Functional Cluster Analysis via Orthonormal Gaussian Basis Expansions and Its Application

Kayano, Mitsunori
Faculty of Mathematics, Kyushu University

Dozono, Koji
Faculty of Mathematics, Kyushu University

Konishi, Sadanori
Faculty of Mathematics, Kyushu University

https://hdl.handle.net/2324/8728

出版情報：2007-12-26. 九州大学大学院数理学研究院
バージョン：
権利関係：

MHF Preprint Series

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

Functional Cluster Analysis via Orthonormalized Gaussian Basis Expansions and Its Application

M. Kayano, K. Dozono
S. Konishi

MHF 2007-19

(Received December 26, 2007)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Functional Cluster Analysis via Orthonormalized Gaussian Basis Expansions and Its Application

Mitsunori Kayano[*], **Koji Dozono**[†] **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

We propose functional cluster analysis (FCA) for multidimensional functional data sets, utilizing orthonormalized Gaussian basis functions. An essential point in FCA is the use of orthonormal bases that yield the identity matrix for the integral of the product of any two bases (identity cross product matrix). We construct orthonormalized Gaussian basis functions using Cholesky decomposition and derive the property of Cholesky decomposition with respect to the Gram-Schmidt orthonormalization. The advantages of the functional clustering approach are that it can be applied to the data observed at different time points for each subject, and the functional structure behind the data can be captured by removing the measurement errors. The proposed method is applied to three-dimensional (3D) protein structural data that determine the 3D arrangement of amino acids in individual protein. In addition, numerical experiments are conducted to investigate the effectiveness of the proposed method with the orthonormalized Gaussian bases, as compared to conventional cluster analysis. The numerical results show that the proposed methodology is superior to the conventional method for noisy data sets with outliers.

KEY WORDS: Cholesky decomposition, clustering, functional data, Gram-Schmidt orthonormalization, protein structure, radial basis functions.

1. Introduction

Cluster analysis is a technique for identifying groups in data, which are often sampled as high-dimensional vectors. It can be thought of as the dual of discriminant analysis, the key distinction being that, in cluster analysis, the group labels are not known a priori. There are several clustering methods, for example, model-based and hierarchical methods

[*] Research Fellow of the Japan Society for the Promotion of Science.

[†] Present address: ONO Pharmaceutical Co., Ltd. 2-1-5 Doshu-Chou, Chuo-ku, Osaka 541-8526, Japan.

and nonhierarchical methods, such as the k-means method (Hartigan and Wong (1978)) and the Self-Organizing Map (SOM, Kohonen (1997)).

A general clustering method assumes that an observational vector can be interpreted as a discretized realization of a function evaluated at possibly different time points for each subject. However, if the number of time points is not exactly the same for each subject, conventional cluster analysis cannot be directly applied to the data. Moreover, in the presence of measurement errors, direct cluster analysis does not take advantage of the functional structure. For these reasons, the present paper considers functional cluster analysis (FCA) that converts each observational vector to a function by a smoothing method and then extracts information from the obtained functional data set by applying concepts from conventional cluster analysis. Functional cluster analysis can also use information for the derivatives of the functional data.

In modeling with functional approaches such as FCA, many studies employ basis expansions that assume that functional data may be expressed as linear combinations of known basis functions. Simple functional clustering methods are given by applying conventional cluster analysis to the coefficient vectors in the basis expansions (Abraham *et al.* (2003)). However, the distances among the coefficient vectors in non-orthonormal basis expansions differ from the distances among the functional data. Rossi *et al.* (2004) have reported a functional clustering method that reserves the distances among the functional data. First, the observations are expressed in the basis expansions, and the matrix for the integrals of the products of any two bases (cross product matrix) is evaluated. Conventional cluster analysis is then applied to the transformed coefficient vectors given by the original coefficient vectors and upper-triangular matrix (Cholesky factor). However, their study did not provide any additional mathematical motivation for the use of the transformed coefficient vectors.

Cholesky decomposition is one of the most widely used techniques of matrix decomposition. Other decompositions include LU decomposition and QR decomposition. Ciarlet (1989, §4.5) described the relationship between the QR decomposition and Gram-Schmidt orthonormalization, and Strang and Borre (1997, §11.1) showed that a Gram-Schmidt orthonormalization procedure is equivalent to the Cholesky decomposition on the special type of matrix. In the present paper, we provide the relationship between the orthonormal basis expansions and transformed coefficient vectors and derive its property concerning the Gram-Schmidt orthonormalization.

An important point in functional approaches based on the basis expansions is the evaluation of the cross product matrix. Since the orthonormal property of the Fourier series yields the identity cross product matrix, we need not evaluate the cross product matrix for Fourier series. However, it is known that Fourier series are not appropriate for non-periodic data. In contrast, spline types of bases (see, e.g., Green and Silverman (1994), de Boor (2001)) do not have the orthonormal property, and consequently the cross product matrix must be calculated. The evaluation of the cross product matrix for the splines, however, is complicated, because they are given by piecewise polynomials.

The main aim of the present paper is to introduce FCA for multidimensional functional data sets, utilizing orthonormalized Gaussian basis functions. The advantages of the use of the Gaussian type of basis functions in the functional approaches are that the cross product matrix can be easily calculated by its exponential form and that a much more flexible instrument is created for transforming each individual's observation into a function. The proposed method is applied to the three-dimensional (3D) protein structural data that determine the 3D arrangement of amino acids in individual protein. An objective of the analysis of the protein data is to characterize the features of proteins.

The present paper is organized as follows. Section 2 describes a method of multidimensional functionalization based on the basis expansions. Section 3 introduces functional clustering techniques and shows the details of the relationship between the transformed coefficient vectors and the orthonormal basis expansions, where we employ the SOM as a conventional clustering method to the transformed coefficient vectors. In Section 4, Monte Carlo simulations are conducted to investigate the effectiveness of FCA with the orthonormalized Gaussian bases, as compared to conventional cluster analysis. Section 5 describes the application of the proposed method to the 3D protein structural data. Finally, 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 $(x_{i1j}, \cdots, x_{ipj})$ is the discrete data observed at t_{ij} for p variables $X_1, \cdots, X_p$: for example, $\{t_{ij}, (x_{i1j}, x_{i2j}, x_{i3j})\ ;\ j = 1, \cdots, n_i\}$ $(i = 1, \cdots, 19)$ are the discretized 3D protein structural data, where t_{ij} are the positions in the i-th amino-acid sequence and $(x_{i1j}, x_{i2j}, x_{i3j})$ are the XYZ coordinates values of amino acids that compose the i-th 3D protein structure. The upper graphs in Figure 1 show an example of the discretized 3D protein structural data with $p = 3$ and $n_i = 149$. We convert each discrete data set $\{(t_{ij}, x_{ilj})\ ;\ j = 1, \cdots, n_i\}$ to a functional data element $x^*_{il}(t)$ by a smoothing method, as follows.

We assume that each discrete data set $\{(t_{ij}, x_{ilj})\,;\, j = 1, \cdots, n_i\}$ is generated from the nonlinear regression model:

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

where $u_{il}(t)$ are nonlinear regression functions and 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 basis functions $\phi_m(t)$:

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

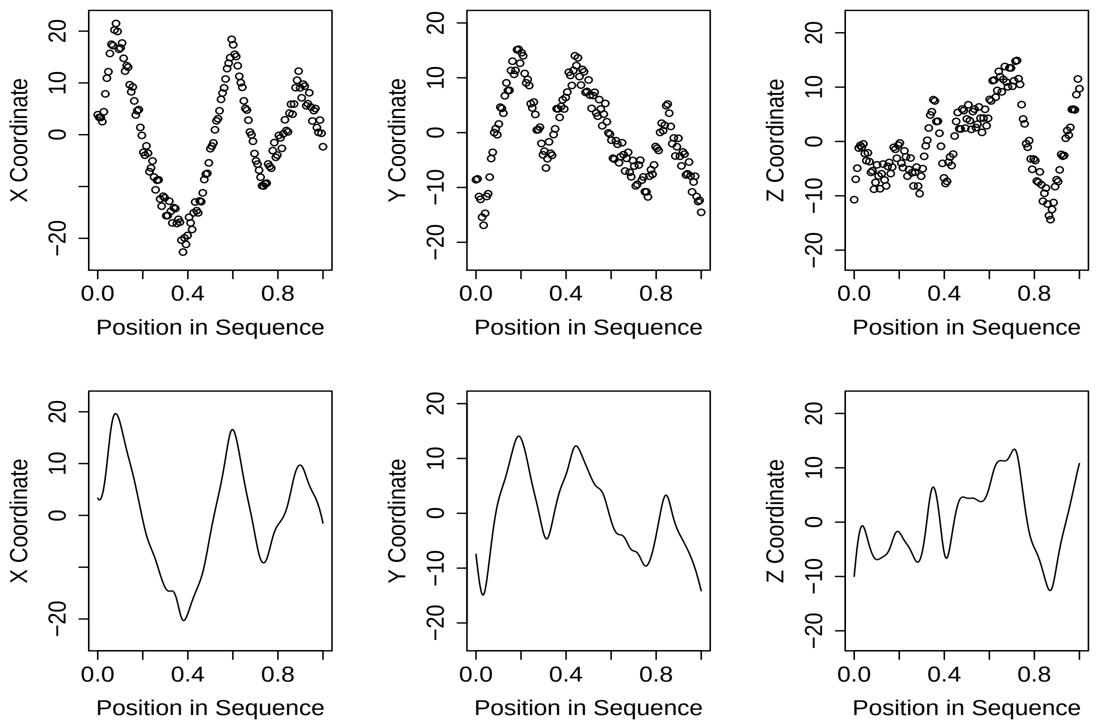

Figure 1: An example of discrete data (upper) and corresponding three-dimensional functional data (lower) for a 3D protein structure ($p = 3$, $n_i = 149$).

The m-th Gaussian basis function $\phi_m(t) = \phi_m(t;\ k_{m+2}, \tau^2)$ has the form

$$\phi_m(t) = \phi_m(t;\ k_{m+2}, \tau^2) = \exp\left\{-\frac{(t - k_{m+2})^2}{2\tau^2}\right\} \qquad (m = 1, \cdots, M)\ ,$$

where $k_1 < \cdots < k_{M+4}$ are the equispaced knots that satisfy $k_4 = \min\{t_{ij}\}$ and $k_{M+1} = \max\{t_{ij}\}$, and $\tau = (k_{m+2} - k_m)/3$ (Kawano and Konishi (2007)). We note that these basis functions $\phi_m(t) = \phi_m(t;\ k_{m+2}, \tau^2)$ have quite similar shapes, such as cubic B-splines (Eilers and Marx (1996), Imoto and Konishi (2003)), and also have exponential forms, such as Gaussian radial basis functions (Moody and Darken (1989), Bishop (1995), Ando *et al.* (2005)).

For each i and l, the coefficient parameter vector $\boldsymbol{c}_{il} = (c_{il1}, \cdots, c_{ilM})'$ and variance parameter σ_{il}^2 are estimated by maximizing the penalized log-likelihood function with a smoothing parameter $\lambda_{il} > 0$ that controls the smoothness of the nonlinear function $u_{il}(t)$. The estimators $\hat{\boldsymbol{c}}_{il} = (\hat{c}_{il1}, \cdots, \hat{c}_{ilM})'$ and $\hat{\sigma}_{il}^2$ depend on the number of basis functions M and the smoothing parameter λ_{il}. These parameters are optimally selected by minimizing the generalized information criterion (GIC) proposed by Konishi and Kitagawa (1996) (see also, Konishi and Kitagawa (2008)).

Thus, we have the estimated nonlinear functions $\hat{u}_{il}(t) = \sum_{m=1}^{M} \hat{c}_{ilm}\phi_m(t)$. The p-dimensional functional data sets $\{(x_{i1}(t), \cdots, x_{ip}(t))\ ;\ t \in \mathcal{T}\}$ $(i = 1, \cdots, N)$ are then given by $x_{il}(t) = \hat{u}_{il}(t)$ for each i and l. An example of the p-dimensional functional data is shown in Figure 1 (lower, $p = 3$), corresponding to the discretized 3D protein structural data in Figure 1 (upper). In the next section, we introduce a functional clustering method for the p-dimensional functional data sets, using orthonormalized Gaussian basis functions.

3. Functional Cluster Analysis

3.1 Functional clustering and orthonormal bases

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 $\{t_{ij}, (x_{i1j}, \cdots, x_{ipj})\ ;\ j = 1, \cdots, n_i\}$ $(i = 1, \cdots, N)$. It is assumed that each functional data element $x_{il}(t)$ can be expressed as a linear combination of the Gaussian basis functions $\{\phi_m(t) = \phi_m(t;\ k_{m+2}, \tau)\}$;

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

where $\boldsymbol{\phi}(t) = (\phi_1(t), \cdots, \phi_M(t))'$. Simple functional clustering methods are given by applying conventional cluster analysis to the estimated coefficient vectors $\hat{\boldsymbol{c}}_{il}$. However, the distances among $\hat{\boldsymbol{c}}_{il}$ in non-orthonormal basis expansions differ from the original distances among the functional data $\{(x_{i1}(t), \cdots, x_{ip}(t))\}$.

Let $\boldsymbol{x}_i(t) = (x_{i1}(t), \cdots, x_{ip}(t))'$ be p-dimensional functional data and let $\hat{\boldsymbol{c}}_i = (\hat{\boldsymbol{c}}_{i1}', \cdots, \hat{\boldsymbol{c}}_{ip}')'$ be the corresponding pM-dimensional coefficient vectors. The squared norms of $\boldsymbol{x}_i(t)$ are given by

$$\|\boldsymbol{x}_i\|_p^2 = \sum_{l=1}^{p} \|x_{il}\|^2 = \sum_{l=1}^{p} \hat{\boldsymbol{c}}_{il}' W^* \hat{\boldsymbol{c}}_{il} = \hat{\boldsymbol{c}}_i' W \hat{\boldsymbol{c}}_i \qquad (i = 1, \cdots, N)\ , \tag{1}$$

where $W^* = \int \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'dt = (W^*_{m,n})_{m,n=1}^{M}$ is the $M \times M$ cross product matrix that has the (m, n)-th component $W^*_{mn} = \int_{\mathcal{T}} \phi_m(t)\phi_n(t)\ dt$, and $W = \mathrm{diag}(W^*, \cdots, W^*)$ is the $pM \times pM$ block diagonal matrix formed from W^*. We adopt a straightforward definition of the norm of a p-dimensional function in (1). The (m, n)-th components of the cross product matrix W^* for the Gaussian basis functions $\{\phi_m(t) = \phi_m(t;\ k_{m+2}, \tau^2)\}$ are given by

$$W^*_{mn} = \sqrt{\pi\tau^2} \exp\left\{-\frac{(k_{m+2} - k_{n+2})^2}{4\tau^2}\right\} \qquad (m, n = 1, \cdots, M).$$

The squared distances between the p-dimensional functional data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are also given by the corresponding coefficient vectors $\hat{\boldsymbol{c}}_i$ and $\hat{\boldsymbol{c}}_j$:

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_p^2 = (\hat{\boldsymbol{c}}_i - \hat{\boldsymbol{c}}_j)' W (\hat{\boldsymbol{c}}_i - \hat{\boldsymbol{c}}_j) \qquad (i, j = 1, \cdots, N)\ . \tag{2}$$

If the matrix W is not the identity matrix, then the clustering methods for the coefficient vectors $\hat{\boldsymbol{c}}_i$ do not preserve the distances among the p-dimensional functional data: $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \neq \|\hat{\boldsymbol{c}}_i - \hat{\boldsymbol{c}}_j\|^2$. Note that orthonormal bases yield the identity cross product matrix, and the clustering methods based on orthonormal bases preserve the distance among the functional data. Rossi *et al.* (2004) have then introduced the functional clustering method that implements the Self-Organizing Map (SOM) on transformed coefficient vectors defined later, although, in a previous study, the k-means method is applied to the estimated coefficient vectors (Abraham *et al.* (2003)). The clustering methods for the transformed coefficient vectors preserve the distances among the functional data. We introduce the following transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$ for the p-dimensional functional data $\boldsymbol{x}_i$, while the previous study by Abraham *et al.* (2003) treated the one-dimensional case.

Let U^* be the $M \times M$ upper triangular matrix given by the Cholesky decomposition of the cross product matrix W^*: $W^* = (U^*)'U^*$, and let $U = \mathrm{diag}(U^*, \cdots, U^*)$ be the $pM \times pM$ block diagonal upper triangular matrix formed from U^*. We then have $W = U'U$. The squared distances (2) can be written as

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_p^2 = (\tilde{\boldsymbol{c}}_i - \tilde{\boldsymbol{c}}_j)'(\tilde{\boldsymbol{c}}_i - \tilde{\boldsymbol{c}}_j) = \|\tilde{\boldsymbol{c}}_i - \tilde{\boldsymbol{c}}_j\|^2 , \qquad (3)$$

where $\tilde{\boldsymbol{c}}_i = (\tilde{\boldsymbol{c}}_{i1}', \cdots, \tilde{\boldsymbol{c}}_{ip}')' = U\hat{\boldsymbol{c}}_i$ are the pM-dimensional transformed coefficient vectors and each element $\tilde{\boldsymbol{c}}_{il}$ of $\tilde{\boldsymbol{c}}_i$ is also given by $\tilde{\boldsymbol{c}}_{il} = U^*\hat{\boldsymbol{c}}_{il}$. Thus, the functional clustering methods based on the transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$ preserve the distances among the p-dimensional functional data $\boldsymbol{x}_i$.

The use of the transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$ corresponds to orthonormal basis expansions of functional data. Let $\psi_1(t), \cdots, \psi_M(t)$ be the basis functions defined by $\boldsymbol{\psi}(t) = (\psi_1(t), \cdots, \psi_M(t))' = U_*^{-1}\boldsymbol{\phi}(t)$, where $U_* = (U^*)'$ is the $M \times M$ lower triangular matrix. The basis functions $\psi_m(t)$ are the orthonormal bases formed by the original bases $\phi_m(t)$: the cross product matrix of $\boldsymbol{\psi}(t)$ is given by $\int \boldsymbol{\psi}(t)\boldsymbol{\psi}(t)'dt = U_*^{-1}W^*(U^*)^{-1} = I_M$, with the identity matrix I_M of size M. We also have

$$\tilde{\boldsymbol{c}}_{il}\boldsymbol{\psi}(t) = \hat{\boldsymbol{c}}_{il}'\boldsymbol{\phi}(t) = x_{il}(t) \qquad (l = 1, \cdots, p,\ i = 1, \cdots, N) .$$

Thus, each element $\tilde{\boldsymbol{c}}_{il}$ of $\tilde{\boldsymbol{c}}_i$ is the coefficient vector in the orthonormal basis expansion of the functional data element $x_{il}(t)$. Note that the use of the transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$ yields the identity cross product matrix in (3) (see also, (2)). Furthermore, we derive the remarkable property of the transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$: these coefficient vectors are equivalent to those of the orthonormal bases given by the Gram-Schmidt orthonormalization of $\phi_1(t), \cdots, \phi_M(t)$. The derivation is shown in the Appendix.

3.2 Self-Organizing Map

The multidimensional functional clustering method applies conventional cluster analysis to the transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$ in (3). As for a clustering method, we employ the

Self-Organizing Map (SOM), which is an unsupervised neural network, and a method of visualizing complex high-dimensional data by drawing a low-dimensional map (see, e.g., Kohonen (1997)).

Let us consider the clustering based on the transformed coefficient vectors $\{\tilde{\boldsymbol{c}}_i \in \mathbb{R}^{pM}\ ;\ i = 1, \cdots, N\}$ into K clusters. The SOM here defines a mapping from the input data space $\mathbb{R}^{pM}$ onto a regular two-dimensional array of nodes. With every node $k \in \{1, \cdots, K\}$, a reference vector $\boldsymbol{p}_k \in \mathbb{R}^{pM}$ is prepared. A coefficient vector $\boldsymbol{c}$ is compared with $\boldsymbol{p}_k$, and the best-matching node k_0 is defined by

$$k_0 = \underset{k}{\operatorname{argmin}} \{\|\boldsymbol{c} - \boldsymbol{p}_k\|\} .$$

The SOM employs useful vectors of $\boldsymbol{p}_k$ as the reference vectors that can be found as convergence limits of the following learning process.

If we have the t-th updated values of $\boldsymbol{p}_k$, the $(t + 1)$-th updated values $\boldsymbol{p}_k^{(t+1)}$ are obtained by

$$\boldsymbol{p}_k^{(t+1)} = \boldsymbol{p}_k^{(t)} + h_{k_0,k}(t) \{\boldsymbol{c}^{(t)} - \boldsymbol{p}_k^{(t)}\} \qquad (k = 1, \cdots, K) , \tag{4}$$

where $h_{k_0,k}(t) = h(\|\boldsymbol{r}_{k_0} - \boldsymbol{r}_k\|, t)$ is the neighborhood kernel with the two-dimensional radius vectors $\boldsymbol{r}_{k_0}$ and $\boldsymbol{r}_k$ of nodes k_0 and k in the array. Each radius vector $\boldsymbol{r}_k$ represents the position of the node k in the two-dimensional plane. The Gaussian type neighborhood kernel is defined by

$$h_{k_0,k}(t) = \alpha(t) \cdot \exp\left\{-\frac{\|\boldsymbol{r}_{k_0} - \boldsymbol{r}_k\|^2}{2\sigma^2(t)}\right\},$$

where $\alpha(t)$ is a scalar-valued learning rate and $\sigma^2(t)$ determines the width of the kernel. Both $\alpha(t)$ and $\sigma^2(t)$ are some monotonically decreasing functions of iteration, and their exact forms are not critical. They could thus be selected to be linear. An algorithm of the SOM is detailed in the following procedure.

1) Set the initial values of reference vectors $\{\boldsymbol{p}_k \in \mathbb{R}^{pM}\ ;\ k = 1, \cdots, K\}$.
2) Find the best-matching node k_0 for the fixed transformed coefficient vector $\tilde{\boldsymbol{c}}_i$.
3) Update the reference vectors $\boldsymbol{p}_k$ by (4).
4) Repeat 2) and 3) for $i = 1, \cdots, N$.
5) Repeat 2) to 4) until convergence.

Resulting clusters C_k $(k = 1, \cdots, K)$ are given by $C_k = \{\boldsymbol{c}_i \in \mathbb{R}^{pM}\ ;\ \operatorname{argmin}_{k'} \|\boldsymbol{c}_i - \boldsymbol{p}_{k'}\| = k\}$, where $\boldsymbol{p}_k$ are the convergence limits of the above procedure.

4. Numerical Experiments

Monte Carlo simulations were conducted to investigate the effectiveness of FCA with the orthonormalized Gaussian bases, as compared to conventional cluster analysis. In the simulation study, we generated a true functional data set $\{x_i^*(t)\,;\, t \in [0, 1],\ i = 1, \cdots, 50\}$, and a new functional data set $\{x_i(t)\ ;\ t \in [0, 1],\ i = 1, \cdots, 50\}$ was constructed by

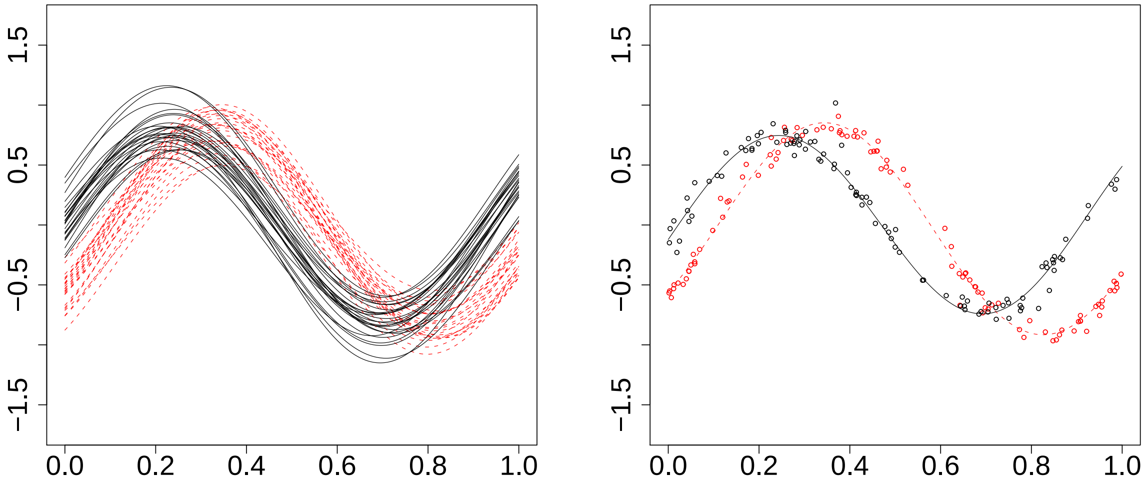

Figure 2: Examples of the generated functional data set (left) and discrete and functional data (right) from 2 clusters. The black (solid) and red (dashed) lines correspond to clusters 1 and 2, respectively.

smoothing the discrete data set $\{x_{ij}\ ;\ i=1,\cdots,50,\ j=1,\cdots,100\}$ that was generated from $x_i^*(t)$. We then performed conventional and functional clustering methods to the discrete data set $\{x_{ij}\}$ and functional data set $\{x_i(t)\}$, respectively, and compared the clustering results. More precisely, we performed clustering methods on the simulated data using the following procedure.

Step 1. Generate a true functional data set $\{x_i^*(t)\ ;\ t\in[0,1],\ i=1,\cdots,50\}$ from the mixed effects models

$$x_i^*(t)=\begin{cases}\mu_1(t)+\displaystyle\sum_{m=1}^{6}\gamma_{1im}\phi_m(t) & (i=1,\cdots,25,\ \ \text{if } x_i^*(t)\in\text{Cluster 1})\\ \mu_2(t)+\displaystyle\sum_{m=1}^{6}\gamma_{2im}\phi_m(t) & (i=26,\cdots,50,\ \ \text{if } x_i^*(t)\in\text{Cluster 2})\end{cases},$$

where

$$\mu_1(t)=0.8\sin(20t/3)\ ,\qquad \mu_2(t)=0.8\sin\{20(t-0.1)/3\}\ ,$$

and $\phi_1(t),\cdots,\phi_6(t)$ indicate the Gaussian basis functions. The random components γ_{1im} and γ_{2im} are assumed to be independently and normally distributed with $\gamma_{1im},\ \gamma_{2im}\overset{iid}{\sim}N(0,\sigma_\gamma^2)$, where σ_γ is set to 0.02 or 0.05.

Step 2. Generate the discrete data set $\{x_{ij}\ ;\ i=1,\cdots,50,\ j=1,\cdots,100\}$ from the nonlinear regression model with the true functions $x_i^*(t)$:

$$x_{ij}(t)=x_i^*(t_{ij})+\varepsilon_{ij}\qquad(i=1,\cdots,50,\ j=1,\cdots,100)\ ,$$

where t_{ij} are the equispaced observational points or points generated from the uniform distribution on $[0, 1]$, and the errors ε_{ij} are assumed to be independently distributed according to a mixture of two normal distributions

$$\varepsilon_{ij} \overset{iid}{\sim} 0.9\, N(0, (\sigma_{\varepsilon_1} R_x)^2) + 0.1\, N(0, (\sigma_{\varepsilon_2} R_x)^2)$$

with R_x being the range of $\{x_i^*(t)\}$ over $t \in [0, 1]$, $\sigma_{\varepsilon_1} = 0.05, 0.1$ and $\sigma_{\varepsilon_2} = 0.2, 0.3$

Step 3. Estimate a functional data set $\{x_i(t)\,;\, t \in [0, 1],\ i = 1, \cdots, 50\}$ by smoothing the generated discrete data set $\{x_{ij}\ ;\ i = 1, \cdots, 50,\ j = 1, \cdots, 100\}$, and a model selection is performed by minimizing GIC. It is assumed that each functional data $x_i(t)$ can be expressed as a linear combination of the Gaussian basis functions.

Step 4. Perform conventional and functional clustering methods on the generated discrete data set $\{x_{ij}\,;\, i = 1, \cdots, 50,\ j = 1, \cdots, 100\}$ and estimated functional data set $\{x_i(t)\,;\, t \in [0, 1],\ i = 1, \cdots, 50\}$, respectively. As for a clustering method, we employed the one-dimensional SOM with the number of clusters being $K = 2$.

Step 5. Compare the clustering results given by **Step 4**, and calculate the misclustering rates for the b-th trial: r_b^{c}, r_b^{f} = (number of misclusterings)/50.

Step 6. Repeat **Steps 2** through **5** for each trial. The means of the misclustering rates are then given by $\bar{r}^{\,\mathrm{c}} = 100^{-1} \sum_{b=1}^{100} r_b^{\mathrm{c}}$ and $\bar{r}^{\,\mathrm{f}} = \sum_{b=1}^{100} r_b^{\mathrm{f}}$, respectively. The means that $\bar{r}^{\,\mathrm{c}}$ and $\bar{r}^{\,\mathrm{f}}$ are the average misclustering rates.

Table 1: Clustering results (average misclustering rates $\bar{r}^{\,\mathrm{c}}$ and $\bar{r}^{\,\mathrm{f}}$).

$\sigma_\gamma = 0.05$	$\sigma_{\varepsilon_1} = 0.05$			
	$\sigma_{\varepsilon_2} = 0.2$		$\sigma_{\varepsilon_2} = 0.3$	
	Conventional	Functional	Conventional	Functional
t_{ij}:equispaced				
Misclustering rates (%)	44.8	9.2	25.6	16.6
	0.0	9.2	1.0	7.1
	0.7	6.1	44.2	17.2
Mean (%)	15.1	8.2	23.9	13.6
t_{ij}:uniform				
Misclustering rates (%)	21.8	10.6	18.7	18.4
	13.2	14.6	11.3	16.8
	9.2	9.2	36.7	24.2
Mean (%)	14.8	11.4	22.2	19.8

The "Mean"s are the mean values of the average misclustering rates.

Figure 2 shows examples of generated discrete and functional data. Steps 1 through 6 were repeated three times for each setting, and we then had the clustering results shown in

Table 2: Clustering results (average misclustering rates $\bar{r}^{\,\mathrm{c}}$ and $\bar{r}^{\,\mathrm{f}}$).

$\sigma_\gamma = 0.05$	$\sigma_{\mathcal{E}_1} = 0.1$			
	$\sigma_{\mathcal{E}_2} = 0.2$		$\sigma_{\mathcal{E}_2} = 0.3$	
	Conventional	Functional	Conventional	Functional
t_{ij}:equispaced				
Misclustering rates (%)	0.6	12.6	0.8	16.0
	0.3	19.8	15.2	19.3
	45.8	13.8	16.6	29.0
Mean (%)	15.6	15.4	10.9	21.4
t_{ij}:uniform				
Misclustering rates (%)	11.8	19.0	38.2	30.4
	4.8	20.2	15.1	27.7
	29.8	20.7	12.9	28.8
Mean (%)	15.5	20.0	22.1	29.0

The "Mean"s are the mean values of the average misclustering rates.

Tables 1 and 2, which represent the average misclustering rates $\bar{r}^{\,\mathrm{c}}$, $\bar{r}^{\,\mathrm{f}}$ and its mean value for each setting with $\sigma_\gamma = 0.05$. The conventional clustering method performed well for the "little individual variation data" ($\sigma_\gamma = 0.02$), while this method was not appropriate for the "large individual variation data" ($\sigma_\gamma = 0.05$, Tables 1 and 2).

In contrast, the functional clustering method performed well for data with a large amount of individual variation data, although this method yielded poor results to the little individual variation data. In particular, if the standard deviations of the error distribution was set to $\sigma_{\mathcal{E}_1} = 0.05$ and $\sigma_{\mathcal{E}_2} = 0.2,\ 0.3$ (Table 1), the average misclustering rates of the functional method were smaller than the corresponding values of the conventional method. The settings $\sigma_{\mathcal{E}_1} = 0.05$ and $\sigma_{\mathcal{E}_2} = 0.2,\ 0.3$ mean that most errors were generated with a small variance. However, a few errors were not. In other words, these settings produce outliers. Therefore, the functional clustering approach is better than the conventional method for data with large individual variation with outliers, through these numerical comparisons. Note that the conventional method cannot be directly applied to data with different n_i for each individual.

5. Real data example

The multidimensional FCA with the orthonormalized Gaussian basis functions is applied to the 3D protein structural data such as that shown in Figure 3. A number of studies have analyzed proteins using statistical methods (Wu *et al.* (1998), Ding and Dubchak (2001), Nguyen and Rajapakse (2003), Green and Mardia (2006), among others).The one-dimensional SOM with the number of clusters being $K = 2$ is applied here to three-

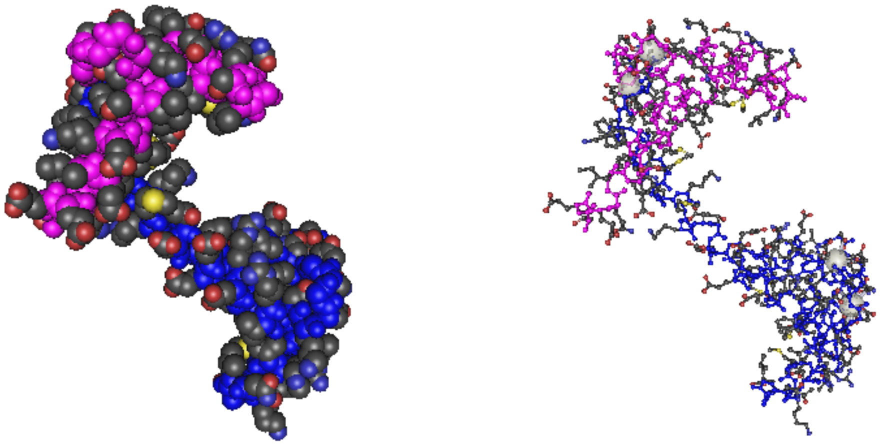

Figure 3: Examples of 3D protein structures. Surface (left) and internal structures (right) of a protein.

Table 3: The 19 proteins from the two classes.

Class	Fold	Protein code (length of the amino-acid sequence)

All-α	Globin-like	2lhb(149), 3sdh-a(145), 1flp(142), 2hbg(147), 2mge(154)
		1eca(136), 2gdm(153), 1bab-b(146), 1ith-a(141), 1ash(147)
		1hlb(157), 1cpc-a(162) [1cpc-a1(127)]

α/β	Flavodoxin-like	3chy(128), 1ntr(124), 1scu-a2(166), 2fcr(173), 2fx2(147)
		1bmt-a2(154), 1gdh-a1(130)

dimensional functional data sets representing 3D protein structures, in order to identify features of the protein structures.

A protein is a class of biomolecules composed of amino-acid sequences and has been hierarchically classified from a biological viewpoint. A set of classified proteins is referred to as a "class" determined by their secondary structures. The present paper treats 19 proteins from the two classes given in Table 3. We selected a protein fold for each class. A protein fold is a lower-level classified protein set than the protein class. The proteins listed in this table were selected from the protein set of Ding and Dubchak (2001). This data set was obtained from the National Center of Biotechnology Information (NCBI, *http://www.ncbi.nlm.nih.gov/*). Note that because the length of amino-acid sequence differs for each protein, conventional cluster analysis cannot be directly applied to the data set. In what follows, it is assumed that we have the XYZ-coordinates of all atoms for each protein in various coordinate systems.

First, the 3D structural data set was converted into discrete data sets using the XYZ-coordinates of the α-carbon atoms, which were typical atoms of amino acids. Each α-carbon atom corresponds to an amino acid. We then have a discrete data set for each coordinate. The smoothing method via the Gaussian basis functions was performed for each discrete data set. We considered values for the number of basis functions M of $25, 26, \cdots, 35$, values for the smoothing parameter λ_{il} of $10^{-10}, 10^{-9}, \cdots, 10^{-1}$, and found optimal values of $M = 35$ and $\lambda_{il} = 10^{-5}$, 10^{-4}. The selected values of M were the modes for all individuals (proteins) and coordinates, although we firstly obtained optimal values of M and λ_{il} for the i-th individual and the l-th ($l = 1$:X, $l = 2$:Y, $l = 3$:Z) coordinate, respectively. To unify the coordinates, we rotated the three-dimensional functional data sets obtained by smoothing, because the coordinate systems differ for each protein. Optimization was performed in rotating each protein to another base protein. Details of the rotation are described by Kayano and Konishi (2007, §5). We then applied the one-dimensional SOM with the number of clusters $K = 2$ to the rotated three-dimensional functional data sets.

Figure 4 shows the classified functional data sets colored by the results of the clustering. In the upper graphs in this figure, the black and green lines represent correctly classified functional data sets for All-α and α/β, respectively. The red line represents the misclassification functional data for the protein 1cpc-a, which is the chain A of the protein 1cpc. The chain A of the protein 1cpc is divided by chain A-1 and other chains. We then applied the SOM to the data set replaced 1cpc-a by 1cpc-a1. The clustered functional data sets is shown in the lower graphs in Figure 4. The protein 1cpc is correctly classified in All-α. Thus, the 3D protein structures could be effectively classified by the proposed functional clustering method.

6. Summary and concluding remarks

We introduced functional cluster analysis (FCA) for multidimensional functional data sets, using orthonormalized Gaussian basis functions. We proved the remarkable property of the transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$ determined by Cholesky decomposition. These coefficient vectors were equivalent to those of orthonormal bases given by the Gram-Schmidt orthonormalization. Numerical experiments were conducted to investigate the effectiveness of FCA with the orthonormalized Gaussian bases, as compared to conventional cluster analysis. The numerical results showed that the proposed method is superior to the conventional method for large individual variation data with outliers.

The proposed method was applied to the 3D protein structural data. We here applied the one-dimensional SOM with the fixed number of clusters $K = 2$ to three-dimensional functional data sets representing 3D protein structures. This paper treated 19 proteins from the two classes, namely, All-α and α/β, and we could effectively classify the 3D protein structures using the proposed functional clustering method. Future research will include 1) the proposal of a model-based FCA and 2) the derivation of model selection criteria from an information-theoretic perspective and also the application of Bayesian approaches.

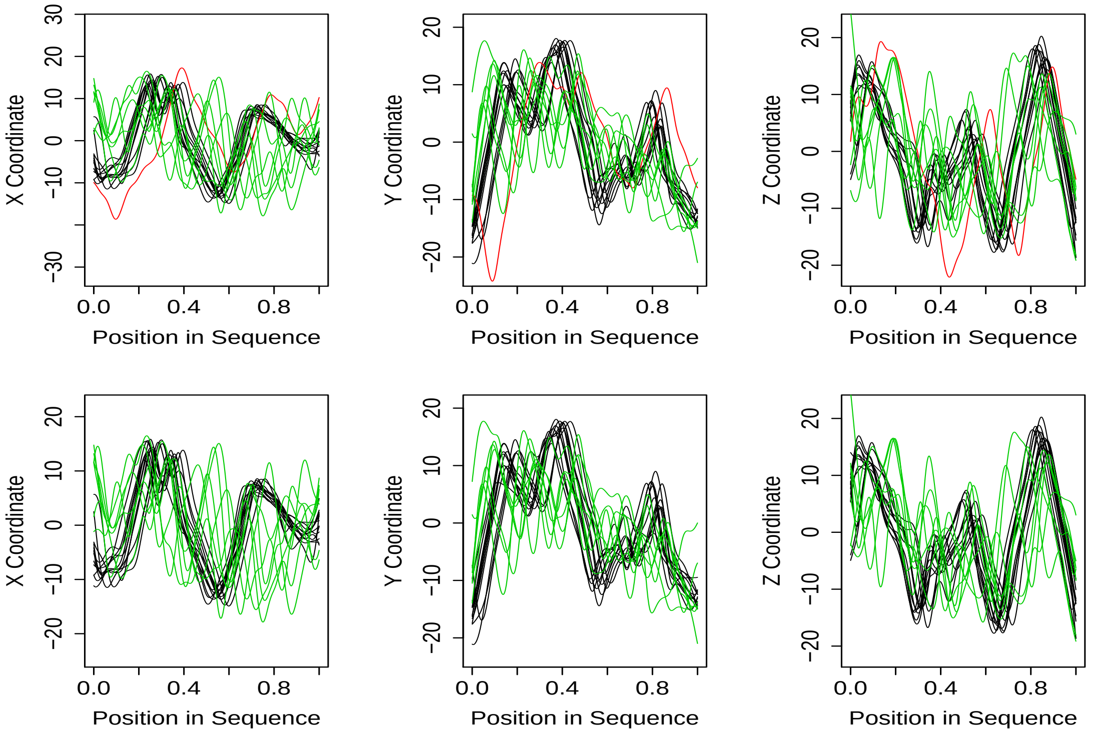

Figure 4: Classified functional data sets. The lines are colored by the results of the clustering. Upper: original data with 1cpc-a, Lower: replaced 1cpc-a by 1cpc-a1.

Appendix. Property of orthonormal bases

This section shows the remarkable property of the transformed coefficient vectors $\tilde{\boldsymbol{c}}_i$ in Section 3.1. These coefficient vectors are equivalent to those of orthonormal bases given by the Gram-Schmidt orthonormalization. Let $x(t)$ $(t \in \mathcal{T})$ be a one-dimensional functional data that may be expressed as a linear combination of any basis functions $\{\phi_m(t)\}$:

$$x(t) = \sum_{m=1}^{M} c_m \phi_m(t) = \boldsymbol{c}'\boldsymbol{\phi}(t) ,$$

with $\boldsymbol{c} = (c_1, \cdots, c_M)'$ and $\boldsymbol{\phi}(t) = (\phi_1(t), \cdots, \phi_M(t))'$.

Let us consider the construction of orthonormal bases $\psi_1(t), \cdots, \psi_M(t)$ using the Gram-Schmidt orthonormalization of $\{\phi_m(t)\}$:

$$\psi_1(t) = \frac{\phi_1(t)}{\|\phi_1\|} , \qquad \psi_m(t) = \frac{\psi_m^*(t)}{\|\psi_m^*\|} \quad (m = 2, \cdots, M) , \tag{5}$$

where $\psi_m^*(t) = \phi_m(t) - \sum_{n=1}^{m-1} \langle \psi_n, \phi_m \rangle \psi_n(t)$ and $\|\psi_m^*\|^2 = \|\phi_m\|^2 - \sum_{n=1}^{m-1} \langle \psi_n, \phi_m \rangle^2$. The functional data $x(t)$ can also be expressed as a linear combination of the orthonormal bases $\psi_m(t)$:

$$x(t) = \sum_{m=1}^{M} \tilde{c}_m \psi_m(t) = \tilde{\boldsymbol{c}}'\boldsymbol{\psi}(t) \ , \tag{6}$$

with $\tilde{\boldsymbol{c}} = (\tilde{c}_1, \cdots, \tilde{c}_M)'$ and $\boldsymbol{\psi}(t) = (\psi_1(t), \cdots, \psi_M(t))'$, since each basis function $\phi_m(t)$ is obtained by the linear combination of the orthonormal bases $\psi_m(t)$. The coefficient vector $\tilde{\boldsymbol{c}} = (\tilde{c}_1, \cdots, \tilde{c}_M)'$ in the orthonormal basis expansion (6) can be written as

$$\begin{aligned}
\tilde{\boldsymbol{c}} &= (\langle \psi_1, \tilde{\boldsymbol{c}}'\boldsymbol{\psi} \rangle, \cdots, \langle \psi_M, \tilde{\boldsymbol{c}}'\boldsymbol{\psi} \rangle)' \\
&= (\langle \psi_1, \boldsymbol{c}'\boldsymbol{\phi} \rangle, \cdots, \langle \psi_M, \boldsymbol{c}'\boldsymbol{\phi} \rangle)' \\
&= \begin{pmatrix} \langle \psi_1, \phi_1 \rangle & \langle \psi_1, \phi_2 \rangle & \cdots & \langle \psi_1, \phi_M \rangle \\ 0 & \langle \psi_2, \phi_2 \rangle & \cdots & \langle \psi_2, \phi_M \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \psi_M, \phi_M \rangle \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{pmatrix} \\
&= U^* \boldsymbol{c}
\end{aligned}$$

with $U^* = (U^*_{mn} = \langle \psi_m, \phi_n \rangle)_{m,n}$, using $x(t) = \boldsymbol{c}'\boldsymbol{\phi}(t) = \tilde{\boldsymbol{c}}'\boldsymbol{\psi}(t)$, $\langle \psi_m, \phi_n \rangle = 0 \ (m > n)$ and the orthonormalities of $\psi_m(t)$. The upper-triangular matrix U^* is equal to the matrix given by Cholesky decomposition of the cross product matrix $W^* = \int_{\mathcal{T}} \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'dt$, as follows.

Let $V = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M)' = (v_{mn})_{m,n}$ be the $M \times M$ upper-triangular matrix given by Cholesky decomposition of W^*. It follows that $V'V = W^*$. From an algorithm of Cholesky decomposition, the (m, n)-th components of V are obtained by

$$\begin{aligned}
v_{1n} &= \frac{\langle \phi_1, \phi_n \rangle}{\sqrt{\langle \phi_1, \phi_1 \rangle}} && (n = 1, \cdots, M) \ , \\
v_{mn} &= \frac{\langle \phi_m, \phi_n \rangle - \sum_{l=1}^{m-1} v_{lm} v_{ln}}{\sqrt{\langle \phi_m, \phi_m \rangle - \sum_{l=1}^{m-1} v_{lm}^2}} && (m, n = 2, \cdots, M, \ m \leq n) \ ,
\end{aligned}$$

where $\langle \phi_m, \phi_m \rangle = \|\phi_m\|^2$ are the (m, m)-th components of the cross product matrix W^*. Using these equations, the $(1, n)$-th components of V can be written as $v_{1n} = \langle \phi_1 / \|\phi_1\|, \phi_n \rangle$ $= \langle \psi_1, \phi_n \rangle$. If we know that the first $m - 1$ row vectors $\boldsymbol{v}'_l \ (l = 1, \cdots, m - 1)$ are given by

$$\begin{array}{cccccccc} & 1 & & l-1 & l & & M & \\ \boldsymbol{v}'_l = & \Big(0 & \cdots & 0 & \langle \psi_l, \phi_l \rangle & \cdots & \langle \psi_l, \phi_M \rangle \Big) \end{array} \tag{7}$$

or $v_{ln} = \langle \psi_l, \phi_n \rangle \ (l \leq n)$, then the m-th row components of V are given by the following equations $(m \leq n)$, using $\psi^*_m(t) = \phi_m(t) - \sum_{n=1}^{m-1} \langle \psi_n, \phi_m \rangle \psi_n(t)$ in (5) and their norms;

$$v_{mn} = \frac{\langle \phi_n, \phi_m \rangle - \sum_{l=1}^{m-1} \langle \psi_l, \phi_m \rangle \langle \psi_l, \phi_n \rangle}{\sqrt{\langle \phi_m, \phi_m \rangle - \sum_{l=1}^{m-1} \langle \psi_l, \phi_m \rangle^2}} = \left\langle \frac{\phi_m - \sum_{l=1}^{m-1} \langle \psi_l, \phi_m \rangle \psi_l}{\|\psi^*_m\|}, \ \phi_n \right\rangle = \langle \psi_m, \phi_n \rangle \ .$$

We then have (7) for $l = m$, that is, for all $l = 1, \cdots, M$. Thus, we have $V = U^*$.

References

Abraham, C., Cornillon, P. A., Matzner-Lober, E. and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, **30**(3), 581-595.

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.*

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford University Press.

Ciarlet, P. G. (1989). *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge University Press.

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.

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.

Hartigan, J. A. and Wong, M. A. (1978). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, **28**, 100-108.

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.

Kawano, S. and Konishi, S. (2007). Nonlinear regression modeling via radial basis functions. *Bulletin of Informatics and Cybernetics*, in Press.

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

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

Kohonen, T. (1997). *Self-Organizing Maps.* Springer.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation.* **1**, 281-294.

Nguyen, M. N. and Rajapakse, J. C. (2003). Multi-class support vector machines for protein secondary structure prediction. *Genome Informatics*, **14**, 218-227.

Rossi, F, Conan-Guez, B. and Golli, A. E. (2004). Clustering functional data with the SOM algorithm. *Proceedings of European Symposium on Artificial Neural Networks Bruges*, Belgium.305-312.

Strang, G. and Borre, K. (1997). *Linear Algebra, Geodesy, and GPS*, Wellesley-Cambridge Press.

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

MHF2007-19 Mitsunori KAYANO, Koji DOZONO & Sadanori KONISHI
Functional Cluster Analysis via Orthonormalized Gaussian Basis Expansions
and Its Application