Journal of Statistical Planning and Inference 139 (2009) 2388 – 2398

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data

Mitsunori Kayano[1], Sadanori Konishi*

Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

ARTICLE INFO

Article history:
Received 15 December 2007
Received in revised form
9 November 2008
Accepted 10 November 2008
Available online 21 November 2008

Keywords:
Functional data analysis
Model selection
Protein structure
Radial basis functions
Regularization
Smoothing parameter
Spline

ABSTRACT

This paper introduces regularized functional principal component analysis for multidimensional functional data sets, utilizing Gaussian basis functions. An essential point in a functional approach via basis expansions is the evaluation of the matrix for the integral of the product of any two bases (cross-product matrix). Advantages of the use of the Gaussian type of basis functions in the functional approach are that its cross-product matrix can be easily calculated, and it creates a much more flexible instrument for transforming each individual's observation into a functional form. The proposed method is applied to the analysis of three-dimensional (3D) protein structural data that can be referred to as unbalanced data. It is shown that our method extracts useful information from unbalanced data through the application. Numerical experiments are conducted to investigate the effectiveness of our method via Gaussian basis functions, compared to the method based on *B*-splines. On performing regularized functional principal component analysis with *B*-splines, we also derive the exact form of its cross-product matrix. The numerical results show that our methodology is superior to the method based on *B*-splines for unbalanced data.

1. Introduction

Multivariate analysis deals with observations on more than one variable, where there is some inherent interdependence between the variables (Mardia et al., 1979), and principal component analysis (PCA) is one of the most widely used multivariate analysis techniques in various fields of natural and social sciences (see, e.g., Jolliffe, 2002). The concepts of PCA are the dimension reduction and visualization of data. However, there are some problems with applying conventional PCA to the longitudinal type of data. For example, if the observational points are not equally spaced and differ among subjects, PCA cannot be directly applied. Accordingly, a number of recent papers have investigated functional principal component analysis (functional PCA) and its regularization methods that reformulate PCA in terms of the functions rather than the discrete observations (Besse and Ramsay, 1986; Rice and Silverman, 1991; Silverman, 1996).

These functional approaches are referred to as functional data analysis (FDA; Ramsay and Silverman, 2002, 2005; Ferraty and Vieu, 2006; Mizuta, 2006). The basic idea behind FDA is the conversion of observational discrete data to functional data by a smoothing method and then extracting information from the obtained functional data set by applying concepts from traditional multivariate analysis. In modeling with FDA, many studies employ a basis expansion which assumes that functional

* Corresponding author.
E-mail addresses: kayano@kuicr.kyoto-u.ac.jp (M. Kayano), konishi@math.kyushu-u.ac.jp (S. Konishi).
[1] Present address: Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.11.002

data and coefficient functions may be expressed as linear combinations of known basis functions. Fourier series are useful if the observations are periodic and have sinusoidal features, whereas splines (Green and Silverman, 1994) and *B*-splines (de Boor, 2001; Eilers and Marx, 1996; Imoto and Konishi, 2003) are utilized to non-periodic data.

An essential point for FDA via basis expansions is the evaluation of the matrix for the integral of the product of any two bases (cross-product matrix). The orthonormal property of Fourier series yields the identity cross-product matrix, and then we need not evaluate the cross-product matrix for Fourier series. In contrast, spline types of bases do not have orthonormal property, and in consequence the cross-product matrix must be calculated. Previous works, however, utilized discrete approximation to evaluate the cross-product matrix for spline types of bases (see, e.g., Ramsay and Silverman, 2002, Section 2). In this paper, we provide the exact form for the integral of the product of any two *B*-spline bases.

The main aim of this paper is to introduce regularized functional PCA for multidimensional (multivariate) functional data sets, utilizing Gaussian basis functions. Advantages of the use of the Gaussian type of basis functions are that its cross-product matrix can be easily calculated, and it creates a much more flexible instrument for transforming each individual's observation into a functional form. Numerical experiments are conducted to investigate the effectiveness of our method via Gaussian basis functions. In addition, the proposed method is applied to functionalized three-dimensional (3D) protein structural data that determine the 3D arrangement of amino-acids in individual protein and also determine proteins that have special structures. An objective of the analysis of the protein structural data is to characterize any features of proteins without relying on their sequence information and physicochemical properties. Our functionalization method permits a low-dimensional visualization of proteins, and provides useful information concerning biological view points.

This paper is organized as follows. Section 2 describes observational discrete data and its functionalization to multidimensional functional data. Section 3 introduces a regularized functional principal component (PC) procedure based on multidimensional functional data sets and gives an outline of its implementation. In Section 4, Monte Carlo simulations are conducted to investigate the effectiveness of the proposed regularized functional PCA based on Gaussian basis functions, in which we compare our procedure to the method based on *B*-splines with the derived exact cross-product matrix. Section 5 describes an application of the proposed method to the 3D protein structural data. Finally, some concluding remarks are presented in Section 6.

2. Discrete and functional data

Suppose we have N independent discrete observations $\{t_{ij}, (x_{i1j}, \ldots, x_{ipj}); j = 1, \ldots, n_i\}$ $(i = 1, \ldots, N)$, where each t_{ij} $(\in \mathscr{T} \subset \mathbb{R})$ is the jth observational point of the ith individual and $(x_{i1j}, \ldots, x_{ipj})$ $(\in \mathbb{R}^p)$ is the discrete data observed at t_{ij} for p variables $X_1, \ldots, X_p$. In particular, the ith discrete data set observed at t_{ij} for X_l is represented by $\{(t_{ij}, x_{ilj}); j = 1, \ldots, n_i\}$. It may be noted that we have the discrete data observed at possibly different observational points $t_{i1}, \ldots, t_{in_i}$ for each subject, and then the discrete observations can be referred to as unbalanced data. For example, $\{t_{ij}, (x_{i1j}, x_{i2j}, x_{i3j}); j = 1, \ldots, n_i\}$ $(i = 1, \ldots, 12)$ are the measurements in *XYZ* coordinates of 3D protein structures, where t_{ij} are the positions in ith amino-acid sequence and $(x_{i1j}, x_{i2j}, x_{i3j})$ are the *XYZ* coordinate values of amino-acids which compose ith 3D protein structure. Fig. 1 (upper) shows an example of discretized 3D protein structural data with $p = 3$ and $n_i = 186$.

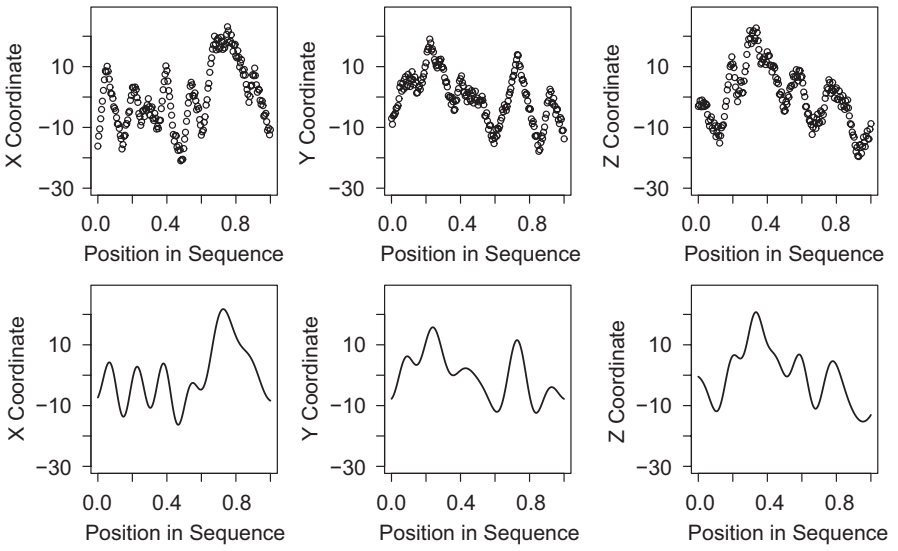

Fig. 1. An example of discrete data (upper) and corresponding three-dimensional functional data (lower) for a 3D protein structure ($p = 3$, $n_i = 186$).

We convert each discrete data set $\{(t_{ij}, x_{ilj}); j = 1, \ldots, n_i\}$ to functional data $x^*_{il}(t)$ using a smoothing method, as follows. It is assumed that each discrete data $\{(t_{ij}, x_{ilj}); j = 1, \ldots, n_i\}$ is generated from the nonlinear regression model

$$x_{ilj} = u_{il}(t_{ij}) + \varepsilon_{ilj} \quad (j = 1, \ldots, n_i),$$

where the errors ε_{ilj} are independently normally distributed with mean 0 and variance σ^2_{il}. We here assume that the observational discrete data x_{ilj} are independent. Yao and Lee (2006) has given the relevance of dealing with the observations sampled from functions as independent data.

The nonlinear functions $u_{il}(t)$ are assumed to be given by linear combinations of Gaussian basis functions $\{\phi_m(t) = \phi_m(t; \nu, \mu_m, \tau^2_m)\}$ with parameters μ_m, τ_m and ν,

$$u_{il}(t) = \sum_{m=1}^{M} c_{ilm}\phi_m(t),$$

where the mth Gaussian basis function $\phi_m(t)$ has the form

$$\phi_m(t) = \phi_m(t; \nu, \mu_m, \tau^2_m) = \exp\left\{-\frac{(t-\mu_m)^2}{2\nu\tau^2_m}\right\} \quad (m = 1, \ldots, M). \tag{1}$$

The parameters μ_m and τ_m express the position and width of the mth basis function, and ν is a hyper-parameter that adjusts the width of basis functions, while the $\{\tau_m\}$ have been determined by a clustering method (Ando et al., 2005).

Each nonlinear function $u_{il}(t)$ is estimated in two steps. First, the parameters μ_m and τ_m are estimated by applying the k-means clustering method with $k = M$ to $\sum_i n_i$ observational points $\{t_{ij}; j = 1, \ldots, n_i,\ i = 1, \ldots, N\}$. The estimated parameters $\hat{\mu}_m$ and $\hat{\tau}^2_m$ are given by the sample mean and variance of $\{t_{ij} \in C_m\}$, where C_m is the mth cluster given by the k-means method. Let $\phi^{\nu}_m(t) = \phi_m(t; \nu, \hat{\mu}_m, \hat{\tau}^2_m)$ be the estimated mth basis function. Next, for each i and l the coefficient parameters $c_{il1}, \ldots, c_{ilM}$ and variance σ^2_{il} are estimated by maximizing the penalized log-likelihood function $p\ell_{\beta_{il}}(\boldsymbol{c}_{il}, \sigma^2_{il})$ with a smoothing parameter $\beta_{il}(>0)$ that controls the smoothness of the nonlinear function $u_{il}(t)$,

$$p\ell_{\beta_{il}}(\boldsymbol{c}_{il}, \sigma^2_{il}) = \sum_{j=1}^{n_i} \log f(x_{ilj}|t_{ij}; \boldsymbol{c}_{il}, \sigma^2_{il}) - \frac{n_i\beta_{il}}{2}\boldsymbol{c}'_{il}D'_2D_2\boldsymbol{c}_{il}, \tag{2}$$

where $f(x_{ilj}|t_{ij}; \boldsymbol{c}_{il}, \sigma^2_{il})$ is the probability density function of x_{ilj}, $\boldsymbol{c}'_{il}D'_2D_2\boldsymbol{c}_{il} = \sum_{m=2}^{M}(\Delta^2 c_{ilm})^2$ is the roughness penalty with difference operator Δ defined by $\Delta c_{ilm} = c_{ilm} - c_{il,m-1}$ and D_2 is the $(M-2) \times M$ matrix representation of the difference operator Δ^2.

The estimators $\hat{\boldsymbol{c}}_{il}$ and $\hat{\sigma}^2_{il}$ depend on the number of basis functions M, hyper-parameter ν in Gaussian basis functions and smoothing parameter β_{il} for each i and l. The parameters are often selected by using an information criterion and cross validation (CV) method. One of the methods that provide us the optimized values of the parameters is to use a grid search with those candidate values. As an information criterion, we here employ the generalized information criterion (GIC), given by Konishi and Kitagawa (1996). The GIC can be applied to evaluate statistical models constructed by various types of estimation procedures such as the maximum penalized likelihood procedure. Notice that one of the most famous information criteria, Akaike information criterion (AIC, Akaike, 1973), evaluates the model estimated by the maximum likelihood method (see also, Konishi and Kitagawa, 2008).

Thus, we have the estimated nonlinear functions $\hat{u}_{il}(t) = \sum_{m=1}^{M}\hat{c}_{ilm}\phi_m(t)$ $(i = 1, \ldots, N, l = 1, \ldots, p)$, where $\phi_m(t) = \phi^{\nu}_m(t) = \phi_m(t; \nu, \hat{\mu}_m, \hat{\tau}^2_m)$ is the mth basis function with the optimal hyper-parameter ν selected by minimizing GIC. The p-dimensional functional data sets $\{x^*_{i1}(t), \ldots, x^*_{ip}(t); t \in \mathcal{T}\}$ are then given by $x^*_{il}(t) = \hat{u}_{il}(t)$ for each i and l. In the next section, we introduce regularized functional PCA for the p-dimensional functional data sets, using the Gaussian basis functions. An example of multidimensional functional data is shown in Fig. 1 (lower), corresponding to the discretized 3D protein structural data in Fig. 1 (upper).

3. Functional PCA

3.1. Model

Let $\{(x^*_{i1}(t), \ldots, x^*_{ip}(t)); t \in \mathcal{T}\}$ $(i = 1, \ldots, N)$ be the p-dimensional functional data sets obtained by smoothing the observational discrete data sets $\{t_{ij}, (x_{i1j}, \ldots, x_{ipj}); j = 1, \ldots, n_i\}$ $(i = 1, \ldots, N)$. A functional PC method is here applied to the p-dimensional functional data sets $\{(x_{i1}(t), \ldots, x_{ip}(t)); t \in \mathcal{T}\}$ $(i = 1, \ldots, N)$, where $x_{il}(t) = x^*_{il}(t) - \bar{x}_l{}^*(t)$ and each $\bar{x}_l{}^*(t)$ is the mean function of the functional data $x^*_{i1}(t), \ldots, x^*_{ip}(t)$. It is assumed that each functional data element $x_{il}(t)$ can be expressed as a linear combination of

the Gaussian basis functions in (1): $x_{il}(t)=\sum_{m=1}^{M}\tilde{c}_{ilm}\phi_m(t)=\tilde{\boldsymbol{c}}_{il}'\boldsymbol{\phi}(t)$ with estimated coefficient vectors $\tilde{\boldsymbol{c}}_{il}$ and basis vector $\boldsymbol{\phi}(t)$ $(i=1,\ldots,N,l=1,\ldots,p)$.

Let f_i be an inner product for a p-dimensional weight function $\boldsymbol{\xi}(t)=(\xi_1(t),\ldots,\xi_p(t))'$ $(t\in\mathscr{T})$ and ith p-dimensional functional data $\boldsymbol{x}_i(t)=(x_{i1}(t),\ldots,x_{ip}(t))'$,

$$f_i=\langle\boldsymbol{\xi},\boldsymbol{x}_i\rangle_p=\sum_{l=1}^{p}\langle\xi_l,x_{il}\rangle=\sum_{l=1}^{p}\int_{\mathscr{T}}\xi_l(t)x_{il}(t)\,\mathrm{d}t\quad(i=1,\ldots,N).\tag{3}$$

We adopt a straightforward definition of an inner product between two p-dimensional functions. It is assumed that the weight functions $\xi_1(t),\ldots,\xi_p(t)$ can be expressed in terms of the same basis functions as the functional data sets $\{(x_{i1}(t),\ldots,x_{ip}(t))\}$,

$$\xi_l(t)=\sum_{m=1}^{M}\theta_{lm}\phi_m(t)=\boldsymbol{\theta}_l'\boldsymbol{\phi}(t)\quad(l=1,\ldots,p)$$

with $\boldsymbol{\theta}_l=(\theta_{l1},\ldots,\theta_{lM})'$. A general functional PC method maximizes the sample variance of the inner products subject to the orthonormal constraints, in order to estimate weight functions. It may be noted that the weight functions correspond to the weight vectors in conventional PCA.

On the other hand, regularized (smoothed) functional PCA (regularized functional PCA) proposed by Rice and Silverman (1991) and Silverman (1996) avoids ill-posed problems from functional PCA and maximizes the PSV instead of the sample variance in functional PCA. In this paper, we estimate the p-dimensional weight function $\boldsymbol{\xi}(t)$ that maximizes the following penalized sample variance (PSV) subject to penalized orthonormal constraints:

$$\mathrm{PSV}_{\boldsymbol{\lambda}}(\boldsymbol{\xi})=\frac{\mathrm{var}(f)}{\|\boldsymbol{\xi}\|_p^2+\boldsymbol{\theta}'Q_{\boldsymbol{\lambda}}\boldsymbol{\theta}},\tag{4}$$

where $\boldsymbol{\theta}=(\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_p')'$, $\|\boldsymbol{\xi}\|_p^2=\sum_l\|\xi_l\|^2=\sum_l\int_{\mathscr{T}}\xi_l^2(t)\,\mathrm{d}t$ is the norm of a p-dimensional weight function $\boldsymbol{\xi}(t)$ and $Q_{\boldsymbol{\lambda}}=\mathrm{diag}(\lambda_1Q^*,\ldots,\lambda_pQ^*)$ is a $pM\times pM$ positive-semidefinite block diagonal matrix with $M\times M$ positive-semidefinite matrix Q^* (roughness penalty matrix) and smoothing parameters $\lambda_l>0$ which control the smoothness of the weight functions $\xi_l(t)$. The roughness penalty matrix Q^* can be taken as $Q^*=D_2'D_2$ with the matrix representation D_2 in (2) of the difference operator Δ^2. The smoothing parameters λ_l can be optimally selected by minimizing a CV score (Section 3.3).

The PC curves are defined by the p-dimensional weight function $\boldsymbol{\xi}(t)$ that maximizes the penalized sample variance $\mathrm{PSV}_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ given by (4) subject to the penalized orthonormal constraints.

First PC curve $\boldsymbol{\xi}_1^{\boldsymbol{\lambda}}(t)=(\xi_{11}^{\boldsymbol{\lambda}}(t),\ldots,\xi_{1p}^{\boldsymbol{\lambda}}(t))'$ the p-dimensional weight function $\boldsymbol{\xi}(t)$ that maximizes $\mathrm{PSV}_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ subject to $\|\boldsymbol{\xi}\|_p^2=1$, $k(\geqslant 2)$th PC curve $\boldsymbol{\xi}_k^{\boldsymbol{\lambda}}(t)=(\xi_{k1}^{\boldsymbol{\lambda}}(t),\ldots,\xi_{kp}^{\boldsymbol{\lambda}}(t))'$: the p-dimensional weight function $\boldsymbol{\xi}(t)$ that maximizes $\mathrm{PSV}_{\boldsymbol{\lambda}}(\boldsymbol{\xi})$ subject to $\|\boldsymbol{\xi}\|_p^2=1$ and $\langle\boldsymbol{\xi},\boldsymbol{\xi}_r^{\boldsymbol{\lambda}}\rangle_p+\boldsymbol{\theta}'Q_{\boldsymbol{\lambda}}\boldsymbol{\theta}_r^{\boldsymbol{\lambda}}=0(r<k)$, where each $\boldsymbol{\theta}_k^{\boldsymbol{\lambda}}$ is the coefficient vector of $\boldsymbol{\xi}_k^{\boldsymbol{\lambda}}(t)$. The kth PC score is defined by $\{f_{ki}^{\boldsymbol{\lambda}}=\langle\boldsymbol{\xi}_k^{\boldsymbol{\lambda}},\boldsymbol{x}_i\rangle_p;$ $i=1,\ldots,N\}$ $(k=1,\ldots,pM)$. We note that there are pM PCs by the assumption of basis expansions.

Ramsay and Silverman (2005, Section 8.5) described a functional PC method to the two-dimensional functional data sets which include the hip and knee angles during a human gait cycle. Another approaches to functional PCA are given by Shi et al. (1996), Rice and Wu (2001), James et al. (2000) and Yao et al. (2005). A main concern in these works is to deal with sparse data that only have small observations from each individual.

In contrast, this paper treats unbalanced data that have usually a lot of observations from each individual and are observed at possibly different points. It is here assumed that the differences among the numbers of observations and the roughness of the curves are not quite large across the p functions: a functional PCA model via sparse data may be useful for the much unbalanced data, and we may put a set of basis functions for each function to the data with different roughness. Our real data example in Section 5 gives rough indication of those points.

3.2. Eigenvalue problem

The PC curves can be estimated by solving an eigenvalue problem. The inner product $f_i=\langle\boldsymbol{\xi},\boldsymbol{x}_i\rangle_p$ for a p-dimensional weight function $\boldsymbol{\xi}(t)$ and ith p-dimensional functional data $\boldsymbol{x}_i(t)$ can be written as

$$f_i=\langle\boldsymbol{\xi},\boldsymbol{x}_i\rangle_p=\sum_{l=1}^{p}\langle\xi_l,x_{il}\rangle=\sum_{l=1}^{p}\int_{\mathscr{T}}\boldsymbol{\theta}_l'\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'\tilde{\boldsymbol{c}}_{il}\,\mathrm{d}t=\sum_{l=1}^{p}\boldsymbol{\theta}_l'W^*\tilde{\boldsymbol{c}}_{il}=\boldsymbol{\theta}'W\tilde{\boldsymbol{c}}_i,$$

where $\tilde{\boldsymbol{c}}_i=(\tilde{\boldsymbol{c}}_{i1}',\ldots,\tilde{\boldsymbol{c}}_{iM}')'$, each $\boldsymbol{\theta}_l$ is the coefficient vectors of $\xi_l(t)$, the $M\times M$ cross-product matrix $W^*=\int_{\mathscr{T}}\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'\,\mathrm{d}t$ has the (m,n)th element $W_{mn}^*=\int_{\mathscr{T}}\phi_m(t)\phi_n(t)\,\mathrm{d}t$, and the $pM\times pM$ matrix $W=\mathrm{diag}(W^*,\ldots,W^*)$ is the block diagonal matrix formed from W^*.

The (m,n)th components of the cross-product matrix W^* for Gaussian basis functions $\phi_l(t)=\phi_l(t;v,\hat{\mu}_l,\hat{\tau}_l^2)$ are given by

$$W^*_{mn}=\frac{\sqrt{2\pi v\hat{\tau}_m^2\hat{\tau}_n^2}}{\sqrt{\hat{\tau}_m^2+\hat{\tau}_n^2}}\exp\left\{-\frac{(\hat{\mu}_m-\hat{\mu}_n)^2}{2v(\hat{\tau}_m^2+\hat{\tau}_n^2)}\right\}\quad(m,n=1,\ldots,M).$$

On performing regularized functional PCA, we need to assume that the cross-product matrix W^* of Gaussian basis functions is positive definite. To satisfy this assumption, we employ Gaussian basis functions $\{\phi_1(t),\ldots,\phi_M(t)\}$ except constant term $\phi_0(t)=1$, which are as flexible as common Gaussian RBF $\{\phi_0(t),\phi_1(t),\ldots,\phi_M(t)\}$.

Now, let $V=N^{-1}\sum_i\tilde{\boldsymbol{c}}_i\tilde{\boldsymbol{c}}_i'$ be the $pM\times pM$ sample variance–covariance matrix of the estimated coefficient vectors $\tilde{\boldsymbol{c}}_i$ of the p-dimensional functional data $\boldsymbol{x}_i(t)$. The sample variance $\mathrm{var}(f)$ of $\{f_i;i=1,\ldots,N\}$ can be written as

$$\mathrm{var}(f)=\frac{1}{N}\sum_{i=1}^N f_i^2=\frac{1}{N}\sum_{i=1}^N\boldsymbol{\theta}'W\tilde{\boldsymbol{c}}_i\tilde{\boldsymbol{c}}_i'W\boldsymbol{\theta}=\boldsymbol{\theta}'WVW\boldsymbol{\theta}.$$

The penalized sample variance $\mathrm{PSV}_{\boldsymbol{\lambda}}(\xi)$ in (4) can then be written as

$$\mathrm{PSV}_{\boldsymbol{\lambda}}(\boldsymbol{\theta})=\frac{\boldsymbol{\theta}'WVW\boldsymbol{\theta}}{\boldsymbol{\theta}'(W+Q_{\boldsymbol{\lambda}})\boldsymbol{\theta}},$$

since the norm of the p-dimensional weight function $\xi(t)$ is expressed as $\|\xi\|_p^2=\sum_{l=1}^p\|\xi_l\|^2=\sum_{l=1}^p\boldsymbol{\theta}_l'W^*\boldsymbol{\theta}_l=\boldsymbol{\theta}'W\boldsymbol{\theta}$. Also let $\boldsymbol{u}=U_{\boldsymbol{\lambda}}\boldsymbol{\theta}$ and $S_{\boldsymbol{\lambda}}=U_{\boldsymbol{\lambda}}^{-1}$, where the $pM\times pM$ non-singular upper triangular matrix $U_{\boldsymbol{\lambda}}$ satisfies $W+Q_{\boldsymbol{\lambda}}=U_{\boldsymbol{\lambda}}'U_{\boldsymbol{\lambda}}$. We then have

$$\mathrm{PSV}_{\lambda}(\boldsymbol{u})=\frac{\boldsymbol{u}'S_{\boldsymbol{\lambda}}'WVWS_{\boldsymbol{\lambda}}\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{u}}.\tag{5}$$

Thus, the maximum problem of the penalized sample variance $\mathrm{PSV}_{\boldsymbol{\lambda}}(\xi)$ is equivalent to the maximum problem of the above quadratic form (5). Therefore we need to solve the eigenvalue problem for the $pM\times pM$ matrix $S_{\boldsymbol{\lambda}}'WVWS_{\boldsymbol{\lambda}}$.

Let $\rho_1\geqslant\cdots\geqslant\rho_{pM}$ be the eigenvalues of $S_{\boldsymbol{\lambda}}'WVWS_{\boldsymbol{\lambda}}$ and $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{pM}$ be the orthonormal eigenvectors corresponding to the eigenvalues $\rho_1,\ldots,\rho_{pM}$, respectively. The estimated coefficient parameter vectors $\hat{\boldsymbol{\theta}}_k^{\boldsymbol{\lambda}}$ are given by

$$\hat{\boldsymbol{\theta}}_k^{\boldsymbol{\lambda}}=\frac{1}{\sqrt{\boldsymbol{e}_k'S_{\boldsymbol{\lambda}}'WS_{\boldsymbol{\lambda}}\boldsymbol{e}_k}}S_{\boldsymbol{\lambda}}\boldsymbol{e}_k\quad(k=1,\ldots,pM).$$

The p-dimensional kth PC curves $\xi_k^{\boldsymbol{\lambda}}(t)$ and PC scores $\{f_{ki}^{\boldsymbol{\lambda}}=\langle\xi_k^{\boldsymbol{\lambda}},\boldsymbol{x}_i\rangle_p;i=1,\ldots,N\}$ can then be obtained by using $\hat{\boldsymbol{\theta}}_k^{\boldsymbol{\lambda}}$.

We can express the p-dimensional functional data sets $\{x_{i1}(t),\ldots,x_{ip}(t);i=1,\ldots,N\}$ as uncorrelated scores, since the sample covariance of the kth and $k'(\neq k)$th PC scores is 0. Furthermore, it is possible to calculate (cumulative) contribution rates of PCs, using the sample variance of $\{f_{k1}^{\boldsymbol{\lambda}},\ldots,f_{kN}^{\boldsymbol{\lambda}}\}$ given by $\rho_k/(\boldsymbol{e}_k'S_{\boldsymbol{\lambda}}'WS_{\boldsymbol{\lambda}}\boldsymbol{e}_k)$. When the cumulative contribution rate for the first few PCs is over 80%, we often consider that the few PCs contain almost complete information about individual variations.

The eigenvalue problem for the $pM\times pM$ matrix $S_{\boldsymbol{\lambda}}'WVWS_{\boldsymbol{\lambda}}$ in (5) must be solved on performing our method that allows us to have pM PCs. Notice that when the value of pM is quite large, we cannot choose a few PCs with almost complete information about individual variations.

3.3. Smoothing parameter selection

The smoothing parameters λ_l in regularized functional PCA can be optimally selected, as follows. Rice and Silverman (1991) and Silverman (1996) selected the optimal smoothing parameter using a CV method.

When we have smoothing parameters λ_l and $k\in\{1,2,\ldots,pM\}$, then ith p-dimensional functional data $\boldsymbol{x}_i(t)$ is projected into the space spanned by the PC curves $\{\xi_r^{\boldsymbol{\lambda},-i}(t);r=1,\ldots,k\}$, where each $\xi_r^{\boldsymbol{\lambda},-i}(t)$ denotes the rth PC curve estimated from the functional data set excluding $\boldsymbol{x}_i(t)$. The projected (reconstructed) functional data $\hat{\boldsymbol{x}}_{ik}^{\boldsymbol{\lambda},-i}(t)$ are given by

$$\hat{\boldsymbol{x}}_{i,k}^{\boldsymbol{\lambda},-i}(t)=\sum_{r=1}^k\sum_{q=1}^k\{(G_k^{\boldsymbol{\lambda},-i})^{-1}\}_{rq}\langle\xi_q^{\boldsymbol{\lambda},-i},\boldsymbol{x}_i\rangle_p\xi_r^{\boldsymbol{\lambda},-i}(t)\quad(i=1,\ldots,N),$$

where the $k\times k$ matrix $G_k^{\boldsymbol{\lambda},-i}$ has (r,q)th components $G_{k,rq}^{\boldsymbol{\lambda},-i}=\langle\xi_r^{\boldsymbol{\lambda},-i},\xi_q^{\boldsymbol{\lambda},-i}\rangle_p$. The CV scores $CV_k(\boldsymbol{\lambda})$ and $CV(\boldsymbol{\lambda})$ are defined by

$$CV_k(\boldsymbol{\lambda})=\sum_{i=1}^N\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_{i,k}^{\boldsymbol{\lambda},-i}\|_p^2,\quad CV(\boldsymbol{\lambda})=\sum_{k=1}^{pM}CV_k(\boldsymbol{\lambda}).$$

The set of optimal smoothing parameters is obtained by minimizing $CV(\boldsymbol{\lambda})$.

4. Numerical experiments

In this section, Monte Carlo experiments are conducted to compare the effectiveness of the proposed method via Gaussian basis functions and via cubic B-splines with equidistant knots. We refer to de Boor (2001) and Imoto and Konishi (2003) for B-splines. We note that Fourier series are orthonormal, while Gaussian basis functions and B-splines are not. The evaluation of the cross-product matrix for Gaussian basis functions was described in the Section 3.2. Then if we perform regularized functional PCA via B-splines, its cross-product matrix W^* must be evaluated. We derived the integral of the product of any two B-spline bases. An outline of the evaluation is shown in Appendix.

Although the main concern of this paper is to introduce regularized functional PCA for multidimensional functional data sets using Gaussian basis functions, our numerical experiments deal with one-dimensional case. This restriction is without loss of generality for comparing the performance of regularized functional PCA with Gaussian basis functions and that with B-splines, because an essential point of our multidimensional model is given by Eq. (3): if any interaction effects among p curves would be taken into account, we should perform the multidimensional simulation.

A true functional data set $\{x_i(t); t \in [0,1], i = 1,\dots,15\}$ was generated in each trial of the Monte Carlo experiments. However, this data set $\{x_i(t)\}$ could not be expressed in terms of a basis expansion, so a discrete data set was generated from $\{x_i(t)\}$, and a new functional data set $\{\tilde{x}_i(t); i = 1,\dots,15\}$ was then obtained by smoothing the generated discrete data set. Applying regularized functional PCA to $\tilde{x}_i(t)$, we calculated the mean square error (MSE) between the true functional data $x_i(t)$ and the functional data $\hat{x}_i^\lambda$ that reconstructed by estimated PC curves. Morepreciselv, we performed the Monte Carlo experiment using the following procedure.

Step 1: Generate a true functional data set $\{x_i(t); i = 1,\dots,15\}$ from mixed effects models (see, e.g., James et al., 2000),

$$x_i(t) = \mu(t) + \sum_{m=1}^{4} \alpha_{im}\xi_m(t) \quad (t \in [0,1],\ i = 1,\dots,15),$$

where the mean function $\mu(t)$ is assumed to be the following functions (see, e.g., Hurvich et al., 1998):

1. $\mu(t) = \mathrm{e}^{-3t}\sin(3\pi t)$,
2. $\mu(t) = 1 - 48t + 218t^2 - 315t^3 + 145t^4$,
3. $\mu(t) = \sin(10\pi t^2)$,

and $\xi_{2r-1}(t) = \sin(2\pi rt)$, $\xi_{2r}(t) = \cos(2\pi rt)$ $(r = 1,2)$. The random components α_{im} are assumed to be independently normally distributed with $\alpha_{im} \overset{\text{iid}}{\sim} \mathrm{N}(0,(0.03R_x)^2)$, where R_x is the range of $\mu(t)$ over $t \in [0,1]$.

Step 2: Generate discrete data $\{x_{ij}; j = 1,\dots,n_i\}$ from the nonlinear regression models with the true functions $x_i(t)$:

$$x_{ij} = x_i(t_{ij}) + \varepsilon_{ij} \quad (j = 1,\dots,n_i,\ i = 1,\dots,15),$$

where the errors ε_{ij} are assumed to be independently normally distributed with $\varepsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0,\sigma_\varepsilon^2)$, where standard deviation (SD) σ_ε is taken as $0.05R_x$, $0.1R_x$, $0.2R_x$. A set of observational points t_{ij} is generated from the uniform distribution on $[0,1]$. The numbers n_i of observational points are taken as $n_i = 100$ or generated from the normal distribution with mean 100 and variance 2^2. We note that the generated discrete data can be referred to as high-dimensional ($n_i \approx 100$) and small sample-size data ($N = 15$).

Step 3: Estimate a functional data set by smoothing the discrete data set $\{x_{ij}; j = 1,\dots,n_i, i = 1,\dots,15\}$. It is assumed that each functional data $x_i(t)$ can be expressed as a linear combination of Gaussian basis functions or B-splines. The number of basis functions M, hyper-parameter ν (for Gaussian basis functions) and smoothing parameters β_i are optimally selected by minimizing GIC. Fig. 2 shows generated true functional data $x_1(t)$ (dashed line), discrete data $\{x_{1j}; j = 1,\dots,n_i\}$ and estimated functional data $\tilde{x}_1(t)$ (solid line) for three mean functions with $\sigma_\varepsilon = 0.1R_x$.

Step 4: Perform regularized functional PCA on the estimated functional data set $\{\tilde{x}_i(t); i = 1,\dots,15\}$ and smoothing parameter selection based on the CV method.

Step 5: Calculate the MSE for the bth trial,

$$\mathrm{MSE}_b = \frac{1}{15}\sum_{i=1}^{15} \|x_i - \hat{x}_i^\lambda\|^2,$$

where λ is the selected smoothing parameter using CV and $\hat{x}_i^\lambda(t) = \sum_{r=1}^{M}\sum_{q=1}^{M}(G_M^\lambda)_{rq}^{-1}\langle\xi_q^\lambda,\tilde{x}_i\rangle\xi_r^\lambda(t)$ are the reconstructed functional data with the $M \times M$ matrix G_M^λ that has (r,q)th element $G_{M,rq}^\lambda = \langle\xi_r^\lambda,\xi_q^\lambda\rangle$.

Step 6.: Repeat Steps 1–5 for each trial. Then the average mean square error (AMSE) is given by $\mathrm{AMSE} = 100^{-1}\sum_{b=1}^{100}\mathrm{MSE}_b$.

Table 1 shows the simulation results with the AMSE and SD of MSE for Gaussian basis functions and B-splines. From this table, when the numbers n_i of observational points were generated from the normal distribution, all AMSEs for Gaussian basis functions would be smaller than the corresponding values for B-splines. Most of the results for Gaussian basis functions to

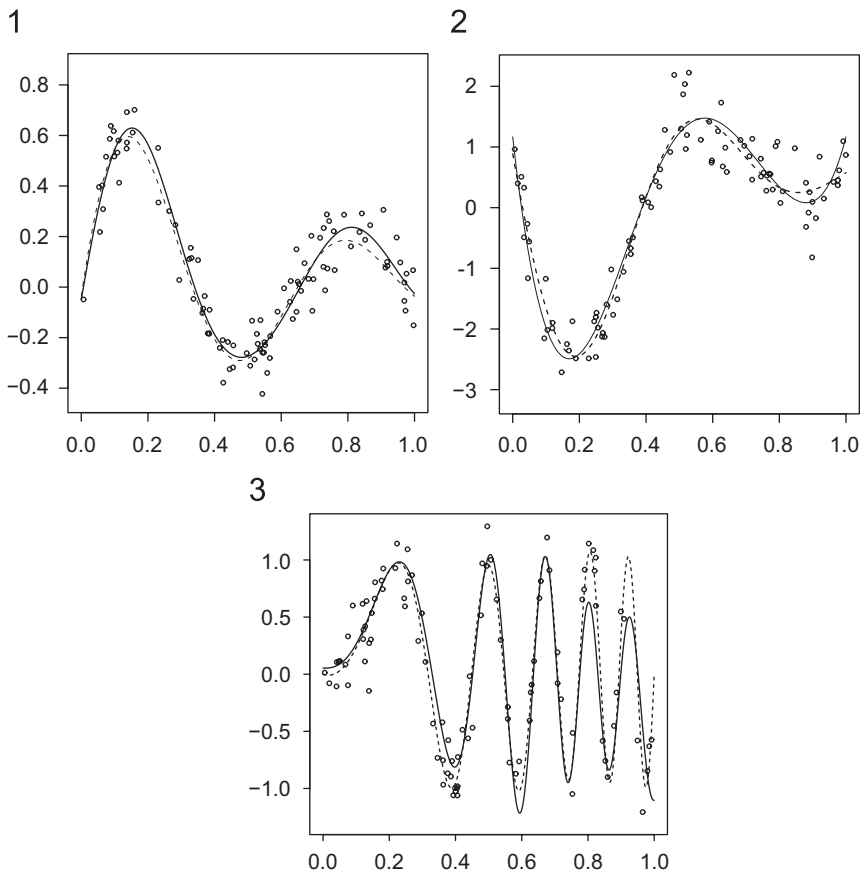

Fig. 2. Examples of simulated data: the dashed lines are the true functional data, while the solid lines are the estimated functional data. $\mu(t)$= (1) $e^{-3t}\sin(3\pi t)$, (2) $1-48t+218t^2-315t^3+145t^4$ and (3) $\sin(10\pi t^2)$.

the unbalanced data (n_i:normal) were better than the results of the same size data ($n_i = 100$), while the results for *B*-splines were not. In addition, when we imposed the centers of Gaussian basis functions on equi-distance, the comparisons of the effectiveness for the methods via these Gaussian basis functions and via *B*-splines led to a similar conclusion from Table 1. As consequence, regularized functional PCA via Gaussian basis functions performed well in the sense of minimizing AMSE through these simulations.

5. Real data example

We apply the proposed regularized functional PCA to 3D protein structures such as that shown in Fig. 3. There have been many studies that have analyzed proteins using statistical methods (Wu et al., 1998; Ding and Dubchak, 2001; Green and Mardia, 2006, among others). Regularized functional PCA is applied here to 3D functional data sets representing 3D protein structures, in order to identify any features of the protein structures.

Proteins have been classified from a biological point of view, and a protein class is referred to as a family. A protein family is a group of evolutionarily related proteins. We treat 12 proteins from the four families given in Table 2. The 3D protein structural data set was obtained from the National Center of Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/). It should be noted that because the length of amino-acid sequence differs for each protein, the conventional multivariate analysis including PCA cannot be directly applied to this unbalanced data set. In what follows, it is assumed that we have the *XYZ*-coordinate values of all atoms for each protein in various coordinate systems.

Firstly, the set of 3D protein structures was converted into discrete data sets using the positions $\{t_{ij}\}$ on the *i*th amino-acid sequence and the *XYZ*-coordinate values $\{x_{ij}, y_{ij}, z_{ij}\}$ of the α-carbon atoms which were typical atoms of amino-acids. Each α-carbon atom corresponds to an amino-acid. We then had a discrete data set for each coordinate, and the smoothing method using Gaussian basis functions was performed for each discrete data set. We considered values for M of $3, 4, \ldots, 20$, values for ν of $1, 2, \ldots, 50$ and values for β_{il} of $10^{-10}, 10^{-9}, \ldots, 10^{-1}$ and found optimal values of $M = 15$, $\nu = 11.6$ and $\beta_{il} = 10^{-8} \sim 10^{-5}$. The selected values of M and ν were the mode and mean of that for all individuals and coordinates, respectively.

Table 1
Simulation results for three mean functions.

	$\sigma_\varepsilon = 0.05R_x$		$\sigma_\varepsilon = 0.1R_x$		$\sigma_\varepsilon = 0.2R_x$	
	Gaussian	B-splines	Gaussian	B-splines	Gaussian	B-splines
$\mu(t) = e^{-3t}\sin(3\pi t)$						
$n_i = 100$						
AMSE ($\times 10^{-3}$)	7.792	7.792	7.962	7.957	7.972	7.960
SD(MSE) ($\times 10^{-4}$)	9.17	9.16	9.77	9.76	9.23	9.26
n_i: normal						
AMSE ($\times 10^{-3}$)	7.682	9.197	7.802	9.230	7.987	9.312
SD(MSE) ($\times 10^{-4}$)	8.43	10.04	9.54	12.15	9.59	12.62
$\mu(t) = 1 - 48t + 218t^2 - 315t^3 + 145t^4$						
$n_i = 100$						
AMSE ($\times 10^{-2}$)	1.472	1.471	1.508	1.504	1.520	1.509
SD(MSE) ($\times 10^{-3}$)	1.72	1.72	1.92	1.92	1.81	1.82
n_i: normal						
AMSE ($\times 10^{-2}$)	1.468	1.786	1.504	1.790	1.518	1.793
SD(MSE) ($\times 10^{-3}$)	1.86	2.03	1.67	1.81	1.85	2.35
$\mu(t) = \sin(10\pi t^2)$						
$n_i = 100$						
AMSE ($\times 10^{-1}$)	2.599	3.353	2.638	3.392	2.913	3.435
SD(MSE) ($\times 10^{-2}$)	3.30	0.29	2.61	0.89	2.75	1.22
n_i: normal						
AMSE ($\times 10^{-1}$)	2.699	3.357	2.782	3.378	2.923	3.449
SD(MSE) ($\times 10^{-2}$)	2.24	0.48	3.05	0.64	2.98	1.73

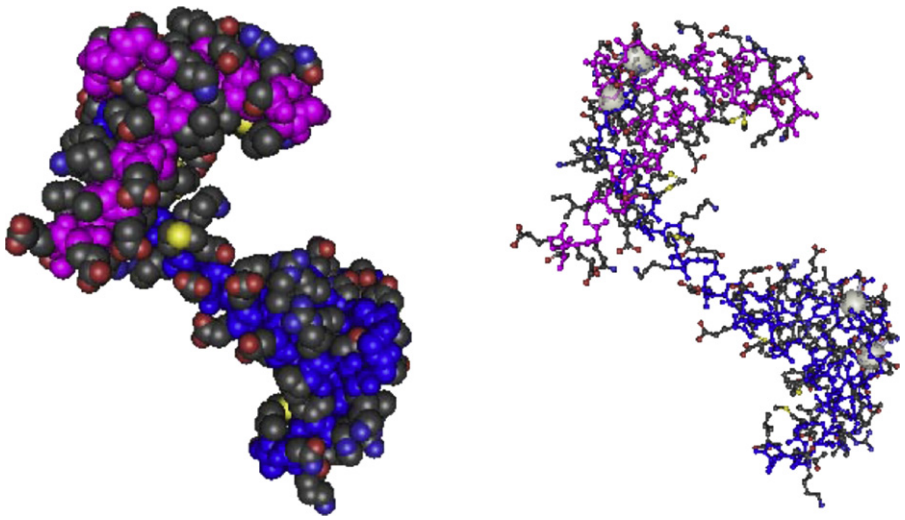

Fig. 3. Examples of 3D protein structures. The space-filling (left) and ball-and-stick models (right) of a protein.

Table 2
The 12 proteins from the four families.

Family code	Family name	Protein code
adk	Nucleotide kinase	1gky (186) 3adk (194)
aza	Azulin/plastocyanin	1azu (125) 1plc (99) 7pcy (98) 1paz (120) 9pcy (92)
cbp	Calcium-binding protein (calmodulin-like)	3cln (142) 4cln (148) 5cln (161)
dhfr	Dehydrofolate reductase	3dfr (162) 8dfr (186)

Each number in the parentheses shows the length of the amino-acid sequence.

To unify the coordinates, we rotated the estimated functional data sets obtained by smoothing, since the coordinate systems differ among the proteins. The rotation serves as a type of functional alignment for the different functions among the individuals. Optimization was performed by rotating each protein to an another base protein. A root mean square deviation (RMSD) for two functional data $\{x(t), y(t), z(t); t \in \mathcal{T}\}$ and $\{x'(t), y'(t), z'(t); t \in \mathcal{T}\}$ was utilized as a criterion for the optimization, and it was here

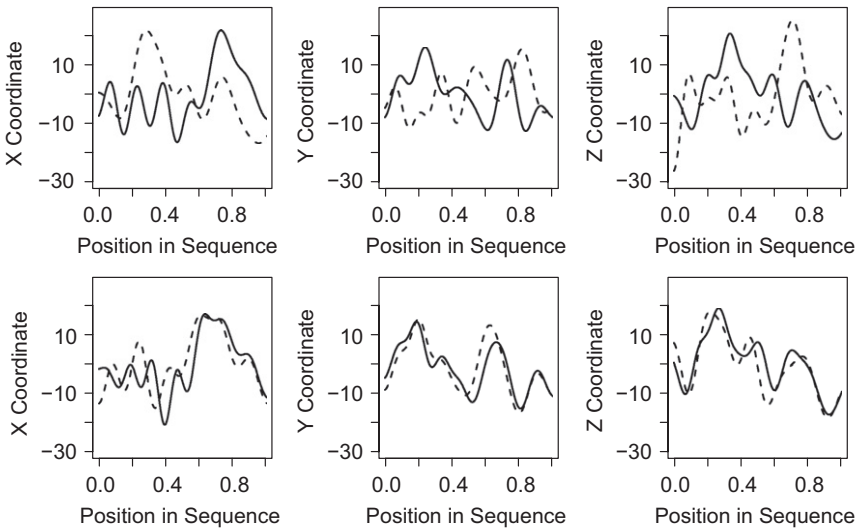

Fig. 4. An example of a rotation of proteins. The three-dimensional functional data (upper) and rotated data (lower) of two proteins.

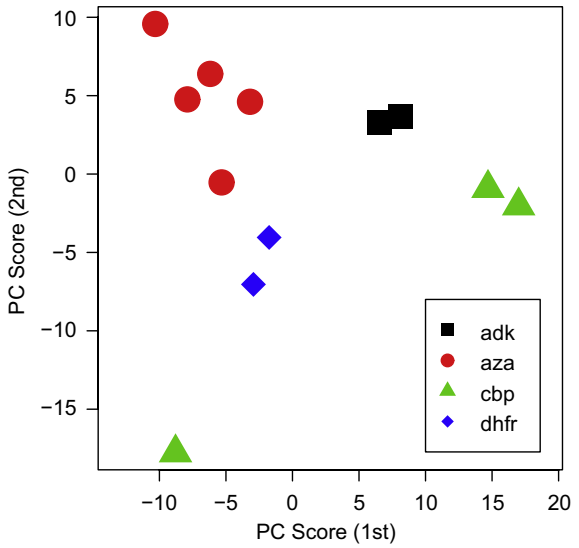

Fig. 5. The principal component (PC) scores for each family. The proteins in the adk, aza and dhfr families are clustered in their respective groups, while the cbp family has an un-clustered protein.

defined by

$$\mathrm{RMSD}_F = \frac{1}{\sqrt{|\mathscr{T}|}}\left[\int_{\mathscr{T}}\{x(t)-x'(t)\}^2\,\mathrm{d}t + \int_{\mathscr{T}}\{y(t)-y'(t)\}^2\,\mathrm{d}t + \int_{\mathscr{T}}\{z(t)-z'(t)\}^2\,\mathrm{d}t\right]^{1/2}.$$

We employed Euler angles $\theta_1, \theta_2, \theta_3$ as a rotation method with step size 10°, and the selected angles $\theta_1^*, \theta_2^*, \theta_3^*$ were then varied with step size 1°. The Euler angles allow us to decompose the rotation into three separate angles. We selected the optimized values of $\theta_1, \theta_2, \theta_3$ by minimizing RMSD. Fig. 4 shows an example of the 3D functional data sets (upper) and the rotated ones (lower). In this figure, we show a rotation of two proteins. The regularized functional PCA was applied to the rotated 3D functional data sets.

Using CV resulted in $\lambda_1 = 6.31 \times 10^{-6}$, $\lambda_2 = 2.51 \times 10^{-6}$ and $\lambda_3 = 3.98 \times 10^{-6}$ as optimal smoothing parameters, where we set the candidate values of λ_l $(l = 1, 2, 3)$ to $\lambda_{li} = 10^{i-11}$ $(i = 1, \ldots, 10)$ and $\lambda_{li} = 10^{(i-71)/10}$ $(i = 1, \ldots, 21)$. We first selected optimal smoothing parameters $\lambda_l = 10^{-6}$ from the candidate values $10^{-10}, 10^{-9}, \ldots, 10^{-1}$, and the parameters were then varied with step size $10^{-0.1}$ around 10^{-6}. With the selected smoothing parameters, we estimated PC curves and PC scores and plotted the PC scores for each family (Fig. 5). The proteins belonging to the adk, aza and dhfr families were clustered in respective family groups; however, the cbp family contained an unclustered protein. This problem may be caused by the "slim" structure of proteins in the

cbp family, while we successfully captured the "ball" structure characteristic of proteins in the adk, aza and dhfr families. We note that when the regularized functional PCA via *B*-splines were applied to the protein data set, the resulting PC plot could not present clear patterns, and the CV score for Gaussian bases was less than the value for *B*-splines.

Thus, using our functionalization method, 3D protein structures can be captured without relying on their sequence information, physicochemical properties and a visual census of an enormous number of proteins. However, we may have to use a robust representation of a 3D protein structure for rotation.

6. Concluding remarks

We introduced regularized functional PCA for multidimensional functional data sets, using Gaussian basis functions. The results of the Monte Carlo experiments showed that our regularized functional PCA based on Gaussian basis functions performed well, and was superior to that based on cubic *B*-splines in the sense of minimizing the MSE and its SD for unbalanced data. The proposed procedure extracted useful information from unbalanced data like the protein structural data. The analysis of the real data set showed that the 3D protein structures could be characterized by our method without relying on their sequence information and physicochemical properties. Future works that remains to be done include derivation of model selection criteria from an information theoretic perspective and also the application of Bayesian approaches instead of CV.

Acknowledgments

The authors would like to thank the editor and reviewers for constructive and helpful comments that improved the quality of the paper considerably. We are also grateful to Professor Satoru Kuhara and Hideki Hirakawa of Kyushu University for their help concerned with the application to protein structural data.

Appendix. Evaluation of the cross-product matrix for cubic *B*-splines

This section shows an outline of the evaluation for the cross-product matrix $W^* = \{W^*_{mn} = \int_{\mathscr{T}} \phi_m(t)\phi_n(t)\,\mathrm{d}t\}_{m,n=1}^{M}$ via cubic *B*-splines $\{\phi_m(t)\}$ with the equispaced knots $k_1 < k_2 < \cdots < k_{M+4}$, where $\mathscr{T} = [k_4, k_{M+1}]$. We refer to de Boor (2001) and Imoto and Konishi (2003) for *B*-splines.

It is known that *B*-splines $\phi_1(t;r), \ldots, \phi_M(t;r)$ with degree $r \in \{1, 2, \ldots\}$ and knots $k_1 < k_2 < \cdots < k_{M+r+1}$ are given by the sequential equation (de Boor, 2001):

$$\phi_m(t;r) = \frac{t - k_m}{k_{m+r} - k_m}\phi_m(t;r-1) - \frac{t - k_{m+r+1}}{k_{m+r+1} - k_{m+1}}\phi_{m+1}(t;r-1),$$

where $\phi_m(t;0) = 1(k_m \leqslant t < k_{m+1}), =0$ (otherwise). The cubic *B*-splines $\{\phi_m(t;3)\}$ are here denoted by $\{\phi_m(t)\}$.

The diagonal components W^*_{mm} of W^* can be evaluated through the integrations $I_1^d = \int_{k_1}^{k_2} \phi_1(t)^2\,\mathrm{d}t$ and $I_2^d = \int_{k_2}^{k_3} \phi_1(t)^2\,\mathrm{d}t$;

$$W^*_{11} = I_1^d (= W^*_{MM}), \quad W^*_{22} = I_1^d + I_2^d (= W^*_{M-1,M-1}),$$
$$W^*_{33} = I_1^d + 2I_2^d (= W^*_{M-2,M-2}), \quad W^*_{mm} = 2I_1^d + 2I_2^d \quad (m = 4, 5, \ldots, M-3).$$

It may be noted that the *B*-splines are symmetric and $\phi_m(t) = 0$ $(t < k_m, k_{m+4} \leqslant t)$. Furthermore, each *B*-spline function $\phi_m(t)$ is given by the parallel translation of the other *B*-splines $\phi_n(t)$ $(n \neq m)$.

The calculation of the non-diagonal components W^*_{mn} $(m < n)$ requires the four integrations $I_1^{nd} = \int_{k_4}^{k_5} \phi_1(t)\phi_2(t)\,\mathrm{d}t$, $I_2^{nd} = \int_{k_4}^{k_5} \phi_1(t)\phi_3(t)\,\mathrm{d}t$, $I_3^{nd} = \int_{k_4}^{k_5} \phi_1(t)\phi_4(t)\,\mathrm{d}t$ and $I_4^{nd} = \int_{k_3}^{k_4} \phi_1(t)\phi_2(t)\,\mathrm{d}t$. We then have the components in the first to third rows;

$$W^*_{12} = I_1^{nd}, \quad W^*_{13} = I_2^{nd}, \quad W^*_{14} = I_3^{nd}, \quad W^*_{15} = \cdots = W^*_{1M} = 0,$$
$$W^*_{23} = I_1^{nd} + I_4^{nd}, \quad W^*_{24} = 2I_2^{nd}, \quad W^*_{25} = I_3^{nd}, \quad W^*_{26} = \cdots = W^*_{2M} = 0,$$
$$W^*_{34} = 2I_1^{nd} + I_4^{nd}, \quad W^*_{35} = 2I_2^{nd}, \quad W^*_{36} = I_3^{nd}, \quad W^*_{37} = \cdots = W^*_{3M} = 0.$$

In a similar way, the components in the fourth, fifth,... rows can be obtained. Especially, the components in the $(M-2)$th and $(M-1)$th rows are given by $W^*_{M-2,M-1} = I_1^{nd} + I_4^{nd} (= W^*_{23})$, $W^*_{M-2,M} = I_2^{nd} (= W^*_{13})$ and $W_{M-1,M} = I_1^{nd} (= W^*_{12})$.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, pp. 267–281.

Ando, T., Konishi, S., Imoto, S., 2005. Nonlinear regression modeling via regularized radial basis function networks. J. Statist. Plann. Inference 138 (11), 3616–3633.

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

de Boor, C., 2001. A Practical Guide to Splines. revised ed. Springer, Berlin.

Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.

Eilers, P., Marx, B., 1996. Flexible smoothing with B-splines and penalties (with discussion). Statist. Sci. 11, 89–121.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer, Berlin.
Green, P.J., Mardia, K.V., 2006. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. Biometrika 93 (2), 235–254.
Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman & Hall, London.
Hurvich, C.M., Simonoff, J.S., Tsai, C.L., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. Roy. Statist. Soc. Ser. B 60, 271–293.
Imoto, S., Konishi, S., 2003. Selection of smoothing parameters in *B*-spline nonparametric regression models using information criteria. Ann. Inst. Statist. Math. 55 (4), 671–687.
James, G., Hastie, T., Sugar, C., 2000. Principal component models for sparse functional data. Biometrika 87, 587–602.
Jolliffe, I.T., 2002. Principal Component Analysis. second ed. Springer, Berlin.
Konishi, S., Kitagawa, G., 1996. Generalized information criteria in model selection. Biometrika 83 (4), 875–890.
Konishi, S., Kitagawa, G., 2008. Information Criteria and Statistical Modeling. Springer, Berlin.
Mizuta, M., 2006. Discrete functional data analysis. In: Proceedings in Computational Statistics 2006.Physica-Verlag, Springer, Berlin, pp. 361–369.
Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. Springer, Berlin.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. second ed. Springer, Berlin.
Rice, J.A., Silverman, B.W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. J. Roy. Statist. Soc. Ser. B 53, 233–243.
Rice, J.A., Wu, C.O., 2001. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57, 253–259.
Shi, M., Weiss, R., Taylor, J., 1996. An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. Appl. Statist. 45, 151–164.
Silverman, B.W., 1996. Smoothed functional principal components analysis by choice of norm. Ann. Statist. 24, 1–24.
Wu, T.D., Hastie, T., Schmidler, S.C., 1998. Regression analysis of multiple protein structures. J. Comput. Biol. 5 (3), 585–596.
Yao, F., Lee, T.C.M., 2006. Penalized spline models for functional principal component analysis. J. Roy. Statist. Soc. Ser. B 68, 3–25.
Yao, F., Müller, H.G., Wang, J.L., 2005. Functional data analysis for sparse longitudinal data. J. Amer. Statist. Assoc. 100, 577–590.