ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Curves clustering with approximation of the density of functional random variables

Julien Jacques and Cristian Preda

Laboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France
INRIA Lille-Nord Europe and Polytech'Lille

Abstract. Model-based clustering for functional data is considered. An alternative to model-based clustering using the functional principal components is proposed by approximating the density of functional random variables. An EM-like algorithm is used for parameter estimation and the maximum a posteriori rule provides the clusters. Real data applications illustrate the interest of the proposed methodology.

1 Introduction

Cluster analysis aims to identify homogeneous groups of data without using any prior knowledge on the group labels of data. Several methods, from k-means to probabilistic model-based clustering [1] have been proposed along the years. A particular type of data for which clustering is a difficult task is the functional data (curves or trajectories [2]). The main difficulty in clustering such data arrises because of the infinite dimensional space data belong. The present paper focuses on model-based clustering which, in addition to providing powerfull clustering algorithm, has interesting interpretability properties. Unlike in the case of finite dimensional data vectors, model-based methods for clustering functional data are not directly available since the notion of probability density function generally does not exist for such data [3]. Consequently, current use of model-based clustering methods on functional data consists usually in a first step of transforming the infinite dimensional problem into a finite dimensional one and in a second step using a model-based clustering method designed for finite dimensional data. The representation of functions in a finite dimensional space can be carried out in several ways: discretizing the time interval, approximating data into a finite basis of functions or using some dimension reduction techniques such as functional principal component analysis (FPCA, [2]). These two-step approaches perform the dimension reduction and the clustering steps separately, and this may lead to a loss of discriminative information. Recently, some new approaches, [4, 5], allow the interaction between the two steps by introducing a stochastic model for the basis coefficients. In this paper, we define a new model for functional data clustering, based on an approximation of the notion of probability density of a functional random variable.

Let X be a functional random variable with values in $L_2([0,T])$, $T > 0$, and X is a L_2-continuous stochastic process, $X = \{X(t),\ t \in [0,T]\}$. Let $\underline{X}(X_1, \ldots, X_n)$ be an i.i.d sample of size n from the same probability distribution as X. Model-based clustering consists in identifying homogeneous groups of data

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

from a mixture densities model. As the notion of probability density is not well defined for functional data, we use a "surrogate density" developped in [3]:

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

where f_{C_j} is the probability density function of the principal components $C_j = \int_0^T (X(t) - \mu(t))\psi_j(t)dt$, $j \geq 1$. This approximation of the density is based on the Karhunen-Loeve expansion: $X(t) = \mu(t) + \sum_{j=1}^{\infty} C_j\psi_j(t)$, in which ψ_j's form an orthonormal system of eigen-functions of the covariance operator of X: $\int_0^T Cov(X(t), X(s))\psi_j(s)ds = \lambda_j\psi_j(t), \forall t \in [0, T]$. The eigen-values λ_j are assumed to be indexed upon the descending order ($\lambda_1 \geq \lambda_2 \geq \ldots$).

2 Model-based clustering for functional data

In the following we suppose that X is a zero-mean gaussian stochastic process. For each $i = 1, \ldots, n$, let associate to X_i the corresponding categorical variable Z_i indicating the group X_i belongs : $Z_i = (Z_{i,1}, \ldots, Z_{i,K}) \in \{0, 1\}^K$ is such that $Z_{i,g} = 1$ if X_i belongs to the cluster g, $1 \leq g \leq K$, and 0 otherwise.

2.1 The mixture model

Let assume that each couple (X_i, Z_i) is an independent realization of the random vector (X, Z) where X has an approximated density depending on its group belonging:

$$f_{X_{|Z_g=1}}^{(q_g)}(x; \Sigma_g) = \prod_{j=1}^{q_g} f_{C_{j\,|Z_g=1}}(c_{j,g}(x); \sigma_{j,g}^2)$$

where q_g is the number of the first principal components retained in the approximation (1) for the group g, $c_{j,g}(x)$ is the jth principal component score of $X_{|Z_g=1}$ for $X = x$, $f_{C_{j\,|Z_g=1}}$ its probability density and Σ_g the diagonal matrix $\text{diag}(\sigma_{1,g}^2, \ldots, \sigma_{q,g}^2)$. Conditionally on the group, the probability density $f_{C_{j\,|Z_g=1}}$ of the jth principal component of X is assumed to be the univariate gaussian density with zero mean (the principal component are centered) and variance $\sigma_{j,g}^2$. This assumption is satisfied when X is a Gaussian process.

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

$$f_X^{(q)}(x; \theta) = \sum_{g=1}^{K} \pi_g \prod_{j=1}^{q_g} f_{C_{j\,|Z_g=1}}(c_{j,g}(x); \sigma_{j,g}^2) \tag{2}$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

where $\theta = (\pi_g, \sigma^2_{1,g}, \ldots, \sigma^2_{q_g,g})_{1 \leq g \leq K}$ have to be estimated and $q = (q_1, \ldots, q_K)$ must be selected. As in the finite dimensional setting, we define an *approximated likelihood* of the sample of curves $\underline{X}$ by:

$$l^{(q)}(\theta; \underline{X}) = \prod_{i=1}^{n} \sum_{g=1}^{K} \pi_g \prod_{j=1}^{q_g} \frac{1}{\sqrt{2\pi}\sigma_{j,g}} \exp\left\{ -\frac{1}{2} \left(\frac{C_{i,j,g}}{\sigma_{j,g}} \right)^2 \right\} \tag{3}$$

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

2.2 Parameter estimation

In the unsupervised context the estimation of the mixture model parameters is not as straightforward as in the supervised context since the groups lables Z_i are unknown. A classical way to maximise a mixture model likelihood when data are missing (here the clusters indicators Z_i) is to use the iterative EM algorithm [6]. 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 and onther one in which the group specific dimension q_g are selected. Our EM-like algorithm consists in maximizing the approximated completed log-likelihood

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

which is known to be easier to maximise than its incomplete version (3), and leads to the same estimate. Let $\theta^{(h)}$ be the current value of the estimated parameter at step h, $h \geq 1$.
E step. As the groups indicators $Z_{i,g}$'s are unknown, the **E** step consists in computing the conditional expectation of the approximated completed log-likelihood:

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

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

$$t_{i,g} = E_{\theta^{(h)}}[Z_{i,g} | \underline{X} = \underline{x}] \simeq \frac{\pi_g \prod_{j=1}^{q_g} f_{C_{j|Z_{i,g}=1}}(c_{i,j,g})}{\sum_{l=1}^{K} \pi_l \prod_{j=1}^{q_l} f_{C_{j|Z_{i,l}=1}}(c_{i,j,l})}. \tag{4}$$

The approximation in (4) is due to the use of an approximation of the density of X.
Principal score update step. The computation of FPCA eigenfunctions and scores within a given cluster follows [2]. In general, this computation needs some approximation. The most usual one is to assume that the curve admits an expansion into a basis of functions $\phi = (\phi_1, \ldots, \phi_L)$. Let Γ be the $n \times L$ expansion

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

coefficients matrix and $W = \int \phi\phi'$ be the matrix of the inner products between the basis functions. Here, the computation of the principal component scores $C_{i,j,g}$ of the curve X_i in the group g is updated depending of the current conditional probability $t_{i,g}$ computed in the previous E step. This computation is carried out by ponderating the importance of each curve in the construction of the principal components with the conditional probabilities $T_g = \text{diag}(t_{1,g}, \ldots, t_{n,g})$. Consequently, the first step consists in centering the curve X^i within the group g by substraction of the mean curve computed using the $t_{i,g}$'s. The principal component scores are then given by $C_{i,j,g} = (\lambda_{j,g})^{-1/2}\gamma_i W\beta_{j,g}$ where γ_i is the ith row of Γ, $\beta_{j,g} = W^{-1/2}\mathbf{u}_{j,g}$, $\mathbf{u}_{j,g}$ and $\lambda_{j,g}$ being the jth eigenvector and respectively eigenvalue of the matrix $n^{-1}W^{1/2}\Gamma' T_g \Gamma W^{1/2}$.

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, classical empirical criteria as the proportion of the explained variance.

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 $\pi_g^{(h+1)} = \frac{1}{n}\sum_{i=1}^{n} t_{i,g}$ and $\sigma_{j,g}^{2\ (h+1)} = \lambda_{j,g}$, for $1 \leq j \leq q_g$ where $\lambda_{j,g}$ is the variance of the jth principal component of the cluster g already computed in the principal score update step.

Numerical considerations. In order to avoid the convergence to a local minimum, our EM-like algorithm is launched 20 times with a small number of iterations (20), and the best solution is used for initializing a larger algorithm, in which the convergence is fixed to a growth in the likelihood lower than 10^{-6} with a maximum number of 1000 iterations.

3 Numerical experiments

Kneading data. This application consists in clustering Danone kneading curves. This dataset comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process. The kneading data set is described in detail in [7]. There are 115 different flours for which the dought 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 *adjustable* quality and 40 *bad* quality.

ECG data. This public dataset is taken from the *UCR Time Series Classification and Clustering* website[1]. It consists of 200 curves from 2 groups sampled at 96 time instants (refer to [8] for more details). The Figure 1 plots both datasets.

Results. The accuracy of our model, called *funclust* in the following, is illustrated with respect to usual clustering methods: *HDDC* [9], *MixtPPCA* [10], *k-means*, Gaussian Mixture Model (*GMM* [1]) and hierarchical clustering (*hclust*). All

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

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

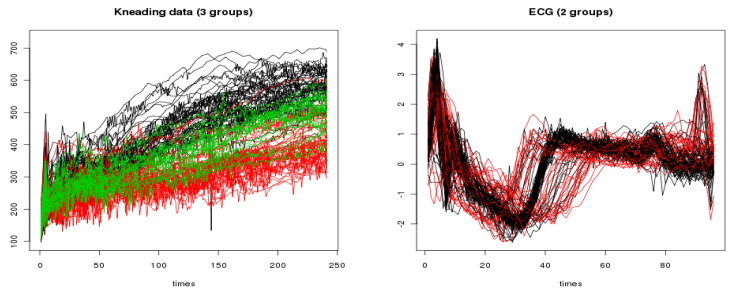

Fig. 1: Kneading data (115 flours observed during 480 seconds) and ECG data (200 curves at 96 time instants).

these methods are successively applied on the discretized data, on the expansion coefficients in a cubic B-spline basis and on the FPCA scores. For *funclust* and the FPCA discretization, the number of dimensions is selected such that at least 95% of the total variance was explained. In order to compare these different clustering procedures, the known class memberships of the data are hidden, a clustering in respectively 3 and 2 groups is performed, and the classifications obtained are then compared to the real (hidden) ones. This is a current way to analyse unsupervised procedures performance.

The clustering results (Table 1) are obtained in about 15 seconds for the ECG dataset and 90 seconds for the Kneading one, on an usual laptop and with a code in R software. Our method *funclust* performs better, on these two datasets, than *fun-HDDC* [5] which similarly to funclust considers group specific subspaces but assume a Gaussian mixture model on the coefficients of the eigen-function expansion, and not on the principal score as *funclust*. The methods from the multivariate finite setting are also outperformed by *funclust* for the Kneading dataset. For the ECG dataset, only *GMM* applied on the FPCA scores is slightly better. Nevertheless, for these two step methods, there is no way to choose between the three discretization techniques, and for the ECG dataset, *GMM* applied on for the two other discretization techniques is not better (according to the clustering accuracy) than *funclust*.

2-steps methods	discretized (241 instants)		spline coeff. (20 splines)		FPCA scores (4 components)		functional methods		
HDDC	66.09	74.5	53.91	73.5	44.35	74.5	fun-HDDC[2]	62.61	75
MixtPPCA	65.22	74.5	64.35	73.5	62.61	74.5	funclust	**67.82**	81
GMM	63.48	81	50.43	80.5	60	**81.5**			
k-means	62.61	74.5	62.61	72.5	62.61	74.5			
hclust	63.48	73	63.48	76.5	63.48	64			

Table 1: Percentage of correct classification for the Kneading dataset (left) and ECG dataset (right)

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4 Conclusion

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

References

[1] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *The Journal of the Pattern Recognition Society*, 28:781–793, 1995.

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

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

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

[5] C. Bouveyron and J. Jacques. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, in press, 2011.

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

[7] C. Lévéder, P.A. Abraham, E. Cornillon, E. Matzner-Lober, and N. Molinari. Discrimination de courbes de pétrissage. In *Chimiométrie 2004*, pages 37–43, Paris, 2004.

[8] R.T. Olszewski. *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2001.

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

[10] M. E. Tipping and C. Bishop. Mixtures of principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.