Computational Statistics and Data Analysis 79 (2014) 133–148

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Functional factorial $K$-means analysis

Michio Yamamoto[a,*], Yoshikazu Terada[b,1]

[a] Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto, 606-8507, Japan

[b] Center for Information and Neural Networks, National Institute of Information and Communications Technology, 1-4 Yamadaoka, Suita-shi, Osaka, 565-0871, Japan

ARTICLE INFO

*Article history:*
Received 9 March 2013
Received in revised form 29 April 2014
Accepted 15 May 2014
Available online 27 May 2014

*Keywords:*
Functional data
Cluster analysis
Dimension reduction
Tandem analysis
$K$-means algorithm

ABSTRACT

A new procedure for simultaneously finding the optimal cluster structure of multivariate functional objects and finding the subspace to represent the cluster structure is presented. The method is based on the $k$-means criterion for projected functional objects on a subspace in which a cluster structure exists. An efficient alternating least-squares algorithm is described, and the proposed method is extended to a regularized method for smoothness of weight functions. To deal with the negative effect of the correlation of the coefficient matrix of the basis function expansion in the proposed algorithm, a two-step approach to the proposed method is also described. Analyses of artificial and real data demonstrate that the proposed method gives correct and interpretable results compared with existing methods, the functional principal component $k$-means (FPCK) method and tandem clustering approach. It is also shown that the proposed method can be considered complementary to FPCK.

## 1. Introduction

In the last few decades, due to technical advances in storing and processing data, we can obtain large amount of data at hand. A particular case of such data is that of variables taking values into an infinite dimensional space, typically a space of functions defined on some set $T$. Such data are represented by curves or functions and thus called as functional data. Recently, it becomes easier to observe functional data in medicine, economics, psychometrics, and many others domains (for example, see Ramsay and Silverman, 2005, for an overview).

In the framework of functional data analysis, many clustering methods have been already proposed in the literature. A common way to proceed is to filter first, that is to approximate each function by a linear combination of a few number of basis functions, and then to apply a classical clustering method to the resulting basis coefficients. For example, the works of Abraham et al. (2003) and Serban and Wasserman (2005) adopt the filtering approach. Another approach is a distance-based method in which clustering algorithms based on specific distances for functional data are used. In Tarpey and Kinateder (2003), the $k$-means algorithm with the usual $L^2$-metric distance is investigated for Gaussian processes, and they prove that the cluster centers are linear combinations of functional principal component analysis (FPCA) eigenfunctions. In addition, Ferraty and Vieu (2006) propose to use a hierarchical clustering algorithm combined with the $L^2$-metric distance with the semi-metric distance. Recent developments of clustering methods for functional data are excellently overviewed in Jacques and Preda (in press).

[*] Corresponding author. Tel.: +81 75 751 4745; fax: +81 75 751 4732.
*E-mail addresses:* michyama@kuhp.kyoto-u.ac.jp (M. Yamamoto), terada@nict.go.jp (Y. Terada).

[1] Tel.: +81 80 9098 3204.

http://dx.doi.org/10.1016/j.csda.2014.05.010

As described in Jacques and Preda (in press), recently, the other clustering methods for functional data have been developed; the new procedure is to identify simultaneously optimal cluster structure of functions and optimal subspaces for clustering. The use of a low-dimensional representation of functions can be of help in providing simpler and more interpretable solutions. Actually, cluster analysis of functional objects is often carried out in combination with dimension reduction (e.g. Illian et al., 2009; Suyundykov et al., 2010). Bouveyron and Jacques (2011) developed a model-based clustering method for functional data that finds cluster-specific functional subspaces. Yamamoto (2012) proposed a method, called functional principal component $k$-means (FPCK) analysis, which attempts to find an optimal common subspace for the clustering of multivariate functional data. The method aims to overcome the problem of *tandem clustering* (Arabie and Hubert, 1994) for functional data, in which first a dimension-reduction technique, such as FPCA (e.g. Ramsay and Silverman, 2005; Besse and Ramsay, 1986; Boente and Fraiman, 2000), is applied and subsequently the ordinary clustering algorithm is used for the principal component scores. Note that Gattone and Rocci (2012) have also developed a subspace clustering procedure that is essentially equivalent to FPCK, though their method deals with univariate functional data.

The methods of Bouveyron and Jacques (2011) and Yamamoto (2012) can be classified into subspace clustering techniques (Timmerman et al., 2010; Vidal, 2011) for functional data. Like subspace clustering techniques for multivariate matrix data, there are two types of methods for functional data: one intends to find a subspace specific to each cluster (Bouveyron and Jacques, 2011), and the other intends to find a subspace that is common to all clusters (Yamamoto, 2012). Here, we focus on the common subspace clustering.

Yamamoto (2012) shows that in various cases the FPCK method can find both an optimal cluster structure and the subspace for the clustering. The FPCK method, however, has a drawback caused by the definition of its loss function; if no substantial correlation is present in the part of functions which is informative on a cluster structure, FPCK fails in obtaining the cluster structure and a subspace for the structure. The drawback will be explained in more detail in the next section. In this paper, to overcome this drawback, we present a new method that simultaneously finds the cluster structure and reduces the dimension of multivariate functional objects. It will be shown that the proposed method has a mutually complementary relationship with the FPCK method.

This paper is organized as follows. Section 2 defines the notation used in this paper and discusses the drawbacks of FPCK analysis. In Section 3, a new clustering and dimension reduction method for functional objects is described, and an algorithm to implement the method is proposed. In Section 4, the performance of the proposed method is studied using artificial data, and an illustrative application to real data is presented in Section 5. Finally, in Section 6, we conclude the paper with a discussion and make recommendations for future research.

## 2. Notation and the drawbacks of the FPCK method

### 2.1. Notation

First we present the notation that we will use throughout this paper. Here, the same notations as Yamamoto (2012) will be used for ease of explanation. Suppose that the $n$th functional object ($n = 1, \dots, N$) with $P$ variables is represented as $x_n(t) = (x_{np}(t) \mid p = 1, \dots, P)$ with a domain $T \subset \mathbb{R}^d$. For simplicity, we write $x_n = (x_n(t) \mid t \in T)$ to denote the $n$th observed function. In the rest of paper, for general understanding of the problem, we consider the single-variable case, i.e., $P = 1$; in this case, the suffix $p$ in the above notation will be omitted. The multivariate case will be described in Appendix A. Let $\mathscr{L} = L^2(T)$, which is the usual Hilbert space of function $f$ from $T$ to $\mathbb{R}$. Here, the inner product for any $x, y \in \mathscr{L}$ is defined as

$$\langle x, y\rangle := \int_T x(t)y(t)dt,$$

and for any $x \in \mathscr{L}$, $\|x\| := \langle x, x\rangle^{1/2} < \infty$.

For simplicity, we shall assume that the mean function of the $x_n$'s has been subtracted, so without loss of generality, we assume that $\sum_{n=1}^N x_n(t) = 0$ for all $t \in T$.

In this paper, we simultaneously find an optimal projection of the data $\boldsymbol{x} = (x_1, \dots, x_N)'$ onto a low-dimensional subspace and a cluster structure. Let $V = \{v_l\}(l = 1, \dots, L < \infty;\; v_l \in \mathscr{L})$ be orthonormal basis functions of the projected low-dimensional subspace. As with Yamamoto (2012), we call $v_l$ a weight function. In addition, let $\mathrm{P}_v$ be an orthogonal projection operator from the functional data space $\mathscr{L}$ onto the subspace $\mathscr{S}_v$, which is spanned by $V$. Let $U = (u_{nk})_{N\times K}$ be cluster assignment parameters, where $u_{nk}$ equals one if subject $n$ belongs to cluster $k$, and zero, otherwise. Let $N_k$ be the number of subjects that are assigned to the $k$th cluster, and for all $k$, $\bar{x}_k := N_k^{-1}\sum_{n=1}^N u_{nk}x_n$, which is the centroid of the $k$th cluster. In this paper, we consider the crisp clustering, in which each object is assigned to only one group.

A basis function expansion approach is used in many functional data analysis models. Let us approximate an object $x_n$ using a basis function, as follows:

$$x_n \approx \sum_{m=1}^M g_{nm}\phi_m = \boldsymbol{\phi}'\boldsymbol{g}_n,$$