MHF Preprint Series

Kyushu University
21st Century COE Program
Development of Dynamic Mathematics with High Functionality

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

M. Kayano& S. Konishi

MHF 2007-18

(Received December 14, 2007)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Functional Principal Component Analysis via Regularized Gaussian Basis Expansions and Its Application to Unbalanced Data

Mitsunori Kayano* **and Sadanori Konishi**

Graduate School of Mathematics, Kyushu University

6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

kayano@math.kyushu-u.ac.jp (M. Kayano)

konishi@math.kyushu-u.ac.jp (S. Konishi)

SUMMARY

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. Numerical experiments are conducted to investigate the effectiveness of our method via Gaussian basis functions, comparing 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 that based on *B*-splines for unbalanced data.

KEY WORDS: functional data analysis, model selection, protein structure, radial basis functions, regularization, smoothing parameter, spline.

1. Introduction

Multivariate analysis deals with observations on more than one variable, where there is some inherent interdependence between the variables (Mardia, Kent and Bibby (1979)),

*Research Fellow of the Japan Society for the Promotion of Science

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 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 the 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, §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 de-

termine 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 a useful information concerning to biological view points.

This paper is organized as follows. Section 2 describes observational discrete data and their functinalization to multidimensional functional data. Section 3 introduces a regularized functional principal component 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 that 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 $\{t_{ij}, (x_{i1j}, \cdots, x_{ipj})\,;\, j = 1, \cdots, n_i\}$ $(i = 1, \cdots, N)$, where each t_{ij} $(\in \mathcal{T} \subset \mathbb{R})$ is the j-th observational point of the i-th individual and $(x_{i1j}, \cdots, x_{ipj})$ $(\in \mathbb{R}^p)$ is the discrete data observed at t_{ij} for p variables $X_1, \cdots, X_p$. In particular, the i-th discrete data set observed at t_{ij} for X_l is represented by $\{(t_{ij}, x_{ilj})\,;\, j = 1, \cdots, n_i\}$. It may be noted that we have the discrete data observed at possibly different observational points $t_{i1}, \cdots, t_{in_i}$ for each subject, and then the discrete observations can be referred to as unbalanced data. For example, $\{t_{ij}, (x_{i1j}, x_{i2j}, x_{i3j})\,;\, j = 1, \cdots, n_i\}$ $(i = 1, \cdots, 12)$ are the measurements in XYZ coordinates of 3D protein structures, where t_{ij} are the positions in i-th amino-acid sequence and $(x_{i1j}, x_{i2j}, x_{i3j})$ are the XYZ coordinates values of amino acids which compose i-th 3D protein structure. Fig1 (upper) shows an example of discretized 3D protein structural data with $p = 3$ and $n_i = 186$.

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

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

where the errors ε_{ilj} are independently normally distributed with mean 0 and variance σ^2_{il}. The nonlinear functions $u_{il}(t)$ are assumed to be given by linear combinations of Gaussian

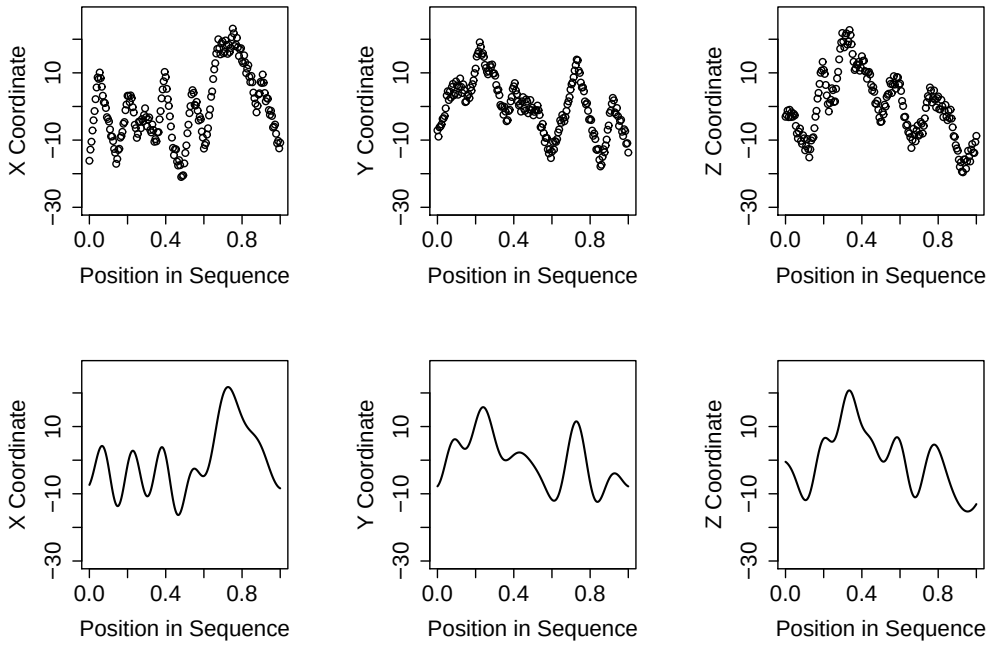

Fig. 1: An example of discrete data (upper) and corresponding three-dimensional functional data (lower) for a 3D protein structure ($p = 3$, $n_i = 186$).

basis functions $\{\phi_m(t) = \phi_m(t;\ \nu, \mu_m, \tau_m^2)\}$ with parameters $\mu_m,\ \tau_m$ and ν,

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

where the m-th Gaussian basis function $\phi_m(t)$ has the form

$$\phi_m(t) = \phi_m(t;\ \nu, \mu_m, \tau_m^2) = \exp\left\{-\frac{(t-\mu_m)^2}{2\nu\tau_m^2}\right\} \qquad (m = 1, \cdots, M)\ . \tag{1}$$

The parameters μ_m and τ_m express the position and width of the m-th basis function and ν is a hyper-parameter that adjusts the degree of overlapping among the basis functions (Ando *et al.* (2005)).

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

$\hat{\sigma}_{il}^2$ 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 selected by minimizing the generalized information criterion (GIC), given by Konishi and Kitagawa (1996) (see also, Konishi and Kitagawa (2008)).

Thus, we have the estimated nonlinear functions $\hat{u}_{il}(t) = \sum_{m=1}^{M} \hat{c}_{ilm}\phi_m(t)$ $((i = 1, \cdots, N,\ l = 1, \cdots, p)$, where $\phi_m(t) = \phi_m^{\nu}(t) = \phi_m(t;\ \nu, \hat{\mu}_m, \hat{\tau}_m^2)$ is the m-th basis function with the optimal hyper-parameter ν selected by minimizing GIC. The p-dimensional functional data sets $\{x_{i1}^*(t), \cdots,\ x_{ip}^*(t)\ ;\ t \in \mathcal{T}\}$ are 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 Gaussian basis functions. An example of multidimensional functional data is shown in Fig1 (lower), corresponding to the discretized 3D protein structural data in Fig1 (upper).

3. Functional Principal Component Analysis

3.1 Model

Let $\{(x_{i1}^*(t), \cdots, x_{ip}^*(t))\ ;\ t \in \mathcal{T}\}$ $(i = 1, \cdots, N)$ be the p-dimensional functional data sets obtained by smoothing the observational discrete data sets $\{t_{ij}, (x_{i1j}, \cdots, x_{ipj})\ ;\ j = 1, \cdots, n_i\}$ $(i = 1, \cdots, N)$. A functional principal component method is here applied to the p-dimensional functional data sets $\{(x_{i1}(t), \cdots, x_{ip}(t))\ ;\ t \in \mathcal{T}\}$ $(i = 1, \cdots, N)$, 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 $x_{i1}^*(t), \cdots, x_{ip}^*(t)$. It is assumed that each functional data element $x_{il}(t)$ can be expressed as a linear combination of Gaussian basis functions $\phi_m(t) = \phi_m^{\nu}(t) = \phi_m(t;\ \nu, \hat{\mu}_m, \hat{\tau}_m^2)$,

$$x_{il}(t) = \sum_{m=1}^{M} \tilde{c}_{ilm}\phi_m(t) = \tilde{\boldsymbol{c}}_{il}'\boldsymbol{\phi}(t) \qquad (i = 1, \cdots, N,\ l = 1, \cdots, p),$$

where $\tilde{\boldsymbol{c}}_{il} = (\tilde{c}_{il1}, \cdots, \tilde{c}_{ilM})'$ and $\boldsymbol{\phi}(t) = (\phi_1(t), \cdots, \phi_M(t))'$.

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

$$f_i = \langle \boldsymbol{\xi}, \boldsymbol{x}_i \rangle_p = \sum_{l=1}^{p} \langle \xi_l, x_{il} \rangle = \sum_{l=1}^{p} \int_{\mathcal{T}} \xi_l(t) x_{il}(t) dt \qquad (i = 1, \cdots, N)\ .$$

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

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

with $\boldsymbol{\theta}_l = (\boldsymbol{\theta}_{l1}, \cdots, \boldsymbol{\theta}_{lM})'$. A general functional principal component 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. Ramsay and Silverman (2005, §8.5) describes the functional principal component method to the 2-dimensional functional data sets which include the hip and knee angles during a human gait cycle.

On the other hand, regularized (smoothed) functional principal component analysis (regularized functional PCA) proposed by Rice and Silverman (1991) and Silverman (1996) avoids ill-posed problems from functional PCA and maximizes the penalized sample variance (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 subject to penalized orthonormal constraints.

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

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \cdots, \boldsymbol{\theta}_p')'$, $\|\boldsymbol{\xi}\|_p^2 = \sum_l \|\xi_l\|^2 = \sum_l \int_{\mathcal{T}} \xi_l^2(t)\, dt$ is the norm of a p-dimensional weight function $\boldsymbol{\xi}(t)$ and $Q_\lambda = \mathrm{diag}(\lambda_1 Q^*, \cdots, \lambda_p Q^*)$ is a $pM \times pM$ positive-semidefinite block diagonal matrix with $M \times M$ positive-semidefinite matrix Q^* and smoothing parameters $\lambda_l > 0$ which control the smoothness of the weight functions $\xi_l(t)$. The smoothing parameters λ_l can be optimally selected by minimizing a cross validation score.

The principal component (PC) curves are defined by the p-dimensional weight function $\boldsymbol{\xi}(t)$ that maximizes the penalized sample variance $\mathrm{PSV}_\lambda(\boldsymbol{\xi})$ given by (2) subject to the penalized orthonormal constraints.

First PC Curve $\boldsymbol{\xi}_1^\lambda(t) = (\xi_{11}^\lambda(t), \cdots, \xi_{1p}^\lambda(t))'$:
the p-dimensional weight function $\boldsymbol{\xi}(t)$ that maximizes $\mathrm{PSV}_\lambda(\boldsymbol{\xi})$
subject to $\|\boldsymbol{\xi}\|_p^2 = 1$,

k (≥ 2)-th PC Curve $\boldsymbol{\xi}_k^\lambda(t) = (\xi_{k1}^\lambda(t), \cdots, \xi_{kp}^\lambda(t))'$:
the p-dimensional weight function $\boldsymbol{\xi}(t)$ that maximizes $\mathrm{PSV}_\lambda(\boldsymbol{\xi})$
subject to $\|\boldsymbol{\xi}\|_p^2 = 1$ and $\langle \boldsymbol{\xi}, \boldsymbol{\xi}_r^\lambda \rangle_p + \boldsymbol{\theta}' Q_\lambda \boldsymbol{\theta}_r^\lambda = 0$ $(r < k)$,

where the l-th element $\xi_{kl}^\lambda(t)$ of $\boldsymbol{\xi}_k^\lambda(t)$ may be expressed as the basis expansion $\xi_{kl}^\lambda(t) = \sum_m \theta_{klm}^\lambda \phi_m(t) = (\boldsymbol{\theta}_{kl}^\lambda)' \boldsymbol{\phi}(t)$, and $\boldsymbol{\theta}_k^\lambda = ((\boldsymbol{\theta}_{k1}^\lambda)', \cdots, (\boldsymbol{\theta}_{kp}^\lambda)')'$. The k-th principal component score is defined by $\{f_{ki}^\lambda = \langle \boldsymbol{\xi}_k^\lambda, \boldsymbol{x}_i \rangle_p \,;\, i = 1, \cdots, N\}$ $(k = 1, \cdots, pM)$. We note that there are pM principal components by the assumption of basis expansions.

3.2 Eigenvalue Problem

The PC curves $\boldsymbol{\xi}_k^\lambda(t) = (\xi_{k1}^\lambda(t), \cdots, \xi_{kp}^\lambda(t))'$ 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 i-th p-dimensional functional data $\boldsymbol{x}_i(t)$ can be written as

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

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

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

$$W_{mn}^* = \frac{\sqrt{2\pi\nu\hat{\tau}_m^2\hat{\tau}_n^2}}{\sqrt{\hat{\tau}_m^2 + \hat{\tau}_n^2}} \exp\left\{ -\frac{(\hat{\mu}_m - \hat{\mu}_n)^2}{2\nu(\hat{\tau}_m^2 + \hat{\tau}_n^2)} \right\} \qquad (m, n = 1, \cdots, M).$$

We assume that the cross-product matrix W^* of Gaussian basis functions is positive definite. From this assumption, it follows that the condition for a norm is satisfied and regularized functional PCA can be applied. In addition, to satisfy the assumption, we employ Gaussian basis functions except constant term (1), which are as flexible as common Gaussian RBF.

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 $\{f_i \,;\, i = 1, \cdots, N\}$ can be written as

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

The penalized sample variance $\mathrm{PSV}_\lambda(\boldsymbol{\xi})$ in (2) can then be written as

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

since the norm of the p-dimensional weight function $\boldsymbol{\xi}(t)$ is expressed as

$$\|\boldsymbol{\xi}\|_p^2 = \sum_{l=1}^{p} \|\xi_l\|^2 = \sum_{l=1}^{p} \boldsymbol{\theta}_l' W^* \boldsymbol{\theta}_l = \boldsymbol{\theta}' W \boldsymbol{\theta} \; .$$

Also let $\boldsymbol{u} = U_\lambda \boldsymbol{\theta}$ and $S_\lambda = U_\lambda^{-1}$, where the $pM \times pM$ non-singular upper triangular matrix U_λ satisfies $W + Q_\lambda = U_\lambda' U_\lambda$. We then have

$$\mathrm{PSV}_\lambda(\boldsymbol{u}) = \frac{\boldsymbol{u}' S_\lambda' W V W S_\lambda \boldsymbol{u}}{\boldsymbol{u}' \boldsymbol{u}} \; . \tag{3}$$

Thus, the maximum problem of the penalized sample variance $\mathrm{PSV}_\lambda(\boldsymbol{\xi})$ is equivalent to the maximum problem of the above quadratic form (3). Therefore we need to solve the eigenvalue problem for the $pM \times pM$ matrix $S_\lambda' WVWS_\lambda$.

Let $\rho_1 \geq \cdots \geq \rho_{pM}$ be the eigenvalues of $S_\lambda' WVWS_\lambda$ and $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{pM}$ be the orthonormal eigenvectors corresponding to the eigenvalues $\rho_1, \cdots, \rho_{pM}$, respectively. The estimated coefficient parameter vectors $\hat{\boldsymbol{\theta}}_k^\lambda = ((\hat{\boldsymbol{\theta}}_{k1}^\lambda)', \cdots, (\hat{\boldsymbol{\theta}}_{kp}^\lambda)')'$ are given by

$$\hat{\boldsymbol{\theta}}_k^\lambda = \frac{1}{\sqrt{\boldsymbol{e}_k' S_\lambda' W S_\lambda \boldsymbol{e}_k}} S_\lambda \boldsymbol{e}_k \qquad (k = 1, \cdots, pM) \ .$$

The p-dimensional k-th PC curves $\boldsymbol{\xi}_k^\lambda(t)$ and PC scores $\{f_{ki}^\lambda = \langle \boldsymbol{\xi}_k^\lambda, \boldsymbol{x}_i \rangle_p \ ; \ i = 1, \cdots, N\}$ can then be obtained by using $\hat{\boldsymbol{\theta}}_k^\lambda$. Furthermore, we can express the p-dimensional functional data sets $\{x_{i1}(t), \cdots, x_{ip}(t) \, ; \, i = 1, \cdots, N\}$ as uncorrelated scores, since the sample covariance of the k-th and k' $(\neq k)$-th PC scores is 0.

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 cross validation (CV) method.

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

$$\hat{\boldsymbol{x}}_{i,k}^{\lambda,-i}(t) = \sum_{r=1}^{k} \sum_{q=1}^{k} \left(G_k^{\lambda,-i} \right)_{rq}^{-1} \left\langle \boldsymbol{\xi}_q^{\lambda,-i}, \boldsymbol{x}_i \right\rangle_p \ \boldsymbol{\xi}_r^{\lambda,-i}(t) \qquad (i = 1, \cdots, N) \ ,$$

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

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

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

4. Numerical experiments

In this section, Monte Carlo experiments are conducted to compare the effectiveness of the proposed method via Gaussian basis functions and 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 subsection 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.

A true functional data set $\{x_i(t); t \in [0,1],\ i = 1, \cdots, 15\}$ was generated in each trial of 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 $\{\tilde{x}_i(t)\,;\, i = 1, \cdots, 15\}$ 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 that reconstructed by estimated PC curves. More precisely, we performed the Monte Carlo experiment using the following procedure.

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

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

where the mean function $\mu(t)$ is assumed to be the following functions

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

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 $\alpha_{im} \overset{iid}{\sim} N(0, (0.03R_x)^2)$, where R_x is the range of $\mu(t)$ over $t \in [0,1]$.

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

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

where the errors ε_{ij} are assumed to be independently normally distributed with $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$, where standard deviation 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

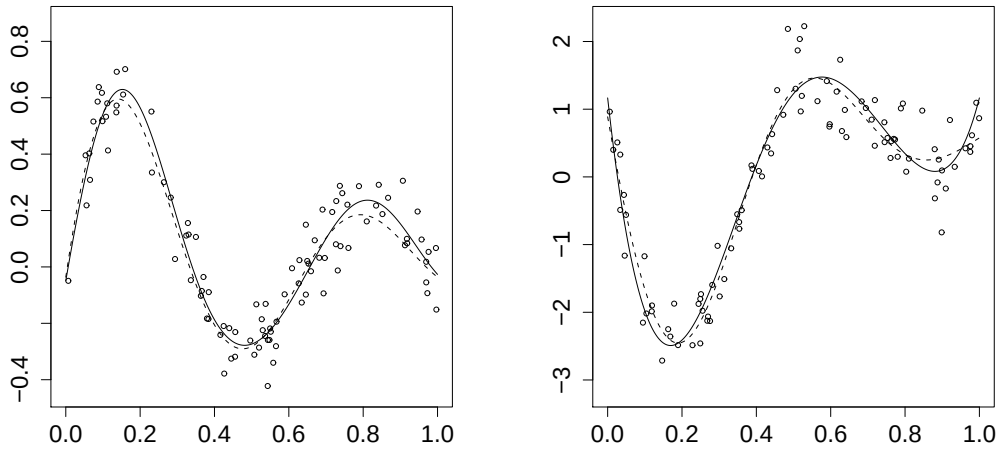

Fig. 2: Examples of simulated data: The dashed lines are the true functional data, while the solid lines are the estimated functional data. $\mu(t) =$ (a) $e^{-3t}\sin(3\pi t)$ and (b) $1-48t+218t^2-315t^3+145t^4$.

with mean 100 and variance 2^2. We note that the generated data can be referred to as high-dimensional and small sample-size data.

Step 3. Estimate a functional data set by smoothing the discrete data set $\{x_{ij}\ ;\ j = 1,\cdots,n_i,\ i = 1,\cdots,12\}$. 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 M of basis functions, hyper-parameter ν (for Gaussian basis functions) and smoothing parameters β_i are optimally selected by minimizing GIC (Konishi and Kitagawa (1996, 2008)). Fig2 shows generated true functional data $x_1(t)$ (dashed line), discrete data $\{x_{1j}; j = 1,\cdots,n_i\}$ and estimated functional data $\tilde{x}_1(t)$ (solid line) for 2 mean functions with $\sigma_\varepsilon = 0.1R_x$.

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

Step 5. Calculate the mean square error for the b-th 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 cross validation and $\hat{x}_i^{\lambda}(t) = \sum_{r=1}^{4}\sum_{q=1}^{4}(G_4^{\lambda})_{rq}^{-1}\left\langle \xi_q^{\lambda}, \tilde{x}_i\right\rangle \xi_r^{\lambda}(t)$ are the reconstructed functional data with the 4×4 matrix G_4^{λ} that has (r,q)-th element $G_{4,rq}^{\lambda} = \langle \xi_r^{\lambda}, \xi_q^{\lambda}\rangle$.

Step 6. Repeat **Steps 1** to **5** for each trial. Then the average mean square error (AMSE)

Table1: Simulation results for 2 mean functions.

$\mu(t) = e^{-3t}\sin(3\pi t)$

	$\sigma_\varepsilon = 0.05R_x$		$\sigma_\varepsilon = 0.1R_x$		$\sigma_\varepsilon = 0.2R_x$	
	Gaussian	B-splines	Gaussian	B-splines	Gaussian	B-splines
$n_i = 100$						
AMSE $\times 10^2$	7.792	7.792	7.962	7.957	7.972	7.960
SD(MSE) $\times 10^3$	9.17	9.16	9.77	9.76	9.23	9.26
n_i : normal						
AMSE $\times 10^2$	7.682	9.197	7.802	9.230	7.987	9.312
SD(MSE) $\times 10^3$	8.43	10.04	9.54	12.15	9.59	12.62

$\mu(t) = 1 - 48t + 218t^2 - 315t^3 + 145t^4$

	$\sigma_\varepsilon = 0.05R_x$		$\sigma_\varepsilon = 0.1R_x$		$\sigma_\varepsilon = 0.2R_x$	
	Gaussian	B-splines	Gaussian	B-splines	Gaussian	B-splines
$n_i = 100$						
AMSE	1.472	1.471	1.508	1.504	1.520	1.509
SD(MSE) $\times 10$	1.72	1.72	1.92	1.92	1.81	1.82
n_i : normal						
AMSE	1.468	1.786	1.504	1.790	1.518	1.793
SD(MSE) $\times 10$	1.86	2.03	1.67	1.81	1.85	2.35

is given by $\text{AMSE} = 100^{-1}\sum_{b=1}^{100}\text{MSE}_b$.

Table1 shows the simulation results with the AMSE and standard deviation (SD) of MSE for Gaussian basis functions and B-splines. From this table, if the numbers n_i of observational points were generated from the normal distribution, all AMSE and SD(MSE) for Gaussian basis functions were smaller than the corresponding values for B-splines. Moreover, most of results for Gaussian basis functions to the unbalanced data (n_i:normal) are better than the that to the equispaced data ($n_i = 100$), while the results for B-splines were not. In consequence, regularized functional PCA via Gaussian basis functions performs well to unbalanced data in the sense of minimizing AMSE and SD(MSE) through these simulations.

5. Real data example

We apply the proposed regularized functional PCA to 3D protein structures such as that shown in Fig3. There have been many studies that have analyzed proteins using statistical methods (Wu, Hastie and Schmidler (1998), Ding and Dubchak (2001), Green

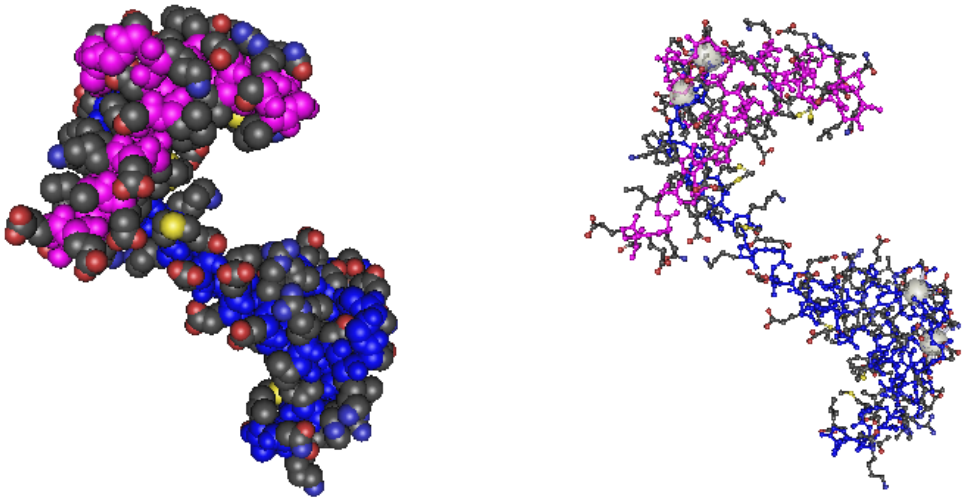

Fig. 3: Examples of 3D protein structures. The surface (left) and internal structures (right) of a protein.

Table2: The 12 proteins from the 4 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	Dhydrofolate reductase	3dfr (162) 8dfr (186)

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

and Mardia (2006), among others). Regularized functional PCA is applied here to 3-dimensional 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 4 families given in Table2. 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-coordinates values of all atoms for each protein in various coordinate systems.

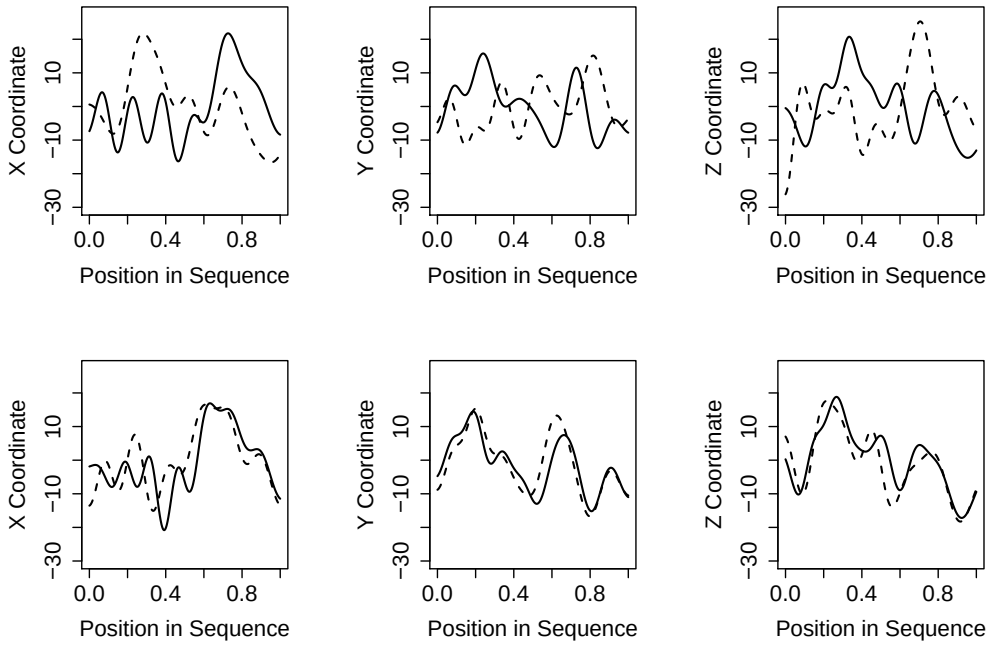

Fig. 4: An example of a rotation of proteins. The 3-dimensional functional data (upper) and rotated data (lower) of two proteins.

Firstly, the 3D protein structural data set was converted into discrete data sets using the XYZ-coordinates values 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 $3, 4, \cdots, 20$, values for ν of $1, 2, \cdots, 50$ and values for β_{il} of $10^{-10}, 10^{-9}, \cdots, 10^{-1}$ 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.

To unify the coordinates, we rotated the estimated functional data sets obtained by smoothing, since the coordinate systems differ for each protein. Optimization was performed in 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 defined by

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

We employed Euler's angle $\theta_1, \theta_2, \theta_3$ as a rotation method with step size 10 degree, and the

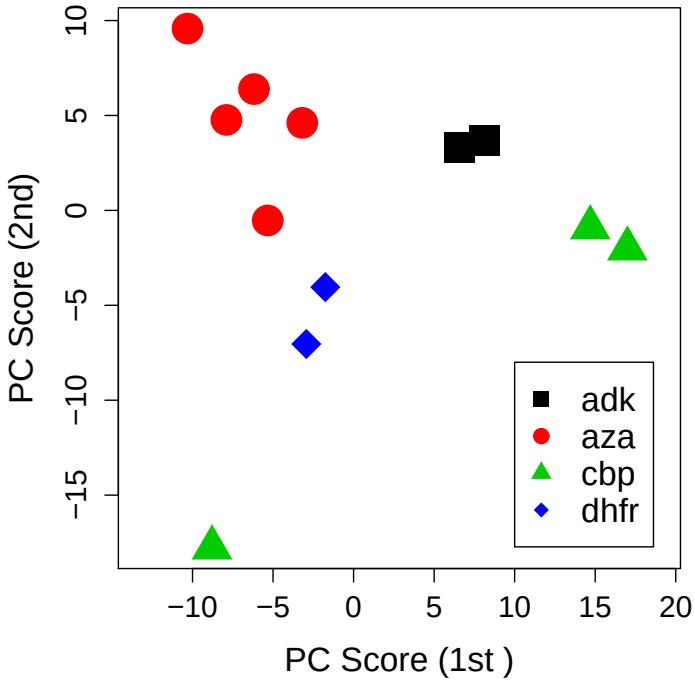

Fig. 5: The principal component (PC) scores for each family. The proteins in the adk, aza and dhfr families are clustered in their respective groups, while the cbp family has an un-clustered protein.

selected angles $\theta_1^*, \theta_2^*, \theta_3^*$ were then varied with step size 1 degree. Fig4 shows an example of the 3-dimensional 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 3-dimensional functional data sets.

Using cross validation 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 $\lambda_l (l = 1, 2, 3)$ to $\lambda_{li} = 10^{11-i}$ $(i = 1, \cdots, 10)$ and $\lambda_{li} = 10^{(i-71)/10}$ $(i = 1, \cdots, 21)$. With the selected smoothing parameters, we estimated PC curves and PC scores and plotted the PC scores for each family (Fig5). 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. 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 mean square error and its standard deviation 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 cross validation.

Acknowledgements

The authors would like to thank Professor Satoru Kuhara and Assistant Professor Hideki Hirakawa of Kyushu University for their help concerning 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 $W^* = \{W^*_{mn} = \int_{\mathcal{T}} \phi_m(t)\phi_n(t)\, dt\}_{m,n=1}^{M}$ via cubic B-splines $\{\phi_m(t)\}$ with the equispaced knots $k_1 < k_2 < \cdots < k_{M+4}$, where $\mathcal{T} = [k_4, k_{M+1}]$. We refer to De Boor (2001) and Imoto and Konishi (2003) for B-splines.

It is known that B-splines $\phi_1(t;r), \cdots, \phi_M(t;r)$ with degree $r \in \{1, 2, \cdots\}$ 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 $\phi_m(t;0) = 1\ (k_m \le t < k_{m+1}),\ = 0$ (otherwise). The cubic B-splines $\{\phi_m(t;3)\}$ are here denoted by $\{\phi_m(t)\}$. Fig6 shows an example of the cubic B-splines with $\mathcal{T} = [0, 1]$ and $M = 9$.

The diagonal components W^*_{mm} of W^* can be evaluated through the integrations $I_1^d = \int_{k_1}^{k_2} \phi_1(t)^2 dt = h/252$ and $I_2^d = \int_{k_2}^{k_3} \phi_1(t)^2 dt = 33h/140$ with the width h of the

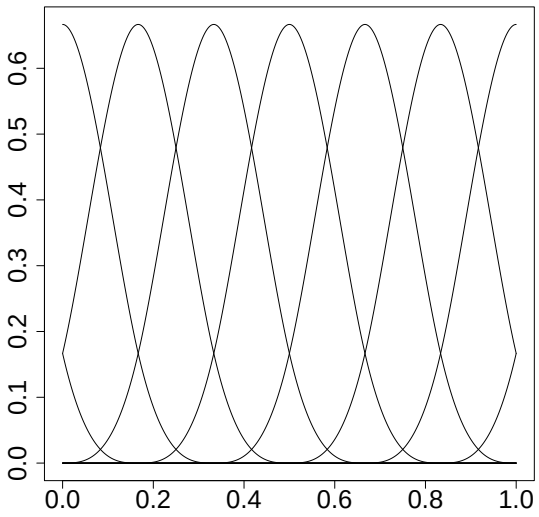

Fig. 6: Cubic B-splines ($\mathfrak{T} = [0, 1]$, $M = 9$).

equidistant knots sequence;

$$W^*_{11} = I^d_1 \,(= W^*_{MM})\,, \qquad W^*_{22} = I^d_1 + I^d_2 \,(= W^*_{M-1,M-1})\,,$$
$$W^*_{33} = I^d_1 + 2I^d_2 \,(= W^*_{M-2,M-2})\,, \quad W^*_{mm} = 2I^d_1 + 2I^d_2 \quad (m = 4, 5, \cdots, M-3)\,.$$

It may be noted that the B-splines are symmetric and $\phi_m(t) = 0$ ($t < k_m$, $k_{m+4} \leq t$). 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 4 integrations $I^{nd}_1 = \int_{k_4}^{k_5} \phi_1(t)\phi_2(t)\, dt = h/210$, $I^{nd}_2 = \int_{k_4}^{k_5} \phi_1(t)\phi_3(t)\, dt = h/84$, $I^{nd}_3 = \int_{k_4}^{k_5} \phi_1(t)\phi_4(t)\, dt = h/5040$ and $I^{nd}_4 = \int_{k_3}^{k_4} \phi_1(t)\phi_2(t)\, dt = 311h/1680$. We then have the components in the 1st to 3rd rows;

$$W^*_{12} = I^{nd}_1\,, \quad W^*_{13} = I^{nd}_2\,, \quad W^*_{14} = I^{nd}_3\,, \quad W^*_{15} = \cdots = W^*_{1M} = 0\,,$$
$$W^*_{23} = I^{nd}_1 + I^{nd}_4\,, \quad W^*_{24} = 2I^{nd}_2\,, \quad W^*_{25} = I^{nd}_3\,, \quad W^*_{26} = \cdots = W^*_{2M} = 0\,,$$
$$W^*_{34} = 2I^{nd}_1 + I^{nd}_4\,, \quad W^*_{35} = 2I^{nd}_2\,, \quad W^*_{36} = I^{nd}_3\,, \quad W^*_{37} = \cdots = W^*_{3M} = 0\,.$$

In a similar way, the components in the 4th,5th,$\cdots$ rows can be obtained. Especially, the components in the $(M-2)$-th and $(M-1)$-th rows are given by $W^*_{M-2,M-1} = I^{nd}_1 + I^{nd}_4 (= W^*_{23})$, $W^*_{M-2,M} = I^{nd}_2 (= W_{13})$ and $W_{M-1,M} = I^{nd}_1 (= W_{12})$.

References

Ando, T., Konishi, S. and Imoto, S., 2005. Nonlinear regression modeling via regularized radial basis function networks. To appear in *Journal of Statistical Planning and Inference*.

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

De Boor, C., 2001. *A Practical Guide to Splines (Revised Edition)*. Springer.

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

Eilers, P. and Marx, B., 1996. Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science* **11**, 89-121.

Ferraty, F. and Vieu, P., 2006. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.

Green, P. J. and Mardia, K. V., 2006. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* **93**(2), 235-254.

Green, P. J. and Silverman, B. W., 1994. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Imoto, S. and Konishi, S., 2003. Selection of smoothing parameters in *B*-spline nonparametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics* **55**(4), 671-687.

James, G., Hastie, T. and Sugar, C., 2000. Principal component models for sparse functional data. *Biometrika* **87**, 587-602.

Jolliffe, I. T., 2002. *Principal Component Analysis (2nd Edition)*. Springer.

Konishi, S. and Kitagawa, G., 1996. Generalized information criteria in model selection. *Biometrika* **83**(4), 875-890.

Konishi, S. and Kitagawa, G., 2008. *Information Criteria and Statistical Modeling*. Springer.

Mizuta, M., 2006. Discrete functional data analysis. *Proceedings in Computational Statistics 2006*, Physica-Verlag/Springer, 361-369.

Ramsay, J. O. and Silverman, B. W., 2002. *Applied Functional Data Analysis*. Springer.

Ramsay, J. O. and Silverman, B. W., 2005. *Functional Data Analysis (2nd Edition)*. Springer.

Rice, J. A. and Silverman, B. W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society*, Series B **53**, 233-243.

Silverman, B. W., 1996. Smoothed functional principal components analysis by choice of norm. *Annals of Statistics* **24**, 1-24.

Wu, T. D., Hastie, T. and Schmidler, S. C., 1998. Regression analysis of multiple protein structures. *Journal of Computational Biology* **5**(3), 585-596.

List of MHF Preprint Series, Kyushu University

21st Century COE Program
Development of Dynamic Mathematics with High Functionality

MHF2005-1 Hideki KOSAKI
Matrix trace inequalities related to uncertainty principle

MHF2005-2 Masahisa TABATA
Discrepancy between theory and real computation on the stability of some finite element schemes

MHF2005-3 Yuko ARAKI & Sadanori KONISHI
Functional regression modeling via regularized basis expansions and model selection

MHF2005-4 Yuko ARAKI & Sadanori KONISHI
Functional discriminant analysis via regularized basis expansions

MHF2005-5 Kenji KAJIWARA, Tetsu MASUDA, Masatoshi NOUMI, Yasuhiro OHTA & Yasuhiko YAMADA
Point configurations, Cremona transformations and the elliptic difference Painlevé equations

MHF2005-6 Kenji KAJIWARA, Tetsu MASUDA, Masatoshi NOUMI, Yasuhiro OHTA & Yasuhiko YAMADA
Construction of hypergeometric solutions to the q - Painlevé equations

MHF2005-7 Hiroki MASUDA
Simple estimators for non-linear Markovian trend from sampled data:
I. ergodic cases

MHF2005-8 Hiroki MASUDA & Nakahiro YOSHIDA
Edgeworth expansion for a class of Ornstein-Uhlenbeck-based models

MHF2005-9 Masayuki UCHIDA
Approximate martingale estimating functions under small perturbations of dynamical systems

MHF2005-10 Ryo MATSUZAKI & Masayuki UCHIDA
One-step estimators for diffusion processes with small dispersion parameters from discrete observations

MHF2005-11 Junichi MATSUKUBO, Ryo MATSUZAKI & Masayuki UCHIDA
Estimation for a discretely observed small diffusion process with a linear drift

MHF2005-12 Masayuki UCHIDA & Nakahiro YOSHIDA
AIC for ergodic diffusion processes from discrete observations

MHF2005-13 Hiromichi GOTO & Kenji KAJIWARA
Generating function related to the Okamoto polynomials for the Painlevé IV equation

MHF2005-14 Masato KIMURA & Shin-ichi NAGATA
Precise asymptotic behaviour of the first eigenvalue of Sturm-Liouville problems with large drift

MHF2005-15 Daisuke TAGAMI & Masahisa TABATA
Numerical computations of a melting glass convection in the furnace

MHF2005-16 Raimundas VIDŪNAS
Normalized Leonard pairs and Askey-Wilson relations

MHF2005-17 Raimundas VIDŪNAS
Askey-Wilson relations and Leonard pairs

MHF2005-18 Kenji KAJIWARA & Atsushi MUKAIHIRA
Soliton solutions for the non-autonomous discrete-time Toda lattice equation

MHF2005-19 Yuu HARIYA
Construction of Gibbs measures for 1-dimensional continuum fields

MHF2005-20 Yuu HARIYA
Integration by parts formulae for the Wiener measure restricted to subsets in $\mathbb{R}^d$

MHF2005-21 Yuu HARIYA
A time-change approach to Kotani's extension of Yor's formula

MHF2005-22 Tadahisa FUNAKI, Yuu HARIYA & Mark YOR
Wiener integrals for centered powers of Bessel processes, I

MHF2005-23 Masahisa TABATA & Satoshi KAIZU
Finite element schemes for two-fluids flow problems

MHF2005-24 Ken-ichi MARUNO & Yasuhiro OHTA
Determinant form of dark soliton solutions of the discrete nonlinear Schrödinger equation

MHF2005-25 Alexander V. KITAEV & Raimundas VIDŪNAS
Quadratic transformations of the sixth Painlevé equation

MHF2005-26 Toru FUJII & Sadanori KONISHI
Nonlinear regression modeling via regularized wavelets and smoothing parameter selection

MHF2005-27 Shuichi INOKUCHI, Kazumasa HONDA, Hyen Yeal LEE, Tatsuro SATO, Yoshihiro MIZOGUCHI & Yasuo KAWAHARA
On reversible cellular automata with finite cell array

MHF2005-28 Toru KOMATSU
Cyclic cubic field with explicit Artin symbols

MHF2005-29 Mitsuhiro T. NAKAO, Kouji HASHIMOTO & Kaori NAGATOU
A computational approach to constructive a priori and a posteriori error estimates for finite element approximations of bi-harmonic problems

MHF2005-30 Kaori NAGATOU, Kouji HASHIMOTO & Mitsuhiro T. NAKAO
Numerical verification of stationary solutions for Navier-Stokes problems

MHF2005-31 Hidefumi KAWASAKI
A duality theorem for a three-phase partition problem

MHF2005-32 Hidefumi KAWASAKI
A duality theorem based on triangles separating three convex sets

MHF2005-33 Takeaki FUCHIKAMI & Hidefumi KAWASAKI
An explicit formula of the Shapley value for a cooperative game induced from the conjugate point

MHF2005-34 Hideki MURAKAWA
A regularization of a reaction-diffusion system approximation to the two-phase Stefan problem

MHF2006-1 Masahisa TABATA
Numerical simulation of Rayleigh-Taylor problems by an energy-stable finite element scheme

MHF2006-2 Ken-ichi MARUNO & G R W QUISPEL
Construction of integrals of higher-order mappings

MHF2006-3 Setsuo TANIGUCHI
On the Jacobi field approach to stochastic oscillatory integrals with quadratic phase function

MHF2006-4 Kouji HASHIMOTO, Kaori NAGATOU & Mitsuhiro T. NAKAO
A computational approach to constructive a priori error estimate for finite element approximations of bi-harmonic problems in nonconvex polygonal domains

MHF2006-5 Hidefumi KAWASAKI
A duality theory based on triangular cylinders separating three convex sets in R^n

MHF2006-6 Raimundas VIDŪNAS
Uniform convergence of hypergeometric series

MHF2006-7 Yuji KODAMA & Ken-ichi MARUNO
N-Soliton solutions to the DKP equation and Weyl group actions

MHF2006-8 Toru KOMATSU
Potentially generic polynomial

MHF2006-9 Toru KOMATSU
Generic sextic polynomial related to the subfield problem of a cubic polynomial

MHF2006-10 Shu TEZUKA & Anargyros PAPAGEORGIOU
Exact cubature for a class of functions of maximum effective dimension

MHF2006-11 Shu TEZUKA
On high-discrepancy sequences

MHF2006-12 Raimundas VIDŪNAS
Detecting persistent regimes in the North Atlantic Oscillation time series

MHF2006-13 Toru KOMATSU
Tamely Eisenstein field with prime power discriminant

MHF2006-14 Nalini JOSHI, Kenji KAJIWARA & Marta MAZZOCCO
Generating function associated with the Hankel determinant formula for the solutions of the Painlevé IV equation

MHF2006-15 Raimundas VIDŪNAS
Darboux evaluations of algebraic Gauss hypergeometric functions

MHF2006-16 Masato KIMURA & Isao WAKANO
New mathematical approach to the energy release rate in crack extension

MHF2006-17 Toru KOMATSU
Arithmetic of the splitting field of Alexander polynomial

MHF2006-18 Hiroki MASUDA
Likelihood estimation of stable Lévy processes from discrete data

MHF2006-19 Hiroshi KAWABI & Michael RÖCKNER
Essential self-adjointness of Dirichlet operators on a path space with Gibbs measures via an SPDE approach

MHF2006-20 Masahisa TABATA
Energy stable finite element schemes and their applications to two-fluid flow problems

MHF2006-21 Yuzuru INAHAMA & Hiroshi KAWABI
Asymptotic expansions for the Laplace approximations for Itô functionals of Brownian rough paths

MHF2006-22 Yoshiyuki KAGEI
Resolvent estimates for the linearized compressible Navier-Stokes equation in an infinite layer

MHF2006-23 Yoshiyuki KAGEI
Asymptotic behavior of the semigroup associated with the linearized compressible Navier-Stokes equation in an infinite layer

MHF2006-24 Akihiro MIKODA, Shuichi INOKUCHI, Yoshihiro MIZOGUCHI & Mitsuhiko FUJIO
The number of orbits of box-ball systems

MHF2006-25 Toru FUJII & Sadanori KONISHI
Multi-class logistic discrimination via wavelet-based functionalization and model selection criteria

MHF2006-26 Taro HAMAMOTO, Kenji KAJIWARA & Nicholas S. WITTE
Hypergeometric solutions to the q-Painlevé equation of type $(A_1 + A_1')^{(1)}$

MHF2006-27 Hiroshi KAWABI & Tomohiro MIYOKAWA
The Littlewood-Paley-Stein inequality for diffusion processes on general metric spaces

MHF2006-28 Hiroki MASUDA
Notes on estimating inverse-Gaussian and gamma subordinators under high-frequency sampling

MHF2006-29 Setsuo TANIGUCHI
The heat semigroup and kernel associated with certain non-commutative harmonic oscillators

MHF2006-30 Setsuo TANIGUCHI
Stochastic analysis and the KdV equation

MHF2006-31 Masato KIMURA, Hideki KOMURA, Masayasu MIMURA, Hidenori MIYOSHI, Takeshi TAKAISHI & Daishin UEYAMA
Quantitative study of adaptive mesh FEM with localization index of pattern

MHF2007-1 Taro HAMAMOTO & Kenji KAJIWARA
Hypergeometric solutions to the q-Painlevé equation of type $A_4^{(1)}$

MHF2007-2 Kouji HASHIMOTO, Kenta KOBAYASHI & Mitsuhiro T. NAKAO
Verified numerical computation of solutions for the stationary Navier-Stokes equation in nonconvex polygonal domains

MHF2007-3 Kenji KAJIWARA, Marta MAZZOCCO & Yasuhiro OHTA
A remark on the Hankel determinant formula for solutions of the Toda equation

MHF2007-4 Jun-ichi SATO & Hidefumi KAWASAKI
Discrete fixed point theorems and their application to Nash equilibrium

MHF2007-5 Mitsuhiro T. NAKAO & Kouji HASHIMOTO
Constructive error estimates of finite element approximations for non-coercive elliptic problems and its applications

MHF2007-6 Kouji HASHIMOTO
A preconditioned method for saddle point problems

MHF2007-7 Christopher MALON, Seiichi UCHIDA & Masakazu SUZUKI
Mathematical symbol recognition with support vector machines

MHF2007-8 Kenta KOBAYASHI
On the global uniqueness of Stokes' wave of extreme form

MHF2007-9 Kenta KOBAYASHI
A constructive a priori error estimation for finite element discretizations in a non-convex domain using singular functions

MHF2007-10 Myoungnyoun KIM, Mitsuhiro T. NAKAO, Yoshitaka WATANABE & Takaaki NISHIDA
A numerical verification method of bifurcating solutions for 3-dimensional Rayleigh-Bénard problems

MHF2007-11 Yoshiyuki KAGEI
Large time behavior of solutions to the compressible Navier-Stokes equation in an infinite layer

MHF2007-12 Takashi YANAGAWA, Satoshi AOKI and Tetsuji OHYAMA
Human finger vein images are diverse and its patterns are useful for personal identification

MHF2007-13 Masahisa TABATA
Finite element schemes based on energy-stable approximation for two-fluid flow problems with surface tension

MHF2007-14 Mitsuhiro T. NAKAO & Takehiko KINOSHITA
Some remarks on the behaviour of the finite element solution in nonsmooth domains

MHF2007-15 Yoshiyuki KAGEI & Takumi NUKUMIZU
Asymptotic behavior of solutions to the compressible Navier-Stokes equation in a cylindrical domain

MHF2007-16 Shuichi INOKUCHI, Yoshihiro MIZOGUCHI, Hyen Yeal LEE & Yasuo KAWAHARA
Periodic Behaviors of Quantum Cellular Automata

MHF2007-17 Makoto HIROTA & Yasuhide FUKUMOTO
Energy of hydrodynamic and magnetohydrodynamic waves with point and continuous spectra

MHF2007-18 Mitsunori KAYANO & Sadanori KONISHI
Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data