NONLINEAR REGRESSION MODELING VIA REGULARIZED GAUSSIAN BASIS FUNCTIONS

Kawano, Shuichi
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Graduate School of Mathematics, Kyushu University

NONLINEAR REGRESSION MODELING VIA REGULARIZED GAUSSIAN BASIS FUNCTIONS

by

Shuichi KAWANO and **Sadanori K**ONISHI

NONLINEAR REGRESSION MODELING VIA REGULARIZED GAUSSIAN BASIS FUNCTIONS

By

Shuichi KAWANO* and **Sadanori** KONISHI†

Abstract

Nonlinear regression modeling based on basis expansions has been widely used to explore data with complex structure. There are various types of basis functions to capture complex nonlinear phenomena. In this paper we introduce nonlinear regression models with Gaussian basis functions, for which new Gaussian bases are constructed, taking advantages of *B*-spline bases. In order to choose adjusted parameters, we derive model selection and evaluation criteria from information-theoretic and Bayesian viewpoints. Monte Carlo simulations and real data analysis show that our proposed modeling strategy performs well in various situations.

Key Words and Phrases: Basis expansion, Bayes approach, Information criterion, Nonlinear regression, Regularization.

1. Introduction

Nonlinear regression models based on basis expansions provide a useful tool to analyze data with complex structure. The essential idea behind basis expansions is to express a regression function as a linear combination of prescribed functions, called basis functions (Hastie *et al.* (2001), Konishi and Kitagawa (2008)). In the model building process, the basis functions are chosen according to data structure under consideration. The Fourier series is useful if observed data are periodic. For non-periodic data, splines (Green and Silverman (1994)), *B*-splines (de Boor (2001), Imoto and Konishi (2003)) and radial basis functions (Bishop (1995), Ripley (1996)) have been widely used to construct nonlinear regression models. Especially Gaussian basis function models with a hybrid learning method that combines self-organized and supervised learning have advantages over multilayer perceptrons, namely faster convergence, and are free from identification problem (Moody and Darken (1989), Bishop (1995), Konishi *et al.* (2004), Ando *et al.* (2006)).

Gaussian basis functions have center and width parameters that have to be determined from observed data. In particular, the width parameters adjust the amount of overlapping among basis functions and play the essential role to capture the structure in the data over the region of the input space. Moody and Darken (1989) used the k-means clustering algorithm to determine the width parameters. Ando *et al.* (2006) proposed Gaussian bases with hyper-parameter that controls the amount of overlapping among basis functions and showed the efficiency of the nonlinear regression modeling based on the proposed Gaussian bases.

* Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. s-kawano@math.kyushu-u.ac.jp

† Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. konishi@math.kyushu-u.ac.jp

The problem, however, still remains in constructing Gaussian bases from a finite and noisy data. First, the *k*-means clustering algorithm depends on initial values and consequently yields different Gaussian basis functions corresponding to each set of initial values. Second, the computational time required becomes enormous, since the value of hyper-parameter has to be chosen in addition to the number of basis functions and the value of a regularization parameter. In order to overcome these problems, we introduce new Gaussian basis functions with a beneficial property of *B*-spline bases and present a nonlinear regression modeling strategy that captures the structure in the data over the region of the input space. Our Gaussian basis function regression models can be easily applied to analyze the continuous and also discontinuous data, and yield stable estimated regression functions. We also note that adjusted parameters included in our proposed regression models are only the number of basis functions and a regularization parameter. This provides the lower computational cost than that of the models proposed previously. To choose these parameters, we derive model selection criteria from information-theoretic and Bayesian viewpoints. The proposed nonlinear modeling procedure is investigated through the analysis of real data and Monte Carlo simulations.

This paper is organized as follow. Section 2 describes a framework of basis expansions and Gaussian basis functions proposed previously. In Section 3 we present new Gaussian basis functions, taking advantages of *B*-spline bases. Section 4 provides nonlinear regression modeling strategies based on proposed Gaussian basis functions and model selection criteria for evaluating statistical models estimated by a regularization method. In Section 5 we investigate the performance of the proposed nonlinear regression modeling techniques, using real data example and Monte Carlo simulations. Some concluding remarks are described in Section 6.

2. Preliminaries

Suppose that we have *n* independent observations $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, \cdots, n\}$, where y_α are random response variables and $\boldsymbol{x}_\alpha$ are vectors of *p*-dimensional explanatory variables. We consider a regression model

$$y_\alpha = u(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \qquad \alpha = 1, \cdots, n, \tag{1}$$

where $u(\cdot)$ is a true smooth function and ε_α $(\alpha = 1, \cdots, n)$ are error terms. It is assumed that the function $u(\cdot)$ is expressed as a linear combination of basis functions $\phi_i(\boldsymbol{x})$ $(i = 1, \cdots, m)$ in the form

$$u(\boldsymbol{x}; \boldsymbol{w}) = \sum_{i=1}^{m} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \cdots, \phi_m(\boldsymbol{x}))^T$ is a vector of basis functions and $\boldsymbol{w} = (w_1, \cdots, w_m)^T$ is an unknown coefficient parameter vector.

For a *p*-dimensional vector of explanatory variables $\boldsymbol{x} = (x_1, \cdots, x_p)^T$, Gaussian basis functions are given by

$$\phi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, h_i^2) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_i||^2}{2h_i^2}\right), \quad i = 1, \cdots, m, \tag{3}$$

where $\boldsymbol{\mu}_i$ is a *p*-dimensional parameter vector determining the center of the basis function, h_i^2 is a parameter that determines the width and $||\cdot||$ is the Euclidian norm.

Unknown parameters contained in the nonlinear regression model based on Gaussian basis functions are the coefficient parameters w_i ($i = 1, \cdots, m$), the centers $\boldsymbol{\mu}_i$ and the width parameters h_i^2. These parameters are generally estimated by a two-stage procedure in order to avoid the problem of local minimum and the identification problem (Moody and Darken (1989)). In the first stage, we determine the centers $\boldsymbol{\mu}_i$ and widths h_i^2, for which we use the k-means clustering algorithm. This algorithm divides the data set of explanatory variables $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ into m clusters $\{C_1, \cdots, C_m\}$ that correspond to the number of basis functions. The centers $\boldsymbol{\mu}_i$ and widths h_i^2 are then determined by

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \sum_{\boldsymbol{x}_\alpha \in C_i} \boldsymbol{x}_\alpha, \qquad \hat{h}_i^2 = \frac{1}{n_i} \sum_{\boldsymbol{x}_\alpha \in C_i} ||\boldsymbol{x}_\alpha - \hat{\boldsymbol{\mu}}_i||^2, \tag{4}$$

where n_i is the number of observations which belongs to the i-th cluster C_i. Replacing $\boldsymbol{\mu}_i$ and h_i^2 in equation (3) by $\hat{\boldsymbol{\mu}}_i$ and $\hat{h}_i^2$, respectively, we have a set of m basis functions

$$\phi_i(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_i, \hat{h}_i^2) = \exp\left(-\frac{||\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i||^2}{2\hat{h}_i^2}\right), \quad i = 1, \cdots, m. \tag{5}$$

In the second stage, we estimate the coefficient parameters w_i ($i = 1, \cdots, m$) by maximizing a penalized log-likelihood function. The details are described in Section 4.

It is, however, noted that the basis functions constructed by the above procedure cannot fully capture beneficial information from the data. The cause is the lack of the amount of overlapping among basis functions. To overcome this problem, Ando *et al.* (2006) proposed Gaussian basis functions with hyper-parameter ν (> 0) in the form

$$\phi_i(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_i, \hat{h}_i^2, \nu) = \exp\left(-\frac{||\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i||^2}{2\nu\hat{h}_i^2}\right), \quad i = 1, \cdots, m, \tag{6}$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{h}_i^2$ are given by (4). Ando *et al.* (2006) showed that nonlinear regression models using these basis functions could capture the information from the data very well.

Basis functions in equations (5) and (6) constructed by the k-means clustering algorithm, however, have some drawbacks. Figure 1 shows the comparison of the position of basis functions based on the same data set of size 40 generated from a uniform distribution on $[0, 1]$. This implies that the constructed basis functions depend on initial values in the k-means clustering algorithm and are not uniquely determined. In consequence, the estimated regression functions vary according to initial values in the clustering algorithm. To overcome this difficulty, we propose new Gaussian basis function with the help of B-spline bases in the next section. Namely we introduce Gaussian bases with centers and width parameters determined by preassigned knots without using the clustering algorithms.

3. Gaussian basis functions

3.1. Curve fitting

Suppose that we have n independent observations $\{(y_\alpha, x_\alpha); \alpha = 1, \cdots, n\}$, where y_α are random response variables and x_α are univariate explanatory variables. It is assumed that the observations $\{x_\alpha; \alpha = 1, \cdots, n\}$ are arranged by magnitude as $x_1 < x_2 < \cdots < x_n$. We set up the knots t_i ($i = 1, \cdots, m + 4$) as follows:

$$t_1 < t_2 < t_3 < t_4 = x_1 < t_5 < \cdots < t_m < t_{m+1} = x_n < t_{m+2} < t_{m+3} < t_{m+4}, \tag{7}$$
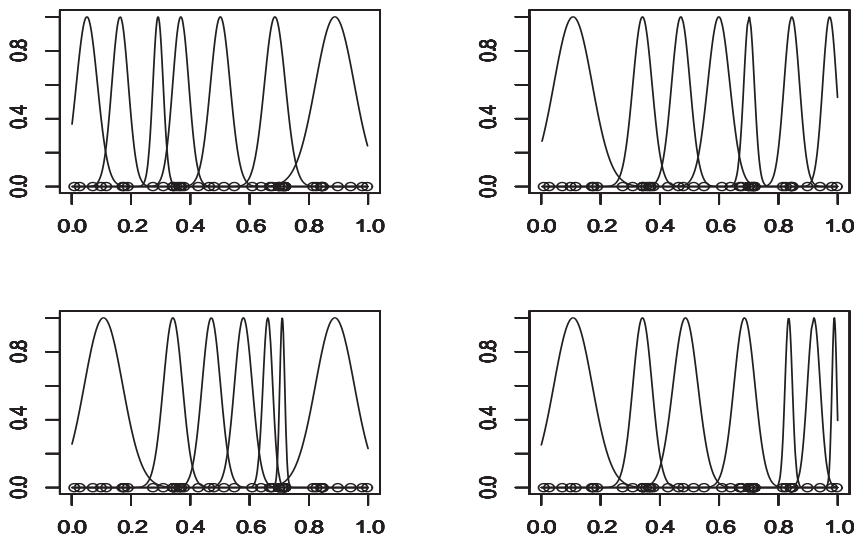
Figure 1: Comparison of the position of Gaussian basis functions for the same data set.

where the knots are equally spaced. By setting the knots in this way the n observations are divided into $(m-3)$ intervals:

$$[t_4, t_5], [t_5, t_6], \cdots, [t_m, t_{m+1}]. \tag{8}$$

Next, we set up basis functions such that each interval is covered by four basis functions as B-spline basis functions of degree 3. We therefore define each basis function which has a center t_i and a width $h = (t_i - t_{i-2})/3$ for $i = 3, \cdots, m+2$. Under this setting, Gaussian basis functions are given by

$$\begin{aligned} \phi_i(x; t_i, h^2) &= \exp\left\{-\frac{(x-t_i)^2}{2h^2}\right\}, \\ h &= \frac{t_i - t_{i-2}}{3}, \qquad i = 3, \cdots, m+2. \end{aligned} \tag{9}$$

It might be noticed that the width parameter h^2 is determined such that each interval is covered by four Gaussian basis functions (see Figure 2). Hence we obtain m Gaussian basis functions for the knots $t_1, \cdots, t_{m+4}$.

3.2. Surface fitting in two-dimensional case

Suppose that we have n observations $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, \cdots, n\}$, where $\boldsymbol{x}_\alpha = (x_{1\alpha}, x_{2\alpha})^T$ are observed values of two-dimensional explanatory variables. First, along x_1 axis, we set equispaced knots as

$$\begin{aligned} t_1 < t_2 < t_3 < t_4 &= \min(x_{1\alpha}) < t_5 < \cdots \\ &< t_r < t_{r+1} = \max(x_{1\alpha}) < t_{r+2} < t_{r+3} < t_{r+4}, \end{aligned} \tag{10}$$
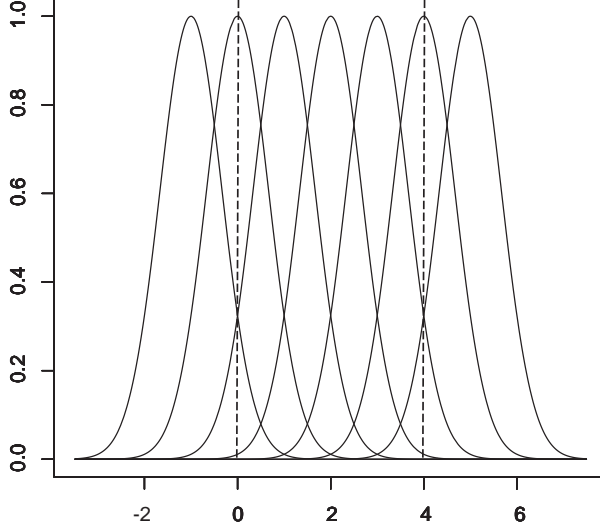
Figure 2: Spline-based Gaussian basis functions over the region of the input space $[0,4]$.

and along x_2 axis the knots are equally set as

$$
\begin{aligned}
s_1 < s_2 < s_3 < s_4 \quad = \quad & \min(x_{2\alpha}) < s_5 < \cdots \\
& < s_l < s_{l+1} = \max(x_{2\alpha}) < s_{l+2} < s_{l+3} < s_{l+4}. \quad (11)
\end{aligned}
$$

Secondly, by taking centers of basis functions as $\boldsymbol{\mu}_{ij} = (t_i, s_j)^T$ $(i = 3, \cdots, r+2; \; j = 3, \cdots, l+2)$, we obtain the following two-dimensional Gaussian basis functions;

$$
\phi_k(\boldsymbol{x}; \boldsymbol{\mu}_k, H) = \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T H^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) \right\}, \quad (12)
$$

where $k = l \times (i-3) + (j-2)$, and H is

$$
H = \left[\begin{array}{cc} h_1^2 & 0 \\ 0 & h_2^2 \end{array} \right]. \quad (13)
$$

Here $h_1 = (t_i - t_{i-2})/3$ and $h_2 = (s_j - s_{j-2})/3$, determined by the same argument as discussed in Subsection 3.1. Thus we obtain $(r+l)$ Gaussian basis functions for the preassigned knots $t_1, \cdots, t_{r+4}$ and $s_1, \cdots, s_{l+4}$.

4. Estimation

For n independent observations $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, \cdots, n\}$, a nonlinear regression model based on spline-based Gaussian basis functions $\phi_i(\boldsymbol{x})$ $(i = 1, \cdots, m)$ given in Section 3 is

$$
\begin{aligned}
y_\alpha \quad &= \quad \sum_{i=1}^{m} w_i \phi_i(\boldsymbol{x}_\alpha) + \varepsilon_\alpha \\
&= \quad \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \qquad \alpha = 1, \cdots, n, \quad (14)
\end{aligned}
$$

where $\boldsymbol{w} = (w_1, \cdots, w_m)^T$, $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \cdots, \phi_m(\boldsymbol{x}))^T$ and ε_α are error terms. If the error terms ε_α are independently and normally distributed with mean 0 and variance σ^2, a probability density of response y_α is given by

$$f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\{y_\alpha - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}^2}{2\sigma^2}\right]. \tag{15}$$

In estimating unknown parameters $\boldsymbol{\theta} = (\boldsymbol{w}^T, \sigma^2)^T$, the maximum likelihood method yields unstable parameter estimates and leads to overfitting. In such cases the unknown parameters are estimated by maximizing the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{\theta}) = \sum_{\alpha=1}^n \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{\theta}) - \frac{n\lambda}{2}\boldsymbol{w}^T K \boldsymbol{w}, \tag{16}$$

where λ (>0) is a regularization parameter that controls the smoothness of the fitted model and K is a known $m \times m$ nonnegative definite matrix (for details, Imoto and Konishi (2003), Konishi and Kitagawa (2008, Chapter 5)). The typical form of K is given by $K = I_m$ for the identity matrix or $K = D_2^T D_2$ for a second-order difference matrix.

The maximum penalized likelihood estimators $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{w}}^T, \hat{\sigma}^2)^T$ are given by

$$\hat{\boldsymbol{w}} = (\Phi^T\Phi + n\lambda\hat{\sigma}^2 K)^{-1}\Phi^T\boldsymbol{y}, \quad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}})^T(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}}), \tag{17}$$

where $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_n))^T$ and $\boldsymbol{y} = (y_1, \cdots, y_n)^T$.

Then we obtain a statistical model by replacing the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{w}^T, \sigma^2)^T$ with the corresponding estimators $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{w}}^T, \hat{\sigma}^2)^T$ in the form

$$f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{\{y_\alpha - \hat{\boldsymbol{w}}^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}^2}{2\hat{\sigma}^2}\right]. \tag{18}$$

The estimators depend upon the number of basis functions m and a regularization parameter λ, and we need to select suitable values of these adjusted parameters. We give criteria for evaluating constructed models from information-theoretic and Bayes approaches in order to select the adjusted parameters.

Konishi and Kitagawa (1996) introduced evaluation criteria of statistical models that can be applied to the evaluation of statistical models estimated by various types of estimation procedures such as the robust and penalized likelihood procedures. By using the result, the model selection criterion for evaluating the statistical model constructed by spline-based Gaussian basis functions is given by

$$\text{GIC} = n\{\log(2\pi) + 1\} + n\log\hat{\sigma}^2 + 2\text{tr}\{Q^{-1}R\}, \tag{19}$$

where Q and R are $(m+1) \times (m+1)$ matrices and are, respectively, given by

$$Q = \frac{1}{n\hat{\sigma}^2}\begin{bmatrix} \Phi^T\Phi + n\lambda\hat{\sigma}^2 K & \dfrac{1}{\hat{\sigma}^2}\Phi^T\Lambda\mathbf{1}_n \\[2mm] \dfrac{1}{\hat{\sigma}^2}\mathbf{1}_n^T\Lambda\Phi & \dfrac{n}{2\hat{\sigma}^2} \end{bmatrix}, \tag{20}$$

$$R = \frac{1}{n\hat{\sigma}^2}\begin{bmatrix} \dfrac{1}{\hat{\sigma}^2}\Phi^T\Lambda^2\Phi - \lambda K\hat{\boldsymbol{w}}\mathbf{1}_n^T\Lambda\Phi & \dfrac{1}{2\hat{\sigma}^4}\Phi^T\Lambda^3\mathbf{1}_n - \dfrac{1}{2\hat{\sigma}^2}\Phi^T\Lambda\mathbf{1}_n \\[2mm] \dfrac{1}{2\hat{\sigma}^4}\mathbf{1}_n^T\Lambda^3\Phi - \dfrac{1}{2\hat{\sigma}^2}\mathbf{1}_n^T\Lambda\Phi & \dfrac{1}{4\hat{\sigma}^6}\mathbf{1}_n^T\Lambda^4\mathbf{1}_n - \dfrac{n}{4\hat{\sigma}^2} \end{bmatrix} \tag{21}$$

with $\mathbf{1}_n = (1, 1, \cdots, 1)^T$ and $\Lambda = \mathrm{diag}[y_1 - \hat{\boldsymbol{w}}^T\boldsymbol{\phi}(\boldsymbol{x}_1), \cdots, y_n - \hat{\boldsymbol{w}}^T\boldsymbol{\phi}(\boldsymbol{x}_n)]$.

Konishi *et al.* (2004) extended Schwarz's (Schwarz (1978)) Bayesian information criterion (BIC) to evaluate statistical models estimated by the maximum penalized likelihood method. By using the result, the model selection criterion for evaluating the statistical model in equation (18) is given by

$$
\begin{aligned}
\mathrm{GBIC} \quad = \quad & n\log\hat{\sigma}^2 + n\lambda\hat{\boldsymbol{w}}^T K\hat{\boldsymbol{w}} + n + n\log(2\pi) + (d+1)\log n \\
& + \log|R| - \log|K|_+ - (m-2)\log\lambda - (d+1)\log(2\pi),
\end{aligned} \tag{22}
$$

where $|K|_+$ is the product of the positive eigenvalues of K with rank d and R is an $m \times m$ matrix given by

$$
R = \frac{1}{n\hat{\sigma}^2}
\begin{bmatrix}
\Phi^T\Phi + n\lambda\hat{\sigma}^2 K & \dfrac{1}{\hat{\sigma}^2}\Phi^T\boldsymbol{e} \\[2mm]
\dfrac{1}{\hat{\sigma}^2}\boldsymbol{e}^T\Phi & \dfrac{n}{2\hat{\sigma}^2}
\end{bmatrix} \tag{23}
$$

with $\boldsymbol{e} = (y_1 - \hat{\boldsymbol{w}}^T\boldsymbol{\phi}(\boldsymbol{x}_1), \cdots, y_n - \hat{\boldsymbol{w}}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^T$ being an n-dimensional residual vector.

We select the optimal value of the number of basis functions and a regularization parameter that minimize either GIC or GBIC.

5. Numerical examples

5.1. Curve fitting

We investigate the performance of our proposed method through the analysis of the motorcycle impact data (Silverman (1985), Härdle (1990)). The motorcycle impact data are a series of measurements of head acceleration in units of gravity and times in millisecond after impact. The data have been widely used in order to illustrate smoothing techniques.

A nonlinear regression model with spline-based Gaussian basis functions was applied to the motorcycle impact data, in which the model is estimated by maximizing the penalized log-likelihood function in equation (16) with second-order difference matrix $D_2^T D_2$. The number of basis functions m and a regularization parameter λ were selected by using the criteria GIC in equation (19) and GBIC in equation (22). The criterion GIC selected $m = 17$ and $\lambda = 3.98 \times 10^{-6}$, while GBIC $m = 17$ and $\lambda = 3.16 \times 10^{-6}$. The left panel of Figure 3 shows the fitted curve by GBIC.

Next we show an instability of nonlinear regression models based on Gaussian basis functions with hyper-parameter, constructed by using the k-means clustering algorithm. Constructing the models was implemented for 10 times against the same motorcycle impact data, where the number of basis functions m was fixed as 12, and a regularization parameter and hyper-parameter ν were chosen by minimizing GBIC. The right panel in Figure 3 shows 10 estimated smooth curves. We observe that the models constructed by using the k-means clustering method are unstable because the k-means clustering method estimates difference centers of basis functions depending on a set of initial values.

Furthermore, to investigate the stability of our nonlinear regression models with spline-based Gaussian basis functions, 100 sets of bootstrap samples were generated from the motorcycle impact data, and the models with spline-based Gaussian basis functions and with basis functions using the k-means method were fitted to each bootstrap sample. Figure 4 shows the curves estimated by 100 bootstrap replications, where the number of basis functions m, a regularization parameter λ and hyper-parameter ν in the models based on the k-means method were
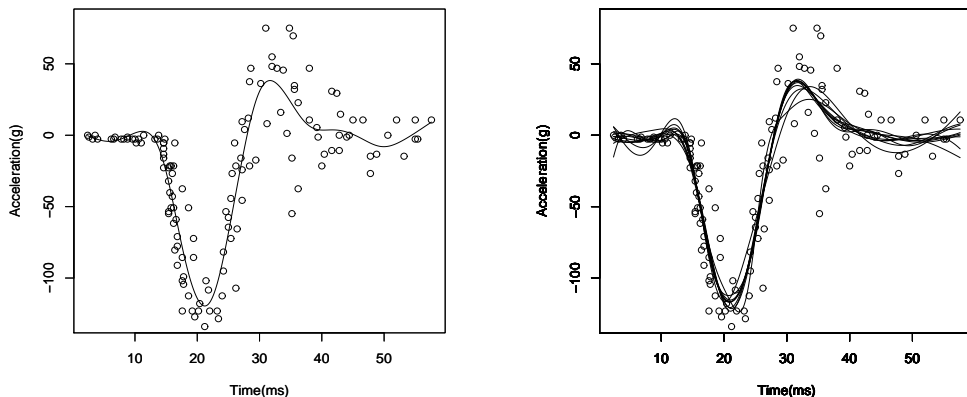
Figure 3: The motorcycle impact data and estimated curves using our proposed procedure (left). Smoothed curves based on nonlinear regression models constructed by the *k*-means clustering algorithm for the same data set (right).

selected by GBIC. It might be seen from Figure 4 that our proposed modeling procedure yields stable estimated curves, and decreases the boundary bias considerably because spline-based Gaussian basis functions are appropriately allocated in the boundary regions.

5.2. Surface fitting

We applied a nonlinear regression modeling with spline-based Gaussian basis functions to data $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, \cdots, 400\}$ generated from

$$y_\alpha = \sin(2\pi x_{1\alpha}) + 2\cos(2\pi x_{2\alpha}) + \varepsilon_\alpha, \qquad \varepsilon_\alpha \sim N(0,1), \tag{24}$$

where $x_{1\alpha}$ and $x_{2\alpha}$ were generated independently from a uniform distribution on $[0,2]$ respectively. The model was estimated by maximizing penalized log-likelihood functions given by (16), and the number of basis functions m and a regularization parameter λ were chosen by using GIC and GBIC. The criterion GIC gave $m = 56$ and $\lambda = 7.94 \times 10^{-5}$, while GBIC gave $m = 63$ and $\lambda = 7.94 \times 10^{-4}$. In Figure 5 the left panel shows the true surface, and the right panel the estimated surface evaluated by GBIC. We observe that our modeling procedure captures the true structure effectively.

5.3. Monte Carlo simulation

In this section, Monte Carlo simulations are conducted to investigate the effectiveness of our proposed nonlinear regression modeling. We consider two simulation examples to illustrate the properties of our proposed method. For the first simulation study, we repeatedly generated random samples $\{(y_\alpha, x_\alpha); \alpha = 1, \cdots, 100\}$ using the true regression model $y_\alpha = u(x_\alpha) + \varepsilon_\alpha$. The design points are uniformly distributed in $[0,1]$ and the noises ε_α are independently, normally distributed with mean 0 and standard deviation σ, where the standard deviation is taken as $\sigma = 0.1R_u$ or $0.2R_u$ with R_u being the range of $u(x)$ over $x \in [0,1]$. We consider the following
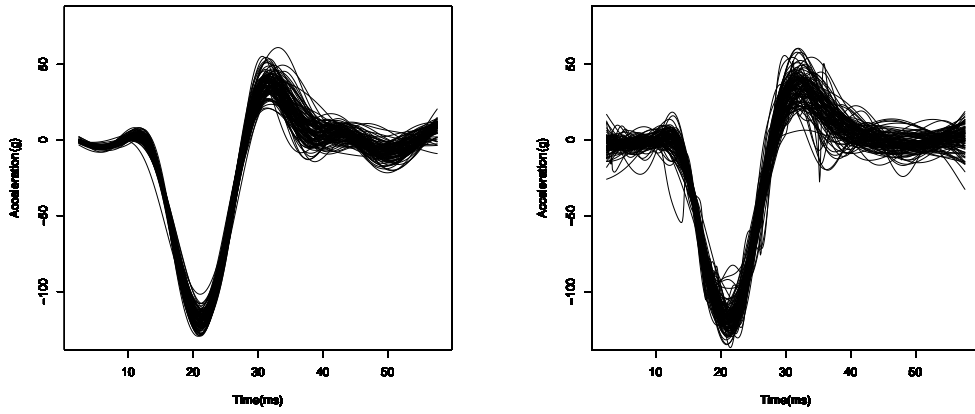
Figure 4: Estimated smoothed curves for 100 bootstrap replications. The left panel shows the result for our proposed method. The right panel shows the result for the models constructed by the *k*-means clustering method.
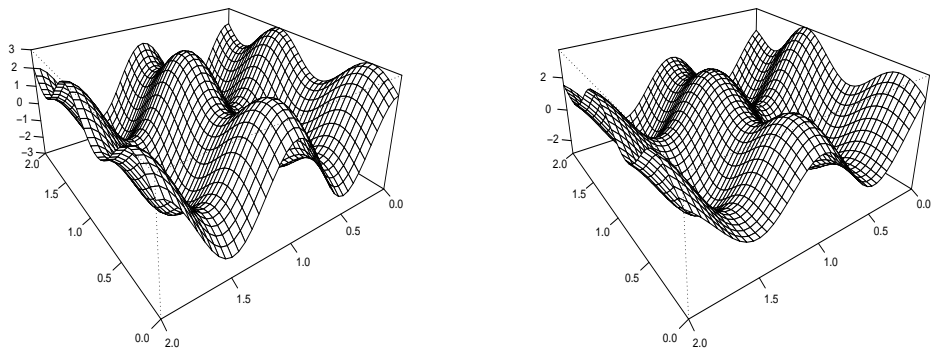

Figure 5: Comparison of the true surface and the smoothed surface. The left panel shows the true surface while the right panel the estimated surface evaluated by GBIC.

functions as true regression model.

$$
\begin{array}{rcll}
(a) & u(x) & = & 10\exp(-10x), \\
(b) & u(x) & = & \sin(2\pi x^3), \\
(c) & u(x) & = & \sin(2\pi x)\exp(-5x).
\end{array}
\tag{25}
$$

We compare the performance of nonlinear regression models based on spline-based Gaussian basis functions (SGB) with that of various models based on basis functions determined by another five methods and B-spline basis functions (BS). As another methods, we used basis functions with hyper-parameter ν given by (6) (HGB), those proposed by Moody and Darken (1989) and those based on "P-nearest neighbor" heuristic. Moody and Darken (1989) constructed basis functions based on the centers defined by (4) and the width parameters determined by the averaged Euclidian distance of its P-nearest neighbor of each basis function. We especially consider two procedures that use the first nearest neighbor of each center (MD1) and the second nearest neighbor of each center (MD2). "P-nearest neighbor" method replaces the averaged Euclidian distance by the Euclidian distance in determining the width parameters. We consider $P = 1$ (PN1) and $P = 2$ (PN2) in our simulation settings.

We fitted regression models based on Gaussian basis functions to the simulated data. The number of basis functions m and the regularization parameter λ were selected by using the criteria, GIC and GBIC. Furthermore in the models based on the basis functions given by (6), the hyper-parameter ν was also chosen as minimizers of the criteria.

We calculated the average squared error (ASE) defined by $\text{ASE} = \sum_{\alpha=1}^{100}\{u(x_\alpha)-\hat{y}_\alpha\}^2/100$ in order to assess the goodness of fit. Table 1 shows the summaries for the simulation results in the case of using the GBIC, while Table 2 shows those in the case of using the GIC. The values of SD indicate standard deviations for the ASE. The simulation results were obtained by averaging over 100 Monte Carlo trials. It might be seen from Tables 1 and 2 that our proposed models are superior to the others in almost all cases in the sense that our models have a minimum of ASE and SD.

In the second experiment we created random samples $\{(y_\alpha, x_{1\alpha}, x_{2\alpha}); \alpha = 1, \cdots, 100\}$ from the true model $y_\alpha = u(x_{1\alpha}, x_{2\alpha}) + \varepsilon_\alpha$, where

$$
\begin{array}{rl}
(1) & u(x_1, x_2) = \sin(2\pi x_1) + \cos(2\pi x_2), \\
\\
(2) & u(x_1, x_2) = 3\sin(\pi x_1 x_2) + 2\sin(\pi x_1 + x_2)
\end{array}
\tag{26}
$$

and the design points are uniformly distributed in $[-1, 1] \times [-1, 1]$. The noises ε_α are assumed to be independently distributed according to the normal distributions with mean 0 and standard deviation σ, where the standard deviation is taken as $\sigma = 0.2R_u$ with R_u being the range of $u(x_1, x_2)$ over $(x_1, x_2) \in [-1, 1] \times [-1, 1]$.

We compare the performance of our proposed modeling procedure with that of six models used in the first simulation. The model selection criteria GIC and GBIC were used by choosing the number of basis functions m and a regularization parameter λ, and hyper-parameter ν only if the models given by (6) were used. To evaluate the performance of each model we calculated the average squared error $\text{ASE} = \sum_{\alpha=1}^{100}\{\hat{y}_\alpha - u(x_{1\alpha}, x_{2\alpha})\}^2/100$, and described boxplots of these values for the 100 trials of Monte Carlo simulations. Figures 6 and 7 show the boxplots of the average squared errors. In almost all situations, it might be seen that our proposed modeling procedure performs well; it yields a relatively small median with small variance.

Table 1: Comparison of the average squared errors for GBIC.

function	σ/R_u		SGB	BS	HGB	MD1	MD2	PN1	PN2
(a)	0.1	ASE $(\times 10^1)$	0.800	0.797	1.400	1.774	1.667	1.613	1.271
		SD $(\times 10^2)$	2.738	2.728	4.928	5.970	6.137	4.569	4.490
	0.2	ASE $(\times 10^1)$	2.880	2.927	5.427	6.353	5.613	6.326	5.545
		SD $(\times 10^1)$	1.316	1.354	2.538	2.399	2.396	2.960	2.238
(b)	0.1	ASE $(\times 10^3)$	3.868	2.927	4.576	5.617	5.659	4.883	4.652
		SD $(\times 10^3)$	1.604	1.624	1.969	2.278	2.490	2.255	2.100
	0.2	ASE $(\times 10^2)$	1.491	1.568	1.690	1.771	1.782	1.775	1.663
		SD $(\times 10^3)$	7.472	7.568	8.599	8.784	8.577	8.552	7.965
(c)	0.1	ASE $(\times 10^4)$	1.730	1.803	2.620	2.655	2.672	2.855	2.595
		SD $(\times 10^4)$	0.746	0.773	1.113	1.002	0.982	1.061	1.059
	0.2	ASE $(\times 10^4)$	5.872	6.191	8.938	8.101	7.762	9.344	8.262
		SD $(\times 10^4)$	2.539	2.655	3.993	3.443	3.256	4.201	3.737

Table 2: Comparison of the average squared errors for GIC.

| function	σ/R_u		SGB	BS	HGB	MD1	MD2	PN1	PN2
(a)	0.1	ASE $(\times 10^1)$	0.865	0.879	1.337	1.605	1.514	1.566	1.150
		SD $(\times 10^2)$	3.925	4.066	4.958	4.454	5.014	5.371	4.893
	0.2	ASE $(\times 10^1)$	2.815	2.757	4.095	4.279	3.797	4.628	3.724
		SD $(\times 10^1)$	1.217	1.177	1.608	1.333	1.433	1.487	1.420
(b)	0.1	ASE $(\times 10^3)$	4.214	4.353	5.704	5.584	5.255	5.841	5.359
		SD $(\times 10^3)$	2.154	2.257	2.796	2.328	2.300	2.593	2.417
	0.2	ASE $(\times 10^2)$	1.725	1.760	2.221	1.888	1.750	2.133	1.851
		SD $(\times 10^2)$	0.760	0.728	1.077	0.765	0.686	0.954	0.811
(c)	0.1	ASE $(\times 10^4)$	1.873	1.903	2.510	2.395	2.103	2.567	2.312
		SD $(\times 10^4)$	0.892	0.897	1.060	0.988	0.828	1.093	0.991
	0.2	ASE $(\times 10^4)$	7.441	7.483	9.753	8.669	8.104	9.132	8.852
		SD $(\times 10^4)$	3.851	3.798	4.614	3.974	3.485	3.875	4.195
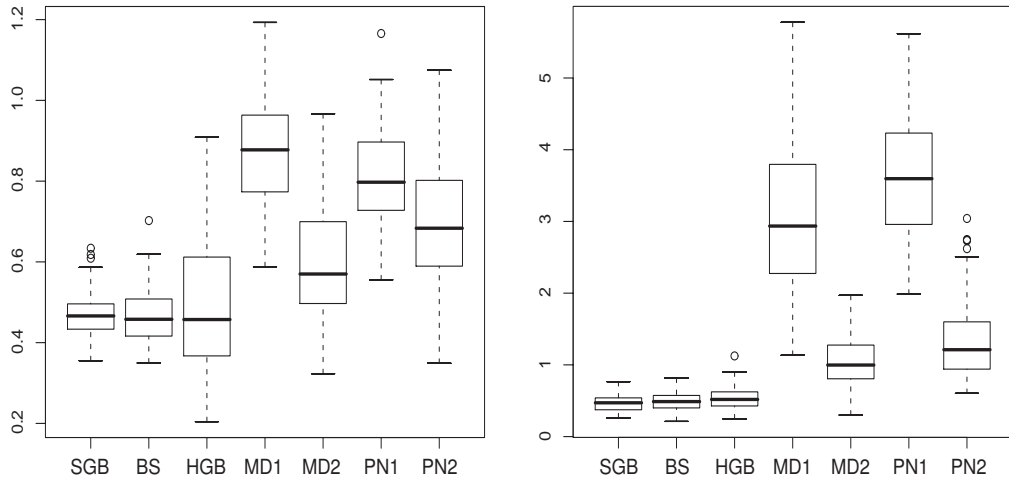
Figure 6: Boxplots of the average squared errors. The left panel shows the result for the function (1) while the right panel shows the result for the function (2). The models were evaluated by using GBIC.

6.   Concluding remarks

In this paper we proposed nonlinear regression modeling procedures with the method of regularization, using Gaussian basis functions whose construction method is based on that of *B*-spline bases. In order to choose optimal values of adjusted parameters, we presented the model selection criteria from the information-theoretic and Bayes approaches. The Gaussian basis function regression model constructed by the hybrid learning method has been widely used to draw information from data with complex structure. Gaussian bases produced previously, however, have some drawbacks. Our Gaussian bases yield the suitable amount of overlapping among basis functions so that the estimated regression function capture the complex structure in the data over the region of the input space.

In recent years, functional data analysis has received considerable attention in various fields of application (Ramsay and Silverman (2002, 2005)). A crucial point in the functional data analysis is to transfer the vector-valued observations to a set of functions. Kayano *et al.* (2006) and Araki *et al.* (2007) have reported that a functionalization by Gaussian basis functions has several advantages compared to using the others such as Fourier series, spline, *B*-spline. We believe that our proposed nonlinear modeling strategy provides the useful tools in several fields of research including the functional data analysis, since it can be easily applied to analyze time-course data with complex structure, and yields stable estimated functions.
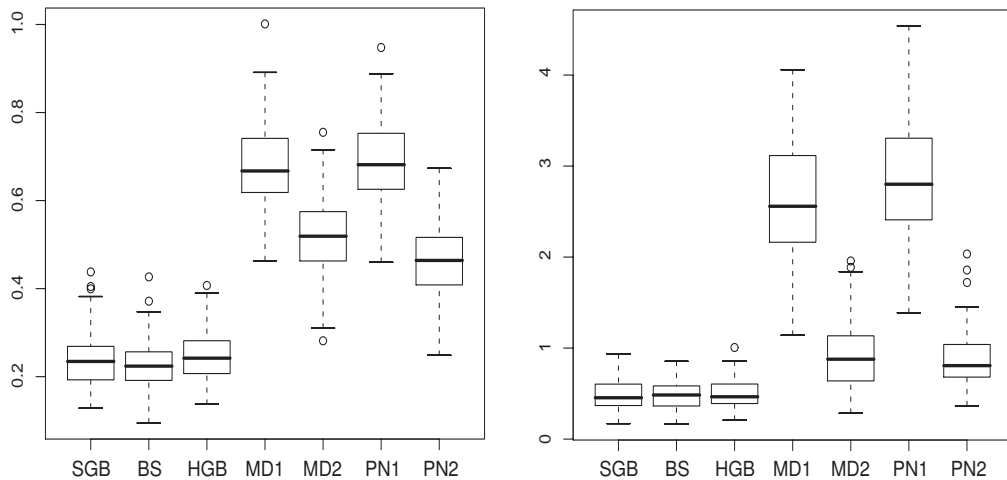
Figure 7: Boxplots of the average squared errors. The left panel shows the result for the function (1) while the right panel shows the result for the function (2). The models were evaluated by using GIC.

References

Ando, T., Konishi, S. and Imoto, S. (2006). Nonlinear regression modeling via regularized radial basis function networks. *Journal of Statistical Planning Inference*, in press.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2007). Functional regression modeling via regularized Gaussian basis expansions. *Annals of the Institute of Statistical Mathematics*, to appear.

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

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

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

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

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

Kayano, M., Konishi, S., Hirakawa, H. and Kuhara, S. (2006). Functional principal component analysis via regularized basis expansion and its application (in Japanese). *Japanese Journal of Applied Statistics* **35**, 1–16.

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

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

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

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

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

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

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistic Society B* **36**, 1–52.