ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

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, France

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 5 December 2012
Accepted 9 December 2012
Available online xxxx

Keywords:
Multivariate functional data
Density approximation
Model-based clustering
Multivariate functional principal component analysis
EM-algorithm

ABSTRACT

The first model-based clustering algorithm for multivariate functional data is proposed. After introducing multivariate functional principal components analysis (MFPCA), a parametric mixture model, based on the assumption of normality of the principal component scores, 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.

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 analyzed 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,

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

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 curves 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 pioneering 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,

* Corresponding author at: Laboratoire Paul Painlevé, UMR CNRS 8524, University Lille I, Lille, France. Tel.: +33 320 436 760; fax: +33 320 434 302.
E-mail addresses: julien.jacques@polytech-lille.fr (J. Jacques), cristian.preda@polytech-lille.fr (C. Preda).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.12.004

Please cite this article in press as: Jacques, J., Preda, C., Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis (2013), doi:10.1016/j.csda.2012.12.004

the main difficulty is due to the infinite dimensional space that the data belong to. Consequently, most clustering algorithms for functional data consist 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 consider 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 (in press) build a model-based clustering based on an approximation of the notion of density for functional variables.

The case of multivariate functional data is more rarely considered in literature: Singhal and Seborg (2005) and Ieva et al. (2012) 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 an orthonormalized Gaussian basis expansions. Tokushige et al. (2007) extend crisp and fuzzy k-means algorithms for multivariate functional data by considering a 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 variables (Delaigle and Hall, 2010). Consequently, model-based clustering algorithms assume a parametric distribution on some finite sets of coefficients characterizing the curves. In Jacques and Preda (in press), the authors use the density surrogate defined in Delaigle and Hall (2010) to build a model-based clustering algorithm for univariate functional data. This density surrogate, based on the truncation of the Karhunen–Loeve expansion, relies on the probability density of the principal component scores of the curves (Ramsay and Silverman, 2005), which is assumed to be Gaussian.

In this paper we propose an extension of the Jacques and Preda (in press) approach to multivariate functional data. For this, we firstly introduce principal component analysis for multivariate functional data and then 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 the 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 variables. 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 Ramsay and Silverman (2005). 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 their 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 component 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 functional space of reduced dimension (principal component analysis), or in clustering, by determining an optimal partition of the observed curves with respect to some distances or homogeneity criteria. 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:

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

Please cite this article in press as: Jacques, J., Preda, C., Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis (2013), doi:10.1016/j.csda.2012.12.004