Regularized functional regression modeling for functional response and predictors

Hidetoshi Matsui, Shuichi Kawano and Sadanori Konishi

Received on March 12, 2009

Abstract. We consider the problem of constructing a functional regression modeling with functional predictors and a functional response. Discretely observed data for each individual are expressed as a smooth function, using Gaussian basis functions. The functional regression model is estimated by the maximum penalized likelihood method, assuming that the coefficient parameters are transformed into a functional form. A crucial issue in constructing functional regression models is the selection of regularization parameters involved in the regularization method. We derive information-theoretic and Bayesian model selection criteria for evaluating the estimated model. Monte Carlo simulations and real data analysis are conducted to examine the performance of our functional regression modeling strategy.

Keywords. Basis expansion, Functional data, Model selection criteria, Regularization.

1. Introduction

Functional data analysis provides a useful tool for analyzing a data set observed at possibly differing time points for each individual, and its effectiveness has been reported in various fields of applications such as ergonomics, meteorology and chemometrics (see e.g., Ramsay and Silverman, 2002; 2005; Ferraty and Vieu, 2006). We consider the problem of constructing a functional regression model which is the functional version of the ordinary regression model.

Various model building procedures have been considered for functional regression models with functional predictors and scalar responses. Rossi *et al.* (2005) described a neural network approach and James (2002) extended the model to the generalized linear model. Furthermore, Araki *et al.* (2008) proposed the use of Gaussian basis functions along with the technique of regularization. Matsui *et al.* (2008) extended the model to the functional version of the multivariate regression model.

On the other hand, Ramsay and Dalzell (1991) considered a functional regression model which both predictor and response are given as functions, and thereafter Ramsay and Silverman (2005) considered its modeling strategy. They estimated the model by the least squares method, and then evaluated it by the squared correlation, R^2, in the framework of the functional regression model. Yao *et al.* (2005) also applied the modeling strategy to the analysis of sparse longitudinal data. Furthermore, Malfait and Ramsay (2003) and Harezlak *et al.* (2007) considered historical functional linear models which are used to model such dependencies of the response on the history of the predictor values. The models estimated by the least squares method yield unstable and/or unfavorable estimates. Moreover, R^2 considered as the goodness of fit a model is not appropriate to the prediction of newly observed data. Yamanishi and Tanaka (2003) estimated it by the weighted least squares method and evaluated it by the cross-validation.

We develop estimation and evaluation methods for functional regression models where both multiple predictors and the response are functions. Discretized observations are converted into functions, using a Gaussian basis expansion along with the technique of regularization. Advantages of Gaussian basis functions are that it can provide a useful instrument for transforming discrete observations into functional form and also be applied to analyze a set of surface fitting data. In order to obtain stable parameter estimates, a functional regression model is estimated by the maximum penalized likelihood method. Our modeling strategy yields more flexible results in terms of prediction ability.

A crucial issue in the functional regression modeling is the choice of regularization parameters involved in the method of regularization. We derive model selection criteria from an information-theoretic and Bayesian approach in order to select regularization parameters effectively. The proposed modeling strategy is applied to the analysis of meteorology data. We predict the fluctuation of annual precipitation using the information of weather data.

This paper is organized as follows. In Section 2 we introduce a functional regression model with functional predictors and a functional response. Section 3 devotes how to estimate the model. Firstly we describe the conventional estimation method and secondly propose new estimation method. We derive model selection criteria for evaluating the model constructed by our estimation procedure in Section 4. In Section 5, Monte Carlo simulations are conducted to investigate the efficiency of our modeling procedure. In Section 6 we also apply the proposed modeling strategy to the analysis of real data. Summary and concluding remarks are given in Section 7.

2. Functional regression model with functional predictors and response

Suppose we have n sets of M functional predictors and a functional response $\{(x_{\alpha m}(s),\ y_\alpha(t));\ s \in \mathcal{S}_m,\ t \in \mathcal{T},\ \alpha = 1, \ldots, n,\ m = 1, \ldots, M\}$, where $\mathcal{S}_m \subset \mathbb{R}$ and $\mathcal{T} \subset \mathbb{R}$ are, respectively, ranges of variables X_m and Y given as functions. It is assumed that both functional predictors $x_{\alpha m}(s)$ and functional responses $y_\alpha(t)$ can be, respectively, expressed as smooth functions

$$(1)\quad \begin{aligned} x_{\alpha m}(s) &= \sum_{j=1}^{J_m} \tilde{c}_{\alpha mj}\phi_{mj}(s) = \tilde{\boldsymbol{c}}_{\alpha m}^T \boldsymbol{\phi}_m(s), \\ y_\alpha(t) &= \sum_{k=1}^{K} \tilde{d}_{\alpha k}\psi_k(t) = \tilde{\boldsymbol{d}}_\alpha^T \boldsymbol{\psi}(t), \end{aligned}$$

where $\tilde{\boldsymbol{c}}_{\alpha m} = (\tilde{c}_{\alpha m1}, \ldots, \tilde{c}_{\alpha mJ_m})^T$ and $\tilde{\boldsymbol{d}}_\alpha = (\tilde{d}_{\alpha 1}, \ldots, \tilde{d}_{\alpha K})^T$ are coefficient vectors, $\boldsymbol{\phi}_m(s) = (\phi_{m1}(s),\ \ldots,\ \phi_{mJ_m}(s))^T$ and $\boldsymbol{\psi}(t) = (\psi_1(t),\ \ldots,\ \psi_K(t))^T$ are vectors of basis functions. Here we use Gaussian basis functions, due to Kawano and Konishi (2007), given as follows:

$$(2)\quad \begin{aligned} \phi_{mj}(s) &= \exp\left\{-\frac{(s-\tau_{j+2}^{(m)})^2}{2h_m^2}\right\}, \\ \psi_k(t) &= \exp\left\{-\frac{(t-\tau_{k+2})^2}{2h^2}\right\}, \end{aligned}$$

where $\tau_j^{(m)}$ and τ_k are equally spaced knots so that the $\tau_j^{(m)}$ satisfy $\tau_1^{(m)} < \ldots < \tau_4^{(m)} = \min(s) < \ldots < \tau_{J+2}^{(m)} = \max(s) < \ldots < \tau_{J+4}^{(m)}$ and τ_k similarly, $h_m = (\tau_{j+2}^{(m)} - \tau_j^{(m)})/3$ and $h = (\tau_{k+2} - \tau_k)/3$. Coefficients $\tilde{\boldsymbol{c}}_{\alpha m}$ and $\tilde{\boldsymbol{d}}_\alpha$ are obtained by smoothing techniques described in Appendix A.

In order to model the relationship between predictors and a response, we consider the following functional regression model (Ramsay and Silverman, 2005; Shimokawa *et al.*, 2000):

$$(3)\ y_\alpha(t) = \beta_0(t) + \sum_{m=1}^{M} \int_{\mathcal{S}_m} x_{\alpha m}(s)\beta_m(s,t)ds + \varepsilon_\alpha(t),$$

where $\beta_0(t)$ is a parameter function, $\beta_m(s,t)$ are bivariate coefficient functions which impose varying weights on $x_{\alpha m}(s)$ at arbitrary time $t \in \mathcal{T}$, and $\varepsilon_\alpha(t)$ are error functions. Using the same basis functions as those used for the predictor and response functions, we express the coefficient functions $\beta_m(s,t)$ as follows:

$$(4)\ \ \beta_m(s,t) = \sum_{j,k} \phi_{mj}(s) b_{mjk}\psi_k(t) = \boldsymbol{\phi}_m^T(s) B_m \boldsymbol{\psi}(t),$$

where $B_m = (b_{mjk})_{j,k}$ are $J_m \times K$ coefficient matrices.

The function $\beta_0(t)$ plays the role of a constant term in the standard regression model. Here, we eliminate it by centering the functional regression model (3) for the subsequent estimation procedure. Centered predictors $x_{\alpha m}^*(s)$ and responses $y_\alpha^*(t)$ are, respectively, obtained by

$$\begin{aligned} x_{\alpha m}^*(s) &= x_{\alpha m}(s) - \bar{x}_m(s) \\ &= \tilde{\boldsymbol{c}}_{\alpha m}^T \boldsymbol{\phi}_m(s) - \bar{\boldsymbol{c}}_m^T \boldsymbol{\phi}_m(s) \\ &= \boldsymbol{c}_{\alpha m}^T \boldsymbol{\phi}_m(s), \\ y_\alpha^*(t) &= y_\alpha(t) - \bar{y}(t) \\ &= \tilde{\boldsymbol{d}}_\alpha^T \boldsymbol{\psi}(t) - \bar{\boldsymbol{d}}^T \boldsymbol{\psi}(t) \\ &= \boldsymbol{d}_\alpha^T \boldsymbol{\psi}(t), \end{aligned}$$

where $\boldsymbol{c}_{\alpha m} = \tilde{\boldsymbol{c}}_{\alpha m} - \bar{\boldsymbol{c}}_m$ and $\boldsymbol{d}_\alpha = \tilde{\boldsymbol{d}}_\alpha - \bar{\boldsymbol{d}}$ with $\bar{\boldsymbol{c}}_m = \sum_\alpha \tilde{\boldsymbol{c}}_{\alpha m}/n$ and $\bar{\boldsymbol{d}} = \sum_\alpha \tilde{\boldsymbol{d}}_\alpha/n$. Then (3) can be rewritten in the form

$$(5)\quad y_\alpha^*(t) = \sum_{m=1}^{M} \int_{\mathcal{S}_m} x_{\alpha m}^*(s)\beta_m(s,t)ds + \varepsilon_\alpha^*(t),$$

where $\varepsilon_\alpha^*(t) = \varepsilon_\alpha(t) - \bar{\varepsilon}(t)$. It follows from equations (1) and (4) that the functional regression model (5) can be expressed as

$$(6)\quad \begin{aligned} \boldsymbol{d}_\alpha^T \boldsymbol{\psi}(t) &= \sum_{m=1}^{M} \boldsymbol{c}_{\alpha m}^T W_{\phi_m} B_m \boldsymbol{\psi}(t) + \varepsilon_\alpha^*(t) \\ &= \boldsymbol{z}_\alpha^T B \boldsymbol{\psi}(t) + \varepsilon_\alpha^*(t), \end{aligned}$$

where $\boldsymbol{z}_\alpha = (\boldsymbol{c}_{\alpha 1}^T W_{\phi_1}, \ldots, \boldsymbol{c}_{\alpha M}^T W_{\phi_M})^T$ with $W_{\phi_m} = \int \boldsymbol{\phi}_m(s) \boldsymbol{\phi}_m^T(s)ds$ and $B = (B_1^T, \ldots, B_M^T)^T$. When we use the Gaussian basis functions given in (2), (j,k)-th elements of W_{ϕ_m} can be easily calculated and are given by

$$W_{\phi_m}^{(j,k)} = \sqrt{\pi h_m^2} \exp\left\{-\frac{(\tau_{j+2}^{(m)} - \tau_{k+2}^{(m)})^2}{4h_m^2}\right\}.$$

From equation (6), the problem of estimating the coefficient functions $\beta_m(s,t)$ in (3) is replaced by the problem of estimating the parameter matrix B.

3. Estimation

We consider the problem of estimating the parameter matrix B in the functional regression model (6). First we describe the least squares method, and then propose the maximum likelihood and maximum penalized likelihood method.

3.1. Least squares method

Ramsay and Silverman (2005) and Shimokawa *et al.* (2000) estimated B in the model (6) by minimizing the integrated residual sum of squares given by

(7)
$$\begin{aligned}&\sum_{\alpha=1}^{n}\int_{\mathcal{T}}\left[y_\alpha^*(t)-\sum_{m=1}^{M}\int_{\mathcal{S}_m}x_{\alpha m}^*(s)\beta_m(s,t)ds\right]^2 dt\\&=\int_{\mathcal{T}}\mathrm{tr}\left\{\left(D\boldsymbol{\psi}(t)-ZB\boldsymbol{\psi}(t)\right)\left(D\boldsymbol{\psi}(t)-ZB\boldsymbol{\psi}(t)\right)^T\right\}dt\\&=\mathrm{tr}\left\{\left(D-ZB\right)W_\psi\left(D-ZB\right)^T\right\},\end{aligned}$$

where $D=(\boldsymbol{d}_1^*,\ldots,\boldsymbol{d}_n^*)^T$, $Z=(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)^T$ and $W_\psi=\int_{\mathcal{T}}\boldsymbol{\psi}(t)\boldsymbol{\psi}^T(t)dt$. The least squares estimator $\hat{B}$ is then given by

$$\mathrm{vec}(\hat{B})=(W_\psi\otimes Z^TZ)^{-1}\mathrm{vec}(Z^TDW_\psi),$$

where vec$(\cdot)$ is an operator that transforms the columnwise vectors of a matrix into a vector and $\otimes$ represents a Kronecker product. When we use Gaussian basis functions (2), W_ψ is nonsingular and $\hat{B}$ can be expressed as

$$\hat{B}=(Z^TZ)^{-1}Z^TD. \tag{8}$$

This estimate has the same form as a least squares estimator for ordinary multivariate regression models with a design matrix Z and a response matrix D.

3.2. Maximum likelihood method

We consider estimating the functional regression model (6) by the maximum likelihood method. Suppose error functions $\varepsilon_\alpha^*(t)$ are represented by a linear combination of basis functions $\psi_k(t)$, which are the same as those for the response functions $y_\alpha^*(t)$, that is,

$$\varepsilon_\alpha^*(t)=\sum_{k=1}^{K}e_{\alpha k}\psi_k(t)=\boldsymbol{e}_\alpha^T\boldsymbol{\psi}(t), \tag{9}$$

where the K-dimensional vectors $\boldsymbol{e}_\alpha=(e_{\alpha 1},\ldots,e_{\alpha K})^T$ are assumed to be independent and identically normally distributed with mean vector $\mathbf{0}$ and variance-covariance matrix Σ. Then the functional regression model (6) can be represented as

$$\boldsymbol{d}_\alpha^T\boldsymbol{\psi}(t)=\boldsymbol{z}_\alpha^TB\boldsymbol{\psi}(t)+\boldsymbol{e}_\alpha^T\boldsymbol{\psi}(t),\quad \boldsymbol{e}_\alpha\overset{i.i.d}{\sim}N_K(\mathbf{0},\Sigma). \tag{10}$$

By multiplying the equation (10) by $\boldsymbol{\psi}^T(t)$ and then integrating with respect to $\mathcal{T}$, it can be rewritten as

$$\boldsymbol{d}_\alpha^TW_\psi=\boldsymbol{z}_\alpha^TBW_\psi+\boldsymbol{e}_\alpha^TW_\psi. \tag{11}$$

Since W_ψ is nonsingular, we obtain

$$\boldsymbol{d}_\alpha=B^T\boldsymbol{z}_\alpha+\boldsymbol{e}_\alpha,\quad \boldsymbol{e}_\alpha\overset{i.i.d}{\sim}N_K(\mathbf{0},\Sigma), \tag{12}$$

which has the same form as a multivariate regression model with predictors $\boldsymbol{z}_\alpha$ and responses $\boldsymbol{d}_\alpha$.

From (12) the model for a functional response y_α given a functional predictor $\boldsymbol{x}_\alpha$ can be expressed as a probability density function as follows:

(13)
$$\begin{aligned}f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta})&=\frac{1}{(2\pi)^{K/2}|\Sigma|^{1/2}}\\&\quad\times\exp\left\{-\frac{1}{2}(\boldsymbol{d}_\alpha-B^T\boldsymbol{z}_\alpha)^T\Sigma^{-1}(\boldsymbol{d}_\alpha-B^T\boldsymbol{z}_\alpha)\right\},\end{aligned}$$

where $\boldsymbol{\theta}=\{B,\Sigma\}$ is a parameter vector. Therefore, maximum likelihood estimators of B and Σ are, respectively, given by

$$\hat{B}=(Z^TZ)^{-1}Z^TD,\quad \hat{\Sigma}=\frac{1}{n}(D-Z\hat{B})^T(D-Z\hat{B}).$$

Comparing this result with (8), we find that the maximum likelihood estimator of B coincides with the least squares estimator.

3.3. Maximum penalized likelihood method

Since least squares or maximum likelihood method often results in unstable estimators, we consider estimating the functional regression model, using the regularization method. It follows from (13) that the penalized log-likelihood function is given by

$$\begin{aligned}l_\lambda(\boldsymbol{\theta})=\sum_{\alpha=1}^{n}&\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta})\\&-\frac{n}{2}\mathrm{tr}\left\{B^T(\Lambda_M\odot\Omega)B\right\},\end{aligned} \tag{14}$$

where Λ_M is a $(\sum_m J_m)\times(\sum_m J_m)$ matrix of regularization parameters $\lambda_1,\ldots,\lambda_M$ that control a variation of B, that is, $\Lambda_M=\boldsymbol{\lambda}_M\boldsymbol{\lambda}_M^T$ with $\boldsymbol{\lambda}_M=(\sqrt{\lambda_1}\mathbf{1}_{J_1}^T,\ldots,\sqrt{\lambda_M}\mathbf{1}_{J_M}^T)^T$. The notation $\odot$ represents the Hadamard product and Ω is a $(\sum_m J_m)\times(\sum_m J_m)$ positive semi-definite matrix. Maximizing the function (14), maximum penalized likelihood estimators $\hat{B}$, $\hat{\Sigma}$ are respectively given by

(15)
$$\begin{aligned}\mathrm{vec}(\hat{B})&=\left(\hat{\Sigma}^{-1}\otimes Z^TZ+nI_K\otimes(\Lambda_M\odot\Omega)\right)^{-1}\\&\qquad\times(\hat{\Sigma}^{-1}\otimes Z^T)\mathrm{vec}(D),\\\hat{\Sigma}&=\frac{1}{n}(D-Z\hat{B})^T(D-Z\hat{B}).\end{aligned}$$

Since $\hat{B}$ and $\hat{\Sigma}$ depend on each other, we provide an initial value for the variance covariance matrix; then they are updated until convergence. Therefore, the maximum penalized likelihood estimator of D is given by

$$\begin{aligned}\mathrm{vec}(\hat{D})&=\mathrm{vec}(Z\hat{B})\\&=S_\lambda\mathrm{vec}(D),\end{aligned} \tag{16}$$

where $S_\lambda=(I_K\otimes Z)(\hat{\Sigma}^{-1}\otimes Z^TZ+nI_K\otimes(\Lambda_M\odot\Omega))^{-1}(\hat{\Sigma}^{-1}\otimes Z^T)$ is a hat matrix for vec(D). Substituting the maximum penalized likelihood estimator $\hat{\boldsymbol{\theta}} = \{\hat{B}, \hat{\Sigma}\}$ into (13) we obtain the statistical model

$$f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{K/2}|\hat{\Sigma}|^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{d}_\alpha - \hat{B}^T\boldsymbol{z}_\alpha)^T\hat{\Sigma}^{-1}(\boldsymbol{d}_\alpha - \hat{B}^T\boldsymbol{z}_\alpha)\right\}. \tag{17}$$

4. Model selection criteria

Since the statistical model (17) estimated by the regularization method depends on the regularization parameters $\lambda_1, \dots, \lambda_M$, selection of these values is an important issue. Although cross-validation is widely used for the regularization parameter selection, the computational time is very large and high variability and tendency to undersmooth are not negligible in the analysis of functional data. We derive model selection criteria for evaluating the functional regression model. We select the model that minimizes the values of these criteria and then consider the corresponding model to be the optimal model. In Section 5 Monte Carlo simulations are conducted to compare the proposed criteria.

(1) Generalized cross validation

Generalized cross validation (GCV; Craven and Wahba, 1979) for evaluating the functional regression model (17) is obtained by applying the hat matrix S_λ given in (16), that is,

$$\text{GCV} = \frac{\text{tr}\left\{(D - ZB)^T(D - ZB)\right\}}{nK\left(1 - \text{tr}(S_\lambda)/(nK)\right)^2}.$$

(2) Modified AIC

Hastie and Tibshirani (1990) modified the AIC (Akaike, 1973) for evaluating the model estimated by the regularization method by substituting a trace of the hat matrix for the number of degrees of freedom, since the hat matrix can be viewed as a measure of the complexity of the model estimated by the regularization method. Using this result, the modified AIC for evaluating (17) is given by

$$\text{mAIC} = -2\sum_{\alpha=1}^{n}\log f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}) + 2\text{tr}(S_\lambda).$$

A problem may arise in the theoretical justification for the use of the bias-correction terms in MAIC, since AIC covers only models estimated by the maximum likelihood method.

(3) Generalized information criterion

Imoto and Konishi (2003) derived an information criterion GIC (Konishi and Kitagawa, 1996) for evaluating a statistical model estimated by the maximum penalized likelihood method. Using this result, the GIC for evaluating the model (17) is given by

$$\text{GIC} = -2\sum_{\alpha=1}^{n}\log f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}) + 2\text{tr}\{R_\lambda(\hat{\boldsymbol{\theta}})^{-1}Q_\lambda(\hat{\boldsymbol{\theta}})\},$$

where $R_\lambda(\boldsymbol{\theta})$, $Q_\lambda(\boldsymbol{\theta})$ are, respectively, given by

$$R_\lambda(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\{\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) - \frac{1}{2}\text{tr}\left\{B^T(\Lambda_M\odot\Omega)B\right\}\right\}, \tag{18}$$

$$Q_\lambda(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\{\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) - \frac{1}{2}\text{tr}\left\{B^T(\Lambda_M\odot\Omega)B\right\}\right\}\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}).$$

(4) Generalized Bayesian information criterion

The Bayesian information criterion (BIC) has been proposed by Schwarz (1978), from the viewpoint of Bayesian inference, based on the idea of maximizing the posterior probability of candidate models. However, the BIC only covers models estimated by the maximum likelihood method. Konishi *et al.* (2004) extended the BIC so that it could be used for evaluating models fitted by the maximum penalized likelihood method, thus deriving GBIC. We derive the GBIC for evaluating the model (17) fitted by the maximum penalized likelihood method, which is given by

$$\begin{aligned}\text{GBIC} = &-2\sum_{\alpha=1}^{n}\log f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}) + n\text{tr}\left\{B^T(\Lambda_M\odot\Omega)B\right\} \\ &+ (r + Kq)\log n - (r + Kq)\log(2\pi) \\ &- K\log|\Lambda_M\odot\Omega|_+ + \log|R_\lambda(\hat{\boldsymbol{\theta}})|,\end{aligned} \tag{19}$$

where $q = p - \text{rank}(\Omega)$, $p = \sum_m J_m$, $r = K(K+1)/2$ and $|\cdot|_+$ denotes the product of the non-zero eigenvalues of a matrix. The derivation of GBIC in (19) is given in Appendix B.

5. Numerical examples

Monte Carlo simulations are conducted to investigate the effectiveness of the proposed modeling strategy. We simulated n sets of a functional predictor and a response $\{(x_\alpha(s), y_\alpha(t));\ s \in \mathcal{S},\ t \in \mathcal{T},\ \alpha = 1, \dots, n\}$, then applied the functional regression modeling. Data sets are generated in two steps; First, $x_{\alpha i}$ $(i = 1, \dots, 50)$ corresponding to the predictor X at observational points s_i are generated by the following rule:

$$x_{\alpha i} = u_\alpha(s_i) + \varepsilon_{\alpha i}, \quad \varepsilon_{\alpha i} \sim N(0,1), \quad s_i \sim U(-1,1).$$

We assume $u_\alpha(s)$ as following settings:

$$\begin{aligned}&\text{(a)} \quad u_\alpha(s) = \exp(a_{1\alpha}s) + a_{2\alpha}s, \\ &\qquad a_{1\alpha} \sim N(2, 0.2^2), \quad a_{2\alpha} \sim N(-3, 0.3^2), \\ &\text{(b)} \quad u_\alpha(s) = b_{1\alpha} + b_{2\alpha}s + b_{3\alpha}s^2 + b_{4\alpha}s^3, \\ &\qquad b_{1\alpha} \sim N(0.2, 0.1^2), \quad b_{2\alpha} \sim N(0.4, 0.2^2), \\ &\qquad b_{3\alpha} \sim N(0.1, 0.08^2), \quad b_{4\alpha} \sim N(0.4, 0.1^2).\end{aligned}$$

Table 1: Comparisons of the average mean squared errors (AMSE) based on various criteria for the simulation (a).

$n = 25$		GCV	mAIC	GIC	GBIC
$\rho = 1.5$	AMSE ($\times 10^{-1}$)	2.418	3.032	2.638	**2.301**
	SD ($\times 10^{-1}$)	1.013	1.337	1.315	0.943
$\rho = 1$	AMSE ($\times 10^{-1}$)	1.650	2.032	1.759	**1.623**
	SD	7.267	9.804	8.299	7.260
$\rho = 0.5$	AMSE ($\times 10^{-1}$)	1.153	1.324	1.183	**1.121**
	SD	4.023	4.336	4.201	3.981
$\rho = 0.1$	AMSE	6.371	**4.915**	5.795	6.571
	SD	2.216	1.284	1.807	2.225
$n = 50$		GCV	mAIC	GIC	GBIC
$\rho = 1.5$	AMSE ($\times 10^{-1}$)	1.391	1.428	1.561	**1.368**
	SD	4.990	5.369	6.435	4.757
$\rho = 1$	AMSE ($\times 10^{-1}$)	1.063	**1.034**	1.120	1.042
	SD	3.545	3.659	4.024	3.553
$\rho = 0.5$	AMSE	8.337	**7.540**	7.659	8.248
	SD	2.054	1.763	1.915	2.065
$\rho = 0.1$	AMSE	5.883	**3.673**	4.114	5.739
	SD	1.562	1.102	1.575	1.354

Table 2: Comparisons of the average mean squared errors (AMSE) based on various criteria for the simulation (b).

$n = 25$		GCV	mAIC	GIC	GBIC
$\rho = 1.5$	AMSE ($\times 10^{-1}$)	2.110	2.812	2.192	**1.955**
	SD ($\times 10^{-1}$)	1.123	1.516	1.039	0.957
$\rho = 1$	AMSE ($\times 10^{-1}$)	1.409	1.803	1.671	**1.337**
	SD	6.119	9.208	9.821	5.554
$\rho = 0.5$	AMSE ($\times 10^{-1}$)	0.868	1.095	0.973	**0.826**
	SD	3.695	5.043	4.279	3.393
$\rho = 0.1$	AMSE	2.732	3.427	2.908	**2.637**
	SD ($\times 10^{1}$)	8.043	11.08	9.357	6.572
$n = 50$		GCV	mAIC	GIC	GBIC
$\rho = 1.5$	AMSE ($\times 10^{-1}$)	1.019	1.047	1.271	**0.985**
	SD	4.572	4.375	6.102	4.087
$\rho = 1$	AMSE	7.403	7.858	9.945	**7.303**
	SD	2.927	3.193	3.750	2.802
$\rho = 0.5$	AMSE	4.573	4.634	5.443	**4.408**
	SD	1.543	1.531	1.826	1.405
$\rho = 0.1$	AMSE	2.074	2.098	2.194	**2.045**
	SD ($\times 10^{1}$)	4.654	4.992	5.105	4.439

Second, $y_{\alpha j}$ $(j = 1, \ldots, 50)$ corresponding to the functional response Y at design points t_j are generated as follows:

$$\begin{aligned} y_{\alpha j} =&\ v_\alpha(t_j) + \varepsilon_{\alpha j}, \quad \varepsilon_{\alpha j} \sim N(0,1), \quad t_j \sim U(-1,1), \\ v_\alpha(t) &= g_\alpha(t) + \varepsilon_\alpha(t), \\ g_\alpha(t) &= \int_{\mathcal{S}} u_\alpha(s)\beta(s,t)ds, \quad \varepsilon_\alpha(t) = \boldsymbol{e}_\alpha^T \boldsymbol{\psi}(t), \end{aligned}$$

where $\mathcal{S} = [-1, 1]$, $\boldsymbol{\psi}(t)$ are Gaussian basis functions and the coefficients $\boldsymbol{e}_\alpha$ are assumed to be independently distributed according to multivariate normal distributions $N(\mathbf{0}, \Sigma)$ with $\Sigma = (0.5^{|k-l|}\rho)_{k,l}$. The coefficient functions $\beta(s,t)$ are given by

$$\text{(a) } \beta(s,t) = s^2 + t^2, \quad \text{(b) } \beta(s,t) = s + t^3.$$

As a first step of the analysis, we converted $x_{\alpha i}$ and $y_{\alpha j}$ into functional data $x_\alpha(s)$ and $y_\alpha(t)$ respectively by the smoothing method. The number of basis functions is supposed to be 10. Next, we constructed a functional regression model and then estimated the model by the maximum penalized likelihood method. Maximum likelihood failed to provide estimates in these cases because of the degeneracy. In order to compare the effectiveness of our modeling procedures, four model selection criteria, GCV, mAIC, GIC and GBIC, are used for evaluating the estimated model. We repeated this strategy for 100 times, then derived 100 averages of mean squared errors AMSE $= \sum_\alpha \sum_i (g_\alpha(t_i) - \hat{y}_\alpha(t_i))^2/n$.

Table 1 and 2 show results of simulation examples, where bold numbers indicate the minimum AMSEs among four criteria. The values of SD indicate standard deviations for the AMSE. It may be seen from these tables that the models evaluated by GBIC are superior to those evaluated by other model selection criteria in most situations in the sense of minimizing AMSEs, especially when the variance parameter ρ is large.

6. Real data example

In this section we apply the proposed functional regression modeling strategy to the analysis of Japanese weather data, predicting the variation of monthly precipitation.

Weather data, available on Chronological Scientific Tables 2005, are recorded from January to December at 79 weather stations in Japan, including the annual monthly average temperature, monthly total times of daylight and monthly total precipitation. These data are averaged over the values obtained from 1971 to 2000. We consider predicting monthly total precipitation using the temperature and times of daylight. For daylight and precipitation data we used the logarithms of observed data.

We performed some pre-processings before applying functional regression modeling. First, we obtained functional data sets by smoothing the data via regularized Gaussian basis function expansion. The resulting functional data sets are shown in Figure 1. Next, the 79 observed data sets were randomly divided into 45 training data sets and 34 test data sets. The training data were centered by subtracting the sample average. We treated temperature and daylight functions as predictors and the precipitation function as a response, thereby constructing a functional regression model.

The model was estimated by the maximum likelihood and maximum penalized likelihood method; four model selection criteria were then used to evaluate the model for

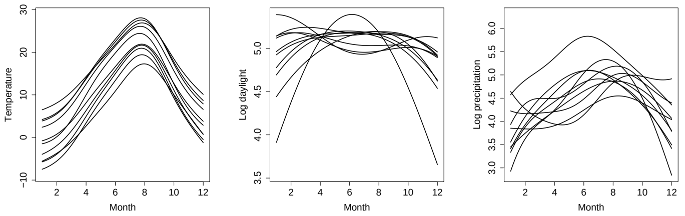

Figure 1: Examples of weather data converted into functions.

Table 3: Results on the analysis of weather data. λ_1 and λ_2 are regularization parameters selected by each model selection criterion.

	MLE	GCV	mAIC	GIC	GBIC
λ_1 $(\times 10^2)$	—	2.51	3.98	5.01	89.1
λ_2 $(\times 10^1)$	—	1.26	1.78	1.78	8.91
Test error $(\times 10^2)$	7.25	5.98	5.90	5.83	**5.37**

maximum penalized likelihood estimates. We used the average squared errors between the smoothed test data and the predicted functional data at 100 time points as the test error.

Table 3 shows regularization parameters for temperature (λ_1) and daylight (λ_2) selected by each model selection criterion and test errors of corresponding models. From these results we observe that the maximum penalized likelihood method is superior to the maximum likelihood method in prediction accuracy. In particular, for the four model selection criteria, GBIC minimized the test error.

Figure 3 shows the results of fitting eight weather stations with the test set. These figures reveal that the predicted functions captured the original data well. The estimated coefficient functions of each predictor are shown in Figure 2. This figure shows that while the temperature around January and the times of daylight around October have negative weights, the temperature at the end of the year and the times of daylight around March have a positive weight for predicting the precipitation. Therefore, if the former values increase the precipitation decreases, and if the latter values increase the precipitation increases.

7. Summary and concluding remarks

We proposed a functional regression modeling with functional predictors and a functional response, using Gaussian basis functions along with the technique of regularization. First discretely observed data for individuals were transformed to a set of smooth functions. Second the functional regression model was constructed, using the method of regularization and also the property that the integral of the product of any two Gaussian basis functions can be directly calculated. We applied the proposed modeling strategy to the analysis of weather data, predicting response functions rather than scalars. The simulation results and the analysis

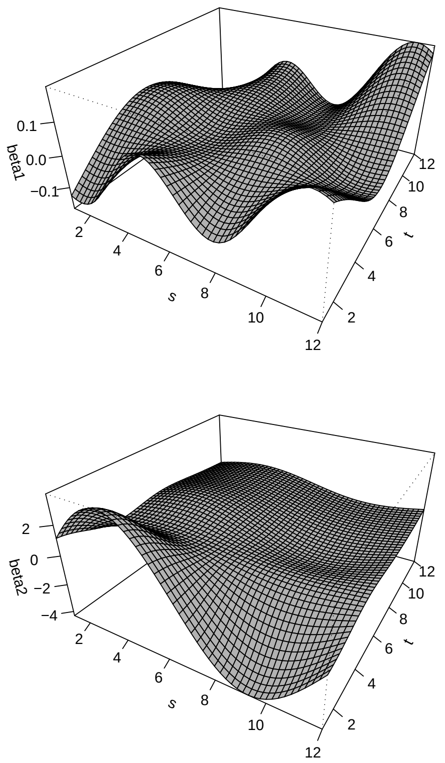

Figure 2: Estimated coefficient functions corresponding to temperature (top) and daylight (bottom).

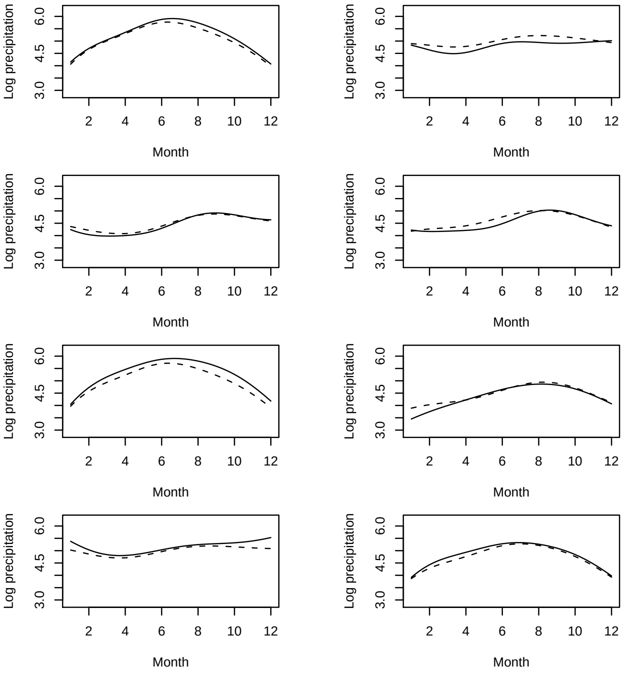

Figure 3: Results on fitting the test data for 8 stations. Solid lines show smoothed test data and dashed lines show the predicted functions by the functional regression model.

of real data showed that our modeling procedure performs well, especially in terms of its flexibility and stability.

Recently electronic measurement technologies enable us to collect large amounts of various types of data in the fields of natural science. In order to extract useful information from such data with complex structure, nonlinear modeling techniques are required. Further work remains to be done towards constructing nonlinear functional regression modeling.

Acknowledgement

The authors would like to thank the referee for his helpful comments and suggestions.

Appendix

A. Converting discrete data to functional data

Since data are generally obtained discretely, we need to explain these data as functions. We use a smoothing method via regularized basis expansions for converting raw data into functional data. In this section we only refer to the predictor, however, same is true of the response.

Suppose we have n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where each $\boldsymbol{x}_\alpha$ are vectors of N_α observations $\{x_{\alpha 1}, \ldots, x_{\alpha N_\alpha};\ \alpha = 1, \ldots, n\}$ at $\{s_{\alpha 1}, \ldots, s_{\alpha N_\alpha};\ s_{\alpha i} \in \mathcal{S},\ i = 1, \ldots, N_\alpha\}$, where $\mathcal{S} \subset \mathbb{R}$ is a range of design points $s_{\alpha 1}, \ldots, s_{\alpha N_\alpha}$. It is assumed that $x_{\alpha i}$s are given by adding Gaussian noises $\varepsilon_{\alpha i}$ to unknown smooth functions $u_\alpha(s)$ at $s_{\alpha i}$, that is,

$$x_{\alpha i} = u_\alpha(s_{\alpha i}) + \varepsilon_{\alpha i}, \quad i = 1, \ldots, N_\alpha, \tag{20}$$

where $\varepsilon_{\alpha i}$ are independently normally distributed with mean 0 and variance $\sigma^2_{x\alpha}$.

Nonlinear functions $u_\alpha(s)$ are supposed to be represented by the basis expansion such as

$$u_\alpha(s) = \sum_{j=1}^{J} c_{\alpha j}\phi_j(s) = \boldsymbol{c}_\alpha^T \boldsymbol{\phi}(s), \tag{21}$$

where $\boldsymbol{c}_\alpha = (c_{\alpha 1}, \ldots, c_{\alpha J})^T$ are vectors of coefficient parameters and $\boldsymbol{\phi}(s) = (\phi_1(s), \ldots, \phi_J(s))^T$ are vectors of basis functions. We assume that basis functions $\phi_j(s)$ $(j = 1, \ldots, J)$ are Gaussian basis functions defined in (2). From these results the regression model (20) has a probability density function

$$f(x_{\alpha i}|s_{\alpha i}; \boldsymbol{c}_\alpha, \sigma^2_{x\alpha}) = \frac{1}{\sqrt{2\pi\sigma^2_{x\alpha}}} \exp\left\{-\frac{(x_{\alpha i} - \boldsymbol{c}_\alpha^T\boldsymbol{\phi}(s_{\alpha i}))^2}{2\sigma^2_{x\alpha}}\right\}.$$

The parameters $\boldsymbol{c}_\alpha$ and $\sigma^2_{x\alpha}$ are estimated by using the maximum penalized likelihood method, which maximizes a penalized log-likelihood function

$$l_{\zeta_\alpha}(\boldsymbol{c}_\alpha, \sigma^2_{x\alpha}) = \sum_{i=1}^{N_\alpha} \log f(x_{\alpha i}|s_{\alpha i}; \boldsymbol{c}_\alpha, \sigma^2_{x\alpha}) - \frac{N_\alpha \zeta_\alpha}{2}\boldsymbol{c}_\alpha^T \Omega \boldsymbol{c}_\alpha,$$

where ζ_α are smoothing parameters which adjust the smoothness of the estimated function, and Ω is a $J \times J$ positive semi-definite matrix. The maximum penalized likelihood estimators $\hat{\boldsymbol{c}}_\alpha$ and $\hat{\sigma}^2_{x\alpha}$ are, respectively, given by

$$\begin{aligned} \hat{\boldsymbol{c}}_\alpha &= (\Phi_\alpha^T\Phi_\alpha + N_\alpha\zeta_\alpha\hat{\sigma}^2_{x\alpha}\Omega)^{-1}\Phi_\alpha^T\boldsymbol{x}_\alpha, \\ \hat{\sigma}^2_{x\alpha} &= \frac{1}{N_\alpha}(\boldsymbol{x}_\alpha - \Phi_\alpha\hat{\boldsymbol{c}}_\alpha)^T(\boldsymbol{x}_\alpha - \Phi_\alpha\hat{\boldsymbol{c}}_\alpha), \end{aligned} \tag{22}$$

where $\Phi_\alpha = (\boldsymbol{\phi}(s_{\alpha 1}), \ldots, \boldsymbol{\phi}(s_{\alpha N_\alpha}))^T$.

The maximum penalized likelihood estimates based on Gaussian basis functions depend on the regularization parameters ζ_α and the number of basis functions J. For the choice of these parameters some model selection criteria are used. Details are referred to Konishi and Kitagawa (2008). Selecting appropriate values of ζ_α and J, leading to appropriate estimates $\hat{u}_\alpha(s)$. Therefore we obtain functional data

$$x_\alpha(s) \equiv \hat{u}_\alpha(s) = \hat{\boldsymbol{c}}_\alpha^T\boldsymbol{\phi}(s). \tag{23}$$

We use a set of functions $\{x_\alpha(s);\ s \in \mathcal{S},\ \alpha = 1, \ldots, n\}$ as data instead of observed data set $\{(s_{\alpha i}, x_{\alpha i});\ i = 1, \ldots, N_\alpha,\ \alpha = 1, \ldots, n\}$.

B. Derivation of GBIC

We show the derivation of the model selection criterion GBIC in (19) for evaluating the functional regression model estimated by the regularization method.

The penalized log-likelihood function (14) is rewritten as

$$\begin{aligned} l_\Lambda(B, \Sigma) &= \log\left\{f(\boldsymbol{y}|\boldsymbol{\theta})\exp\left[-\frac{n}{2}\mathrm{tr}\{B^T(\Lambda_M \odot \Omega)B\}\right]\right\} \\ &= \log\left\{f(\boldsymbol{y}|\boldsymbol{\theta})\prod_{k=1}^{K}\exp\left[-\frac{n}{2}\boldsymbol{b}_{(k)}^T(\Lambda_M \odot \Omega)\boldsymbol{b}_{(k)}\right]\right\}, \end{aligned} \tag{24}$$

where $\log f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_\alpha \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{\theta})$. We set the prior density of $\boldsymbol{\theta}$ as a product of K multivariate normal distribution, that is,

$$\begin{aligned} \pi(\boldsymbol{\theta}|\Lambda_M) = \prod_{k=1}^{K} & \frac{n^{(p-q)/2}|\Lambda_M \odot \Omega|_+^{1/2}}{(2\pi)^{(p-q)/2}} \\ & \times \exp\left[-\frac{n}{2}\boldsymbol{b}_{(k)}^T(\Lambda_M \odot \Omega)\boldsymbol{b}_{(k)}\right]. \end{aligned} \tag{25}$$

Then the marginal likelihood of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ with prior distribution (25) can be expressed as

$$\begin{aligned} p(\boldsymbol{y}|\Lambda_M) &= \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\Lambda_M)d\boldsymbol{\theta} \\ &= \int \exp\left[n \times \frac{1}{n}\log\{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\Lambda_M)\}\right]d\boldsymbol{\theta} \\ &= \int \exp\{nq(\boldsymbol{\theta}|\Lambda_M)\}\,d\boldsymbol{\theta}, \end{aligned} \tag{26}$$

where $q(\boldsymbol{\theta}|\Lambda_M) = \log\{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\Lambda_M)\}/n$.

A Taylor series expansion of $q(\boldsymbol{\theta}|\Lambda_M)$ around $\hat{\boldsymbol{\theta}}$, the maximum penalized likelihood estimator of $\boldsymbol{\theta}$, is given by

(27)
$$q(\boldsymbol{\theta}|\Lambda_M) = q(\hat{\boldsymbol{\theta}}|\Lambda_M) - \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T R_\lambda(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \cdots$$

since $\partial q(\hat{\boldsymbol{\theta}}|\Lambda_M)/\partial\boldsymbol{\theta} = \mathbf{0}$. Substituting (27) into (26), we obtain the following Laplace approximation

(28)
$$\begin{aligned}
&\int \exp\{nq(\boldsymbol{\theta}|\Lambda_M)\}\, d\boldsymbol{\theta} \\
&= \int \exp\left[n\left\{q(\hat{\boldsymbol{\theta}}|\Lambda_M) - \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T R_\lambda(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \cdots\right\}\right] d\boldsymbol{\theta} \\
&\approx \exp\left\{nq(\hat{\boldsymbol{\theta}}|\Lambda_M)\right\} \int \exp\left\{-\frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T R_\lambda(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \\
&= \frac{(2\pi)^{(r+Kp)/2}}{n^{(r+Kp)/2}|R_\lambda(\hat{\boldsymbol{\theta}})|^{1/2}} \exp\left\{nq(\hat{\boldsymbol{\theta}}|\Lambda_M)\right\}.
\end{aligned}$$

Therefore, the GBIC evaluating the multivariate functional regression model estimated by the maximum penalized likelihood method is given by

(29)
$$\begin{aligned}
&-2\log p(\boldsymbol{y}|\Lambda_M) \\
&= -2\log\left\{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\Lambda_M)d\boldsymbol{\theta}\right\} \\
&\approx -2\sum_{\alpha=1}^{n} f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}) + n\mathrm{tr}\{\hat{B}^T(\Lambda_M \odot \Omega)\hat{B}\} \\
&\quad + (r+Kq)\log n - (r+Kq)\log(2\pi) \\
&\quad - \sum_{k=1}^{K} \log|\Lambda_M \odot \Omega|_+ + \log|R_\lambda(\hat{\boldsymbol{\theta}})|.
\end{aligned}$$

References

[1] Akaike, H.: Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (Petrov, B. N. and Csaki, F. eds.) (1973) 267–281, Akademiai Kiado, Budapest.

[2] Araki, Y., Konishi, S., Kawano, S. and Matsui, H.: Functional regression modeling via regularized Gaussian basis expansions. To appear in *Ann. Inst. Statist. Math* (2008).

[3] Craven, P. and Wahba, G.: Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numer. Math.* **31** (1979) 317–403.

[4] Ferraty, F. and Vieu, P.: *Nonparametric Functional Data Analysis.* Springer, New York, 2006.

[5] Harezlak, J., Coull, B., Laird, N., Magari, S. and Christiani, D.: Penalized solutions to functional regression problems. *Comput. Statist. Data Anal.* **51** (2007) 4911–4925.

[6] Hastie, T. and Tibshirani, R.: *Generalized Additive Models.* Chapman and Hall, London, 1990.

[7] Imoto, S. and Konishi, S.: Selection of smoothing parameters in B-spline nonparametric regression models using information criteria, *Ann. Inst. Statist. Math.* **55** (2003) 671–687.

[8] James, G. M.: Generalized linear models with functional predictor variables. *J. Roy. Statist. Soc. Ser. B* **64** (2002) 411–432.

[9] Kawano, S. and Konishi, S.: Nonlinear regression modeling via regularized Gaussian basis functions. *Bull. Inform. Cybern.* **39** (2007) 83–96.

[10] Konishi, S., Ando, T. and Imoto, S.: Bayesian information criteria and smoothing parameter selection in radial basis function network, *Biometrika* **91** (2004) 27–43.

[11] Konishi, S. and Kitagawa, G.: Generalised information criteria in model selection. *Biometrika* **83** (1996) 875–890.

[12] Konishi, S. and Kitagawa, G.: *Information Criteria and Statistical Modeling.* Springer, New York, 2008.

[13] Malfait, N. and Ramsay, J. O.: The historical functional linear model. *Canad. J. Statist.* **31** (2003) 115–128.

[14] Matsui, H., Araki, Y. and Konishi, S.: Multivariate regression modeling for functional data. *J. Data Sci.* **6** (2008) 313–331.

[15] Ramsay, J. O. and Dalzell, C. J.: Some tools for functional data analysis (with discussion). *J. Roy. Stat. Soc. Ser. B* **58** (1991) 495–508.

[16] Ramsay, J. O. and Silverman, B. W.: *Applied Functional Data Analysis.* Springer, New York, 2002.

[17] Ramsay, J.O. and Silverman, B.W.: *Functional Data Analysis.* 2nd ed. Springer, New York, 2005.

[18] Rossi, F., Delannay, N., Conan-Guez, B. and Verleysen, M.: Representation of functional data in neural networks, *Neurocomputing* **64** (2005) 183–210.

[19] Schwarz, G.: Estimating the dimension of a model. *Ann. Statist.* **6** (1978) 461–464.

[20] Shimokawa, M., Mizuta, M. and Sato, Y.: An expansion of functional regression analysis. *J. Jap. App. Statist. (in Japanese)* **29** (2000) 27–39.

[21] Yamanishi, Y. and Tanaka, Y.: Geographically weighted functional multiple regression analysis: a numerical investigation, *J. Jap. Soc. Comp. Statist.* **15** (2003) 307–317.

[22] Yao, F., Müller, H. G. and Wang, J. L.: Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** (2005) 2873–2903.

Hidetoshi Matsui
Graduate School of Mathematics, Kyushu University, Fukuoka
812-8581, Japan.
E-mail: hmatsui(at)math.kyushu-u.ac.jp

Shuichi Kawano
Graduate School of Mathematics, Kyushu University, Fukuoka
812-8581, Japan.
E-mail: s-kawano(at)math.kyushu-u.ac.jp

Sadanori Konishi
Faculty of Mathematics, Kyushu University, Fukuoka 812-
8581, Japan.
E-mail: konishi(at)math.kyushu-u.ac.jp