Sparse functional principal component analysis via regularized basis expansions and its application

Kayano, Mitsunori
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Faculty of Mathematics, Kyushu University

出版情報：MI Preprint Series. 2009-19, 2010-07-15. Taylor & Francis
バージョン：
権利関係：

MI Preprint Series
Kyushu University
The Grobal COE Program
Math-for-Industry Education & Research Hub

Sparse functional principal component analysis via regularized basis expansions and its application

M. Kayano & S. Konishi

MI 2009-19

(Received May 19, 2009)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Sparse functional principal component analysis via regularized basis expansions and its application

Mitsunori Kayano* **and Sadanori Konishi**

Graduate School of Mathematics, Kyushu University

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

kayano@kuicr.kyoto-u.ac.jp (M. Kayano)

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

SUMMARY

This paper introduces principal component analysis for multidimensional sparse functional data sets, utilizing Gaussian basis functions. Our multidimensional model is estimated by maximizing a penalized log-likelihood function, while previous mixed-type models were estimated by maximum likelihood methods for one-dimensional sparse functional data set. The penalized estimation performs well for our multidimensional model, while maximum likelihood methods yield unstable parameter estimates and some of the parameter estimates are often infinite. Numerical experiments are conducted to investigate the effectiveness of our method via the Gaussian bases for some types of missing data. The proposed method is applied to handwriting data, which consist of the XY coordinates values in handwritings.

KEY WORDS: EM algorithm, handwriting, multivariate functional data, missing data, mixed model.

1 Introduction

A general multivariate analysis such as principal component analysis (PCA) is obtained by assuming that an observational vector (discrete data) can be interpreted as a discretized realization of a function evaluated at possibly different time points for each subject. However, there are some problems with applying conventional multivariate analysis to the longitudinal type of data: if the number of the time points are not exactly the same for each subject, the conventional multivariate techniques cannot be directly applied to the data. Moreover, in the presence of measurement errors, the multivariate techniques do not take advantage of the functional structure. Accordingly, a number of recent papers have investigated functional data analysis (FDA) that reformulate the methods of multivariate analysis in terms of the functions rather than the discrete observations (Ramsay and Silverman (2002, 2005), Ferraty and Vieu (2006)).

On the other hand, we often have only a few and much irregularly spaced observations for each individual. Those types of data have been called sparse data and treated by mixed-type

*Present address: Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.

models (Laird and Ware (1982), Shi *et al.* (1996), Rice and Wu (2001), James *et al.* (2000)). Shi *et al.* (1996) and Rice and Wu (2001) proposed the use of the mixed effects model to solve functional principal component problems. These models provide some advantages that each functional data $x_i(t)$ can be estimated by all observations not only i-th individual, and the coefficient vector in fixed effect and the variance-covariance matrix of random vectors are estimated by the maximum likelihood method. However, many local maximum points exist in these models, and parameters cannot be uniquely estimated, since the number of parameters is more than the number of observations. James *et al.* (2000) has then proposed the reduced rank model that avoids those ill-posed problems from the mixed effects model.

In this paper, we present a principal component approach based on the reduced rank model for multidimensional sparse functional data, and our model is estimated by maximizing a penalized log-likelihood function. We refer to Yao *et al.* (2005) for sparse FDA and note that Zhou *et al.* (2008) has proposed two-dimensional sparse PC model via a penalized estimation and splines. There are some differences between our model and their model in 1) type of basis functions and 2) simplicity. Our model is simpler than their model, although their model takes into consideration the covariance structure of paired functional data.

This paper is organized as follows. Section 2 describes details of principal component analysis in one-dimensional case and multidimensional case, respectively. Section 3 introduces estimation methods such as the maximum likelihood method and its penalized version, and Section 4 describes model selections for those models, for example AIC, BIC and cross-validation approaches. In Section 5, numerical experiments are conducted to reveal the possibility of the proposed method to some types of incomplete data. Section 6 describes an application of the proposed method to handwriting data. Finally, some concluding remarks are presented in Section 7.

2 Principal component models

2.1 one-dimensional model

Let $x_i(t)$ $(i = 1, \cdots, N,\ t \in \mathcal{T} \subset \mathbb{R})$ be the values of i-th functional data at observational point t, $\mu(t)$ be an overall mean function, $\xi_j(t)$ $(j = 1, \cdots, k,\ t \in \mathcal{T})$ be j-th principal component (PC) curves, and $\boldsymbol{\xi}(t) = (\xi_1(t), \cdots, \xi_k(t))'$ is a k-dimensional PC curve. It is assumed that the following principal component model for $x_i(t)$ (reduced rank model, James *et al.* (2000)):

$$\begin{aligned} x_i(t) &= \mu(t) + \sum_{j=1}^{k} \xi_j(t)\alpha_{ij} + \varepsilon_i(t) \\ &= \mu(t) + \boldsymbol{\xi}(t)'\boldsymbol{\alpha}_i + \varepsilon_i(t) \qquad (i = 1, \cdots, N)\ , \end{aligned} \tag{1}$$

with random vectors $\boldsymbol{\alpha}_i = (\alpha_{i1}, \cdots, \alpha_{ik})'$ and observational errors $\varepsilon_i(t)$ that are independently normally distributed with the mean vector $\mathbf{0}$ and diagonal variance-covariance matrix D, the mean 0 and variance σ^2, respectively: $\boldsymbol{\alpha}_i \overset{iid}{\sim} N_k(\mathbf{0}, D)$ and $\varepsilon_i(t) \overset{iid}{\sim} N(0, \sigma^2)$. The random vectors $\boldsymbol{\alpha}_i = (\alpha_{i1}, \cdots, \alpha_{ik})'$ give weights of the PC curves to i-th individual. The PC curves $\xi_j(t)$ satisfy

the orthonormal constraints $\int_{\mathcal{T}} \boldsymbol{\xi}(t)\boldsymbol{\xi}(t)'dt = I_k$, that is,

$$\langle \xi_j, \xi_{j'} \rangle = \int_{\mathcal{T}} \xi_j(t)\xi_{j'}(t)\, dt = \delta_{jj'} \qquad (j, j' = 1, \cdots, k)\ , \tag{2}$$

where I_k is the $k \times k$ identity matrix and $\delta_{jj'} = 1\ (j = j'),\ 0\ (j \neq j')$.

We assume basis expansions with our orthonormalized Gaussian basis functions for the mean function and PC curves:

$$\mu(t) = \sum_{m=1}^{M} \theta_{\mu m}\psi_m^\nu(t) = \boldsymbol{\theta}_\mu'\boldsymbol{\psi}^\nu(t)\ , \quad \xi_j(t) = \sum_{m=1}^{M} \theta_{jm}\psi_m^\nu(t) = \boldsymbol{\theta}_j'\boldsymbol{\psi}^\nu(t) \quad (j = 1, \cdots, k)\ ,$$

where $\boldsymbol{\psi}^\nu(t) = (\psi_1^\nu(t), \cdots, \psi_M^\nu(t))'$ is the vector of our orthonormalized Gaussian basis functions and $\boldsymbol{\theta}_\mu = (\theta_{\mu 1}, \cdots, \theta_{\mu M})'$ and $\boldsymbol{\theta}_j = (\theta_{j1}, \cdots, \theta_{jM})'$ are the M-dimensional coefficient vectors in the basis expansions. The orthonormalized Gaussian basis functions are here constructed as follows. Let W_ϕ^ν and U_ν be the $M \times M$ cross-product matrix and upper triangular matrix given by $W_\phi^\nu = \int_{\mathcal{T}} \boldsymbol{\phi}^\nu(t)\boldsymbol{\phi}^\nu(t)'dt = U_\nu'U_\nu$ (Cholesky decomposition) with Gaussian basis functions $\boldsymbol{\phi}^\nu(t)' = (\phi_1^\nu(t), \cdots, \phi_M^\nu(t))$ by Ando *et al.* (2008):

$$\phi_m(t) = \phi_m(t;\, \nu, \mu_m, \tau_m^2) = \exp\{-(t - \mu_m)^2/(2\nu\tau_m^2)\} \qquad (m = 1, \cdots, M)\ ,$$

where the parameters μ_m and τ_m express the position and width of the m-th basis function respectively, and ν is a hyper-parameter that adjusts the width of basis functions, while the $\{\tau_m\}$ have been determined by a clustering method. Our orthonormalized bases are then defined by

$$\boldsymbol{\psi}^\nu(t) = (\psi_1^\nu(t), \cdots, \psi_M^\nu(t))' = U_\nu^{-1}\boldsymbol{\phi}^\nu(t)\ .$$

It follows that $W_\psi^\nu = \int_{\mathcal{T}} \boldsymbol{\psi}^\nu(t)\boldsymbol{\psi}^\nu(t)'dt = I_M$ with identity matrix I_M of size M. The cross product matrix W_ϕ^ν is here intended by W. We refer to Moody and Darken (1989), Powell (1987) for Gaussian type radial basis functions.

The principal component model (1) can then be written as

$$x_i(t) = \boldsymbol{\psi}^\nu(t)'\boldsymbol{\theta}_\mu + \boldsymbol{\psi}^\nu(t)'\Theta\boldsymbol{\alpha}_i + \varepsilon_i(t) \qquad (i = 1, \cdots, N)\ .$$

The orthonormal constraints in (2) of the principal component curves $\xi_j(t)$ are also written as

$$\Theta'\Theta = I_k\ , \qquad W = \int_{\mathcal{T}} \boldsymbol{\psi}^\nu(t)\boldsymbol{\psi}^\nu(t)'dt = I_M\ ,$$

with $\Theta = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k)$ and $\boldsymbol{\xi}(t) = \Theta'\boldsymbol{\psi}^\nu(t)$.

Although the previous paragraphs described the modeling for the functional data $x_i(t)$, we only have discrete data set $\{(t_{ij}, x_{ij})\ ;\ j = 1, \cdots, n_i\}\ (i = 1, \cdots, N)$, where each $t_{ij}\ (\in \mathcal{T} \subset \mathbb{R})$ is the j-th observational point of the i-th individual and each x_{ij} is the discrete data observed at t_{ij} for a variable X. Let $\boldsymbol{x}_i = (x_i(t_{i1}), \cdots, x_i(t_{in_i}))'$ be the observational vectors, $\Psi_i^\nu = (\boldsymbol{\psi}^\nu(t_{i1}), \cdots, \boldsymbol{\psi}^\nu(t_{in_i}))'$ be the basis matrices for i-th individual and $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \cdots, \varepsilon_i(t_{in_i}))'$ be the error vectors. We then have the mixed-type models

$$\boldsymbol{x}_i = \Psi_i^\nu\boldsymbol{\theta}_\mu + \Psi_i^\nu\Theta\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i \qquad (i = 1, \cdots, N)\ , \tag{3}$$

where $\boldsymbol{\varepsilon}_i \overset{iid}{\sim} N_{n_i}(\mathbf{0}, \sigma^2 I_{n_i})$, $\boldsymbol{\alpha}_i \overset{iid}{\sim} N_k(\mathbf{0}, D)$ and the coefficient matrix Θ is orthogonal. The unknown parameters $\boldsymbol{\theta}_\mu$, Θ, D, σ^2 are estimated by the maximum likelihood and penalized likelihood methods described in §3. We note that the optimal number of the basis functions M, principal components k and hyper-parameter ν in the basis functions are optimally selected by a model selection.

2.2 Multidimensional model

We here present an approach for the extension of the one-dimensional principal component model to a multidimensional model. Suppose that we have p-dimensional N functional observations $\{x_{i1}(t), \cdots, x_{ip}(t)\,;\, t \in \mathcal{T} \subset \mathbb{R}\}$ $(i = 1, \cdots, N)$, where $x_{il}(t)$ are the values of the i-th functional data for the l-th variable X_l at t. We assume the following model for each $x_{il}(t)$:

$$\begin{aligned} x_{il}(t) &= \mu_l(t) + \sum_{j=1}^{k} \xi_{lj}(t)\alpha_{ij} + \varepsilon_{il}(t) \\ &= \mu_l(t) + \boldsymbol{\xi}_l(t)'\boldsymbol{\alpha}_i + \varepsilon_{il}(t) \qquad (i = 1, \cdots, N,\ l = 1, \cdots, p)\ , \end{aligned} \tag{4}$$

where the PC curves $(\xi_{1j}(t), \cdots, \xi_{pj}(t))$ satisfy the orthonormal constraints $\sum_{l=1}^{p} \int_{\mathcal{T}} \boldsymbol{\xi}_l(t)\, \boldsymbol{\xi}_l(t)' dt = I_k$, that is,

$$\sum_{l=1}^{p} \int_{\mathcal{T}} \xi_{lj}(t)\xi_{lj'}(t)dt = \delta_{jj'} \qquad (j, j' = 1, \cdots, k)\ ,$$

and the random vectors $\boldsymbol{\alpha}_i = (\alpha_{i1}, \cdots, \alpha_{ik})'$ and error functions $\varepsilon_{il}(t)$ are independently normally distributed with the mean vector $\mathbf{0}$ and diagonal variance-covariance matrix D, the mean 0 and variance σ^2, respectively: $\boldsymbol{\alpha}_i \overset{iid}{\sim} N_k(\mathbf{0}, D)$ and $\varepsilon_{il}(t) \overset{iid}{\sim} N(0, \sigma^2)$. It may be noted that the random components α_{ij} are independent from l.

Let $\boldsymbol{x}_i(t) = (x_{i1}(t), \cdots, x_{ip}(t))'$, $\boldsymbol{\mu}(t) = (\mu_1(t), \cdots, \mu_p(t))'$ and $\boldsymbol{\varepsilon}_i(t) = (\varepsilon_{i1}(t), \cdots, \varepsilon_{ip}(t))'$ be the p-dimensional functional data for i-th individual, mean function and error function, respectively, and $\Xi(t) = (\boldsymbol{\xi}_1(t), \cdots, \boldsymbol{\xi}_p(t))$ be the $k \times p$ matrix formed from the PC curves. Then the principal component model for the multidimensional functional data sets given by (4) can be written as

$$\boldsymbol{x}_i(t) = \boldsymbol{\mu}(t) + \Xi(t)'\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i(t)\ , \tag{5}$$

where $\Xi(t) = (\boldsymbol{\xi}_1(t), \cdots, \boldsymbol{\xi}_p(t))$ satisfies the orthonormal constraint

$$\int_{\mathcal{T}} \Xi(t)\Xi(t)'dt = \sum_{l=1}^{p} \int_{\mathcal{T}} \boldsymbol{\xi}_l(t)\boldsymbol{\xi}_l(t)'dt = I_k\ .$$

We also assume basis expansions to the mean functions $\mu_l(t)$ and PC curves $\xi_{lj}(t)$.

$$\mu_l(t) = \sum_{m=1}^{M} \theta_{\mu lm}\psi_m^\nu(t) = \boldsymbol{\theta}'_{\mu l}\boldsymbol{\psi}^\nu(t) \qquad (l = 1, \cdots, p)\ ,$$

$$\xi_{lj}(t) = \sum_{m=1}^{M} \theta_{ljm}\psi_m^\nu(t) = \boldsymbol{\theta}'_{lj}\boldsymbol{\psi}^\nu(t) \qquad (l = 1, \cdots, p,\ j = 1, \cdots, k)\ ,$$

where $\boldsymbol{\theta}_{\mu l} = (\theta_{\mu l1}, \cdots, \theta_{\mu lM})'$ and $\boldsymbol{\theta}_{lj} = (\theta_{lj1}, \cdots, \theta_{ljM})'$. Let $\boldsymbol{\theta}_{\mu} = (\boldsymbol{\theta}'_{\mu 1}, \cdots, \boldsymbol{\theta}'_{\mu p})'$ and $\Psi^{\nu}(t) = \mathrm{diag}(\boldsymbol{\psi}^{\nu}(t), \cdots, \boldsymbol{\psi}^{\nu}(t))$ be the pM-dimensional coefficient vector and $pM \times p$ block diagonal matrix formed from $\boldsymbol{\psi}^{\nu}(t)$, respectively. For example, the first and p-th columns of $\Psi^{\nu}(t)$ are given by $(\boldsymbol{\psi}^{\nu}(t)', \mathbf{0}', \cdots, \mathbf{0}')'$ and $(\mathbf{0}', \cdots, \mathbf{0}', \boldsymbol{\psi}^{\nu}(t)')'$.

The p-dimensional mean function $\boldsymbol{\mu}(t)$ and $k \times p$ matrix function of the PC curves $\Xi(t) = (\boldsymbol{\xi}_1(t), \cdots, \boldsymbol{\xi}_p(t))$ are then given by $\boldsymbol{\mu}(t) = \Psi^{\nu}(t)'\boldsymbol{\theta}_{\mu}$ and $\Xi(t) = \Theta'\Psi^{\nu}(t)$ respectively, where $\Theta = (\Theta'_1, \cdots, \Theta'_p)'$ is $pM \times k$ coefficients matrix with coefficients matrices $\Theta_l = (\boldsymbol{\theta}_{l1}, \cdots, \boldsymbol{\theta}_{lk})$ $(l = 1, \cdots, p)$. Thus the equation (5) can be expressed as

$$\boldsymbol{x}_i(t) = \Psi^{\nu}(t)'\boldsymbol{\theta}_{\mu} + \Psi^{\nu}(t)'\Theta\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i(t) \ .$$

The orthonormal constraint is also given by

$$\int_{\mathcal{T}} \Xi(t)\Xi(t)'dt = \Theta' \int_{\mathcal{T}} \Psi^{\nu}(t)\Psi^{\nu}(t)'dt \, \Theta = I_k \ ,$$

that is, $\Theta'\Theta = I_k$ and $\int_{\mathcal{T}} \Psi^{\nu}(t)\Psi^{\nu}(t)'dt = I_{pM}$, where the second condition $\int_{\mathcal{T}} \Psi^{\nu}(t)\Psi^{\nu}(t)'dt = I_{pM}$ is equivalent to $\int_{\mathcal{T}} \boldsymbol{\psi}^{\nu}(t)\boldsymbol{\psi}^{\nu}(t)'dt = I_M$.

The principal component model for the multidimensional functional data sets can also be discretized, as follows. Let $t_{i1}, \cdots, t_{in_i}$ be the observational points, $\boldsymbol{x}_i = (\boldsymbol{x}_i(t_{i1})', \cdots, \boldsymbol{x}_i(t_{in_i})')'$ and $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_i(t_{i1})', \cdots, \boldsymbol{\varepsilon}_i(t_{in_i})')'$ be the pn_i-dimensional observational and error vectors and $\Psi_i^{\nu} = (\Psi^{\nu}(t_{i1}), \cdots, \Psi^{\nu}(t_{in_i}))'$ be the $pn_i \times pM$ basis matrices. The multidimensional principal component model is also given by the following equations:

$$\boldsymbol{x}_i = \Psi_i^{\nu}\boldsymbol{\theta}_{\mu} + \Psi_i^{\nu}\Theta\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i \qquad (i = 1, \cdots, N) \ ,$$

where $\boldsymbol{\varepsilon}_i \overset{iid}{\sim} (\mathbf{0}, \sigma^2 I_{pn_i})$, $\boldsymbol{\alpha}_i \overset{iid}{\sim} (\mathbf{0}, D)$, Θ and $\Psi^{\nu}(t)$ are orthogonal and orthonormal, respectively: $\Theta'\Theta = I_k$, $\int_{\mathcal{T}} \Psi^{\nu}(t)\Psi^{\nu}(t)'dt = I_{pM}$. In the next section, we estimate the unknown parameters in the model, and section 4 shows methods of selecting the optimal number of basis functions M, principal components k and hyper-parameter ν in the basis functions.

3 Estimation

3.1 Maximum likelihood method

The first goal of functional principal component models is the estimation of a mean function $\mu(t)$ and PC curves $\xi_j(t)$. The second goal is the inference of random vectors $\boldsymbol{\alpha}_i$. These goals are equivalent to estimate the coefficients $\boldsymbol{\theta}_{\mu}$ and Θ and variances σ^2 and D, which are all unknown parameters. In this section, we first describe the maximum likelihood method for the one-dimensional principal component model, and a penalized maximum likelihood method is introduced latter. The multidimensional model can be estimated by the same procedures.

The random vectors $\{\boldsymbol{\alpha}_i \, ; \, i = 1, \cdots, N\}$ and observational error vectors $\{\boldsymbol{\varepsilon}_i \, ; \, i = 1, \cdots, N\}$ in the principal component model (3) are independently normally distributed, and the observational

vectors $\{\boldsymbol{x}_i\ ;\ i=1,\cdots,N\}$ are also independently normally distributed with the mean vector $\Psi_i^\nu\boldsymbol{\theta}_\mu$ and variance-covariance matrix $\sigma^2 I_{n_i}+\Psi_i^\nu\Theta D\Theta'(\Psi_i^\nu)'$:

$$\boldsymbol{x}_i \overset{iid}{\sim} N\left(\Psi_i^\nu\boldsymbol{\theta}_\mu,\ \sigma^2 I_{n_i}+\Psi_i^\nu\Theta D\Theta'(\Psi_i^\nu)'\right) \quad (i=1,\cdots,N)\ .$$

The likelihood function of the joint distribution of $\{\boldsymbol{x}_i\ ;\ i=1,\cdots,N\}$ is then given by

$$\prod_{i=1}^{N}\frac{1}{(2\pi)^{n_i/2}|\sigma^2 I_{n_i}+\Psi_i^\nu\Theta D\Theta'(\Psi_i^\nu)'|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\Psi_i^\nu\boldsymbol{\theta}_\mu)'(\sigma^2 I_{n_i}+\Psi_i^\nu\Theta D\Theta'(\Psi_i^\nu)')^{-1}(\boldsymbol{x}_i-\Psi_i^\nu\boldsymbol{\theta}_\mu)\right\}\ .$$

James *et al.* (2000) described that the maximization of the likelihood function for the joint distribution of $\{\boldsymbol{x}_i\ ;\ i=1,\cdots,N\}$ with respect to $\boldsymbol{\theta}_\mu,\Theta,\sigma^2,D$ is a non-convex optimization problem, and to solve the maximization problem, they used the EM algorithm (Dempster *et al.* (1977)) that considers $\{\boldsymbol{\alpha}_i\ ;\ i=1,\cdots,N\}$ as missing values. If we observed the random vectors $\boldsymbol{\alpha}_i$, the joint distribution would be given by the following simple form:

$$\prod_{i=1}^{N}\frac{1}{(2\pi)^{(n_i+k)/2}\sigma^{n_i}|D|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{x}_i-\Psi_i^\nu\boldsymbol{\theta}_\mu-\Psi_i^\nu\Theta\boldsymbol{\alpha}_i)'(\boldsymbol{x}_i-\Psi_i^\nu\boldsymbol{\theta}_\mu-\Psi_i^\nu\Theta\boldsymbol{\alpha}_i)-\frac{1}{2}\boldsymbol{\alpha}_i'D^{-1}\boldsymbol{\alpha}_i\right\}\ .$$

Appendix A shows procedures of the EM algorithm to obtain the maximum likelihood estimators. The procedures base on the that introduced by James *et al.* (*http://www-rcf.usc.edu/~gareth*).

3.2 Penalized maximum likelihood method

However, it is well known that maximum likelihood methods yield unstable parameter estimates and some of the parameter estimates are often infinite. We then estimate unknown parameters $\boldsymbol{\theta}_\mu,\Theta,\sigma^2,D$ by maximizing a penalized log-likelihood function

$$p\ell_\lambda(\boldsymbol{\theta}_\mu,\Theta,\sigma^2,D)=\sum_{i=1}^{N}\log f(\boldsymbol{x}_i|\boldsymbol{t}_i;\boldsymbol{\theta}_\mu,\Theta,\sigma^2,D) - \frac{N\lambda}{2}\ (\text{roughness penalty for } \mu(t))\ , \tag{6}$$

where $f(\boldsymbol{x}_i|\boldsymbol{t}_i;\boldsymbol{\theta}_\mu,\Theta,\sigma^2,D)$ is the density of $\boldsymbol{x}_i$ and λ is a smoothing parameter that controls the smoothness of the mean function $\mu(t)$. It is assumed that the roughness penalty for $\mu(t)$ is given by $\boldsymbol{\theta}_\mu'K\boldsymbol{\theta}_\mu$, where K is a $M\times M$ positive semidefinite matrix. We use the roughness penalties given by $\sum_{m=2}^{M}(\Delta^2\theta_{\mu m})^2=\boldsymbol{\theta}_\mu'D_2'D_2\boldsymbol{\theta}_\mu$, where Δ is the difference operator defined by $\Delta\theta_{\mu m}=\theta_{\mu m}-\theta_{\mu,m-1}$ and D_2 is a $(M-2)\times M$ matrix representation of the difference operator Δ^2.

The unknown parameters $\boldsymbol{\theta}_\mu,\Theta,\sigma^2,D$ are estimated by the penalized maximum likelihood method, using the EM algorithm which maximizes the Q function defined by the following

equation with observational data $\boldsymbol{D}$ and missing data $\boldsymbol{Z}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{Z}}\left[\log f(\boldsymbol{D}, \boldsymbol{Z}; \boldsymbol{\theta})|\boldsymbol{D}; \boldsymbol{\theta}^{(t)}\right] ,$$

where $\boldsymbol{\theta}^{(t)}$ is the t-th updated value of unknown parameter $\boldsymbol{\theta}$. The penalized Q function $Q_p(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is here defined by the conditional expectation of $\log f(\boldsymbol{D}, \boldsymbol{Z}; \boldsymbol{\theta}) - g_\lambda(\boldsymbol{\theta}; \boldsymbol{D})$, where $g_\lambda(\boldsymbol{\theta}; \boldsymbol{D})$ is a penalty term with a smoothing parameter λ that controls the smoothness of fitted curve.

In the principal component model in this paper, the observational data $\boldsymbol{D}$, missing data $\boldsymbol{Z}$ and unknown parameter vector $\boldsymbol{\theta}$ are given by $\boldsymbol{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, $\boldsymbol{Z} = \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_N\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\mu, \Theta, \sigma^2, D\}$, respectively. The penalty term $g_\lambda(\boldsymbol{\theta}; \boldsymbol{D})$ for this model is given by $g_\lambda(\boldsymbol{\theta}; \boldsymbol{D}) = N\lambda/2\ \boldsymbol{\theta}_\mu' K \boldsymbol{\theta}_\mu$. We obtain the penalized maximum likelihood estimators of the unknown parameters $\sigma^2, D, \boldsymbol{\theta}_\mu, \Theta$, maximizing the penalized Q function $Q_p(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. We also have the estimators $\{\hat{\boldsymbol{\alpha}}_i\ ;\ i = 1, \cdots, N\}$ of $\{\boldsymbol{\alpha}_i\ ;\ i = 1, \cdots, N\}$ from the E-step of the EM algorithm. If we have t-th updated values of Θ, σ^2 and $\boldsymbol{\alpha}_i$, then the $(t+1)$-th updated value of $\boldsymbol{\theta}_\mu$ is obtained by

$$\hat{\boldsymbol{\theta}}_\mu^{(t+1)} = \left(\sum_{i=1}^N (\Psi_i^\nu)'\Psi_i^\nu + N\lambda\hat{\sigma}^{(t)\,2} K\right)^{-1} \sum_{i=1}^N (\Psi_i^\nu)'(\boldsymbol{x}_i - \Psi_i^\nu \hat{\Theta}^{(t)} \hat{\boldsymbol{\alpha}}_i^{(t)}) .$$

Thus, the effects of the penalty directly appears in the estimation of the mean function.

4 Model selection

The joint distribution of the observational vectors $\{\boldsymbol{x}_i\ ;\ i = 1, \cdots, N\}$ depends on the number of the basis functions M, principal components k, smoothing parameter λ in the penalized log-likelihood function (6) and hyper-parameters included in the basis functions. In this section, we introduce a cross validation method and Akaike and Bayesian information criteria for selecting optimal values of these parameters, and the optimal values are obtained by minimizing a criterion.

A cross validation method can be directly applied in the principal component models. Shi *et al.* (1996) have selected optimal values of parameters by the cross validation score

$$\mathrm{CV} = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^{n_i} \{x_{ij} - \hat{x}_{ij}^{(-ij)}\}^2 ,$$

where $\hat{x}_{ij}^{(-ij)}$ are estimated functional data $x_{ij} = x_i(t_{in_i})$ from the observations excluding the (i, j)-th observation. We here introduce a cross validation score to reduce calculation amount:

$$\mathrm{CV} = \frac{1}{N}\sum_{i=1}^N \left\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i^{(-i)}\right\|^2 = \frac{1}{N}\sum_{i=1}^N \sum_{l=1}^p \int_{\mathcal{T}} \left\{x_{il}(t) - \hat{x}_{il}^{(-i)}(t)\right\}^2 dt , \tag{7}$$

where $\hat{\boldsymbol{x}}_i^{(-i)}$ are estimated functional data $\boldsymbol{x}_i$ from the observations excluding the i-th observational vector. The cross validation methods can be utilized as the model selection criteria for evaluating the model estimated by maximum and penalized maximum likelihood methods.

On the other hand, if a principal component model would be estimated by the maximum likelihood method, Akaike information criterion (AIC) and Bayesian information criterion (BIC) would be utilized as model selection criteria:

$$\begin{aligned}\mathrm{AIC} &= -2\sum_{i=1}^{N}\log f(\boldsymbol{x}_i|\boldsymbol{t}_i;\hat{\boldsymbol{\theta}}_\mu,\hat{\Theta},\hat{\sigma}^2,\hat{D}) + 2P\ ,\\ \mathrm{BIC} &= -2\sum_{i=1}^{N}\log f(\boldsymbol{x}_i|\boldsymbol{t}_i;\hat{\boldsymbol{\theta}}_\mu,\hat{\Theta},\hat{\sigma}^2,\hat{D}) + P\log N\ .\end{aligned} \tag{8}$$

where $\hat{\boldsymbol{\theta}}_\mu, \hat{\Theta}, \hat{\sigma}^2, \hat{D}$ are the maximum likelihood estimators, $\sum_{i=1}^{N}\log f(\cdot)$ is the maximum log-likelihood function and $P = (pM+1)(k+1)$ is the number of the parameters. Optimal values of the parameters are given by minimizing a criterion.

5 Numerical experiments

In this section, Monte Carlo experiments are conducted to reveal the possibility of the proposed method to some types of incomplete data. The first step of our experiments generated two-dimensional true functional data from the following models:

$$x_{il}(t) = \mu_l(t) + \sum_{m=1}^{4}\alpha_m\zeta_m(t) \qquad (t\in[0,1],\ i=1,\cdots,15,\ l=1,2)\ , \tag{9}$$

where the mean functions $\mu_1(t), \mu_2(t)$ were set to be the following functions:

1. $\mu_1(t) = \sin(2\pi t)$, $\mu_2(t) = \cos(2\pi t)$,
2. $\mu_1(t) = \sin(10\pi t)$, $\mu_2(t) = \cos(10\pi t)$,
3. $\mu_1(t) = \sin(10\pi t)$, $\mu_2(t) = \cos(2\pi t)$,

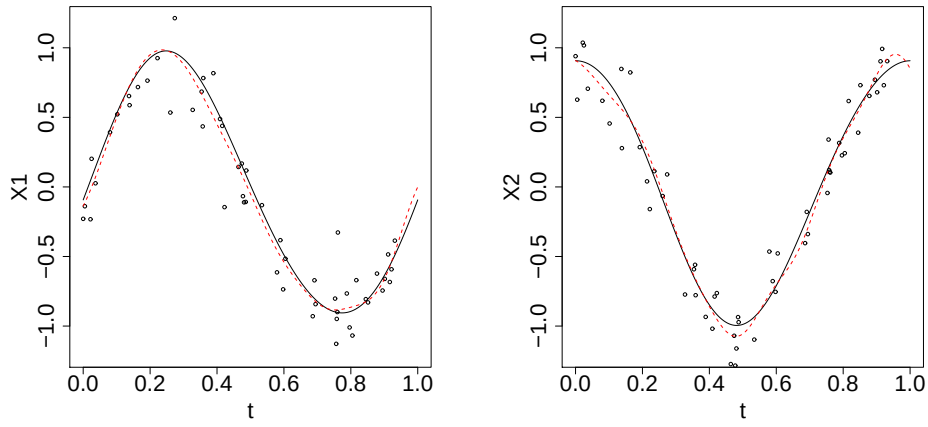
and $\zeta_{2r-1}(t) = \sin(2\pi r t)$, $\zeta_{2r}(t) = \cos(2\pi r t)$. The random components α_{im} were assumed to be independently normally distributed: $\alpha_{im} \overset{iid}{\sim} N(0, (5-m)(0.02\bar{Rx})^2)$, where $\bar{Rx} = (Rx_1 + Rx_2)/2$ and Rx_l were the range of $\mu_l(t)$ over $t\in[0,1]$ $(l=1,2)$.

We then generated discrete data from the nonlinear regression models with the true functions $x_{il}(t)$:

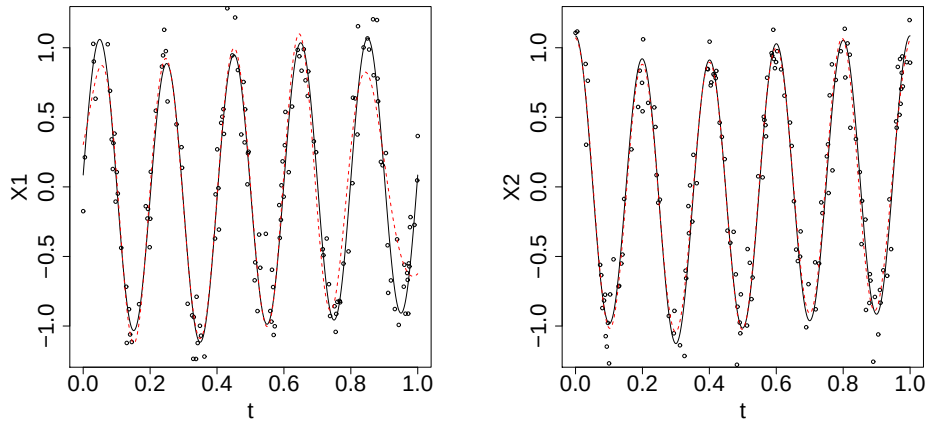
$$x_{ilj} = x_{il}(t_{ij}) + \varepsilon_{ilj} \qquad (i=1,\cdots,15,\ l=1,2,\ j=1,\cdots,n_i)\ , \tag{10}$$

where the errors ε_{ilj} were assumed to be independently normally distributed with mean 0 and variance $(0.1Rx_j)^2$. The t_{ij} were created through $t_j^* = j/600$ $(j=1,\cdots,600)$ as follows:

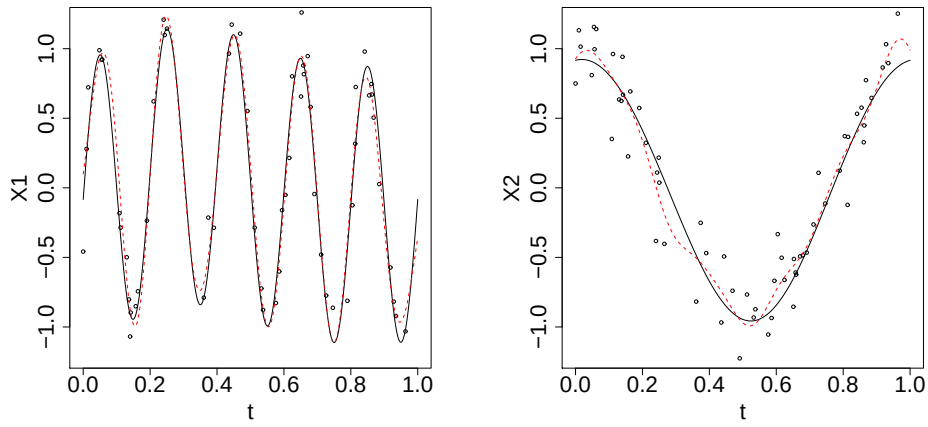
1. Set a missing rate p,
2. For each j, the t_j^* is randomly eliminated with probability p.
3. We then have $\{t_{ij}\ :\ j=1,\cdots,n_i\}$ with $n_i \simeq 600(1-p)$.



1) $\mu_1(t) = \sin(2\pi t)$, $\mu_2(t) = \cos(2\pi t)$



2) $\mu_1(t) = \sin(10\pi t)$, $\mu_2(t) = \cos(10\pi t)$



3) $\mu_1(t) = \sin(10\pi t)$, $\mu_2(t) = \cos(2\pi t)$

Figure 1: Examples of a generated two-dimensional true functional data (solid lines, black) with individual noise function and a estimated functional data (dashed lines, red) with the missing rate= 0.9 for each setting of the mean functions.

Table 1: Simulation results of three sets of the mean functions.

	Missing Rate				
	0	0.25	0.5	0.75	0.90
1) $\mu_1(t) = \sin(2\pi t)$, $\mu_2(t) = \cos(2\pi t)$					
MSE_1 ($\times 10^{-3}$)	0.111	0.137	0.189	0.325	0.774
MSE_2 ($\times 10^{-3}$)	0.253	0.292	0.349	0.532	1.101
MSE ($\times 10^{-3}$)	0.365	0.428	0.539	0.856	1.875
2) $\mu_1(t) = \sin(10\pi t)$, $\mu_2(t) = \cos(10\pi t)$					
MSE_1 ($\times 10^{-3}$)	0.386	0.505	0.757	1.470	3.699
MSE_2 ($\times 10^{-3}$)	0.385	0.513	0.749	1.518	3.037
MSE ($\times 10^{-3}$)	0.771	1.018	1.506	2.988	6.842
3) $\mu_1(t) = \sin(10\pi t)$, $\mu_2(t) = \cos(2\pi t)$					
MSE_1 ($\times 10^{-2}$)	0.153	0.232	0.328	0.799	1.465
MSE_2 ($\times 10^{-2}$)	0.085	0.122	0.165	0.355	1.787
MSE ($\times 10^{-2}$)	0.238	0.354	0.493	1.155	3.252

The missing rate p were set to be $0, 0.25, 0.5, 0.75, 0.90$ for each setting of the mean functions. Our two-dimensional principal component method was applied to the data and estimated two-dimensional functions $\hat{\boldsymbol{x}}_i(t)$ by maximizing the penalized log-likelihood and minimizing the cross-validation score. We finally calculated the mean-square error (MSE) between the true functional data and the estimated functions:

$$\text{MSE}_l = \frac{1}{15}\sum_{i=1}^{15} \|x_{il} - \hat{x}_{il}\|^2 , \qquad \text{MSE} = \frac{1}{15}\sum_{i=1}^{15} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 .$$

Figure 1 shows examples of a generated two-dimensional true functional data with individual noise function and a estimated functional data with the missing rate= 0.9 for each setting of the mean functions. Table 1 shows the results of the expreriments. Our method actually performed well to data with missing values such as these synthetic data even if they had a high missing rate.

6 Real data examples

6.1 Human gait data

We firstly apply a principal component model to human gait data which consist of the hip and knee angles for $N = 39$ subjects, using the maximum likelihood method. We note that the data have common set of the 20 time points (non-sparse data). We refer to Olshen *et al.* (1989) for details of these human gait data.

The one-dimensional principal component model via Gaussian basis functions with hyper-parameter was applied to the hip angles. The model selection was performed with the candidate values of the number of basis functions $M = 3, 4 \cdots, 9$, hyper-parameter $\nu = 5, 10, \cdots, 50$ and

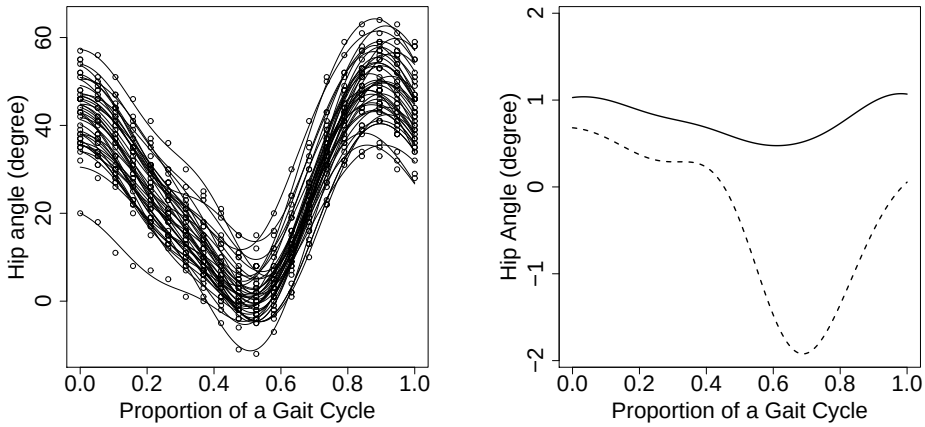


Figure 2: Left: the hip angles and estimated functional data ($N = 39,\ n = n_i = 20$). Right: The PC curves for the hip angles. The solid and dashed lines represent the 1st and 2nd PC curves, respectively.

principal components $k = 2, 3, \cdots, 9$. The values of $(M, \nu, k) = (9, 35, 7)$ and $(7, 20, 6)$ minimized the AIC and BIC, respectively. We selected the simpler model with $M = 7,\ \nu = 20,\ k = 6$ as the optimal ones.

Figure 2 shows the estimated functional data and the PC curves by the optimal model. It is shown that structures of the hip angles were successfully estimated by the principal component model, and the estimated PC curves coincided with the ones estimated by the ordinary FPCA. The contribution rates of the 1st and 2nd PCs were 75.9% and 10.5% in the $k = 6$ PCs: the 1st PCs had most of information for the variation among the data. The principal component model performed well to these equispaced data.

6.2 Handwriting data

We next apply our two-dimensional principal component model via the maximum likelihood method and the penalized maximum likelihood method to the handwriting data such as that shown in Figure 3. In this figure, the left panel shows an original script sample for one of the Japanese characters, and the middle and right panels show the corresponding XY coordinates values, respectively. Our approach treats script samples on 2-dimensional writing surface, and the script samples were recorded by a simple device: to obtain the script samples, we composed a simple device using the programming language VisualBasic.NET and using the pen tablet intuos 3 made in WACOM. The XY coordinate values while the pen lifts off the writing surface cannot be obtained. We then consider the XY coordinate values of handwritings as 2-dimensional sparse functional data. We refer to Ramsay (2000) which has treated handwriting data recorded as XYZ coordinates values using the OPTOTRACK tracking system.

The script samples were written by the 12 students in our statistical laboratory (male, right handedness, 21 to 26 years old). Scaling was performed to the XY coordinates values and

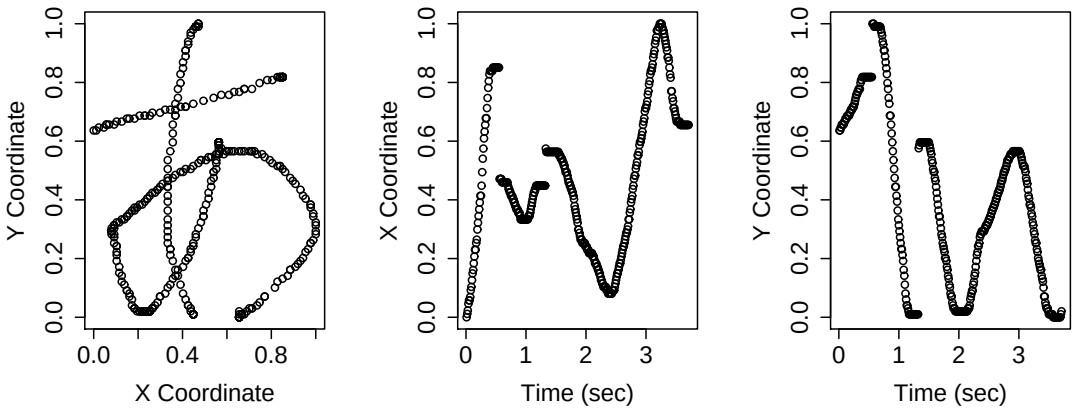


Figure 3: An example of the handwriting data. (left): original script sample, (middle, right): time series data of XY-coordinates. We treat XY-coordinates values in the middle and right panels as handwriting data.

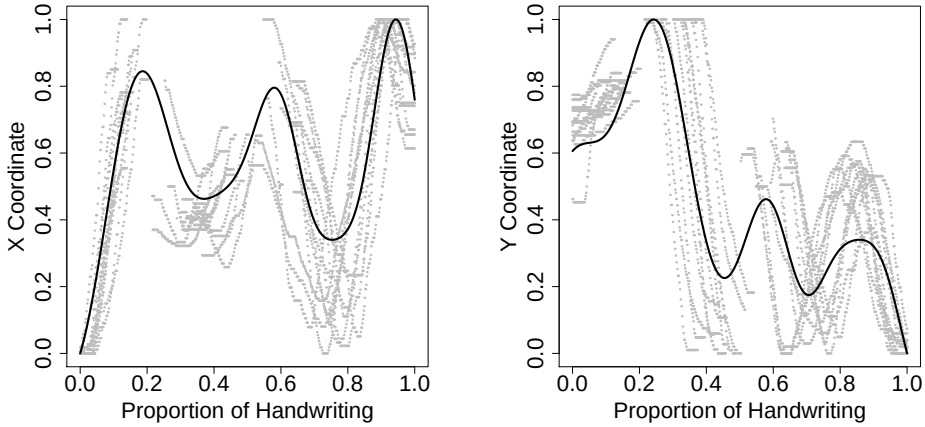


Figure 4: The scaled handwriting data (dots) and the estimated mean functions (heavy lines).

writing times ($\simeq 6 \sim 8$ seconds): the time interval and ranges of XY coordinates were to be $[0, 1]$. The dots in Figure 4 show the scaled data for all individuals.

We tried to perform the model selections to the scaled data by minimizing AIC and BIC, where the model was estimated by the maximum likelihood method. The candidate values of M, k, nu were $M = 3, 4, \cdots, 12$ and $k = 2, 3, \cdots, 10$ ($\leq M$), $\nu = 5, 10, \cdots, 50$. However, AIC and BIC could not be calculated, since the log-likelihood function infinitely exuded for these data. The model was then estimated by maximizing the penalized likelihood function, and we performed the model selection by minimizing the cross validation score (7). We considered values for M of $6, 7, \cdots, 15$, values for ν of $5, 10, \cdots, 30$, values for k of $3, 4, \cdots, 10$ and values for λ of $0, 10^{-10}, 10^{-9}, \cdots, 10^{-1}$ and found optimal values of $M = 12$, $\nu = 25$, $k = 10$ and $\lambda = 10^{-1}$. The upper bound of k was set to 10, since $k > 10$ caused unstable and impractical parameter estimation.

The heavy lines in Figure 4 show the estimated mean functions. Figure 5 shows the estimated PC curves. From the estimated PC curves, we give some interpretations of the PCs. The optimal number of principal components was selected as $k = 10$, and the cumulative contribution rate to the first 3 PCs was 70.8 % in the 10 principal components. The contribution rates of the X and Y coordinates for each set of PCs were 39.2 and 60.8 % for the 1st PCs, 47.1 and 52.9 % for the 2nd PCs, and 53.0 and 47.0 % for the 3rd PCs. We then consider the Y coordinate for the 1st PCs and both coordinates for the 2nd and 3rd PCs. Most of PCs represented the time shift component, while 2nd PCs had the effect of the angle and length of the 1st strokes (XY values on $t = 0 \sim 0.2$). In addition, the 3rd PCs indicated an effect of starting point on the 3rd strokes.

7 Concluding remarks

In this paper, we addressed the issue of extension of sparse functional PCA to multidimensional case, based on mixed models. Our model based on the orthonormalized Gaussian basis functions and a penalized estimation. Our Gaussian bases were constructed by the Cholesky decomposion of the cross-product matrix W_ϕ^ν, and the penalized estimation performed well, while maximum likelihood methods yielded unstable parameter estimates and some of the parameter estimates were infinite.

Numerical experiments were conducted to reveal the possibility of the proposed method to some types of incomplete data. Our method actually performed well to data with missing values such as our three types of synthetic data even if they had a high missing rate. The proposed method is then applied to the handwriting data. To obtain the data, we composed a simple device using the programming language VisualBasic.NET and using a pen tablet. Our device treated script samples on 2-dimensional writing surface as the time series handwriting data. The two-dimesional proposed method was applied to the scaled data by maximizing the penelized likelihood function and minimizing the cross-validation score. We then gave interpletation of the first 3 PCs: most of PCs represented the time shift component, while 2nd PCs had the effect of the angle and length of the 1st strokes and the 3rd PCs indicated an effect of starting point

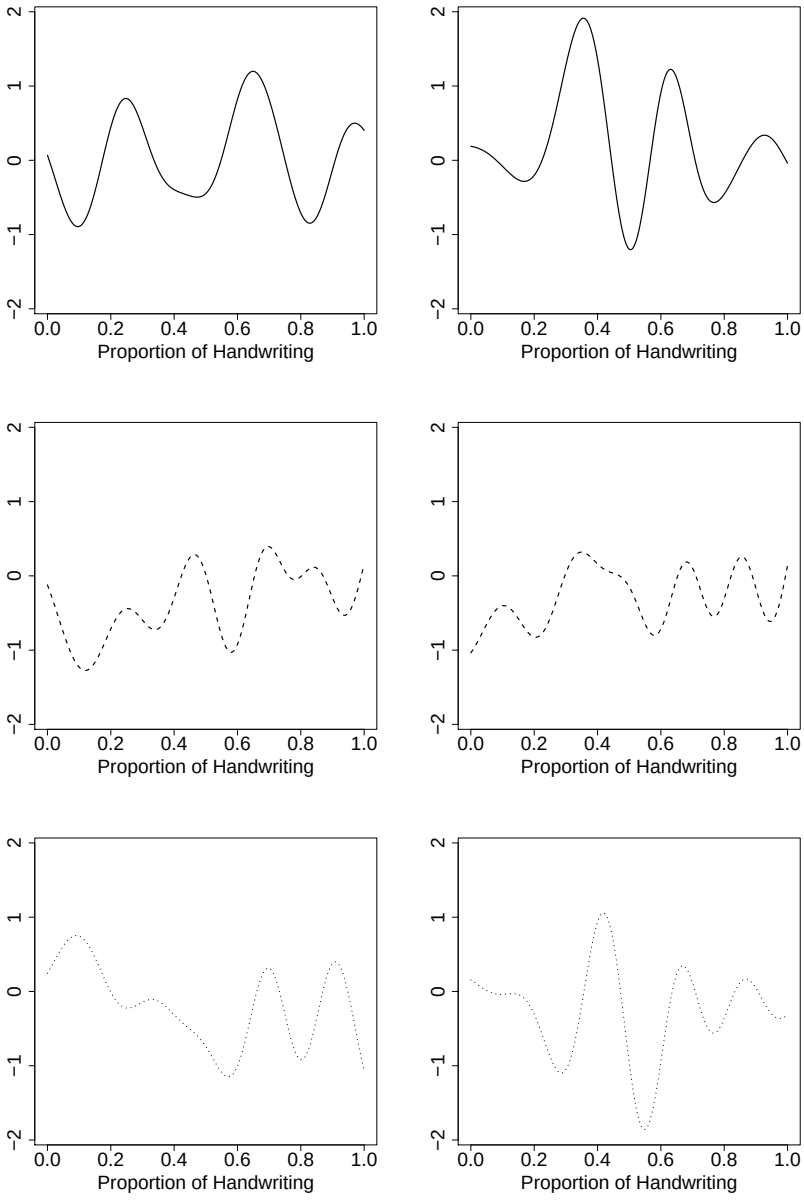


Figure 5: The estimated PC curves. The solid, dashed and broken lines represent the 1st, 2nd and 3rd PC curves respectively.

on the 3rd strokes.

Future works that remain to be done include 1) derivation of model selection criteria from an information-theoretic perspective and also the application of Bayesian approaches for the penalized model (see, e.g., Konishi and Kitagawa (2008)), 2) taking into account of other script information such as the pressure of the pen and treating registered handwriting data (see, e.g., Ramsay and Li (1998)), and 3) investigating the possibility of our and other functional data approaches for applying to huge types of real data such as in life science, image processing, marketing and so on.

A Details of EM algorithm

We have summarized methods to estimate unknown parameters in the principal component models (§3), using the EM algorithm. This section shows details of the EM algorithm and derives its estimation procedure.

A.1 Derivation

In this section, we derive the procedure of the EM algorithm to obtain the penalized maximum likelihood estimators in the principal component models. An EM algorithm for maximum likelihood methods maximizes the Q function defined by the following equation with observational data $\boldsymbol{D}$ and missing data $\boldsymbol{Z}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{Z}}\left[\log f(\boldsymbol{D}, \boldsymbol{Z}; \boldsymbol{\theta})|\boldsymbol{D}; \boldsymbol{\theta}^{(t)}\right] ,$$

where $\boldsymbol{\theta}$ is an unknown parameter vector and $\boldsymbol{\theta}^{(t)}$ is the t-th updated value of $\boldsymbol{\theta}$. The finding the Q function, that is, the replacing missing data with conditional expectation of the complete log-likelihood, is called *E-step* (expectation step). The $(t+1)$-th updated value $\boldsymbol{\theta}^{(t+1)}$ of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) .$$

The maximization of the Q function is called *M-step* (maximization step). The maximum likelihood estimator of $\boldsymbol{\theta}$ can be obtained by repeating the E and M-steps until convergence.

The penalized maximum likelihood method maximizes the penalized Q functions

$$Q_p(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_Z\left[\log f(D, Z; \boldsymbol{\theta}) - g_\lambda(\boldsymbol{\theta};\, \boldsymbol{D})|D; \boldsymbol{\theta}^{(t)}\right] , \tag{11}$$

where $g_\lambda(\boldsymbol{\theta};\, \boldsymbol{D})$ is a penalty term with a smoothing parameter $\lambda\ (>0)$ that controls the degree of the penalty. The maximization of $Q_p(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ corresponds to the that of the penalized log-likelihood function. In the principal component model, the observational data $\boldsymbol{D}$, missing data $\boldsymbol{Z}$ and unknown parameter $\boldsymbol{\theta}$ are given by $\boldsymbol{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, $\boldsymbol{Z} = \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_N\}$ and $\boldsymbol{\theta} = \{\sigma^2, D, \boldsymbol{\theta}_\mu, \Theta\}$ respectively, and the penalty term $g_\lambda(\boldsymbol{\theta};\, \boldsymbol{D})$ is also given by $g_\lambda(\boldsymbol{\theta};\, \boldsymbol{D}) = N\lambda/2 g(\boldsymbol{\theta}_\mu) = N\lambda/2 \cdot \boldsymbol{\theta}_\mu' K \boldsymbol{\theta}_\mu$, where K is a positive semidefinite matrix. We calculate the penalized Q function $Q_p(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in (11).

The joint distribution of the complete data $\{\boldsymbol{D}, \boldsymbol{Z}\} = \{(\boldsymbol{x}_i, \boldsymbol{\alpha}_i)\ ;\ i = 1, \cdots, N\}$ is given by $f(\boldsymbol{D}, \boldsymbol{Z}\ ;\ \boldsymbol{\theta}) = f(\boldsymbol{x}_i, \cdots, \boldsymbol{x}_N,\ \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_N\ ;\ \boldsymbol{\theta}) = \prod_{i=1}^{N} f(\boldsymbol{x}_i, \boldsymbol{\alpha}_i\ ;\ \boldsymbol{\theta})$, then it follows that

$$\begin{aligned} Q_p(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_Z\left[\log f(D, Z; \boldsymbol{\theta})|D; \boldsymbol{\theta}^{(t)}\right] - g_\lambda(\boldsymbol{\theta};\ \boldsymbol{D}) \\ &= \sum_{i=1}^{N} E_Z\left[\log f(\boldsymbol{x}_i, \boldsymbol{\alpha}_i; \boldsymbol{\theta})|D; \boldsymbol{\theta}^{(t)}\right] - g_\lambda(\boldsymbol{\theta};\ \boldsymbol{D})\ . \end{aligned} \tag{12}$$

We also have the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{D}$: $f(\boldsymbol{Z}|\boldsymbol{D}; \boldsymbol{\theta}^{(t)}) = f(\boldsymbol{D}, \boldsymbol{Z}; \boldsymbol{\theta}^{(t)}) / f(\boldsymbol{D}; \boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{N} f(\boldsymbol{x}_i, \boldsymbol{\alpha}_i; \boldsymbol{\theta}^{(t)}) / \prod_{i=1}^{N} f(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{N} f(\boldsymbol{\alpha}_i|\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)})$. The conditional expectation in (12) can then be written as

$$E_Z\left[\log f(\boldsymbol{x}_i, \boldsymbol{\alpha}_i; \boldsymbol{\theta})|D; \boldsymbol{\theta}^{(t)}\right] = E_{\boldsymbol{\alpha}_i}\left[\log f(\boldsymbol{x}_i, \boldsymbol{\alpha}_i\ ;\ \boldsymbol{\theta}) \,\middle|\, \boldsymbol{x}_i\ ;\ \boldsymbol{\theta}^{(t)}\right]\ . \tag{13}$$

Hence, the penalized Q function $Q_p(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is obtained by calculating the conditional expectation (13).

A.2 Algorithm

This section shows the details of the estimation procedure of the principal component model. The unknown parameters $\boldsymbol{\theta}_\mu, \Theta, \sigma^2, D$ and missing values $\boldsymbol{\alpha}_i$ are estimated by the following steps. The maximum likelihood estimators are also given by these steps with $\lambda = 0$.

Step 0. Let $\boldsymbol{\theta}^{(0)} = \{\sigma^2_{(0)}, D_{(0)}, \boldsymbol{\theta}_{\mu,(0)}, \Theta_{(0)}\}$ be an initial value of the unknown parameter $\boldsymbol{\theta} = \{\sigma^2, D, \boldsymbol{\theta}_\mu, \Theta\}$. An initial value of $\boldsymbol{\alpha}_i$ is also given by $\hat{\boldsymbol{\alpha}}_{i,(0)} = E[\boldsymbol{\alpha}_i|\boldsymbol{x}_i; \boldsymbol{\theta}^{(0)}]$.

Step 1. When we have t-th updated values of $\boldsymbol{\theta}_\mu, \Theta, \boldsymbol{\alpha}_i$, the error variance σ^2 and diagonal components of the variance covariance matrix $D = \mathrm{diag}(D_{11}, \cdots, D_{kk})$ are updated by

$$\begin{aligned} \hat{\sigma}^2_{(t+1)} &= \frac{1}{\sum n_i} \sum_{i=1}^{N} E_{\boldsymbol{\alpha}_i}[\|\boldsymbol{x}_i - \Psi_i^\nu \hat{\boldsymbol{\theta}}_{\mu,(t)} - \Psi_i^\nu \hat{\Theta}_{(t)} \boldsymbol{\alpha}_i\|^2 | \boldsymbol{x}_i\ ;\ \boldsymbol{\theta}^{(t)}] \\ &= \frac{1}{\sum n_i} \sum_{i=1}^{N} \Big[\|\boldsymbol{x}_i - \Psi_i^\nu \hat{\boldsymbol{\theta}}_{\mu,(t)} - \Psi_i^\nu \hat{\Theta}_{(t)} \hat{\boldsymbol{\alpha}}_{i,(t)}\|^2 \\ &\qquad + \mathrm{tr}\left\{ \Psi_i^\nu \hat{\Theta}_{(t)} \left(\hat{D}^{-1}_{(t)} + \hat{\Theta}'_{(t)} (\Psi_i^\nu)' \Psi_i^\nu \hat{\Theta}_{(t)} / \hat{\sigma}^2_{(t)} \right)^{-1} \hat{\Theta}'_{(t)} (\Psi_i^\nu)' \right\} \Big]\ , \\ \hat{D}_{jj,(t+1)} &= \frac{1}{N} \sum_{i=1}^{N} E_{\boldsymbol{\alpha}_i}[\alpha^2_{ij} | \boldsymbol{x}_i\ ;\ \boldsymbol{\theta}^{(t)}] \\ &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \hat{\alpha}^2_{ij,(t)} + \left(\hat{D}^{-1}_{(t)} + \hat{\Theta}'_{(t)} (\Psi_i^\nu)' \Psi_i^\nu \hat{\Theta}_{(t)} / \hat{\sigma}^2_{(t)} \right)^{-1}_{jj} \right\}\ . \end{aligned}$$

Step 2. With t-th updated values of $\sigma^2, D, \boldsymbol{\alpha}_i$, the coefficient vector $\boldsymbol{\theta}_\mu$ and matrix Θ are

updated by

$$\hat{\boldsymbol{\theta}}_\mu = \left(\sum_{i=1}^{N} (\Psi_i^\nu)' \Psi_i^\nu + N\lambda \hat{\sigma}^2_{(t)} K \right)^{-1} \sum_{i=1}^{N} (\Psi_i^\nu)' (\boldsymbol{x}_i - \Psi_i^\nu \hat{\Theta} \hat{\boldsymbol{\alpha}}_{i,(t)}) \ , \tag{14}$$

$$\mathrm{vec}\, \hat{\Theta} = \left(\sum_{i=1}^{N} \widehat{\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i'} \otimes (\Psi_i^\nu)' \Psi_i^\nu \right)^{-1} \mathrm{vec} \left(\sum_{i=1}^{N} (\Psi_i^\nu)' (\boldsymbol{x}_i - \Psi_i^\nu \hat{\boldsymbol{\theta}}_\mu) \hat{\boldsymbol{\alpha}}_{i,(t)}' \right) \ , \tag{15}$$

where equations (14) and (15) are repeated until convergence. The $(t+1)$-th updated value $\hat{\boldsymbol{\theta}}_{\mu,(t+1)}$ of $\boldsymbol{\theta}_\mu$ is obtained by the converged value of $\hat{\boldsymbol{\theta}}_\mu$. Let $\hat{\Theta}_*$ be the converged value of $\hat{\Theta}$. We then have the $(t+1)$-th updated value $\hat{\Theta}_{(t+1)}$ of $\hat{\Theta}$, using the first k eigen vectors of $\hat{\Gamma} = \hat{\Theta}_* \hat{D}_{(t)} \hat{\Theta}_*'$, where $\hat{\Gamma}$ is the reduced rank estimator of the variance covariance matrix Γ of $\Theta \boldsymbol{\alpha}_i$.

Step 3. When we have the t-th updated value $\boldsymbol{\theta}^{(t)} = \{\sigma^2_{(t)}, D_{(t)}, \boldsymbol{\theta}_{\mu,(t)}, \Theta_{(t)}\}$ of $\boldsymbol{\theta} = \{\sigma^2, D, \boldsymbol{\theta}_\mu, \Theta\}$, the random vectors $\boldsymbol{\alpha}_i$ and its cross product $\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i'$ are updated by

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_{i,(t+1)} &= E(\boldsymbol{\alpha}_i | \boldsymbol{x}_i \ ; \ \hat{\boldsymbol{\theta}}^{(t)}) \\
&= (\hat{\sigma}^2_{(t)} \hat{D}^{-1}_{(t)} + \hat{\Theta}'_{(t)} (\Psi_i^\nu)' \Psi_i^\nu \hat{\Theta}_{(t)})^{-1} \hat{\Theta}'_{(t)} (\Psi_i^\nu)' (\boldsymbol{x}_i - \Psi_i^\nu \hat{\boldsymbol{\theta}}_{\mu,(t)}) \ , \\
\widehat{\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i'}_{(t+1)} &= E(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i' | \boldsymbol{x}_i \ ; \ \hat{\boldsymbol{\theta}}^{(t)}) \\
&= \hat{\boldsymbol{\alpha}}_{i,(t+1)} \hat{\boldsymbol{\alpha}}'_{i,(t+1)} + (\hat{D}^{-1}_{(t)} + \hat{\Theta}'_{(t)} (\Psi_i^\nu)' \Psi_i^\nu \hat{\Theta}_{(t)} / \hat{\sigma}^2_{(t)})^{-1} \ .
\end{aligned}$$

Step 4. Repeat the **Steps 1-4** until convergence.

References

Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks. *Journal of Statistical Planning and Inference*, **138**(11), 3616-3633.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 1-38.

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

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

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

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-74.

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

Olshen, R. A., Biden, E. N., Wyatt, M. P. and Sutherland, D. H. (1989). Gait analysis and the bootstrap. *Annals of Statistics*, **17**, 1419-1440.

Powell, M. J. D. (1987). Radial basis functions for multivariable interpolation: A review. *Technical Report* DAMTP, Department of Applied Mathematics and Theoretical Physics, University of Campridge.

Ramsay, J. O. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association*, **95**, 9-15.

Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society, Series B*, **60**(2), 351-363.

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

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

Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

Shi, S. Y. M. and Suganthan, P. N. (2003). Feature analysis and classification of protein secondary structure data. *in Lecture Notes in Computer Science (Artificial Neural Networks and Neural Information Proceeding ICANN/ICONIP 2003, Istanbul, Turkey)*, Ed. by O.Kaynak et al. (Springer, Berlin, 2003), **2714**, 1151-1158.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577-590.

Zhou, L., Huang, J. Z., Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, **95**(3), 601-619.

List of MI Preprint Series, Kyushu University

The Grobal COE Program
Math-for-Industry Education & Research Hub

MI

MI2008-1 Takahiro ITO, Shuichi INOKUCHI & Yoshihiro MIZOGUCHI
Abstract collision systems simulated by cellular automata

MI2008-2 Eiji ONODERA
The intial value problem for a third-order dispersive flow into compact almost Hermitian manifolds

MI2008-3 Hiroaki KIDO
On isosceles sets in the 4-dimensional Euclidean space

MI2008-4 Hirofumi NOTSU
Numerical computations of cavity flow problems by a pressure stabilized characteristic-curve finite element scheme

MI2008-5 Yoshiyasu OZEKI
Torsion points of abelian varieties with values in nfinite extensions over a p-adic field

MI2008-6 Yoshiyuki TOMIYAMA
Lifting Galois representations over arbitrary number fields

MI2008-7 Takehiro HIROTSU & Setsuo TANIGUCHI
The random walk model revisited

MI2008-8 Silvia GANDY, Masaaki KANNO, Hirokazu ANAI & Kazuhiro YOKOYAMA
Optimizing a particular real root of a polynomial by a special cylindrical algebraic decomposition

MI2008-9 Kazufumi KIMOTO, Sho MATSUMOTO & Masato WAKAYAMA
Alpha-determinant cyclic modules and Jacobi polynomials

MI2008-10 Sangyeol LEE & Hiroki MASUDA
Jarque-Bera Normality Test for the Driving Lévy Process of a Discretely Observed Univariate SDE

MI2008-11 Hiroyuki CHIHARA & Eiji ONODERA
A third order dispersive flow for closed curves into almost Hermitian manifolds

MI2008-12 Takehiko KINOSHITA, Kouji HASHIMOTO and Mitsuhiro T. NAKAO
On the L^2 a priori error estimates to the finite element solution of elliptic problems with singular adjoint operator

MI2008-13 Jacques FARAUT and Masato WAKAYAMA
Hermitian symmetric spaces of tube type and multivariate Meixner-Pollaczek polynomials

MI2008-14 Takashi NAKAMURA
Riemann zeta-values, Euler polynomials and the best constant of Sobolev inequality

MI2008-15 Takashi NAKAMURA
Some topics related to Hurwitz-Lerch zeta functions

MI2009-1 Yasuhide FUKUMOTO
Global time evolution of viscous vortex rings

MI2009-2 Hidetoshi MATSUI & Sadanori KONISHI
Regularized functional regression modeling for functional response and predictors

MI2009-3 Hidetoshi MATSUI & Sadanori KONISHI
Variable selection for functional regression model via the L_1 regularization

MI2009-4 Shuichi KAWANO & Sadanori KONISHI
Nonlinear logistic discrimination via regularized Gaussian basis expansions

MI2009-5 Toshiro HIRANOUCHI & Yuichiro TAGUCHII
Flat modules and Groebner bases over truncated discrete valuation rings

MI2009-6 Kenji KAJIWARA & Yasuhiro OHTA
Bilinearization and Casorati determinant solutions to non-autonomous 1+1 dimensional discrete soliton equations

MI2009-7 Yoshiyuki KAGEI
Asymptotic behavior of solutions of the compressible Navier-Stokes equation around the plane Couette flow

MI2009-8 Shohei TATEISHI, Hidetoshi MATSUI & Sadanori KONISHI
Nonlinear regression modeling via the lasso-type regularization

MI2009-9 Takeshi TAKAISHI & Masato KIMURA
Phase field model for mode III crack growth in two dimensional elasticity

MI2009-10 Shingo SAITO
Generalisation of Mack's formula for claims reserving with arbitrary exponents for the variance assumption

MI2009-11 Kenji KAJIWARA, Masanobu KANEKO, Atsushi NOBE & Teruhisa TSUDA
Ultradiscretization of a solvable two-dimensional chaotic map associated with the Hesse cubic curve

MI2009-12 Tetsu MASUDA
Hypergeometric τ-functions of the q-Painlevé system of type $E_8^{(1)}$

MI2009-13 Hidenao IWANE, Hitoshi YANAMI, Hirokazu ANAI & Kazuhiro YOKOYAMA
A Practical Implementation of a Symbolic-Numeric Cylindrical Algebraic Decomposition for Quantifier Elimination

MI2009-14 Yasunori MAEKAWA
On Gaussian decay estimates of solutions to some linear elliptic equations and its applications

MI2009-15 Yuya ISHIHARA & Yoshiyuki KAGEI
Large time behavior of the semigroup on L^p spaces associated with the linearized compressible Navier-Stokes equation in a cylindrical domain

MI2009-16 Chikashi ARITA, Atsuo KUNIBA, Kazumitsu SAKAI & Tsuyoshi SAWABE
Spectrum in multi-species asymmetric simple exclusion process on a ring

MI2009-17 Masato WAKAYAMA & Keitaro YAMAMOTO
Non-linear algebraic differential equations satisfied by certain family of elliptic functions

MI2009-18 Me Me NAING & Yasuhide FUKUMOTO
Local Instability of an Elliptical Flow Subjected to a Coriolis Force

MI2009-19 Mitsunori KAYANO & Sadanori KONISHI
Sparse functional principal component analysis via regularized basis expansions and its application