Functional Factorial K-means Analysis

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

[a] *Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan. TEL:+81-75-751-4745, FAX:+81-75-751-4732.*

[b] *Center for Information and Neural Networks, National Institute of Information and Communications Technology, 1-4 Yamadaoka, Suita-shi, Osaka, 565-0871, Japan. TEL:+81-80-9098-3204.*

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 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.

Keywords: Functional data, Cluster analysis, Dimension reduction, Tandem analysis, K-means algorithm
2000 MSC: 62H30, 91C20

*Corresponding author.
Email addresses: michyama@kuhp.kyoto-u.ac.jp (Michio Yamamoto), terada@nict.go.jp (Yoshikazu Terada)

1. Introduction

In the last few decades, due to technical advances in storing and processing data, we can obtain the 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).

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 nth functional object ($n = 1, \ldots, N$) with P variables is represented as $x_n(t) = (x_{np}(t) \mid p = 1, \ldots, P)$ with a domain $T \subset \mathbb{R}^d$. For simplicity, we write $x_n = (x_n(t) \mid t \in T)$ to denote the nth 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 $\boldsymbol{x} = (x_1, \cdots, x_N)'$ onto a low-dimensional subspace and a cluster structure. Let $V = \{v_l\}$ ($l = 1, \ldots, L < \infty$; $v_l \in \mathscr{L}$) be orthonormal basis functions of the projected low-dimensional subspace. As with Yamamoto (2012), we call v_l a weight function. In addition, let $\mathtt{P}_v$ 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 kth 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 kth 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

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

where ϕ_m's ($m = 1, \ldots, M$) are basis functions (e.g., Fourier or B-spline basis functions) and g_{nm} is a coefficient corresponding to (x_n, ϕ_m), and we write $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_M)'$ and $\boldsymbol{g}_n = (g_{n1}, \cdots, g_{nM})'$. Then, we have

$$(x_1, \ldots, x_N)' \approx (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N)'\boldsymbol{\phi} = \mathbf{G}\boldsymbol{\phi}. \tag{1}$$

Similarly, the weight functions described above are expanded by the same basis functions,

$$v_l \approx \sum_{m=1}^{M} a_{lm}\phi_m = \boldsymbol{\phi}'\boldsymbol{a}_l,$$

where $\boldsymbol{a}_l = (a_{l1}, \cdots, a_{lM})'$. Then, we also have

$$(v_1, \ldots, v_L)' \approx (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_L)'\boldsymbol{\phi} = \mathbf{A}'\boldsymbol{\phi}. \tag{2}$$

Let $\mathbf{H}$ be an $M \times M$ matrix that has $\langle \phi_i, \phi_j \rangle$ for the ijth element. Furthermore, let $\mathbf{G}_H = \mathbf{G}\mathbf{H}^{\frac{1}{2}}$ and $\mathbf{A}_H = \mathbf{H}^{\frac{1}{2}}\mathbf{A}$.

2.2. Drawbacks of the FPCK method

As described in Introduction, the clustering method with dimension reduction can produce useful information about the cluster structure that exists in functional data. To attain this purpose, FPCK has been proposed (Yamamoto, 2012), and this method succeeds in extracting a cluster structure that provides useful information. However, the FPCK method has a drawback. A typical example in which the FPCK analysis does not perform well is given as follows:

Example 1. *Consider that a* 100×10 *coefficient matrix* $\mathbf{G}_H$ *consists of two parts,* $\mathbf{G}_H = (\mathbf{G}_1, \mathbf{G}_2)$*, where* $\mathbf{G}_1$ *is a* 100×2 *matrix which defines a cluster structure and* $\mathbf{G}_2$ *is a* 100×8 *matrix whose elements are generated randomly independent of the cluster structure.* $\mathbf{G}_1$ *is shown in the middle panel of Figure 1, and the left panel of the figure shows functional data of* 100 *objects generated through the basis function expansion with the fourth-order B-spline basis functions using* $\mathbf{G}_H$ *as its coefficient. If the FPCK method is applied to the data, we obtain the result shown in the right panel of Figure 1. As seen in the figure, the FPCK method fails to recover the true cluster structure, since there are many misclustered objects.*

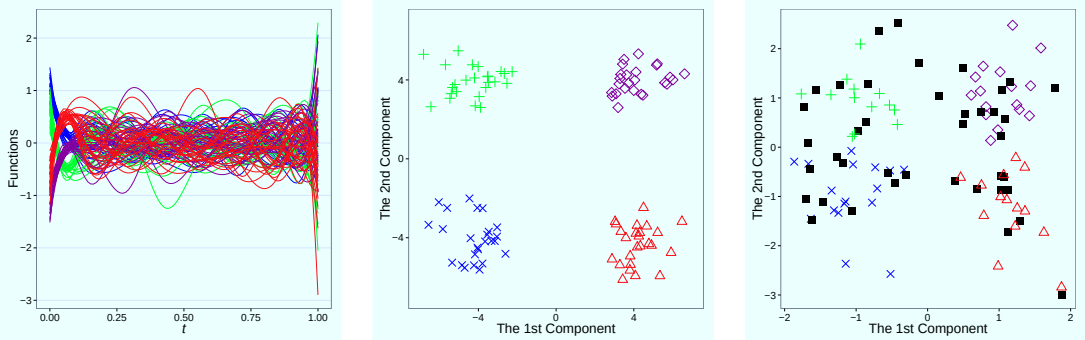

Figure 1: Curves of 100 functional objects (left), the true cluster structure in a two-dimensional subspace (middle), and estimated cluster structure by the FPCK method with two dimensions and four clusters (right). The colors and symbols indicate the true cluster in which each object is grouped. In the right panel, a black square denotes a misclustered object.

This failure of the FPCK method can be explained through the decomposition of its loss function. The loss function L_{fpck} of the FPCK method has the following decomposition:

$$L_{fpck}(U,V) = \sum_{n=1}^{N} \|x_n - \mathtt{P}_v x_n\|^2 + \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}\|\mathtt{P}_v x_n - \mathtt{P}_v \bar{x}_k\|^2. \qquad (3)$$

If we use basis function expansions of the data and weight functions, L_{fpck} is approximated as

$$L_{fpck}(U,V) \approx \|\mathbf{G}_H - \mathbf{G}_H\mathbf{A}_H\mathbf{A}_H'\|^2 + \|\mathbf{G}_H\mathbf{A}_H\mathbf{A}_H' - \mathbf{P}_U\mathbf{G}_H\mathbf{A}_H\mathbf{A}_H'\|^2,$$

where $\mathbf{P}_U$ is a projection matrix onto the space spanned by the columns of $\mathbf{U} = (u_{nk})$. The first term of the right-hand side measures the distance between the coefficient matrix $\mathbf{G}_H$ and the projection of $\mathbf{G}_H$ onto the subspace spanned by the columns of $\mathbf{A}_H$. That is, this term determines the degree of the dimension reduction of the data. On the other hand, the second term measures the distance between the projection of $\mathbf{G}_H$ and the centroid of clusters in the subspace. Based on this formulation, it is found that there are some cases where FPCK analysis does not work well. We illustrate this using a concrete example.

As with Example 1, consider that an $N \times M$ coefficient matrix $\mathbf{G}_H$ consists of two parts, $\mathbf{G}_H = (\mathbf{G}_1, \mathbf{G}_2)$, where $\mathbf{G}_1$ is an $N \times M_1$ matrix that is related to the cluster structure, and $\mathbf{G}_2$ is an $N \times M_2$ matrix ($M = M_1 + M_2$) that is independent of the cluster structure. Usually, N denotes the sample size, and M is the number of basis functions. If $\mathbf{G}_1$ has no substantial correlations, then FPCK analysis is likely to provide a different subspace from that spanned by the true $\mathbf{A}_H$. This is mainly because $\mathbf{G}_1$ is full rank, and the first term of the decomposition may be minimized by weight functions which are different from true ones. It can be inferred that when $\mathbf{G}_1$ is full rank, the FPCK method gets worse with an increase in the column size of $\mathbf{G}_2$. Evidently, it can be seen that, if the contributing part $\mathbf{G}_1$ to the cluster structure has no substantial correlations and the masking part $\mathbf{G}_2$ substantially exists, the FPCK method may fail to find the true cluster structure.

3. Proposed Method

3.1. Criterion of the functional factorial k-means method

To overcome the drawback of FPCK analysis discussed above, we propose a new clustering method with dimension reduction. The notation and settings were explained in Section 2. For ease of explanation, we first consider the case in which there is only one variable, i.e., $P = 1$. Thus, in this section, the suffix p is omitted from the notation. An extension to the multivariate model is straightforward and is described in Appendix A.

A least-squares objective function for the proposed approach, in which the first few principal components of the data are defined to be the most informative about the cluster structure, is

$$L_{ffkm}(U, V) = \sum_{n=1}^{N} \sum_{k=1}^{K} u_{nk} \|\mathtt{P}_v x_n - \mathtt{P}_v \bar{x}_k\|^2. \tag{4}$$

This loss function is optimized over the cluster parameter U and the projected space V.

Here, a component score f_{nl} of subject n for the lth component is defined as $f_{nl} = \langle x_n, v_l \rangle$ using the estimated weight function v_l. Analysis for the first few estimated component scores $\{f_{nl}\}$ ($l = 1, \ldots, L$), where L is two or three, seems to be helpful for the interpretation of a cluster structure in functional data.

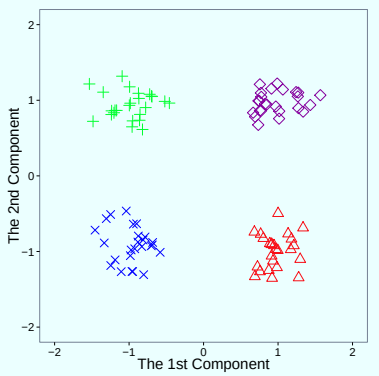

Figure 2: Estimated cluster structures by the FFKM method with two dimensions and four clusters. There were no misclustered objects.

This approach, minimizing the objective function in (4) with respect to U and V simultaneously, is called the functional factorial k-means (FFKM) method because this method is a direct extension of the factorial k-means method (Vichi and Kiers, 2001) to the method for the functional setting. The loss function (4) is equivalent to the second term of the decomposition (3) of the loss function of FPCK. It might be expected that we can resolve the problem of FPCK by ruling out the first term in Eq. (3). Note that this loss function (4) was shortly referred in Yamamoto (2012).

Example 2. *The FFKM method was applied to the data in Example 1. Figure 2 shows the two-dimensional representation of the data given by FFKM. It is found that FFKM recovered the true cluster structure completely.*

3.2. Algorithm for optimizing the proposed criterion

We now present an efficient algorithm for this approach. As in the FPCK method, the loss function (4) can be optimized using the alternating least-squares (ALS) approach, as follows.

STEP1. Initialize parameter V subject to the restriction mentioned above.
STEP2. Minimize the loss function in Eq. (4) for fixed V over U.
STEP3. Minimize the loss function in Eq. (4) for fixed U over V.
STEP4. Go to *STEP2*, or stop.

There are two parts to the algorithm. The first part of the above ALS algorithm is to minimize L_{ffkm} for fixed V over U. To solve the optimization problem, we use a basis function expansion technique described in Section 2. If a projected object $\mathtt{P}_v x_n$ is expanded using some basis function, that is, $\mathtt{P}_v x_n = \boldsymbol{d}_n' \boldsymbol{\phi}$ where $\boldsymbol{d}_n = (d_{n1}, \ldots, d_{nM})'$, then the loss function (4) can be written as

$$\sum_{n=1}^{N} \sum_{k=1}^{K} u_{nk} \|\mathtt{P}_v x_n - \mathtt{P}_v \bar{x}_k\|^2 = \sum_{n=1}^{N} \sum_{k=1}^{K} u_{nk} \|\boldsymbol{d}_n - \bar{\boldsymbol{d}}_k\|_{\mathbf{H}}^2, \tag{5}$$

where $\bar{\boldsymbol{d}}_k$ is a coefficient vector corresponding to the basis function expansion of the projected mean function $\bar{x}_k$ of the kth cluster, and $\|\cdot\|_{\mathbf{H}}$ means the Euclidean norm with the metric $\mathbf{H}$, i.e., for $\mathbf{y} \in \mathbb{R}^M$, $\|\boldsymbol{y}\|_{\mathbf{H}}^2 = \boldsymbol{y}'\mathbf{H}\boldsymbol{y}$. Thus, Eq. (5) can be minimized using the usual k-means algorithm (Lloyd, 1982) for $\mathbf{H}^{\frac{1}{2}} \boldsymbol{d}_n$. Using the expansions in Eq. (2), it is found that $\mathbf{H}^{\frac{1}{2}} \boldsymbol{d}_n = \mathbf{A}_H \mathbf{A}_H' \boldsymbol{g}_{Hn}$.

The second part is to minimize L_{ffkm} regarding V. The loss function in Eq. (4) can be written as (see Yamamoto, 2012, p.246)

$$L_{ffkm}(U, V) = -\sum_{l=1}^{L} \langle v_l, \mathtt{F} v_l \rangle, \tag{6}$$

where $\mathtt{F}$ is an integral operator defined as, for any $y \in \mathscr{L}$,

$$(\mathtt{F}y)(t) := -\sum_{n=1}^{N} \sum_{k=1}^{K} u_{nk} \langle x_n - \bar{x}_k, y \rangle (x_n(t) - \bar{x}_k(t)).$$

Note that it is easily verified that the integral operator $\mathtt{F}$ is a Hilbert-Schmidt integral operator. Thus, $\mathtt{F}$ is a compact operator. In addition, $\mathtt{F}$ is clearly self-adjoint. Minimizing the loss function is, therefore, equivalent to solving the following eigenvalue equation (see, for example, Dunfort and Schwartz, 1988),

$$\mathtt{F}\xi_l = \rho_l \xi_l, \quad \text{subject to} \quad \langle \xi_l, \xi_{l'} \rangle = \delta_{ll'} \tag{7}$$

for $l = 1, \ldots, L$, where $\delta_{ll'}$ is the Kronecker delta. Each eigenfunction $\{\xi_l\}$ $(l = 1, \ldots, L)$ corresponds to a weight function $\{v_l\}$ $(l = 1, \ldots, L)$, which is to be estimated. As with the first part of the ALS algorithm, to solve this eigenvalue problem, we use the basis function expansion. Then, $\mathtt{F}$ operates on a function ξ_l as

$$(\mathtt{F}\xi_l)(t) = \boldsymbol{\phi}'(t) \mathbf{G}' (\mathbf{P}_U - \mathbf{I}_N) \mathbf{G} \mathbf{H} \boldsymbol{a}_l.$$

Eventually, solving the eigenvalue problem (7) amounts to solving the eigenvalue problem

$$\mathbf{G}_H'(\mathbf{P}_U - \mathbf{I}_N)\mathbf{G}_H\boldsymbol{a}_{Hl} = \rho\boldsymbol{a}_{Hl},$$

where $\boldsymbol{a}_{Hl} = \mathbf{H}^{\frac{1}{2}}\boldsymbol{a}_l$. The eigenfunction ξ_l is given by the estimated eigenvector $\boldsymbol{a}_{Hl}$ as the approximation in Eq. (2) using $\boldsymbol{a}_l = \mathbf{H}^{-\frac{1}{2}}\boldsymbol{a}_{Hl}$.

The above ALS algorithm monotonically decreases the loss function L_{ffkm} and the loss function is bounded from below. Then this algorithm guarantees the convergence to a certain point; but it may not be the global minimum. Also, in general, the k-means algorithm, which is utilized in the ALS algorithm, is sensitive to local optima (Steinley, 2003). Thus, to safeguard against those local minima, the proposed algorithm needs to be repeated with a number of random initial starts for V.

3.3. Regularized method

In this section, we propose a smoothing method for the FFKM model. If functional data can be assumed to be sufficiently smooth, taking into account their smoothness often provides better results (Ramsay and Silverman, 2005). To take into account such smoothness in FFKM, we may assume that functions exist in some smooth functional space such as Sobolev space (Silverman, 1996). It can be achieved by using the following inner product instead of the usual inner product defined earlier, for $x, y \in \mathscr{L}$,

$$\langle x, y\rangle_\lambda := \langle x, y\rangle + \lambda\langle \mathtt{D}^2 x, \mathtt{D}^2 y\rangle,$$

where $\mathtt{D}^2$ denotes the second-order differential operator and λ is a roughness penalty. For any $x \in \mathscr{L}$, $\|x\|_\lambda := \langle x, x\rangle_\lambda^{1/2}$. Let $\mathtt{S}_\lambda^2$ be the usual spline smoothing operator (Green and Silverman, 1994). Then, to obtain smooth weight function v_l, the following loss function is minimized over U and V

$$L_{ffkm}(U, V) = \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}\|\mathtt{P}_v\mathtt{S}_\lambda^2 x_n - \mathtt{P}_v\mathtt{S}_\lambda^2\bar{x}_k\|_\lambda^2.$$

The parameters U and V, which minimize $L_{ffkm}(U, V)$, are estimated using an ALS algorithm similar to that for the non-regularized FFKM method, though there are two differences between the two methods: in the regularized method, the inner product $\langle\cdot,\cdot\rangle_\lambda$ is used and the smoothed data $\mathtt{S}_\lambda^2 x_n$ is expanded. Let $\boldsymbol{g}_{\lambda,n}$ be a vector with length M containing coefficients corresponding to the basis function expansion of $\mathtt{S}_\lambda^2 x_n$, and let $\mathbf{H}_\lambda$ be an $M \times M$

matrix in which the ijth element is $\langle \phi_i, \phi_j \rangle_\lambda$. Furthermore, let $\mathbf{A}_{H_\lambda} = \mathbf{H}_\lambda^{\frac{1}{2}} \mathbf{A}$ and $\mathbf{W} = \mathbf{H}_\lambda^{\frac{1}{2}} \mathbf{A}_{H_\lambda} \mathbf{A}'_{H_\lambda} \mathbf{H}_\lambda^{\frac{1}{2}}$. Then, in *STEP2* of the ALS algorithm, the optimal U is obtained by minimizing the following loss function for fixed V over U:

$$\sum_{n=1}^{N} \sum_{k=1}^{K} u_{nk} \| \boldsymbol{g}_{\lambda,n} - \bar{\boldsymbol{g}}_k \|_{\mathbf{W}}^2,$$

where $\| \cdot \|_{\mathbf{W}}$ is the Euclidean norm with metric $\mathbf{W}$. Thus, as with the non-regularized method, this loss function will be optimized using the usual k-means algorithm for $\mathbf{A}_{H_\lambda} \mathbf{A}'_{H_\lambda} \boldsymbol{g}_{H_\lambda n}$.

Next, we describe how to estimate the weight functions V in *STEP3*. Using the above basis function expansion to estimate the optimal V, the following eigenvalue problem is considered:

$$\mathbf{G}'_{H_\lambda} (\mathbf{P}_U - \mathbf{I}_n) \mathbf{G}_{H_\lambda} \boldsymbol{a}_{H_\lambda l} = \rho \boldsymbol{a}_{H_\lambda l},$$

where $\boldsymbol{a}_{H_\lambda l}$ is the lth column of $\mathbf{A}_{H_\lambda}$. Then, as in the non-regularized method, the smoothed weight function v_l is approximated as $v_l \approx \boldsymbol{\phi}' \mathbf{H}_\lambda^{-\frac{1}{2}} \boldsymbol{a}_{H_\lambda l}$.

A component score f_{nl} can be defined as that for the FFKM method, $f_{nl} = \langle x_n, v_l \rangle$. The determination of the value of λ is presented in the next section.

3.4. Model selection

Prior to applying the above algorithm, we need to determine the values of parameters: the smoothness of the basis functions λ, the number of clusters K, and the dimensionality of the subspace L. We adopt the same procedure as those recommended in Yamamoto (2012). The value of λ is determined by generalized cross-validation (GCV) criterion for individual functions. The value of K is chosen according to usual decision procedures, such as those described in Milligan and Cooper (1985) and Hardy (1996). Then, for the selection of L, it is recommended to first take $L = K - 1$ and then check the adequacy of the dimensionality. For instance, it may be useful to check whether the cluster centroids appear to lie in a lower-dimensional plane, in which case it is advised to refit the FFKM model with fewer components. By thus verifying the solutions for different numbers of clusters, one can select the solution that gives the most interpretable results.

4. Analyses of Artificial Data

4.1. Data and evaluation procedures

To investigate the performance of the FFKM method, artificial data, which included a known low-dimensional cluster structure, were analyzed by four different methods: (i) the FFKM method, (ii) the two-step FFKM method (FFKMts) (iii) the FPCK method, and (iv) tandem analysis (TA) that consisted of FPCA using a basis function expansion (Ramsay and Silverman, 2005) followed by a standard k-means cluster analysis of the principal component scores on the first L principal components. Note that the loss function of FFKM is bounded above by the squared norm of the projected functional data as follows:

$$\sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}\|\mathtt{P}_v x_n - \mathtt{P}_v \bar{x}_k\|^2 \leq \sum_{n=1}^{N} \|\mathtt{P}_v x_n\|^2. \tag{8}$$

Thus, when an empirical covariance operator of functional data has excessively small eigenvalues compared with the others, the subspace spanned by eigenfunctions corresponding to the small eigenvalues provides the smallest values of loss function of FFKM regardless of cluster assignments. In fact, when the smallest eigenvalue of an empirical covariance operator is zero, using the corresponding eigenfunction as a weight function for $\mathtt{P}_v$ sets the value of right-hand side of (8) to zero, and then the loss of FFKM is also zero. That is, if there exist trivial dimensions of functional data, FFKM may fail to find the optimal cluster structure. Thus, to avoid such trivial solutions of FFKM, here we introduce a two-step approach, called two-step FFKM. The two-step FFKM method is a two-step approach in which first we eliminate trivial dimensions from the data and then apply the FFKM algorithm to the reduced data. This two-step approach can improve the efficiency of the FFKM method when the coefficient matrix $\mathbf{G}_H$ has some correlations. This two-step approach is described in Appendix B in more detail. The artificial functional data had a structure of four clusters in a two-dimensional subspace, i.e., $L = 2$ and $K = 4$.

As described in Section 2.2, we suppose that the coefficient matrix $\mathbf{G}_H$ consists of two parts, $\mathbf{G}_H = (\mathbf{G}_1, \mathbf{G}_2)$, where $\mathbf{G}_1$ is an $N \times M_1$ matrix that is related to the cluster structure and $\mathbf{G}_2$ is an $N \times M_2$ matrix that is independent of the cluster structure. Let an $N \times L$ component score matrix $\mathbf{F}$ have a cluster structure with N objects drawn from four bivariate normal distributions with the same covariance matrices, $\mathbf{I}_2$, and different means. Let $\mathbf{A}_1$

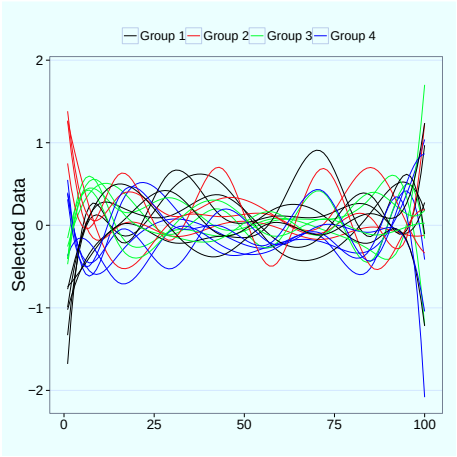

Figure 3: Selected artificial data. The color denotes to which group each functional object was assigned. The proportion of overlap is 0.05.

be an $M_1 \times L$ orthonormal matrix whose elements were randomly generated and subsequently orthonormalized. Using these matrices, the matrix $\mathbf{G}_1$ was calculated as $\mathbf{G}_1 = \mathbf{F}\mathbf{A}_1'$. The elements of $\mathbf{G}_2$ were generated according to a strategy described later.

Let $\boldsymbol{\phi}$ be the fourth-order B-spline basis functions with eight knots, and let $\mathbf{\Phi}$ be a $T \times M$ matrix whose tmth element is $\phi_m(t)$. In this simulation study, we consider 100 sampling points $\boldsymbol{t} = (1, \ldots, 100)$ and 10 basis functions. Then, an artificial data matrix that includes discretized functional data was calculated as $\mathbf{X} = \mathbf{G}_H \mathbf{H}^{-\frac{1}{2}} \mathbf{\Phi}'$. Note that before calculating $\mathbf{X}$, the columns of $\mathbf{G}_H$ were standardized. The artificial data selected are shown in Figure 3.

In this simulation analysis, four factors were manipulated in the experiment: (1) the number of objects (N), (2) the expected proportion of overlap (PO) between clusters in the correct subspace, (3) the ranks of the coefficient matrices $\mathbf{G}_1$ and $\mathbf{G}_2$, and (4) the number of variables which have no information about the true cluster structure (the number of non-informative variables, NN). The number of objects was varied from 100 to 500 in steps of 200. The PO was defined as the proportion of shared density between clusters, as proposed by Steinley and Henson (2005). The PO was set at four levels: 0.0001, 0.05, 0.10, and 0.15. To offer an impression of the effect of the manipulation of the PO, an example of $\mathbf{F}$ for 200 objects in four clus-

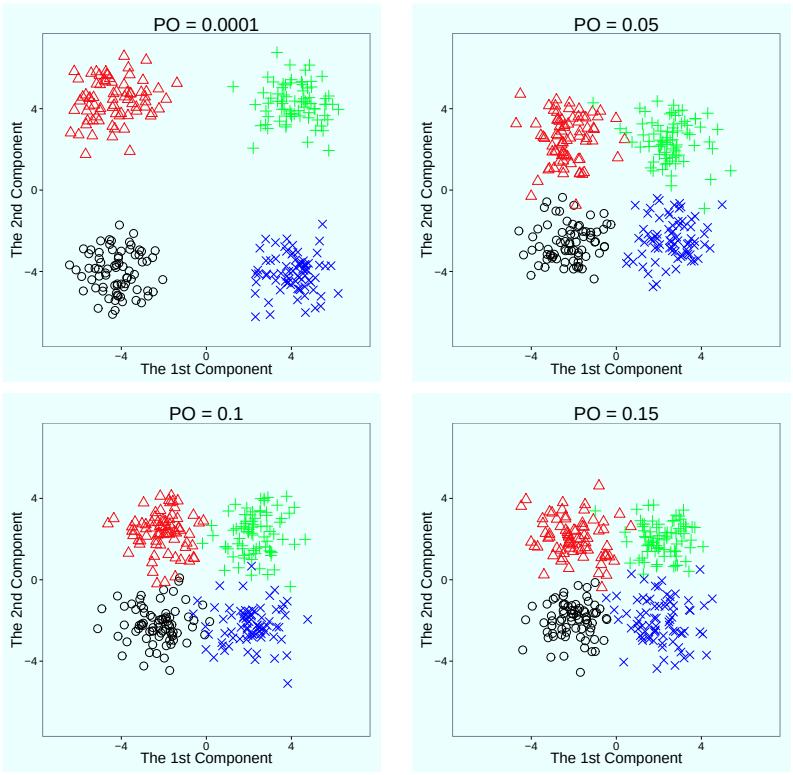

Figure 4: Example of simulated component scores for 300 objects in four clusters in the correct two-dimensional subspace at four levels of proportion of overlap (PO).

ters is depicted in Figure 4, for each of the different levels of the PO. We considered two cases for the ranks of $\mathbf{G}_1$ and $\mathbf{G}_2$: full rank (FR) or rank deficient (RD). The rank of $\mathbf{G}_1$ was controlled by the number of columns M_1, which was set at 2 for an FR case and 5 for an RD case. For an FR case of $\mathbf{G}_2$, the elements of $\mathbf{G}_2$ were independently drawn from a standard normal distribution $N(0, 1)$, while for an RD case, $\mathbf{G}_2$ was calculated as $\mathbf{G}_2 = \mathbf{E}\mathbf{A}_2'$, where $\mathbf{E}$ is an $N \times 2$ matrix and $\mathbf{A}$ is an $M_2 \times 2$ matrix. The elements of $\mathbf{E}$ and $\mathbf{A}_2$ were independently drawn from $N(0, 1)$ and $\mathbf{A}_2$ was subsequently orthonormalized. When $\mathbf{G}_1$ and $\mathbf{G}_2$ are FR, FFKM works well but FPCK does not. On the other hand, when $\mathbf{G}_1$ is RD and $\mathbf{G}_2$ is FR, FPCK works well but FFKM does not. Furthermore, it can be inferred that both FFKM and FPCK are effected negatively by the rank deficiency of $\mathbf{G}_2$. A non-informative variable $\mathbf{Z}$ was also generated through the basis function

expansion $\mathbf{Z} = \mathbf{G}_H^* \mathbf{H}^{-\frac{1}{2}} \mathbf{\Phi}'$ in which elements of a coefficient matrix $\mathbf{G}_H^*$ were independently drawn from $N(0, 1)$ and standardized to have a same variance with the informative data $\mathbf{X}$. In this study, the number of non-informative variables was set at three levels: 0, 1, and 2. The experimental design was fully crossed, with 50 replicates per cell, yielding $3 \times 4 \times 4 \times 3 \times 50 = 7200$ simulated data sets.

The cluster membership recovery was assessed by the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). The ARI has the maximal value of 1 in the case of a perfect recovery of the underlying clustering structure, and a value of 0 in the case where the true membership U and estimated membership $\hat{U}$ coincide no more than would be expected by chance. When the PO is high, the k-means clustering in the true subspace defined by the true $\mathbf{A}_1$ does not work. Thus, in order to calculate the ARI, the k-means clustering with 100 random starts was conducted with the true $\mathbf{F}$, and then the estimated cluster structure was considered to be the true cluster structure.

The FFKM and FPCK methods need initial values for the parameters in the first step of the algorithms. In our limited experience, FFKM is rather sensitive to local optima so that it needs many initial values. Thus, in this simulation, we used 1000 random initial values for FFKM and 100 random initial values for FPCK. For two-step FFKM, a selection of the number R of components in the first step is needed. In this simulation, R was determined in view of cumulative percentage of the total variation (Jolliffe, 2002) in which a selected cut-off provided 90% cumulative variation.

4.2. Results

Boxplots of the ARIs obtained by the four methods are shown in Figure 5, 6, 7, and 8 that are results for the cases of (FR, FR), (RD, FR), (FR, RD), and (RD, RD), respectively, corresponding to the ranks of $(\mathbf{G}_1, \mathbf{G}_2)$. The modified boxplot (Hubert and Vandervieren, 2008) was used for the asymmetry of the distributions of ARIs. In these figures, boxplots of four methods for each sample size are arranged by the proportion of overlap (PO) and the number of non-informative variables (NN). As can be inferred from Figure 5, when both $\mathbf{G}_1$ and $\mathbf{G}_2$ were FR, under all conditions, FFKM and two-step FFKM showed the best result, or at least a result comparable to those of the other two methods. It can be seen that ARIs became worse with an increase in PO and NN, while the indices improved with an increase in the sample size. FPCK also worked well only under the easiest condition where PO was small, N was large, and there was no non-informative variable. This

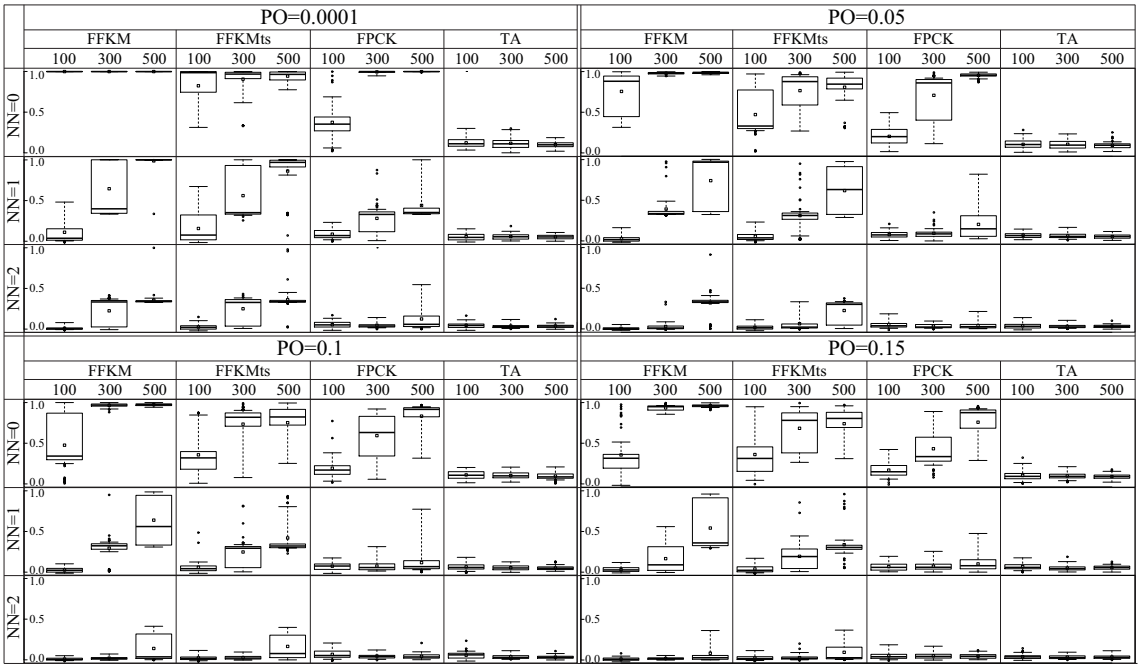

Figure 5: Boxplots of the adjusted Rand indices when $\mathbf{G}_1$ is FR and $\mathbf{G}_2$ is FR. In each case, from the left, the boxplots indicate the results of the FFKM, two-step FFKM, FPCK, and tandem analysis by sample size, respectively. The numbers under the name of each method in abscissa axis denotes sample size.

result shows that the FPCK method provided a poor result if the contributing part $\mathbf{G}_1$ to the cluster structure was FR. We also see that tandem analysis did not work well, regardless of the chosen values of PO, NN, and N.

When $\mathbf{G}_1$ was RD and $\mathbf{G}_2$ was FR (Figure 6), we can see that FPCK showed the best result under all values of PO and NN, while FFKM did not. The two-step FFKM method provided better results when PO = 0.0001 than when PO was large, and the ARI became worse with an increase in PO and NN. Since $\mathbf{G}_2$ was FR and all columns of $\mathbf{G}_1$ contributed to the cluster structure, the optimal subspace obtained from functional principal component analysis are coincident with that obtained from FPCK. This fact explains that tandem analysis worked as well as FPCK in this case.

When $\mathbf{G}_1$ was FR and $\mathbf{G}_2$ was RD (Figure 7), only two-step FFKM recovered the true cluster structure. It can be inferred that FFKM were effected negatively by the correlation of $\mathbf{G}_2$, while two-step FFKM improved

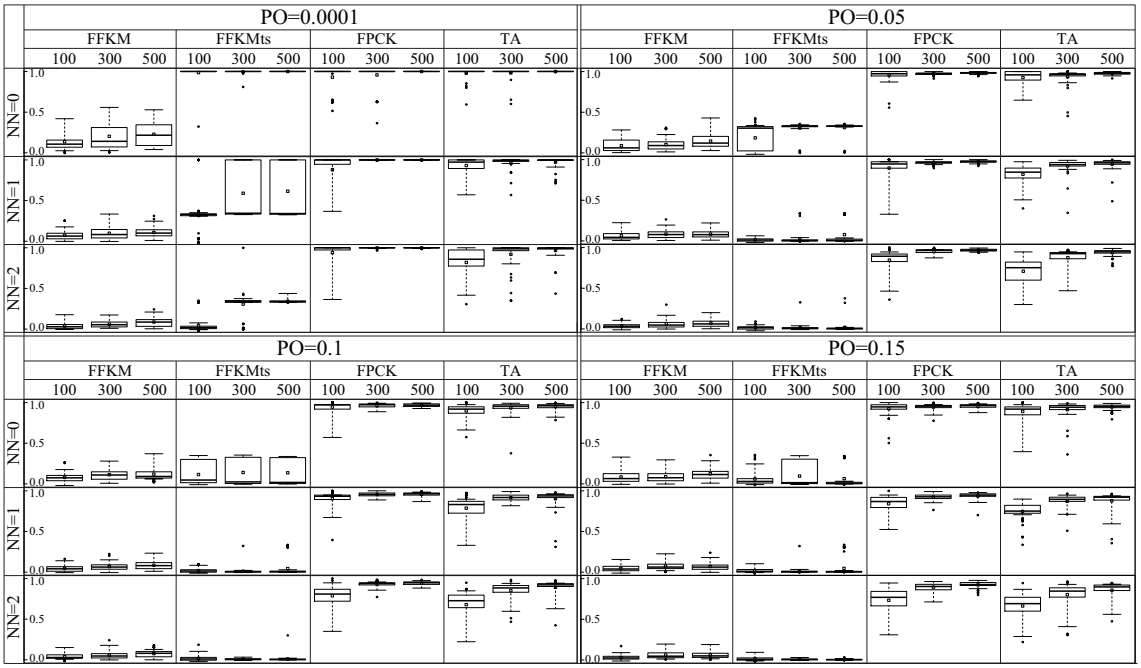

Figure 6: Boxplots of the adjusted Rand indices when $\mathbf{G}_1$ is RD and $\mathbf{G}_2$ is FR. In each case, from the left, the boxplots indicate the results of the FFKM, two-step FFKM, FPCK, and tandem analysis by sample size, respectively. The numbers under the name of each method in abscissa axis denotes sample size.

the performance of FFKM to remove the negative effect of the cumbersome correlation as it had been expected.

When both $\mathbf{G}_1$ and $\mathbf{G}_2$ were RD (Figure 8), FPCK showed the best result, or at least a result comparable to those of other methods. Two-step FFKM also worked well under mild conditions in which both PO and NN were small. FFKM and tandem analysis did not recovered the cluster structure well because of the existence of substantial correlation of $\mathbf{G}_2$.

We used 1000 random starts for the FFKM method. However, even in the case of one of the easiest settings, where PO = 0.0001 and $N = 500$, only 93 initial starts attained the global optimal solution. In addition, more local optimal solutions seem to occur when the overlap is increased. Thus, in practice, it is necessary to check carefully whether the solution is a global optimal solution. If not, more initial random starts may be necessary.

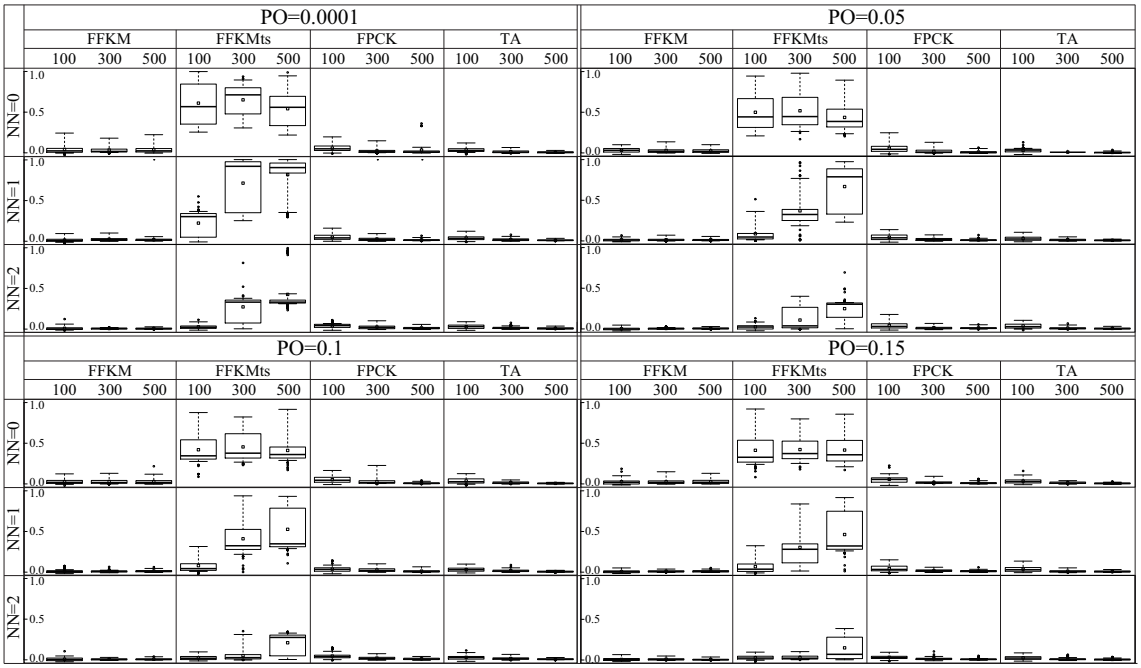

Figure 7: Boxplots of the adjusted Rand indices when $\mathbf{G}_1$ is FR and $\mathbf{G}_2$ is RD. In each case, from the left, the boxplots indicate the results of the FFKM, two-step FFKM, FPCK, and tandem analysis by sample size, respectively. The numbers under the name of each method in abscissa axis denotes sample size.

5. Empirical Example

In this section, we perform an empirical analysis to demonstrate the use of the FFKM method and to compare its performance with that of the existing methods, the FPCK and tandem analysis (TA). We used the well-known phoneme dataset for a speech-recognition problem, as described by Hastie et al. (1995). The data are log-periodograms of 32 ms duration that correspond to five phonemes, as follows: "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". Hastie et al. (1995) applied their penalized discriminant analysis to obtain a discriminant rule for well separation of phonemes. Ferraty and Vieu (2003) analysed this dataset by their nonparametric curves discrimination. Although this dataset is used for an illustration of newly proposed methods in the context of the supervised learning, it is also useful for an illustration of

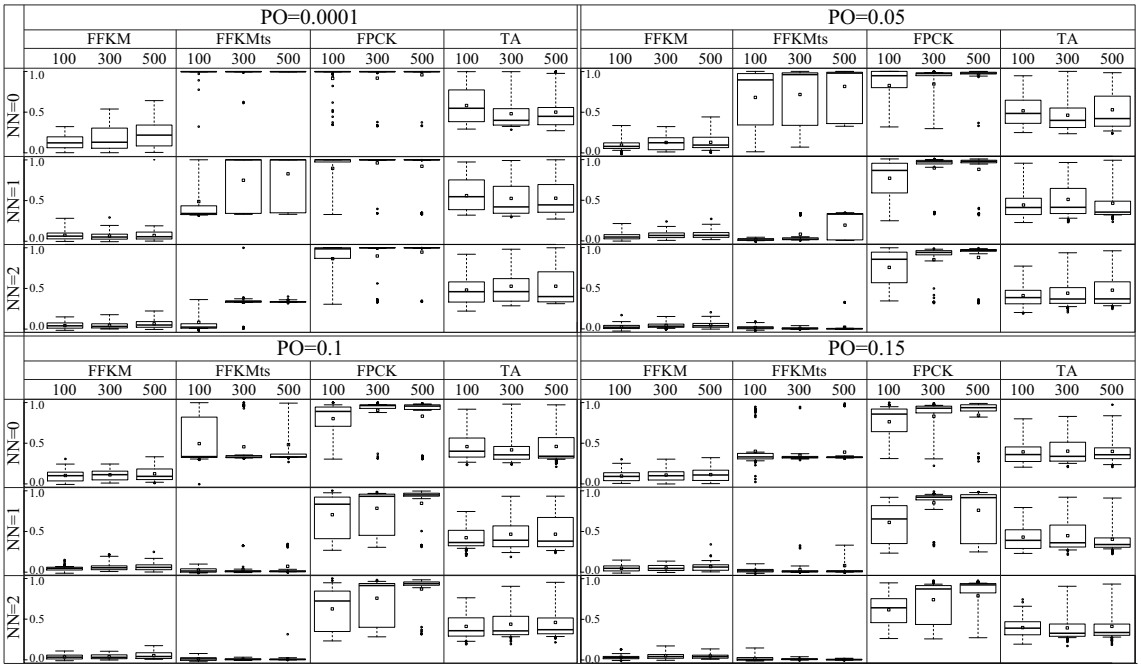

Figure 8: Boxplots of the adjusted Rand indices when $\mathbf{G}_1$ is RD and $\mathbf{G}_2$ is RD. In each case, from the left, the boxplots indicate the results of the FFKM, two-step FFKM, FPCK, and tandem analysis by sample size, respectively. The numbers under the name of each method in abscissa axis denotes sample size.

the methods in the context of the unsupervised learning, e.g., clustering. In the unsupervised learning contexts, since there are no true cluster labels, it is difficult to evaluate the goodness of the clustering results objectively. On the other hand, in the supervised learning contexts, we have the true cluster labels so that we can easily evaluate the goodness of the clustering results by some agreement measures (e.g., the Adjusted Rand Index). Actually, the phoneme data are used to investigate performances of clustering algorithms. For example, Shamir and Tishby (2008) used phoneme data to illustrate their model selection procedure based on cluster stability. In addition, Gattone and Rocci (2012) used the data to investigate the performance of their functional reduced k-means. Then, we used the phoneme data to investigate the performance of the proposed method. Here, we considered only the first 150 frequencies used in Ferraty and Vieu (2003), thus obtaining a dataset of 2000 log-periodograms with the known class-phoneme membership.

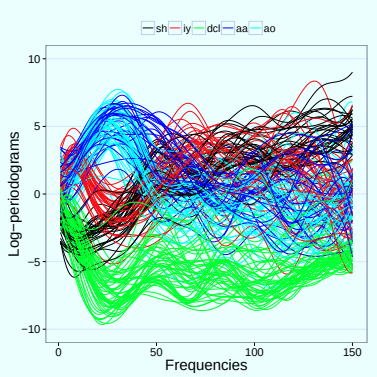

Figure 9: Selected phoneme data of 200 log-periodograms; the color denotes groups of phoneme.

In this example, suppose that we want to find correct clusters with $K = 5$ and obtain a low-dimensional subspace with $L = 2$ for interpreting the cluster structure. For all methods, we used the fourth-order B-spline basis function with ten knots. In this case, the number of basis functions is twelve. The value of λ that gives the minimum of GCV among the different values of λ, varying from 0.1 to 500, was selected: $\lambda = 61.31$. The selected log-periodograms expanded by these basis functions are shown in Figure 9. For the FFKM and FPCK method, the initial random starts with 100 were used.

In general, the coefficient matrix $\mathbf{G}_H$ of the functional data has some correlations between the coefficient vectors corresponding to the discretized basis functions $\phi_m(t)$. In such a case, there often exist small eigenvalues, which may be nearly zero, of $\mathbf{G}'_H\mathbf{G}_H$, so that the FFKM is likely to provide a poor recovery of the true cluster structure. Thus, we used a two-step approach which was investigated in Section 4.

Then, first we conducted FPCA with four components; the number of components was determined by the cumulative percentage of the total variation and the size of the variances of the principal components, as introduced in Jolliffe (2002). In view of cumulative percentage of the total variation, Jolliffe (2002) notes that choosing a cut-off somewhere between 70% and 90% and retaining R components, where R is the number determined by the cut-off, provides a rule that preserves most of the information in the data in the first R components. This is shown in the left panel of Figure 10. Further-

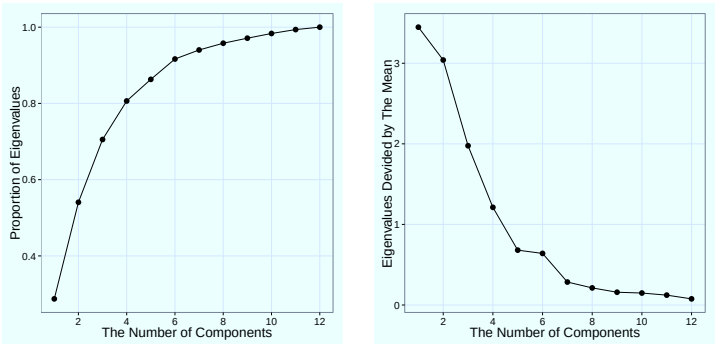

Figure 10: Plots for justification of the number of components; the left denotes the proportion of eigenvalues; the right denotes the eigenvalues divided by their mean value.

more, in view of the size of the variances of the principal components, it is recommended that we take as a cut-off the average value of the eigenvalues. The eigenvalues divided by their mean are shown in the right panel of Figure 10. From these plots, we see that the chosen number, four, is justified. We therefore conducted the FFKM analysis using the first four component scores.

The ARIs obtained by the three methods are shown in parentheses of Figure 11. We can see that the FFKM method can recover the true phoneme clusters well, while the other two methods provide cruder recoveries of the true cluster structure. The estimated component scores with the estimated cluster labels are plotted in the figure. In each plot, the symbols denote the estimated clusters of objects and the colors denote the true cluster structure. From these plots, it is concluded that the FFKM gives the optimal subspace representing the true cluster structure, while the subspaces given by the FPCK method and tandem analysis may not be appropriate for finding the cluster structure. This result may be compared with Figure 4 of Hastie et al. (1995) where log-periodograms are represented in the first two discriminant coordinates and only three groups are clearly identified, while FFKM provided the cluster structure in which four clusters were well identified.

As with the FPCK method described by Yamamoto (2012), it may be beneficial to interpret the estimated subspace using the estimated weight functions v_l. The weight functions estimated by the two-step approach are shown in Figure 12. In the figure, the black and red curves denote the weight functions corresponding to the first and second components, respectively. It

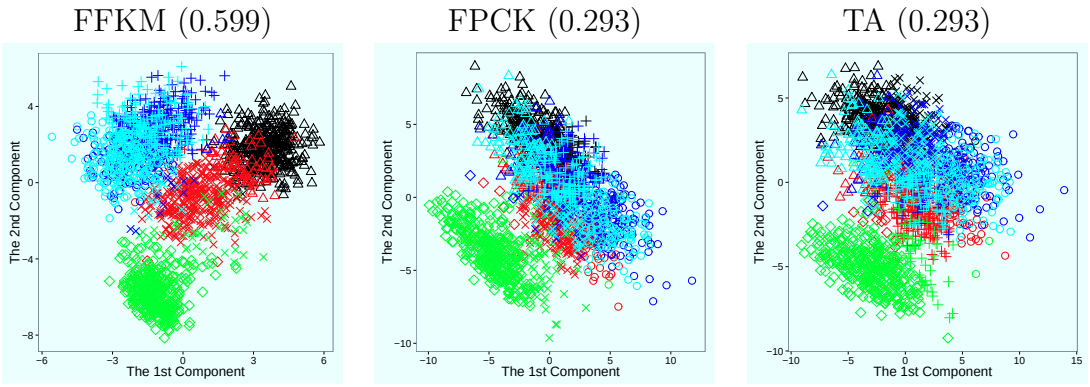

Figure 11: Plots of the component scores estimated by the FFKM, FPCK, and tandem analysis; a value in a parenthesis denotes a value of ARI; symbols of plots denote the estimated clusters of objects, and the colors denote the true cluster structure.

can be seen that the weight functions have large values in the region where the frequency is between 10 and 50 and in the last region. This implies that the cluster structure is determined by the behavior of the data in these regions, and this is reasonable considering the original data that is shown in Figure 9. Note that the component scores shown in the right pannel of Figure 12, calculated using these estimated weight functions, may be a little bit different from the original subspace representation shown in Figure 11. In this case, however, the cluster structure seems to be the same as the original one shown in Figure 11. This difference is due to the method of estimating the weight functions in the two-step approach described in Appendix B.

Note that most of the solutions of FFKM analysis given by initial random starts attained the same values for the loss functions. Thus, in this case, the number of initial random starts is sufficient to obtain the global solution.

6. Discussion

In this article, we explained the drawbacks of the FPCK method and proposed a new method, FFKM analysis, to overcome the problem. The FFKM method aims to simultaneously classify functional objects into optimal clusters and find a subspace that best describes the classification and dimension reduction of the data. The ALS algorithm was proposed to efficiently solve the minimization problem of the least-squares objective function. Analyses

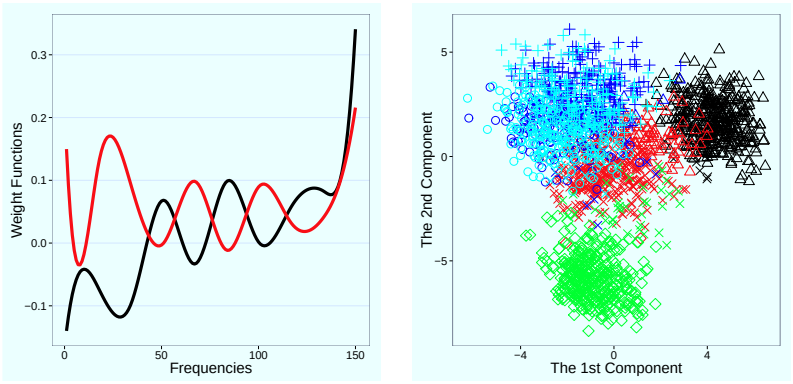

Figure 12: Estimated weight functions (left) by the FFKM method and the plot of corresponding component scores (right); the black and red curves correspond to the first and second components, respectively.

of artificial data reveal that the FFKM method can give an optimal cluster structure when both the coefficient matrix, $\mathbf{G}_1$, which is related to the true cluster structure, and a non-informative part, $\mathbf{G}_2$, have no substantial correlation.

However, the simulation study in Section 4 showed that when either $\mathbf{G}_1$ or $\mathbf{G}_2$ is rank deficient, FFKM failed in providing an optimal cluster structure. To avoid the negative effect of correlation among $\mathbf{G}_H = (\mathbf{G}_1, \mathbf{G}_2)$, the two-step approach to FFKM was also described. Two-step FFKM aims to eliminate trivial dimensions followed by applying the FFKM algorithm to the reduced functional data. The simulation study showed that when $\mathbf{G}_1$ was full rank and $\mathbf{G}_2$ was rank deficient, two-step FFKM recovered a cluster structure well. Furthermore, when $\mathbf{G}_1$ was rank deficient, it worked well under the mild conditions regardless of the rank of $\mathbf{G}_2$. Thus, in practice, it is recommended to use two-step FFKM instead of simple FFKM.

The simulation study also showed that when $\mathbf{G}_1$ was rank deficient, FPCK worked well regardless of the rank of $\mathbf{G}_2$. However, it did not work very well when $\mathbf{G}_1$ was full rank. Specifically, if $\mathbf{G}_1$ was full rank and $\mathbf{G}_2$ was rank deficient, it did not recover the true cluster structure at all. On the other hand, in the situation, only two-step FFKM worked well. This fact shows that FFKM has a mutually complementary relationship with FPCK. In practical situations, $\mathbf{G}_2$ often has a substantial correlation, that is, $\mathbf{G}_2$ is likely to be rank deficient. Therefore, it is recommended that first the

two-step FFKM method is implemented. If the result does not seem to be good, then FPCK is implemented.

Both the FFKM and FPCK methods need several initial random starts for the parameters in order to avoid local optima. In our limited experience, this problem seems to be more serious for the FFKM method. Thus, a more efficient algorithm for this model is needed.

In our approach, the tuning of the smoothing parameter, λ, is done by applying the GCV criterion to each curve, and the real data example introduced in Section 5 shows that this approach works well for finding the cluster structure. Another approach can also be adopted. For example, Gattone and Rocci (2012) proposed an automatic smoothing algorithm in which the smoothing is carried out within the clustering, and the amount of smoothing is determined adaptively. Recently, Wang (2010) has proposed a method based on clustering instability for selecting the number of clusters. These approaches may be applicable to the selection of the model in FFKM. This is an area of future research that we intend to pursue.

Acknowledgment

We thank the Associate Editor and two anonymous reviewers for their constructive comments that helped to improve the quality of this article. This work was supported by JSPS Grant-in-Aid for JSPS Fellows Number 24-2676.

Appendix A: FFKM for multivariate functional data

The method for the univariate case has been described above. Here, we explain our method for the multivariate case. Let $\mathscr{L}^P$ be the Cartesian product of P sets of $\mathscr{L} = L^2(T)$. Then, subject n has P functions, $x_n = (x_{n1}, \ldots, x_{nP}) \in \mathscr{L}^P$, and an inner product for $x, y \in \mathscr{L}^P$ is redefined as

$$\langle x, y \rangle_{P,\lambda} = \sum_{p=1}^{P} \left(\int_T x_p(t) y_p(t) dt + \lambda \int_T \mathtt{D}^2 x_p(t) \mathtt{D}^2 y_p(t) dt \right). \tag{A.1}$$

Note that the norm $\| \cdot \|_{P,\lambda}$ is given by the inner product, i.e., $\|x\|_{P,\lambda} = \langle x, x \rangle_{P,\lambda}^{1/2}$. Then, the objective function in Eq. (4) will be optimized as in the univariate case, with $\mathtt{P}_v^P$ for $\mathtt{P}_v$, where $\mathtt{P}_v^P$ is an orthogonal projection operator from $\mathscr{L}^P$ onto the subspace $\mathscr{S}_v^P$, and the weight functions $v_l^P \in \mathscr{L}^P$ span

$\mathscr{S}_v^P$. That is, the loss function for the multivariate regularized case can be written as

$$L_{ffkm}(U,V) = \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}\|\mathtt{P}_v\mathtt{S}_\lambda^2 x_n - \mathtt{P}_v\mathtt{S}_\lambda^2\bar{x}_k\|_{P,\lambda}^2.$$

Here, we can consider the basis function expansion for $\mathtt{S}_\lambda^2 x_n$ as

$$\mathtt{S}_\lambda^2 x_n = (\boldsymbol{g}_{n1}'\boldsymbol{\phi},\ldots,\boldsymbol{g}_{nP}'\boldsymbol{\phi}), \tag{A.2}$$

where $\boldsymbol{g}_{np}$ is a coefficient vector for the basis function expansion of $\mathtt{S}_\lambda^2 x_{np}$. Then, the criterion L_{ffkm} can be derived by

$$\begin{aligned}
L_{ffkm}(U,V) &= \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}\sum_{p=1}^{P}\|(\mathtt{P}_v\mathtt{S}_\lambda^2 x_n) - (\mathtt{P}_v\mathtt{S}_\lambda^2\bar{x}_k)_p\|_{P,\lambda}^2 \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}\sum_{p=1}^{P}\left\{\sum_{l=1}^{L}\left(\sum_{p'=1}^{P}\boldsymbol{a}_{lp'}'\mathbf{H}_\lambda\boldsymbol{g}_{np'}\right)\boldsymbol{a}_{lp} - \sum_{l=1}^{L}\left(\sum_{p'=1}^{P}\boldsymbol{a}_{lp'}'\mathbf{H}_\lambda\bar{\boldsymbol{g}}_{kp'}\right)\boldsymbol{a}_{lp}\right\}' \\
&\qquad \mathbf{H}_\lambda\left\{\sum_{l=1}^{L}\left(\sum_{p'=1}^{P}\boldsymbol{a}_{lp'}'\mathbf{H}_\lambda\boldsymbol{g}_{np'}\right)\boldsymbol{a}_{lp} - \sum_{l=1}^{L}\left(\sum_{p'=1}^{P}\boldsymbol{a}_{lp'}'\mathbf{H}_\lambda\bar{\boldsymbol{g}}_{kp'}\right)\boldsymbol{a}_{lp}\right\}.
\end{aligned}$$

The algorithm to minimize the objective function for multivariate functional data is the same as that for univariate functional data described above, i.e., the ALS algorithm can be applied, although there are some differences between these cases. In *STEP2*, the basis function expansion of a projected object $\mathtt{P}_v^P\mathtt{S}_\lambda x_n$ can be applied as in the case of univariate data. Thus, the cluster parameters U are estimated using the k-means algorithm for a parameter vector $\boldsymbol{d}^* = (\boldsymbol{d}_1',\ldots,\boldsymbol{d}_N')'$, where $\boldsymbol{d}_{np} = (d_{np1},\ldots,d_{npM})'$ and $\boldsymbol{d}_n = (\boldsymbol{d}_{n1}',\ldots,\boldsymbol{d}_{nP}')'$, which is the parameter vector for the basis function expansion of $\mathtt{P}_v^P\mathtt{S}_\lambda x_n$.

Next, we consider the optimization over V. Let an integral operator $\mathtt{F}^P$ be defined, for any $y \in \mathscr{L}^P$, as

$$\mathtt{F}^P y := (\mathtt{F}^{(1)}y,\ldots,\mathtt{F}^{(P)}y),$$

where

$$(\mathtt{F}^{(p)}y)(t) := -\sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}(x_{np}(t)-\bar{x}_{kp}(t))\sum_{p'=1}^{P}\langle x_{np'}-\bar{x}_{kp'}, y_{p'}\rangle \quad (p=1,\ldots,P).$$

Then, to estimate an optimal V in *STEP3* of the above ALS algorithm, the following optimization problem is considered:

$$\max_{V} \sum_{l=1}^{L} \langle v_l, \mathtt{F}^P v_l \rangle_{P,\lambda}.$$

As with the univariate case, it can be verified that the operator $\mathtt{F}^P$ is self-adjoint and compact. Thus, optimizing the criterion is equivalent to solving the following eigenvalue equation,

$$\mathtt{F}^P \xi_l = \rho_l \xi_l, \quad \text{subject to } \langle \xi_l, \xi_{l'} \rangle_{P,\lambda} = \delta_{ll'}.$$

Let $\mathbf{G}_p = (\boldsymbol{g}_{1p}, \ldots, \boldsymbol{g}_{Np})'$ and $\mathbf{G}_{Hp} = \mathbf{G}_p \mathbf{H}_\lambda^{\frac{1}{2}}$. Let $\mathbf{G}_H^P$ be the block diagonal matrix that has $\mathbf{G}_{Hj}$ for the jth diagonal block, and let $\boldsymbol{a}_{Hl}^P = (\boldsymbol{a}_{Hl1}', \ldots, \boldsymbol{a}_{HlP}')'$. Then, the above eigenvalue equation reduces to

$$\mathbf{G}_H^{P\prime}(\mathbf{1}_P \mathbf{1}_P' \otimes (\mathbf{P}_U - \mathbf{I}_N))\mathbf{G}_H^P \boldsymbol{a}_{Hl}^P = \rho \boldsymbol{a}_{Hl}^P,$$

where $\otimes$ denotes the Kronecker product. Finally, the estimated weight function can be calculated as $v_{lp} = \boldsymbol{a}_{Hlp}' \mathbf{H}^{-\frac{1}{2}} \boldsymbol{\phi}$.

Appendix B: Two-step approach for FFKM

As described in Section 4.1, when an empirical covariance operator of functional data has excessively small eigenvalues compared with the others, the subspace spanned by eigenfunctions corresponding to the small eigenvalues provides the smallest value of loss function of FFKM; this results in poor recovery of the true cluster structure. Actually, this problem also occurs in the factorial k-means (Vichi and Kiers, 2001) for a usual data matrix, and it is recommended that such trivial dimensions could be first eliminated from the data. Thus, it is inferred that the direct use of the FFKM method may fail to find an optimal cluster structure. To overcome this problem, we propose the two-step approach described below. Note that this two-step approach has a completely different aim from that of tandem analysis: tandem analysis finds a low-dimensional subspace regardless of the cluster structure, whereas the two-step approach just eliminates the trivial dimensions and finds a low-dimensional subspace where a cluster structure exists.

First, we conduct FPCA (Ramsay and Silverman, 2005) based on the basis function expansion using the basis function $\{\phi_m\}_{m=1,\ldots,M}$ of the raw data. This gives the principal curves $\{w_r\}_{r=1,\ldots,R}$, where R is the number of principal components. The number R should be selected so that principal components contain sufficiently high variances for the data. The usual selection rules described by Jolliffe (2002) may work well. Let $\mathtt{P}_w$ be the operator that projects functional objects onto the space spanned by the principal curves, and we then obtain the projected functional data $\mathtt{P}_w x_i$. As described in Eq. (1), the FFKM method requires a basis function expansion of the data. Here, using the basis functions ϕ_m used in the FPCA, the reduced functional data can be expressed as

$$(\mathtt{P}_w x_1, \ldots, \mathtt{P}_w x_N)' = \mathbf{G}_H \mathbf{B}_H \mathbf{B}_H' \mathbf{H}^{-\frac{1}{2}} \boldsymbol{\phi}, \tag{B.1}$$

where $\mathbf{B}_H = (\boldsymbol{b}_{H1}, \ldots, \boldsymbol{b}_{HR})$ denotes the coefficient matrix in the basis function expansion of the principal curves, such that $w_r = \boldsymbol{b}_{Hr}' \mathbf{H}^{-1/2} \boldsymbol{\phi}$. In this notation, the principal component score matrix is calculated as $\mathbf{F}_{pca} = \mathbf{G}_H \mathbf{B}_H$.

Thus, the optimization problem of the FFKM method with basis function expansions of the reduced functional data is defined as

$$\min_{\mathbf{A}_H, U} \|\mathbf{F}_{pca} \mathbf{B}_H' \mathbf{A}_H - \mathbf{P}_U \mathbf{F}_{pca} \mathbf{B}_H' \mathbf{A}_H\|^2. \tag{B.2}$$

We can see that $\mathbf{F}_{pca} \mathbf{B}_H'$ corresponds to $\mathbf{G}_H$, which is the coefficient matrix of the basis function expansion of the reduced functional data. Clearly, the rank of $\mathbf{F}_{pca} \mathbf{B}_H'$ is R, i.e., the coefficient matrix is rank deficient. Here, according to the recommendation by Vichi and Kiers (2001), we consider eliminating the trivial dimensions of the coefficient matrix. In the case of FFKM analysis, we can use $\mathbf{F}_{pca}$ as the full-rank (neither singular nor near-singular) matrix to be analyzed.

Therefore, instead of the optimization problem in Eq. (B.2), the following optimization problem is considered,

$$\min_{\mathbf{A}_H^*, U} \|\mathbf{F}_{pca} \mathbf{A}_H^* - \mathbf{P}_U \mathbf{F}_{pca} \mathbf{A}_H^*\|^2, \tag{B.3}$$

where $\mathbf{A}_H^*$ is an $R \times L$ orthogonal matrix that spans an optimal subspace for representing the cluster structure. This optimization problem can be solved by the same algorithm described in Section 3.2. That is, we just have to use $\mathbf{F}_{pca}$ and $\mathbf{A}_H^*$ as $\mathbf{G}_H$ and $\mathbf{A}_H$, respectively, in the algorithm for the FFKM method.

Using this two-step approach, we can obtain the cluster structure in the low-dimensional subspace. However, this procedure does not provide the weight functions v_l that span the subspace of the functional data. The weight functions are often useful to interpret the estimated subspace and cluster structure. Thus, we consider estimating the weight functions from the estimates $\mathbf{A}_H^*$.

To obtain the coefficient matrix $\mathbf{A}_H$ of the weight functions v_l, the following optimization problem is considered,

$$\min_{\mathbf{A}_H} \|\mathbf{F}_{pca}\mathbf{B}_H'\mathbf{A}_H - \mathbf{F}_{pca}\mathbf{A}_H^*\|^2. \tag{B.4}$$

Note that $\mathbf{A}_H$ is an orthogonal matrix. This is the well-known orthogonal Procrustes rotation problem (ten Berge, 1993), and it can be solved easily. The singular value decomposition $\mathbf{B}_H\mathbf{F}_{pca}'\mathbf{F}_{pca}\mathbf{A}_H^* = \mathbf{PDQ}'$ yields $\mathbf{A}_H = \mathbf{PQ}'$ as the optimizing solution, where $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_L$ and $\mathbf{D}$ is a diagonal matrix whose diagonal element is a singular value. Then, using $\mathbf{A}_H = (\boldsymbol{a}_{H1}, \ldots, \boldsymbol{a}_{HL})$, the estimated weight function is calculated as $v_l = \boldsymbol{a}_{Hl}'\mathbf{H}^{-1/2}\boldsymbol{\phi}$. Furthermore, we can obtain the component score matrix $\mathbf{F}$ as $\mathbf{F} = \mathbf{G}_H\mathbf{A}_H$.

References

Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari, N., 2003. Unsupervised curve clustering using B-splines. Scandinavian Journal of Statistics 30, 581–595.

Arabie, P., Hubert, L., 1994. Cluster analysis in marketting research. In: Bagozzi, R.P. (Ed.), Advanced methods of marketing research. Blackwell, Oxford.

Besse, P.C., Ramsay, J.O., 1986. Principal components analysis of sampled functions. Psychometrika 51 (2), 285–311.

Boente, G., Fraiman, R., 2000. Kernel-based functional principal components. Statistics & Probability Letters 48 (4), 335–345.

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

Dunford, N., Schwartz. J.T., 1988. Linear operators, spectral theory, self adjoint operators in Hilbert space, part 2. Interscience, New York.

Ferraty, F., Vieu, P., 2003. Curves discrimination: a non parametric functional approach. Computational Statistics & Data Analysis 44, 161–173.

Ferraty, F., Vieu, P., 2006. Nonparametric functional data analysis. Springer, New York.

Gattone, S.A., Rocci, R., 2012. Clustering curves on a reduced subspace. Journal of Computational and Graphical Statistics 21 (2), 361–379.

Green, P.J., Silverman, B.W., 1994. Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman and Hall, London.

Hardy, A., 1996. On the number of clusters. Computational Statistics & Data Analysis 23, 83–96.

Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. The Annals of Statistics 23 (1), 73–102.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. Computational Statistics & Data Analysis 52, 5186–5201.

Illian, J.B., Prosser. J.I., Baker, K.L., Rangel-Castro, J.I., 2009. Functional principal component data analysis: A new method for analysing microbial community fingerprints. Journal of Microbiological Methods 79 (1), 89–95.

Jacques, J., Preda, C. Functional data clustering: a survey. Advances in Data Analysis and Classification, in press.

Jolliffe, I.T., 2002. Principal component analysis, 2nd Edition. Springer, New York.

Lloyd, S., 1982. Least squares quantization in pem. IEEE Transactions on Information Theory 28 (2), 128–137.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 (2), 159–179.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, 2nd Edition. Springer, New York.

Serban, N., Wasserman, L., 2005. CATS: Clustering after transformation and smoothing. Journal of the American Statistical Association 100, 990–999.

Shamir, O., Tishby, N., 2008. Cluster stability for finite samples. In: Platt, J., Koller, D., Singer, Y., Rowseis, S. (Eds.), Advances in Neural Information Processing Systems (NIPS) 21. MIT Press, Cambridge, MA.

Silverman, B.W., 1996. Smoothed functional principal components analysis by choice of norm. The Annals of Statistics 24 (1), 1–24.

Steinley, D., 2003. Local optima in k-means clustering: what you don't know may hurt you. Psychological Methods 8 (3), 294–304.

Steinley, D., Henson, R., 2005. OCLUS: an analytic method for generating clusters with known overlap. Journal of classification 22, 221–250.

Suyundykov, R., Puechmorel, S., Ferre, L., 2010. Multivariate functional data clusterization by PCA in Sobolev space using wavelets. Hyper Articles en Ligne :inria-00494702.

Tarpey, T., Kinateder, K., 2003. Clustering functional data. Journal of Classification 20, 93–114.

ten Berge, J.M.F., 1993. Least squares optimization in multivariate analysis. DSWO Press, Leiden University, Leiden.

Timmerman, M.E., Ceulemans, E., Kiers, H.A.L., Vichi, M., 2010. Factorial and reduced k-means reconsidered. Computational Statistics & Data Analysis 54 (7), 1858–1871.

Vichi, M., Kiers H.A.L., 2001. Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis 37 (1), 49–64.

Vidal, R., 2011. Subspace clustering. Signal Processing Magazine, IEEE, 52–68.

Wang, J., 2010. Consistent selection of the number of clusters via crossvalidation. Biometrika 97 (4), 893–904.

Yamamoto, M., 2012. Clustering of functional data in a low-dimensional subspace. Advances in Data Analysis and Classification 6, 219–247.