Ann Inst Stat Math (2009) 61:811–833
DOI 10.1007/s10463-007-0161-1

Functional regression modeling via regularized Gaussian basis expansions

Yuko Araki · Sadanori Konishi ·
Shuichi Kawano · Hidetoshi Matsui

Received: 22 September 2006 / Revised: 19 September 2007 / Published online: 31 January 2008

Abstract We consider the problem of constructing functional regression models for scalar responses and functional predictors, using Gaussian basis functions along with the technique of regularization. An advantage of our regularized Gaussian basis expansions to functional data analysis is that it creates a much more flexible instrument for transforming each individual's observations into functional form. In constructing functional regression models there remains the problem of how to determine the number of basis functions and an appropriate value of a regularization parameter. We present model selection criteria for evaluating models estimated by the method of regularization in the context of functional regression models. The proposed functional regression models are applied to Canadian temperature data. Monte Carlo simulations are conducted to examine the efficiency of our modeling strategies. The simulation results show that the proposed procedure performs well especially in terms of flexibility and stable estimates.

Keywords Functional regression modeling · Gaussian basis function · Regularization · Smoothing parameter selection

Y. Araki (✉)
Biostatistics Center, Kurume University, 67 Asahi-Machi, Kurume, Fukuoka 830-0011, Japan
e-mail: araki_yuuko@med.kurume-u.ac.jp

S. Konishi · S. Kawano · H. Matsui
Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku,
Fukuoka 812-8581, Japan
e-mail: konishi@math.kyushu-u.ac.jp

S. Kawano
e-mail: s-kawano@math.kyushu-u.ac.jp

H. Matsui
e-mail: hmatsui@math.kyushu-u.ac.jp

1 Introduction

Recently, functional data analysis has received considerable attention in different fields of application such as criminology, electromyography, signal processing, and a number of successful applications have been reported (see, e.g., Ramsay and Silverman 2002, 2005; Araki and Konishi 2006; Mizuta 2006).

The basic idea behind functional data analysis is to express discrete observations in the form of a function, and then draw information from a collection of functional data by applying concepts from multivariate data analysis. The focus in the present paper will be on the problem of constructing functional regression models, where the observed values can be interpreted as a discretized realization of a function evaluated at possibly differing time points for each subject.

The early works on functional data analysis mainly use Fourier series, spline or B-spline smoothing techniques in transforming vector-valued data into functions. A Fourier series is useful as basis functions if the observed data are periodic and have sinusoidal features. Moreover, a remarkable point is that the orthogonal property of Fourier series basis yields an identity matrix for the integral of the product of any two basis functions in model building process. For non-periodic data, splines and B -splines are employed as a useful tool in transforming discrete data with complex structure into functions. Despite their attractive properties, there is a drawback in modeling the relationship between a response and functional predictors. Spline types of basis functions do not have the orthogonal property, and in consequence the cross-product matrix may not be directly calculated.

James (2002) presented a technique for extending generalized linear models to a situation where some of the predictor variables are observations from a curve or function, in which the spline coefficients for the functional predictor are assumed to be distributed according to a multivariate normal distribution. In contrast we consider a direct generalization of functional regression models that uses Gaussian basis expansions for filtering the predictor functions and weight functions. We propose functional regression modelings for scalar responses, using Gaussian basis functions along with the technique of regularization. We also unified functional regression models in the context of generalized linear models. There are several advantages for the use of Gaussian basis in functional data analysis. First, it creates a much more flexible instrument for transforming each individual's observations into functional form. Second, we can model the coefficient parameter function by using the same Gaussian basis as for the predictors, since the integral of the product of any two Gaussian basis functions can be easily calculated.

In practice, individuals are measured at possibly different time points, and the amount of smoothness imposed on a set of discrete data may differ from each other. Hence a crucial issue in functional regression modeling is the choice of a smoothing parameter and the number of basis functions. Cross-validation (CV) and generalized cross-validation (GCV) are often referred as in the literature. An advantage of these procedures lies in their independence from probabilistic assumptions. The computational time of the procedures is very large, however, and the high variability and tendency to undersmooth in CV and GCV are not negligible in the analysis of functional data, since the selectors are repeatedly applied.

We present an information-theoretic criterion for evaluating models estimated by the method of regularization in the context of functional regression modeling. The criteria are applied to choose smoothing parameters and the number of basis functions. The proposed method is illustrated through real data analyses and numerical studies. It is shown that the proposed functional regression modeling procedures perform well especially in terms of flexibility and stable estimates.

This paper is organized as follows. In Sect. 2 we consider a Gaussian basis expansion for converting the observed discrete data into the functional form. Section 3 describes the problem of constructing functional regression models that directly model the relationship between a response and a functional predictor. In Sect. 4, we introduce the functional logistic regression model with Gaussian bases. In the context of generalized linear models we present a functional regression model in Sect. 5 and derive model selection criteria in Sect. 6. In Sect. 7 Monte Carlo simulations are conducted to investigate the effectiveness of our modeling strategies. We also apply the proposed modeling procedure to Canadian temperature data. Summary and concluding remarks are given in Sect. 8.

2 Functionalization by Gaussian basis functions

Suppose we have n independent observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_\alpha$ are the vectors consisting of the N_α observed values $x_{\alpha 1}, x_{\alpha 2}, \ldots, x_{\alpha N_\alpha}$ at times $t_{\alpha 1}, t_{\alpha 2}, \ldots, t_{\alpha N_\alpha}$, respectively. Our goal here is to express this kind of data $\{(x_{\alpha i}, t_{\alpha i});\ i = 1, 2, \ldots, N_\alpha,\ \ t_{\alpha i} \in \mathcal{T} \subset R\}$ $(\alpha = 1, 2, \ldots, n)$ as a set of smooth functions $\{x_\alpha(t); \alpha = 1, 2, \ldots, n,\ t \in \mathcal{T}\}$ by an appropriate smoothing technique. In this section we drop the notational dependence on the subject $\boldsymbol{x}_\alpha$ and consider a functionalization of the data $\{(x_i, t_i);\ i = 1, \ldots, N\}$.

It is assumed that the observed values $\{(x_i, t_i);\ i = 1, 2, \ldots, N\}$ for a subject are drawn from the regression model

$$x_i = u(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{1}$$

where $u(t)$ is a smooth function to be estimated and the errors ϵ_i are independently, normally distributed with mean zero and variance σ^2. We also assume that the function $u(t)$ can be expressed as a linear combination of basis functions

$$u(t) = \omega_0 + \sum_{k=1}^{m} \omega_k \phi_k(t; \mu_k, \eta_k^2), \tag{2}$$

where $\phi_k(t; \mu_k, \eta_k^2)$ are Gaussian basis functions given by

$$\phi_k(t; \mu_k, \eta_k^2) = \exp\left\{-\frac{(t-\mu_k)^2}{2\eta_k^2}\right\}, \quad k = 1, 2, \ldots, m. \tag{3}$$

Here μ_k are the positions of the centers, η_k are the dispersion parameters and m is the number of basis functions.

The centers and the dispersion parameters are determined first, then the weights are estimated, using the method of regularization. This two-stage learning is reported to solve the problem of convergence and the identification problem (Moody and Darken 1989; Ando et al. 2001, 2005). We use the k-means clustering algorithm to determine the centers μ_k and the dispersion parameters η_k of the Gaussian basis functions. More precisely, the observation points $\{t_i; i = 1, \ldots, N\}$ are divided into m clusters $\{C_1, C_2, \ldots, C_m\}$, where m is the given number of Gaussian basis functions. The centers μ_k and the dispersion parameters η_k of the clusters C_k are then determined by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{t_i \in C_k} t_i, \quad \hat{\eta}_k^2 = \frac{1}{n_k} \sum_{t_i \in C_k} (t_i - \hat{\mu}_k)^2, \tag{4}$$

where n_k represents the number of observations that belongs to the cluster C_k. If the subjects are measured at different times, then all the time points $\{t_{\alpha i}; i = 1, \ldots, N_\alpha, \alpha = 1, \ldots, n\}$ are divided into m clusters. Replacing μ_k and η_k in equation (3) by their sample estimates (4), we have a set of m basis functions

$$\phi_k(t; \hat{\mu}_k, \hat{\eta}_k^2) = \exp\left\{-\frac{(t - \hat{\mu}_k)^2}{2\hat{\eta}_k^2}\right\} \equiv \phi_k(t), \quad k = 1, 2, \ldots, m. \tag{5}$$

It follows from (1) and (5) that the nonlinear regression model based on the Gaussian basis functions for the α-th subject can be written as

$$f(x_i|t_i; \boldsymbol{\omega}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left\{x_i - \boldsymbol{\omega}^T \boldsymbol{\phi}(t_i)\right\}^2}{2\sigma^2}\right], \quad i = 1, 2, \ldots, N, \tag{6}$$

where $\boldsymbol{\omega} = (\omega_0, \omega_1, \ldots, \omega_m)^T$ and $\boldsymbol{\phi}(t) = (1, \phi_1(t), \ldots, \phi_m(t))^T$. The maximum likelihood estimates of the parameters $\boldsymbol{\omega}$ and σ^2 can be easily obtained. However, in the context of functional data analysis, all the individual data that are observed discretely should be smoothed by using the common basis functions. Moreover, it is expected that the amount of smoothness imposed on sets of discrete data will differ between the subjects. To take this into account, the parameters $\boldsymbol{\omega}$ and σ^2 are estimated by using the regularization method instead of the maximum likelihood method.

The regularization method maximizes the penalized log-likelihood function

$$\begin{aligned} \ell_\zeta(\boldsymbol{\omega}, \sigma^2) &= \sum_{i=1}^{N} \log f(x_i|t_i; \boldsymbol{\omega}, \sigma^2) - \frac{N\zeta}{2} \boldsymbol{\omega}^T K \boldsymbol{\omega} \\ &= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\boldsymbol{x} - \Phi\boldsymbol{\omega})^T (\boldsymbol{x} - \Phi\boldsymbol{\omega}) - \frac{N\zeta}{2} \boldsymbol{\omega}^T K \boldsymbol{\omega}, \end{aligned} \tag{7}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)^T$, Φ is an $N \times (m + 1)$ matrix defined by $\Phi = (\boldsymbol{\phi}(t_1), \ldots, \boldsymbol{\phi}(t_N))^T$ and ζ is the regularization or smoothing parameter, which adjusts the amount of smoothness and also avoids ill-posed problems. Typical forms for the regularization term are given as

$$\sum_{j=k+1}^{m} (\Delta^k w_j)^2 = \boldsymbol{w}^T D_k^T D_k \boldsymbol{w}, \quad \sum_{j=0}^{m} w_j^2 = \boldsymbol{w}^T I_{m+1} \boldsymbol{w}, \tag{8}$$

where Δ is a difference operator such as $\Delta w_j = w_j - w_{j-1}$, D_k is an $(m+1-k)\times(m+1)$ matrix that represents the kth order difference operator Δ^k and I_{m+1} is an $(m+1)$ dimensional identity matrix. Hence the regularization term can be represented as a quadratic form $\eta(\boldsymbol{w}) = \boldsymbol{w}^T K \boldsymbol{w}$ by taking appropriate matrix $D_k^T D_k$ or I_{m+1} for the $(m+1) \times (m+1)$ matrix K. The penalized maximum likelihood estimates are given by

$$\hat{\boldsymbol{\omega}} = (\Phi^T \Phi + N\zeta\hat{\sigma}^2 K)^{-1} \Phi^T \boldsymbol{x}, \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left\{ x_i - \hat{\boldsymbol{\omega}}^T \boldsymbol{\phi}(t_i) \right\}^2. \tag{9}$$

Then the observed discrete data $\{(x_i, t_i);\ i = 1, \ldots, N\}$ are converted into the functional form given by

$$\hat{u}(t) = \hat{\omega}_0 + \sum_{k=1}^{m} \hat{\omega}_k \phi_k(t) \equiv x(t). \tag{10}$$

We note here that in the functional regression modeling all the individual data observed discretely should be smoothed by using the common basis functions. The amount of the smoothness imposed on sets of discrete data will differ among the subjects, however. Hence we transfer the issue of the number of basis functions into the choice of the smoothing parameter.

In practice, we first obtain the optimal number of basis functions by using GIC (Ando et al. 2005) for each curve. Then the most frequently selected number of basis functions (m) among n sample is determined. Once m is fixed, then we choose the optimum value of the smoothing parameter ζ for each set of discrete data as the minimizer of the criterion.

$$\mathrm{GIC}(\zeta) = N \log(2\pi\hat{\sigma}^2) + N + 2\mathrm{tr}\{QR^{-1}\}, \tag{11}$$

where $\hat{\sigma}^2$ is given in (9) and the $(m+2) \times (m+2)$ matrices Q and R are respectively given by

$$Q = \frac{1}{N\hat{\sigma}^2} \begin{pmatrix} \frac{1}{\hat{\sigma}^2}\Phi^T \Lambda^2 \Phi - \zeta K \hat{\boldsymbol{\omega}} \mathbf{1}_N^T \Lambda \Phi & \frac{1}{2\hat{\sigma}^4}\Phi^T \Lambda^3 \mathbf{1}_N - \frac{1}{2\hat{\sigma}^2}\Phi^T \Lambda \mathbf{1}_N \\ \frac{1}{2\hat{\sigma}^4}\mathbf{1}_N^T \Lambda^3 \Phi - \frac{1}{2\hat{\sigma}^2}\mathbf{1}_N^T \Lambda \Phi & \frac{1}{4\hat{\sigma}^6}\mathbf{1}_N^T \Lambda^4 \mathbf{1}_N - \frac{N}{4\hat{\sigma}^2} \end{pmatrix}, \tag{12}$$

$$R = \frac{1}{N\hat{\sigma}^2}\begin{pmatrix} \Phi^T\Phi + N\zeta\hat{\sigma}^2 K & \frac{1}{\hat{\sigma}^2}\Phi^T\Lambda\mathbf{1}_N \\ \frac{1}{\hat{\sigma}^2}\mathbf{1}_N^T\Lambda\Phi & \frac{N}{2\hat{\sigma}^2}\end{pmatrix}, \tag{13}$$

where $\mathbf{1}_N = (1, 1, \dots, 1)^T$ and $\Lambda = \mathrm{diag}[\, x_1 - \hat{\boldsymbol{\omega}}^T\boldsymbol{\phi}(t_1), \dots, x_N - \hat{\boldsymbol{\omega}}^T\boldsymbol{\phi}(t_N)]$.

The observed discrete data $\{(x_{\alpha i}, t_{\alpha i});\, t_{\alpha i} \in \mathcal{T},\ i = 1, \dots, N_\alpha\}$ $(\alpha = 1, \dots, n)$ are smoothed by the method described above, producing a functional data set $\{x_\alpha(t);\ \alpha = 1, \dots, n\}$ given by

$$\hat{u}(t) = \hat{\omega}_{\alpha 0} + \sum_{k=1}^{m} \hat{\omega}_{\alpha k}\phi_k(t) \equiv x_\alpha(t), \quad t \in \mathcal{T} \subset R, \tag{14}$$

with the common basis functions $\{\phi_1(t), \dots, \phi_m(t)\}$. In the next section we model the relationship between a response and a functional predictor.

3 Functional regression model with Gaussian noise

Suppose that the n sets of observed discrete data $\{(x_{\alpha i}, t_{\alpha i}); t_{\alpha i} \in \mathcal{T} \subset R,\ i = 1, \dots, N_\alpha\}$ $(\alpha = 1, \dots, n)$ are functionalized by the method described in the previous section, and that we have $\{(x_\alpha(t), y_\alpha);\ t \in \mathcal{T},\ \alpha = 1, 2, \dots, n\}$, where $x_\alpha(t)$ are functional predictors and y_α are time independent scalar responses. Assume that the functional predictor $x_\alpha(t)$ for the αth subject is

$$\begin{aligned} x_\alpha(t) &= w_{\alpha 0} + \sum_{k=1}^{m} w_{\alpha k}\phi_k(t) \\ &= \boldsymbol{w}_\alpha^T\boldsymbol{\phi}(t), \end{aligned} \tag{15}$$

where $\boldsymbol{w}_\alpha = (w_{\alpha 0}, w_{\alpha 1}, \dots, w_{\alpha m})^T$ are the estimated weight vectors and $\boldsymbol{\phi}(t) = (1, \phi_1(t), \dots, \phi_m(t))^T$ is a vector of Gaussian basis functions $\phi_k(t)$ given by Eq. (5).

In order to draw information from the set of functional data, we model the relationship between the response and predictor as follows:

$$y_\alpha = \beta_f + \int_{\mathcal{T}} x_\alpha(t)\beta(t)\mathrm{d}t + \epsilon_\alpha, \quad \alpha = 1, 2, \dots, n, \tag{16}$$

where ϵ_α are independently, normally distributed with mean 0 and variance σ^2 (Ramsay and Silverman 2005). Using the same Gaussian basis functions $\boldsymbol{\phi}(t)$ as in (15), we expand the functional parameter as

$$\begin{aligned} \beta(t) &= \beta_0 + \sum_{k=1}^{m} \beta_k\phi_k(t) \\ &= \boldsymbol{\gamma}^T\boldsymbol{\phi}(t), \end{aligned} \tag{17}$$

where $\boldsymbol{\gamma} = (\beta_0, \beta_1, \ldots, \beta_m)^T$. Substituting Eqs. (15) and (17) into Eq. (16) yields

$$\begin{aligned} y_\alpha &= \beta_f + \boldsymbol{w}_\alpha^T \int \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^T \mathrm{d}t \boldsymbol{\gamma} + \epsilon_\alpha \\ &= \beta_f + \boldsymbol{w}_\alpha^T J \boldsymbol{\gamma} + \epsilon_\alpha \\ &= \boldsymbol{z}_\alpha^T \boldsymbol{\beta} + \epsilon_\alpha, \quad \alpha = 1, 2, \ldots, n, \end{aligned} \tag{18}$$

where $\boldsymbol{z}_\alpha^T = (1, \boldsymbol{w}_\alpha^T J)$, $\boldsymbol{\beta} = (\beta_f, \boldsymbol{\gamma}^T)^T$ and J is an $(m+1) \times (m+1)$ matrix with (j,k)th element

$$J_{jk} = \int \phi_j(t)\phi_k(t)\mathrm{d}t, \quad j, k = 0, 1, \ldots, m. \tag{19}$$

An advantage of the use of the Gaussian type of basis functions is that the integral of the product of any two Gaussian basis functions can be easily calculated. In fact, we have $J_{00} = 1$, $J_{0k} = \sqrt{2\pi\hat{\eta}_k^2} (k = 1, \ldots, m)$, $J_{j0} = \sqrt{2\pi\hat{\eta}_j^2} (j = 1, \ldots, m)$ and

$$J_{jk} = \frac{\sqrt{2\pi}}{\sqrt{\dfrac{1}{\hat{\eta}_j^2} + \dfrac{1}{\hat{\eta}_k^2}}} \exp\left\{-\frac{1}{2(\hat{\eta}_j^2 + \hat{\eta}_k^2)}(\hat{\mu}_j - \hat{\mu}_k)^2\right\}, \quad j, k = 1, \ldots, m, \tag{20}$$

where $\hat{\mu}_j$ and $\hat{\eta}_j$ are given in equation (4). Then it follows from (18) that the likelihood function is given by

$$\begin{aligned} \prod_{\alpha=1}^n f(y_\alpha | x_\alpha; \boldsymbol{\beta}, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{\alpha=1}^n \left(y_\alpha - \boldsymbol{z}_\alpha^T \boldsymbol{\beta}\right)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - Z\boldsymbol{\beta})^T(\boldsymbol{y} - Z\boldsymbol{\beta})\right\}, \end{aligned} \tag{21}$$

where x_α is the functional predictor, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and Z is an $n \times (m+2)$ matrix defined by

$$Z^T = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ J^T \boldsymbol{w}_1 & J^T \boldsymbol{w}_2 & \ldots & J^T \boldsymbol{w}_n \end{pmatrix}. \tag{22}$$

Our Gaussian basis function regression models can also be applied to analyze a set of multidimensional functional data such as surface fitting, since the explicit formula for the J matrix in the Eq. (19) can be easily obtained by calculating the integral of the product of any two Gaussian basis functions.

The maximum likelihood method often gives an unsatisfactory result in estimating the functional regression coefficient $\beta(t)$ in terms of instability and computability. The inverse of $Z^T Z$, which is required for computing the maximum likelihood estimates, tends to be unstable and often yields an ill-posed problem, especially in functional logistic model. Hence we estimate the $(m+2)$-dimensional unknown parameter vector $\boldsymbol{\beta}$ and the error variance σ^2 by the method of regularization.

Instead of maximizing the log-likelihood function, we choose the $\boldsymbol{\beta}$ and σ^2 to maximize the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - Z\boldsymbol{\beta})^T(\boldsymbol{y} - Z\boldsymbol{\beta}) - \frac{n\lambda}{2}\boldsymbol{\beta}^T K\boldsymbol{\beta}, \quad (23)$$

where K is an $(m+2) \times (m+2)$ penalty matrix and λ is a smoothing parameter that controls the smoothness of the functional parameter. For a fixed value of the regularization parameter λ, the penalized maximum likelihood estimates are given by

$$\hat{\boldsymbol{\beta}} = (Z^T Z + n\hat{\sigma}^2\lambda K)^{-1} Z^T \boldsymbol{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - Z\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - Z\hat{\boldsymbol{\beta}}). \quad (24)$$

Adding a positive constant to the elements of $Z^T Z$, the solution $\hat{\boldsymbol{\beta}}$ makes the problem nonsingular even when $Z^T Z$ is not of full rank.

We choose the value of the smoothing parameter which minimizes the information criterion derived in Sect. 6. Then the functional parameter estimate and the predictive values are, respectively, given by

$$\hat{\beta}(t) = \hat{\boldsymbol{\gamma}}^T\boldsymbol{\phi}(t) \quad \text{and} \quad \hat{\boldsymbol{y}} = Z(Z^T Z + n\hat{\sigma}^2\lambda K)^{-1} Z^T \boldsymbol{y}. \quad (25)$$

4 Functional logistic regression model

We consider a functional regression modeling in the case of a binary response variable, resulting in functional logistic regression with regularization parameter estimates.

Suppose that $\{y_\alpha;\ \alpha = 1, 2, \ldots, n\}$ are independent observations of a response Y taking the value 0 or 1, associated with the functional data $\{x_\alpha(t);\ t \in \mathcal{T}, \alpha = 1, \ldots, n\}$ for a predictor where $x_\alpha(t)$ are given by Eq. (15). The conditional probabilities of Y given the functional predictor x_α are assumed to be

$$\Pr(Y_\alpha = 1|x_\alpha) = \pi^{(\alpha)} \quad \text{and} \quad \Pr(Y_\alpha = 0|x_\alpha) = 1 - \pi^{(\alpha)}. \quad (26)$$

We consider the functional logistic regression model in the form

$$\log\left\{\frac{\pi^{(\alpha)}}{1 - \pi^{(\alpha)}}\right\} = \beta_f + \int_{\mathcal{T}} x_\alpha(t)\beta(t)\mathrm{d}t. \quad (27)$$

By using the same Gaussian basis functions $\boldsymbol{\phi}(t) = (1, \phi_1(t), \ldots, \phi_m(t))^T$ as in (15), we expand the functional parameter $\beta(t)$ as

$$\begin{aligned}\beta(t) &= \beta_0 + \sum_{k=1}^{m} \beta_k \phi_k(t) \\ &= \boldsymbol{\gamma}^T \boldsymbol{\phi}(t),\end{aligned} \tag{28}$$

where $\boldsymbol{\gamma} = (\beta_0, \beta_1, \ldots, \beta_m)^T$. Substituting $\beta(t)$ and $x_\alpha(t)$ into Eq. (27), we have

$$\log\left\{\frac{\pi^{(\alpha)}}{1 - \pi^{(\alpha)}}\right\} = \boldsymbol{z}_\alpha^T \boldsymbol{\beta}, \tag{29}$$

where $\boldsymbol{\beta} = (\beta_f, \boldsymbol{\gamma}^T)^T$ and $z_\alpha^T = (1, \boldsymbol{w}_\alpha^T J)$ with $(m+1) \times (m+1)$ matrix $J = (J_{jk})$ given by Eq. (19).

The conditional probabilities can be rewritten as

$$\begin{aligned}\pi^{(\alpha)} &= \frac{\exp\left\{\beta_f + \int_{\mathcal{T}} x_\alpha(t)\beta(t)\mathrm{d}t\right\}}{1 + \exp\left\{\beta_f + \int_{\mathcal{T}} x_\alpha(t)\beta(t)\mathrm{d}t\right\}} \\ &= \frac{\exp\left(\boldsymbol{z}_\alpha^T \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{z}_\alpha^T \boldsymbol{\beta}\right)}.\end{aligned} \tag{30}$$

Then the log-likelihood function for y_α in terms of $\boldsymbol{\beta}$ is

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{\alpha=1}^{n} \left\{y_\alpha \log \pi^{(\alpha)} + (1 - y_\alpha) \log(1 - \pi^{(\alpha)})\right\} \\ &= \sum_{\alpha=1}^{n} \left[y_\alpha \left(\boldsymbol{\beta}^T \boldsymbol{z}_\alpha\right) - \log\left\{1 + \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_\alpha\right)\right\}\right].\end{aligned} \tag{31}$$

Estimates for $\boldsymbol{\beta}$ can be found using a regularization method that maximizes the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{\beta}) = \sum_{\alpha=1}^{n} \left[y_\alpha \left(\boldsymbol{\beta}^T \boldsymbol{z}_\alpha\right) - \log\left\{1 + \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_\alpha\right)\right\}\right] - \frac{n\lambda}{2} \boldsymbol{\beta}^T K \boldsymbol{\beta}, \tag{32}$$

where K is an $(m+2) \times (m+2)$ penalty matrix and λ is a smoothing parameter that controls the smoothness of $\beta(t)$.

When a particular value of λ is given, the following iterative algorithm, Newton–Raphson method, is used to find the parameter estimates,

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} + \left\{ -\frac{\partial^2 \ell_\lambda(\boldsymbol{\beta}^{\text{old}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\}^{-1} \frac{\partial \ell_\lambda(\boldsymbol{\beta}^{\text{old}})}{\partial \boldsymbol{\beta}}. \tag{33}$$

The updated $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}}^{\text{new}} = (Z^T D Z + n\lambda K)^{-1} Z^T D \boldsymbol{\xi}, \tag{34}$$

where Z is given by Eq. (22), $\boldsymbol{\xi} = Z\boldsymbol{\beta}^{\text{old}} + D^{-1}(\boldsymbol{y} - \Pi \mathbf{1})$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\mathbf{1} = (1, \ldots, 1)^T$ and

$$\begin{aligned} D &= \text{diag}\left[\pi^{(1)}\{1 - \pi^{(1)}\}, \ldots, \pi^{(n)}\{1 - \pi^{(n)}\}\right], \\ \Pi &= \text{diag}\left[\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}\right]. \end{aligned}$$

After choosing the value of the smoothing parameter λ that minimizes the information criterion derived in Sect. 6, we have the estimate of the functional parameter and the predicted value given by

$$\hat{\beta}(t) = \hat{\boldsymbol{\gamma}}^T \boldsymbol{\phi}(t) \quad \text{and} \quad \hat{y} = \frac{\exp\left(z_\alpha^T \hat{\boldsymbol{\beta}}\right)}{1 + \exp\left(z_\alpha^T \hat{\boldsymbol{\beta}}\right)}, \tag{35}$$

respectively.

5 Functional generalized linear models

Generalized linear model (GLM) introduced by Nelder and Wedderburn (1972) provides a unified theoretical and computational framework for a class of nonlinear and nonnormal regression models. Green and Silverman (1994) proposed nonparametric GLM with roughness penalty methods. The functional version of GLM was implicitly introduced in the literature by Marx and Eilers (1999) as a penalized splines procedure. This section considers various types of functional regression models based on Gaussian basis functions in the context of generalized linear models.

Suppose that we have n sets of observations $\{(x_\alpha(t), y_\alpha); t \in \mathcal{T}\}$ ($\alpha = 1, 2, \ldots, n$), where y_α is a scalar response and $x_\alpha(t)$ are functional predictors. It is assumed that the functional predictor $x_\alpha(t)$ for the αth subject is functionalized by the method described in Sect. 2 and is given by

$$x_\alpha(t) = w_{\alpha 0} + \sum_{j=1}^{m} w_{\alpha j}\phi_j(t)$$
$$= \boldsymbol{w}_\alpha^T \boldsymbol{\phi}(t), \quad \alpha = 1, 2, \ldots, n, \tag{36}$$

where $\boldsymbol{w}_\alpha = (w_{\alpha 0}, w_{\alpha 1}, \ldots, w_{\alpha m})^T$ and $\boldsymbol{\phi}(t) = (1, \phi_1(t), \phi_2(t), \ldots, \phi_m(t))^T$ is a vector of the Gaussian basis functions.

To draw information from a collection of the functional data, we use the exponential family of densities

$$f(y_\alpha|x_\alpha; \xi_\alpha, \psi) = \exp\left\{\frac{y_\alpha\xi_\alpha - b(\xi_\alpha)}{\psi} + c(y_\alpha, \psi)\right\}, \tag{37}$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are specific functions and ξ_α and ψ are unknown parameters. Under the functional generalized linear model framework, the conditional expectation $E[y_\alpha|x_\alpha] = \mu_\alpha(= b'(\xi_\alpha))$ is related to the predictor η_α by $h(\mu_\alpha) = \eta_\alpha$, where $h(\cdot)$ is a link function. In systematic component, it is assumed that

$$h(\mu_\alpha) = \eta_\alpha = \beta_a + \int_{\mathcal{T}} x_\alpha(t)\beta(t)\mathrm{d}t. \tag{38}$$

Using the same Gaussian basis functions $\phi_1(t), \phi_2(t), \ldots, \phi_m(t)$ as in Eq. (36), we expand the functional parameter as

$$\beta(t) = \beta_0 + \sum_{k=1}^{m} \beta_k\phi_k(t)$$
$$= \boldsymbol{\gamma}^T \boldsymbol{\phi}(t), \tag{39}$$

where $\boldsymbol{\gamma} = (\beta_0, \beta_1, \ldots, \beta_m)^T$. The systematic component can then be expressed as

$$h(\mu_\alpha) = \beta_a + \sum_{j=0}^{m}\sum_{k=0}^{m} w_{\alpha j}\beta_k \int \phi_j(t)\phi_k(t)\mathrm{d}t$$
$$= \beta_a + \boldsymbol{w}_\alpha^T J\boldsymbol{\gamma}$$
$$= \boldsymbol{z}_\alpha^T\boldsymbol{\beta}, \quad \alpha = 1, 2, \ldots, n, \tag{40}$$

where J is an $(m + 1) \times (m + 1)$ cross-product matrix given as in Eq. (20), $\boldsymbol{\beta} = (\beta_a, \boldsymbol{\gamma}^T)^T$ is the $(1 + m)$-dimensional parameter vector and $\boldsymbol{z}_\alpha = (1, \boldsymbol{w}_\alpha^T J)^T$. Combining the random component (37) and the systematic component (40), we have a functional generalized linear model

$$f(y_\alpha|x_\alpha; \boldsymbol{\beta}, \psi) = \exp\left\{\frac{y_\alpha k(\boldsymbol{z}_\alpha^T\boldsymbol{\beta}) - r(\boldsymbol{z}_\alpha^T\boldsymbol{\beta})}{\psi} + c(y_\alpha, \psi)\right\}, \tag{41}$$

where $k(\cdot) = b'^{-1} \circ h^{-1}(\cdot)$ and $r(\cdot) = b \circ b'^{-1} \circ h^{-1}(\cdot)$.

The unknown parameters $\boldsymbol{\beta}$ and ψ are estimated by maximizing the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{\beta}, \psi) = \sum_{\alpha=1}^{n} \left\{ \frac{y_\alpha k(\boldsymbol{z}_\alpha^T \boldsymbol{\beta}) - r(\boldsymbol{z}_\alpha^T \boldsymbol{\beta})}{\psi} + c(y_\alpha, \psi) \right\} - \frac{n\lambda}{2} \boldsymbol{\beta}^T K \boldsymbol{\beta}, \tag{42}$$

where K is an $(m+2)$-dimensional positive semidefininte matrix with rank $m-d$.

The maximum penalized likelihood estimator $\hat{\boldsymbol{\beta}}$ is a solution of the penalized likelihood equation $\partial \ell_\lambda(\boldsymbol{\beta}, \psi)/\partial \boldsymbol{\beta} = \boldsymbol{0}$. This solution in general will be a nonlinear optimization problem, and for fixed values of ψ and λ, the iteration may be expressed as

$$\boldsymbol{\beta}^{\text{new}} = (Z^T W Z + n\lambda\psi K)^{-1} Z^T W \boldsymbol{k}, \tag{43}$$

where Z is given in (22), $\boldsymbol{k}$ is an n dimensional vector with the αth element given by $k_\alpha = (y_\alpha - \mu_\alpha)h'(\mu_\alpha) + \boldsymbol{z}_\alpha^T \boldsymbol{\beta}$ and W is an $n \times n$ diagonal matrix with the αth element given by $w_{\alpha\alpha} = \left\{ b''(\xi_\alpha) h'(\mu_\alpha)^2 \right\}^{-1}$.

In each Fisher scoring step $\boldsymbol{\beta}$ is updated to $\boldsymbol{\beta}^{\text{new}}$ by (43) until a suitable convergence criterion is satisfied. After the estimate $\hat{\boldsymbol{\beta}}$ is obtained, the estimate of the scale parameter ψ is given as a solution of $\partial \ell_\lambda(\hat{\boldsymbol{\beta}}, \psi)/\partial \psi = 0$. Substituting the sample estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\psi}$ into Eq. (41) yields the statistical model

$$f(y_\alpha | x_\alpha; \hat{\boldsymbol{\beta}}, \hat{\psi}) = \exp \left\{ \frac{y_\alpha k(\boldsymbol{z}_\alpha^T \hat{\boldsymbol{\beta}}) - r(\boldsymbol{z}_\alpha^T \hat{\boldsymbol{\beta}})}{\hat{\psi}} + c(y_\alpha, \hat{\psi}) \right\}, \tag{44}$$

which depends on the values of the smoothing parameter λ. In Sect. 6, we derive a model selection criterion for evaluating the functional generalized linear models with Gaussian basis functions from an information theoretic point of view.

The functional generalized linear model can be used with various types of distributions. Here shown are two examples.

Example 1 Suppose that the observations y_α are independently and normally distributed with mean μ_α and common variance σ^2. By taking $b(\xi_\alpha) = \xi_\alpha^2/2$, $\psi = \sigma^2$, $c(y_\alpha, \psi) = -y_\alpha^2/(2\sigma^2) - \log(\sigma\sqrt{2\pi})$ and $h(\mu_\alpha) = \mu_\alpha$ in the exponential family of densities (37), we have a functional regression model with Gaussian noise explained in Sect. 3.

Example 2 Suppose that we have n sets of observations $\{(x_\alpha(t), y_\alpha);\ t \in \mathcal{T}, \alpha = 1, 2, \ldots, n\}$, where $x_\alpha(t)$ are a functional predictor and y_α are independent random variables coded as either 0 or 1. By taking $b(\xi_\alpha) = \log\{1 + \exp(\xi_\alpha)\}$, $\psi = 1$, $c(y_\alpha, \psi) = 0$ and $h(\mu_\alpha) = \log\{\mu_\alpha/(1-\mu_\alpha)\}$ in (37), we have a functional logistic regression model explained in Sect. 4.

6 Model selection criteria

In the functional regression models, we need to determine the appropriate values of the adjusted parameters that include the number of basis functions and a smoothing parameter or a regularization parameter. Choosing these parameters can be viewed as a model selection and evaluation problem; how to choose the best approximating model from the competing models by a suitable criterion. Although there is a large amount of literature regarding model selection (see, for example, Linhart and Zucchini 1986; Rao and Wu 2001), research on model selection for functional data analysis has not yet been developed. Rice and Wu (2001) used the model selection techniques AIC (Akaike information criterion), BIC (Bayesian information criterion) and CV (cross-validation) to select the number of breakpoints for the splines. Ramsay and Silverman (2005) used CV in a functional linear model to choose the smoothing parameter. It might be noticed that AIC and BIC cover only models estimated by the maximum likelihood method. Estimation in our model building process is by regularization. Hence we obtain a model selection criterion for evaluating models estimated by regularization in the context of functional regression modeling.

6.1 Generalized information criterion

Suppose that independent responses $y_1, \ldots, y_n$ are generated from an unknown true distribution $G(y|x)$ having probability density $g(y|x)$. Based on the information contained in the observations, we choose a model which consists of a family of probability distributions $f(y|x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is the p-dimensional vector of unknown parameters. This specified parametric family of densities may or may not contain the true density $g(y|x)$.

The unknown parameter vector $\boldsymbol{\theta}$ in a specified model is estimated by the method of regularization, and we have a statistical model $f(y|x; \hat{\boldsymbol{\theta}})$ with regularized estimator $\hat{\boldsymbol{\theta}}$. Once a statistical model has been estimated, then an overall measure of the divergence of the statistical model from the true density would be assessed by the Kullback-Leibler information (Kullback and Leibler 1951) from the predictive point of view.

Suppose that $\boldsymbol{z} = \{z_1, \ldots, z_n\}$ are future observations for the response variable drawn from $g(y|x)$. Let $f(\boldsymbol{z}|x; \hat{\boldsymbol{\theta}}) = \prod_{\alpha=1}^{n} f(z_\alpha|x_\alpha; \hat{\boldsymbol{\theta}})$ and $g(\boldsymbol{z}|x) = \prod_{\alpha=1}^{n} g(z_\alpha|x_\alpha)$. Then the Kullback–Leibler information is given by

$$\begin{aligned} I\{g, f\} &= E_{G(\boldsymbol{z}|x)}\left[\log \frac{g(\boldsymbol{z}|x)}{f(\boldsymbol{z}|x; \hat{\boldsymbol{\theta}})}\right] \\ &= E_{G(\boldsymbol{z}|x)}\left[\log g(\boldsymbol{z}|x)\right] - E_{G(\boldsymbol{z}|x)}\left[\log f(\boldsymbol{z}|x; \hat{\boldsymbol{\theta}})\right]. \end{aligned} \tag{45}$$

We choose the model that minimizes the Kullback-Leibler information from among different statistical models. A model selection criterion is obtained as an estimator of the Kullback-Leibler information or equivalently minus twice the expected log-likelihood $-2E_{G(z|x)}[\log f(z|x;\hat{\boldsymbol{\theta}})]$ and is, in general, given by

$$\mathrm{IC} = -2\sum_{\alpha=1}^{n} \log f(y_\alpha|x_\alpha;\hat{\boldsymbol{\theta}}) + 2\hat{b}(G), \tag{46}$$

where $\hat{b}(G)$ is an estimator of the bias defined by

$$b(G) = E_{G(\boldsymbol{y}|x)}\left[\sum_{\alpha=1}^{n} \log f(y_\alpha|x_\alpha;\hat{\boldsymbol{\theta}}) - E_{G(z|x)}[\log f(z|x;\hat{\boldsymbol{\theta}})]\right]. \tag{47}$$

Konishi and Kitagawa (1996) obtained an asymptotic bias for a statistical model with functional estimator and approximated the bias by the trace of a matrix for products of the empirical influence function of estimators and the score function of a specified parametric model. The influence function of the regularized estimator $\hat{\boldsymbol{\theta}}$ is given by

$$\boldsymbol{T}^{(1)}(y|x;G) = R(G)^{-1}\left.\frac{\partial\{\log f(y|x;\boldsymbol{\theta}) - (1/2)\boldsymbol{\theta}^T K\boldsymbol{\theta}\}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

where

$$R(G) = -\int \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\left\{\log f(y|x;\boldsymbol{\theta}) - (1/2)\boldsymbol{\theta}^T K\boldsymbol{\theta}\right\}\mathrm{d}G.$$

Then using Theorem 2.1 given by Konishi and Kitagawa (1996, p. 876), we have an information criterion

$$\mathrm{GIC} = -2\sum_{\alpha=1}^{n} \log f(y_\alpha|x_\alpha;\hat{\boldsymbol{\theta}}) + 2\mathrm{tr}\{R(\hat{G})^{-1}Q(\hat{G})\}, \tag{48}$$

where the matrices in the bias correction term are, for the empirical distribution function $\hat{G}$, given by

$$\begin{aligned} R(\hat{G}) &= -\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\left\{\log f(y_\alpha|x_\alpha;\boldsymbol{\theta}) - (1/2)\boldsymbol{\theta}^T K\boldsymbol{\theta}\right\}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \\ Q(\hat{G}) &= \frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial\{\log f(y_\alpha|x_\alpha;\boldsymbol{\theta}) - (1/2)\boldsymbol{\theta}^T K\boldsymbol{\theta}\}}{\partial\boldsymbol{\theta}} \frac{\partial \log f(y_\alpha|x_\alpha;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \end{aligned} \tag{49}$$

Substituting the density (41) into Eqs. (48) and (49) and differentiating the result with respect to $\boldsymbol{\theta}$ and ψ, we have the following result;

$$\mathrm{GIC}(\lambda) = -2\sum_{\alpha=1}^{n}\left\{\frac{y_\alpha k(z_\alpha^T\hat{\boldsymbol{\beta}}) - r(z_\alpha^T\hat{\boldsymbol{\beta}})}{\hat{\psi}} + c(y_\alpha,\hat{\psi})\right\} + 2\mathrm{tr}\left(\hat{Q}\hat{R}^{-1}\right), \tag{50}$$

where $\hat{Q}$ and $\hat{R}$ are the $(m+3) \times (m+3)$ matrices given by

$$\begin{aligned} \hat{Q} &= \frac{1}{n\hat{\psi}} \begin{pmatrix} Z^T \Lambda/\hat{\psi} - \lambda K \hat{\boldsymbol{\beta}} \mathbf{1}_n^T \\ \boldsymbol{p}^T \end{pmatrix} \left(\Lambda Z, \hat{\psi} \boldsymbol{p} \right), \\ \hat{R} &= \frac{1}{n\hat{\psi}} \begin{pmatrix} Z^T \Gamma Z + n\hat{\psi}\lambda K & Z^T \Lambda \mathbf{1}_n/\hat{\psi} \\ \mathbf{1}_n^T \Lambda Z/\hat{\psi} & -\hat{\psi}\boldsymbol{q}^T \mathbf{1}_n \end{pmatrix}. \end{aligned} \tag{51}$$

Here Λ and Γ are the $n \times n$ diagonal matrices with the α-th diagonal elements

$$\begin{aligned} \Lambda_{\alpha\alpha} &= \frac{y_\alpha - \hat{\mu}_\alpha}{b''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha)}, \\ \Gamma_{\alpha\alpha} &= \frac{(y_\alpha - \hat{\mu}_\alpha)\left\{ b'''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha) + b''(\hat{\xi}_\alpha)^2 h''(\hat{\mu}_\alpha) \right\}}{\left\{ b''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha) \right\}^3} + \frac{1}{b''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha)^2}, \end{aligned}$$

and $\boldsymbol{p}$ and $\boldsymbol{q}$ are n-dimensional vectors with α-th elements

$$\begin{aligned} p_\alpha &= -\frac{y_\alpha k(z_\alpha^T \hat{\boldsymbol{\beta}}) - r(z_\alpha^T \hat{\boldsymbol{\beta}})}{\hat{\psi}^2} + \frac{\partial}{\partial \psi} c(y_\alpha, \psi) \bigg|_{\psi = \hat{\psi}}, \\ q_\alpha &= \frac{\partial p_\alpha}{\partial \psi} \bigg|_{\psi = \hat{\psi}}. \end{aligned}$$

We present model selection criteria for evaluating functional regression and functional logistic regression models with multiple predictors constructed by the regularized Gaussian basis functions.

Example 3 Consider the functional regression model given in Sect. 3. The statistical model estimated by the regularization method is given by $f(y_\alpha|x_\alpha; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-1/2} \exp\{-(y_\alpha - z_\alpha^T \hat{\boldsymbol{\beta}})^2/(2\hat{\sigma}^2)\}$. By taking $b(\hat{\xi}_\alpha) = \hat{\xi}_\alpha^2/2$, $\hat{\psi} = \hat{\sigma}^2$, $c(y_\alpha, \hat{\psi}) = -y_\alpha^2/(2\hat{\sigma}^2) - \log(\hat{\sigma}\sqrt{2\pi}), h(\hat{\mu}_\alpha) = \hat{\mu}_\alpha$ in (50) and (51), we have a generalized information criterion for evaluating the statistical model in the following;

$$\mathrm{GIC}(\lambda) = n(\log 2\pi + 1) + n \log \hat{\sigma}^2 + 2\mathrm{tr}\left(\hat{Q} \hat{R}^{-1} \right),$$

where matrices $\hat{Q}$ and $\hat{R}$ are given by

$$\hat{R} = \frac{1}{n\hat{\sigma}^2} \begin{pmatrix} Z^T Z + n\lambda\hat{\sigma}^2 K & \frac{1}{\hat{\sigma}^2} Z^T \hat{\Lambda} \mathbf{1}_n \\ \frac{1}{\hat{\sigma}^2} \mathbf{1}_n^T \hat{\Lambda} Z & \frac{n}{2\hat{\sigma}^2} \end{pmatrix}, \tag{52}$$

$$\hat{Q} = \frac{1}{n\hat{\sigma}^2} \begin{pmatrix} \frac{1}{\hat{\sigma}^2} Z^T \hat{\Lambda}^2 Z - \lambda K \hat{\boldsymbol{\beta}} \mathbf{1}_n^T \hat{\Lambda} Z & \frac{1}{2\hat{\sigma}^4} Z^T \hat{\Lambda}^3 \mathbf{1}_n - \frac{1}{2\hat{\sigma}^2} Z^T \hat{\Lambda} \mathbf{1}_n \\ \frac{1}{2\hat{\sigma}^4} \mathbf{1}_n^T \hat{\Lambda}^3 Z - \frac{1}{2\hat{\sigma}^2} \mathbf{1}_n^T \hat{\Lambda} Z & \frac{1}{4\hat{\sigma}^6} \mathbf{1}_n^T \hat{\Lambda}^4 \mathbf{1}_n - \frac{n}{4\hat{\sigma}^2} \end{pmatrix}, \tag{53}$$

where $\hat{\Lambda} = \mathrm{diag}\{y_1 - \hat{\boldsymbol{\beta}}^T z_1, \ldots, y_n - \hat{\boldsymbol{\beta}}^T z_n\}$.

Example 4 Consider the functional logistic regression model given in Sect. 4. The statistical model estimated by the regularization method is given by $f(y_\alpha|x_\alpha; \hat{\boldsymbol{\beta}}) = \{\hat{\pi}^{(\alpha)}\}^{y_\alpha}\{1 - \hat{\pi}^{(\alpha)}\}^{1-y_\alpha}$, where $\hat{\pi}^{(\alpha)} = \exp(z_\alpha^T \hat{\boldsymbol{\beta}})/\{1 + \exp(z_\alpha^T \hat{\boldsymbol{\beta}})\}$. By taking $b(\hat{\xi}_\alpha) = \log\{1 + \exp(\hat{\xi}_\alpha)\}$, $\psi = 1$, $c(y_\alpha, \psi) = 0$ and $h(\hat{\mu}_\alpha) = \log\{\hat{\mu}_\alpha/(1 - \hat{\mu}_\alpha)\}$ in (50) and (51), we have a model selection criterion for evaluating the statistical model in the following;

$$\mathrm{GIC}(\lambda) = -2\sum_{\alpha=1}^{n}\left[y_\alpha\left(\hat{\boldsymbol{\beta}}^T z_\alpha\right) - \log\left\{1 + \exp\left(\hat{\boldsymbol{\beta}}^T z_\alpha\right)\right\}\right] + 2\mathrm{tr}\left(\hat{Q}\hat{R}^{-1}\right), \quad (54)$$

where the matrices $\hat{Q}$ and $\hat{R}$ are respectively given by

$$\hat{Q} = \frac{1}{n}\left\{Z^T\hat{\Lambda}^2 Z - \lambda K\hat{\boldsymbol{\beta}}\mathbf{1}_n^T\hat{\Lambda}Z\right\} \quad \text{and} \quad \hat{R} = \frac{1}{n}Z^T\hat{\Pi}(I_n - \hat{\Pi})Z + \lambda K, \quad (55)$$

where $\hat{\Lambda} = \mathrm{diag}\left[y_1 - \hat{\pi}^{(1)}, y_2 - \hat{\pi}^{(2)}, \ldots, y_n - \hat{\pi}^{(n)}\right]$ and $\hat{\Pi} = \mathrm{diag}[\hat{\pi}^{(1)}, \hat{\pi}^{(2)}, \ldots, \hat{\pi}^{(n)}]$.

There exist other criteria for selecting the smoothing parameters in the functional regression model with Gaussian noise.

6.2 Modified AIC

Suppose that the fitted value $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ may be expressed as $\hat{\boldsymbol{y}} = S\boldsymbol{y}$, where S is the hat matrix for functional regression model with Gaussian noise, $S = Z(Z^T Z + n\hat{\sigma}^2\lambda K)^{-1}Z^T$. Then Hastie and Tibshirani (1990) and Hurvich et al. (1998) proposed to use the trace of the smoother matrix as an approximation to the model complexity. By replacing the number of parameters in AIC by the trace of the smoother matrix, we have

$$\mathrm{MAIC}(\lambda) = -2\sum_{\alpha=1}^{n}\log f(y_\alpha|x_\alpha; \hat{\boldsymbol{\beta}}, \hat{\psi}) + 2\mathrm{tr}S, \quad (56)$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\psi}$ are regularized estimates. A problem may arise in the theoretical justification for the use of the bias-correction terms in MAIC selector automatically, since AIC covers only models estimated by the maximum likelihood method.

6.3 Cross-validation

Cross-validation method creates the new observation situation from the given data by predicting for each observation based on the remaining data. Let $\hat{y}^{(-\alpha)}$ be a regression response value estimated by the observed data except $(x_\alpha(t), y_\alpha)$. The cross-validation criterion is then

$$\begin{aligned}\mathrm{CV}(\lambda) &= \frac{1}{n}\sum_{\alpha=1}^{n}\left(y_\alpha - \hat{y}^{(-\alpha)}\right)^2 \\ &= \frac{1}{n}\sum_{\alpha=1}^{n}\left(\frac{y_\alpha - \hat{y}_\alpha}{1 - S_{\alpha\alpha}(\lambda)}\right)^2, \end{aligned} \tag{57}$$

where $S_{\alpha\alpha}(\lambda)$ is an αth diagonal element of the hat matrix S.

Generalized cross-validation, a modified form of cross-validation, introduced by Craven and Wahba (1979) replaces $S_{\alpha\alpha}(\lambda)$ in Eq. (57) by the average and is

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\frac{y_\alpha - \hat{y}_\alpha}{1 - \mathrm{tr}S/n}\right)^2.$$

In the next section, we compare GIC with modified AIC, cross-validation and generalized cross-validation through Monte Carlo simulations.

7 Numerical results and practical examples

7.1 Role of a smoothing parameter

In the functional regression modeling all the individual data that are observed discretely should be smoothed by using the common basis functions. The amount of the smoothness imposed on sets of discrete data will differ among the subjects. Therefore, the smoothing parameter plays an important role in adjusting the difference of the individual smoothness. We conduct a Monte Carlo simulation to examine the efficiency of the smoothing parameter ζ for a regression model with the fixed number of basis functions.

Figure 1 plots a set of 100 generated data from the model

$$\text{(a)} \quad y_\alpha = \exp(-2x_\alpha)\sin(5\pi x_\alpha) + \epsilon_\alpha, \quad \text{(b)} \quad y_\alpha = \sin(2\pi x_\alpha^3) + \epsilon_\alpha,$$

where the errors ϵ_α are assumed to be independently distributed according to the normal distribution with means 0 and the standard deviations are taken as $0.2R_y$ with R_y being the range of (a) $\exp(-2x_\alpha)\sin(5\pi x_\alpha)$ and (b) $\sin(2\pi x_\alpha^3)$, over the input space. The independent variable x_α are generated from uniform distribution on [0, 1].

The solid lines are smoothed curves produced by using Gaussian basis functions, estimated by the regularization method. The number of basis functions and the smoothing parameter ζ were selected by GIC in Eq. (11). The selected values were $m = 6,\ \zeta = 10^{-2.9}$, GIC = 46.35 for (a) and $m = 4,\ \zeta = 10^{-2.1}$, GIC = 93.66 for (b).

Next we tried to adjust the smoothness of each curve by changing only the smoothing parameter, fixing the number of basis functions. First, 100 pairs of data were

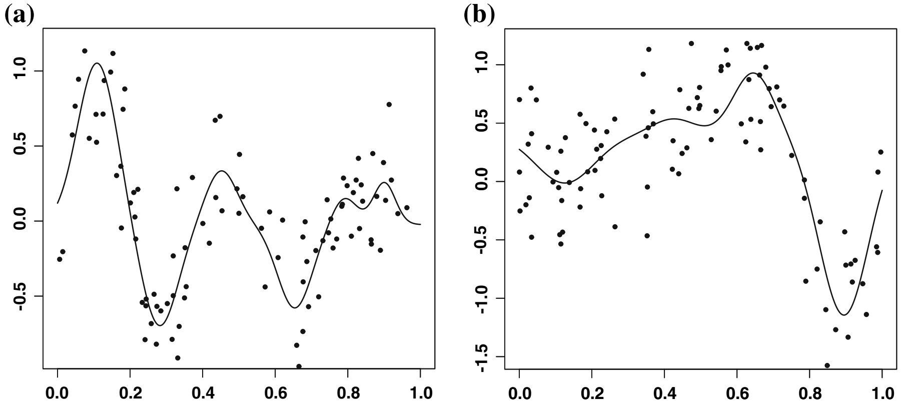

Fig. 1 Gaussian basis function smoothing; the figures (**a**) and (**b**) show the regularized Gaussian basis functions that fit to the data generated from the true model (**a**) $y_\alpha = \exp(-2x_\alpha)\sin(5\pi x_\alpha) + \epsilon_\alpha$, (**b**) $y_\alpha = \sin(2\pi x_\alpha^3) + \epsilon_\alpha$, and $\epsilon \sim N(0, 0.2R_y)$, with R_y being the range of y over the input space

generated from the models (a) and (b), and the optimal number of basis functions and the smoothing parameter were determined by GIC. Then the mean squared errors of $\sum_{\alpha=1}^{100}\{\hat{u}(x_\alpha) - u(x_\alpha)\}^2/100$ were calculated, where $u(x)$ denotes the true function. Second, with a fixed $m = 6$, only the number of the smoothing parameter was selected by the GIC and the mean squared errors were calculated. This process was repeated 100 times and the averages of the 100 simulation were obtained.

When the number of the basis functions and also the smoothing parameter being selected by GIC, the mean squared errors for the models (a) and (b) are 0.02270 and 0.04351, respectively. On the other hand, when we fixed the number of the basis functions as $m = 6$, the resulting values for the models (a) and (b) were 0.02941 and 0.04782, respectively. The differences between the adjusting number of basis and the fixed number of basis for (a) and (b) are 6.71×10^{-3} and 4.31×10^{-3}, respectively. This result shows the effect of the smoothing parameter for the regression model with a fixed number of basis functions. We can conclude that even with a fixed number of basis functions, the smoothing parameter ζ can effectively adjust individual smoothness differences.

7.2 Monte Carlo simulation

Monte Carlo experiments were conducted to investigate the effectiveness of the proposed modeling strategy with GIC comparing to others (modified AIC (MAIC), cross-validation (CV) and generalized cross-validation (GCV)). In the simulation study, we generated a data set $\{(y_\alpha, x_\alpha(t));\ \alpha = 1, \ldots, n, t \in \mathcal{T}\}$ according to the following procedure. Firstly, scalar responses y_α were generated from the true model $y_\alpha = g(u_\alpha) + \epsilon_\alpha$ with $g(u_\alpha) = \int \beta(t)u_\alpha(t)\mathrm{d}t$, where $u(t)$ and $\beta(t)$ are given as follows.

$$\begin{aligned}
&\text{(a)}\quad \beta(t) = t^2, \quad u_\alpha(t) = \exp(a_{1\alpha} t) + a_{2\alpha} t,\\
&\qquad a_1 \sim N(2, 0.3^2), \quad a_2 \sim N(-3, 0.4^2), \quad \mathcal{T} = [-1, 1], \qquad (58)\\
&\text{(b)}\quad \beta(t) = t^2, \quad u_\alpha(t) = b_{1\alpha} + b_{2\alpha} t + b_{3\alpha} t^2 + b_{4\alpha} t^3,\\
&\qquad b_1 \sim N(0.2, 0.1^2), \quad b_2 \sim N(0.4, 0.2^2), \quad b_3 \sim N(0.1, 0.08^2),\\
&\qquad b_4 \sim N(0.4, 0.1^2), \quad \mathcal{T} = [-1, 2]. \qquad (59)
\end{aligned}$$

The errors ϵ_α are assumed to be independently distributed according to the normal distributions with means 0 and the variances are taken as (I) $\sigma^2 = 0.01 R_x$, (II) $\sigma^2 = 0.05 R_x$ with R_x being the range of $g(u_\alpha)$ over the input space. Secondly, functional predictors $x_\alpha(t)$ were generated as following steps:

Step 1. The design points $\{t_i;\ i = 1, \ldots, 100\}$ are generated independently from uniform distribution on $\mathcal{T}$.

Step 2. We obtain a discrete data set $\{(x_{\alpha i}, t_i);\ i = 1, \ldots, 100,\ t_i \in \mathcal{T}\}$, where $x_{\alpha i} = u_\alpha(t_i) + e_{\alpha i}$. The errors $e_{\alpha i}$ are independently, normally distributed with mean 0 and variance 1.

Step 3. The discrete data sets are transformed into a functional data set $\{x_\alpha(t);\ \alpha = 1, \ldots, n\}$ along with the smoothing technique described in Sect. 2.

We fitted the functional regression model. The smoothing parameter λ was determined by GIC, MAIC, CV and GCV. Tables 1 and 2 compare the average squared errors ASE $= \sum_{\alpha=1}^{n} \{g(u_\alpha) - \hat{y}_\alpha\}^2 / n$, and the means and standard deviations of the smoothing parameter λ. The simulation results were obtained by averaging over 100 Monte Carlo trials. It may be seen from the simulation results that the models evaluated by GIC are superior to those based on CV-type or MAIC in all cases in the sense that they give smaller average squared error (ASE). The standard deviations of λ determined by GIC are smaller than the others in many cases.

7.3 Canadian temperature data

We investigate the efficiency of our functional regression modeling technique through the analysis of Canadian temperature data (Ramsay and Silverman 2005), for which we compare models constructed by the cross-validation (CV) and the GIC via the bootstrap method (Efron 1979). The data consist of daily average temperatures and annual rainfall observed at thirty-five Canadian weather stations.

Ramsay and Silverman (2005) transformed the daily average temperature data into a temperature function by Fourier expansion. The number of basis functions was selected subjectively based on the amount of variability in the estimated regression coefficient function. Next, the functional regression model was fitted by minimizing penalized squared error and the smoothing parameter λ was selected by the CV method.

To investigate the stability of the functional regression model evaluated by GIC, 100 sets of bootstrap samples were generated from the 35 temperature functions, and the regression coefficient function $\hat{\beta}(t)$ was estimated for each bootstrap sample. The estimates based on CV created a large variation, while the estimates based on the GIC were stable as we can observe from the Fig. 2a and b. The averages of

Table 1 Comparison of the average squared errors (ASE) based on various criteria for the simulation (a)

	GIC	MAIC	CV	GCV
$n = 25$				
$\sigma^2/R_x = 0.01$				
ASE	2.548×10^{-4}	2.669×10^{-4}	2.697×10^{-4}	2.632×10^{-4}
λ	2.934×10	3.174×10	3.288×10	3.335×10
SD(λ)	2.814×10	2.779×10	2.764×10	2.760×10
$\sigma^2/R_x = 0.05$				
ASE	6.810×10^{-3}	7.371×10^{-3}	7.340×10^{-3}	6.878×10^{-3}
λ	2.469×10	3.012×10	3.241×10	3.565×10
SD(λ)	2.705×10	2.791×10	2.764×10	2.701×10
$n = 50$				
$\sigma^2/R_x = 0.01$				
ASE	2.493×10^{-4}	2.515×10^{-4}	2.591×10^{-4}	2.527×10^{-4}
λ	1.646×10	2.285×10	2.488×10	2.367×10
SD(λ)	2.461×10	2.758×10	2.795×10	2.785×10
$\sigma^2/R_x = 0.05$				
ASE	6.475×10^{-3}	6.851×10^{-3}	6.846×10^{-3}	6.756×10^{-3}
λ	3.327×10	3.432×10	3.487×10	3.487×10
SD(λ)	2.769×10	2.754×10	2.742×10	2.743×10
$n = 100$				
$\sigma^2/R_x = 0.01$				
ASE	2.527×10^{-4}	2.572×10^{-4}	2.574×10^{-4}	2.573×10^{-4}
λ	0.782×10	1.578×10	1.759×10	1.631×10
SD(λ)	1.778×10	2.536×10	2.537×10	2.564×10
$\sigma^2/R_x = 0.05$				
ASE	6.175×10^{-3}	6.199×10^{-3}	6.287×10^{-3}	6.203×10^{-3}
λ	0.784×10	0.851×10	1.072×10	0.906×10
SD(λ)	1.914×10	2.015×10	2.215×10	2.069×10

the squared differences between the predicted values $\hat{y}_\alpha$ and the observed values y_α $(\alpha = 1, \ldots, 35)$ for 100 bootstrap replications were 6.27×10^{-11} by the GIC and 1.21×10^{-10} by CV. For comparison with CV as a model selector, we observe the efficiency of the proposed modeling procedure based on the Gaussian basis function with the GIC.

Table 2 Comparison of the average squared errors (ASE) based on various criteria for the simulation (b)

	GIC	MAIC	CV	GCV
$n = 25$				
$\sigma^2/R_x = 0.01$				
ASE	1.534×10^{-1}	1.562×10^{-1}	1.633×10^{-1}	1.655×10^{-1}
λ	1.028×10^{-2}	1.198×10^{-2}	1.571×10^{-2}	1.669×10^{-2}
SD(λ)	7.266×10^{-3}	7.593×10^{-3}	8.760×10^{-3}	8.891×10^{-3}
$\sigma^2/R_x = 0.05$				
ASE	2.012×10^{-1}	2.017×10^{-1}	2.060×10^{-1}	2.063×10^{-1}
λ	0.723×10^{-2}	0.818×10^{-2}	1.135×10^{-2}	1.158×10^{-2}
SD(λ)	5.782×10^{-3}	6.160×10^{-3}	6.561×10^{-3}	6.892×10^{-3}
$n = 50$				
$\sigma^2/R_x = 0.01$				
ASE	1.762×10^{-1}	1.766×10^{-1}	1.782×10^{-1}	1.789×10^{-1}
λ	6.815×10^{-3}	7.158×10^{-3}	8.469×10^{-3}	8.765×10^{-3}
SD(λ)	4.358×10^{-3}	4.415×10^{-3}	5.136×10^{-3}	5.296×10^{-3}
$\sigma^2/R_x = 0.05$				
ASE	2.127×10^{-1}	2.131×10^{-1}	2.142×10^{-1}	2.144×10^{-1}
λ	6.115×10^{-3}	6.503×10^{-3}	7.282×10^{-3}	7.643×10^{-3}
SD(λ)	3.505×10^{-3}	3.569×10^{-3}	3.791×10^{-3}	3.556×10^{-3}
$n = 100$				
$\sigma^2/R_x = 0.01$				
ASE	1.910×10^{-1}	1.911×10^{-1}	1.916×10^{-1}	1.917×10^{-1}
λ	5.778×10^{-3}	5.942×10^{-3}	6.387×10^{-3}	6.494×10^{-3}
SD(λ)	3.410×10^{-3}	3.496×10^{-3}	3.661×10^{-3}	3.739×10^{-3}
$\sigma^2/R_x = 0.05$				
ASE	2.094×10^{-1}	2.096×10^{-1}	2.102×10^{-1}	2.102×10^{-1}
λ	4.158×10^{-3}	4.327×10^{-3}	4.595×10^{-3}	4.655×10^{-3}
SD(λ)	2.245×10^{-3}	2.401×10^{-3}	2.404×10^{-3}	2.448×10^{-3}

8 Summary and concluding remarks

We introduced functional regression modelings, using Gaussian basis functions with the method of regularization. We first transfer the vector-valued observations to a set of functions. Second, functional regression models are constructed by using the property

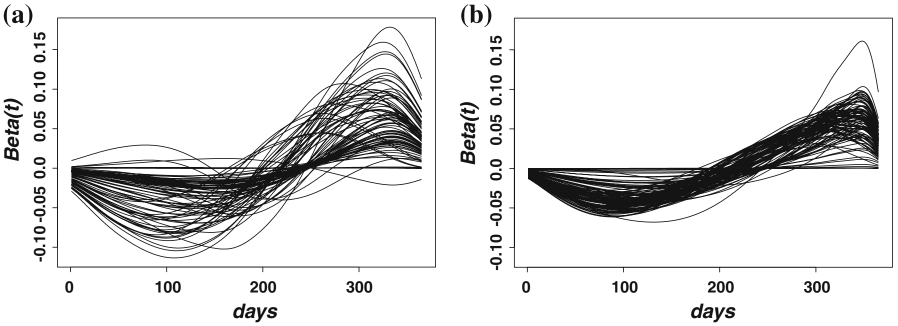

Fig. 2 **a** Estimated regression coefficient functions based on cross-validation (CV) and **b** estimated regression coefficient functions based on GIC for 100 bootstrap replications

that the integral of the product of any two Gaussian basis functions can be directly calculated. In order to select adjusted parameters, we derived model selection criteria within the framework of functional regression modeling from an information-theoretic approach.

In recent years, statistical challenges arise in such areas as genome databases in life science, motion data in robotics, POS data in marketing and economic data. Especially in the analysis of genome science the number of variables is much greater than the number of observations. One way to handle the large number of variables is to employ techniques in the functional data analysis. In practice it is required to use a flexible instrument for transforming each individual's observations into functional form. We observed that Gaussian bases produce a variety of functional forms, using the method of regularization and the model selection criterion GIC given in Sect. 6. Our modeling strategies may be applied to the problem of constructing a discriminant rule based on a collection of functional data, which will be discussed in another paper.

Acknowledgments The authors would like to thank the anonymous reviewers for constructive and helpful comments that improved the quality of the paper considerably.

References

Ando, T., Imoto, S., Konishi, S. (2001). Estimating nonlinear regression models based on Gaussian basis function networks (in Japanese). *Japanese Journal of Applied Statistics*, *30*, 19–35.

Ando, T., Konishi, S., Imoto, S. (2005). Nonlinear regression modeling via regularized Gaussian basis function networks. *Journal of Statistical Planning and Inference* (in press).

Araki, Y., Konishi, S. (2006). Functional supervised and unsupervised classification of gene expression data. In: *Proceedings in computational statistics 2006*, pp. 1105–1112. Physica-Verlag/Springer.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, *31*, 377–403.

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *The Annals of Statistics*, *7*, 1–26.

Green, P. J., Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. London: Chapman & Hall.

Hastie, T., Tibshirani, R. (1990). *Generalized additive models*. London: Chapman & Hall.

Hurvich, C. M., Simonoff, J. S., Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B*, *60*, 359–373.

James, G. (2002). Generalized linear models with functional predictor variables. *Journal of the Royal Statistical Society B*, *64*, 411–432.

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

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Linhart, H., Zucchini, W. (1986). Finite sample selection criteria for multinomial models. *Statistische Hefte*, *27*, 173–178.

Marx, B. D., Eilers, P. H. C. (1999). Generalized linear regression for sampled signals or curves: A P-spline approach. *Technometrics*, *41*, 1–13.

Mizuta, M. (2006). Discrete functional data analysis. In: *Proceedings in computational statistics 2006*, pp. 361–369. Physica-Verlag/Springer.

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

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, *135*, 370–384.

Ramsay, J. O., Silverman, B. W. (2002). *Applied functional data analysis*. New York: Springer-Verlag.

Ramsay, J. O., Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer-Verlag.

Rao, C. R., Wu, Y. (2001). Model selection. In: P. Lahiri (Ed.), *Model selection: IMS lecture notes-monograph series*, pp. 1–18.

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