverman: Functional data analysis, 2nd ed., New York: Springer (2005)