Nonlinear regression modeling via the lasso-type regularization

Tateishi, Shohei
Graduate School of Mathematics, Kyushu University

Matsui, Hidetoshi
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Faculty of Mathematics, Kyushu University

出版情報：MI Preprint Series. 2009-8, 2010-05. Elsevier
バージョン：
権利関係：

MI Preprint Series

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

Nonlinear regression modeling via the lasso-type regularization

S. Tateishi, H. Matsui & S. Konishi

MI 2009-8

(Received February 10, 2009)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Nonlinear regression modeling via the lasso-type regularization

Shohei Tateishi[1], Hidetoshi Matsui[1], Sadanori Konishi[2]

[1] *Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.*

[2] *Faculty of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.*

s-tateishi@math.kyushu-u.ac.jp hmatsui@math.kyushu-u.ac.jp konishi@math.kyushu-u.ac.jp

Abstract

We consider the problem of constructing nonlinear regression models with Gaussian basis functions, using lasso regularization. Regularization with a lasso penalty is an advantageous in that it reduces some unknown parameters in linear regression models toward exactly zero. We propose imposing a weighted lasso penalty on a nonlinear regression model and thereby selecting the number of basis functions effectively. In order to select tuning parameters in the regularization method, we use model selection criteria derived from information-theoretic and Bayesian viewpoints. Simulation results demonstrate that our methodology performs well in various situations.

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

1 Introduction

The regularization or shrinkage method has been used widely for the solution of ill-posed problems arising in the use of maximum likelihood or least squares procedures, and it has been proved successful in several fields, including regression analysis and machine learning (see, e.g., Bishop, 2006; Hastie *et al.*, 2001). It imposes a penalty with respect to parameters of objective functions that are utilized in optimization problems, and various kinds of penalties have been proposed (Candes and Tao, 2007; Fan and Li, 2001; Frank and Friedman, 1993).

One of the most commonly used penalty methods is ridge regression (Hoerl and Kennard, 1970), which imposes an L_2 norm penalty on regression coefficients. The Ridge regression achieves good prediction performance through a bias-variance trade-off. While

on the other hand, Tibshirani (1996) proposed imposing an L_1 norm on parameters, called a lasso. Owing to the nature of the L_1 penalty, the lasso encourages both shrinkage and automatic variable selection simultaneously in linear regression models, and it has been extended and applied in various fields, particularly in bioscience (Zou, 2006; Zou and Hastie, 2005).

In this paper, we consider applying lasso regularization to nonlinear regression models involving basis expansions, which are useful tools for analyzing data with complex structures. The essential idea behind basis expansions is to express a regression functions as a linear combination of known functions, called basis functions (Bishop, 2006; Konishi and Kitagawa, 2008). Although Fourier series or B-splines are widely used as basis functions (de Boor, 2001; Imoto and Konishi, 2003), we employ Gaussian basis functions (Bishop, 1995) because they can be expressed in a simple form and can be easily applied to analyze a set of high-dimensional data including surface fitting data. Furthermore, Ando *et al.* (2008) have proposed using Gaussian basis functions with hyperparameter that controls the amount of overlapping among basis functions. They also discussed the efficiency of such nonlinear regression modeling.

When Gaussian basis functions are constructed for unequally spaced data, narrow basis functions with very small dispersions may be constructed, which can lead to unsmooth or unstable results if we employ these bases. We therefore apply lasso estimation, which allows us to avoid these effects on the basis functions by estimating their coefficients towards exactly zero. In this regard we adjust the weight of the lasso penalty so that the proper penalties are imposed on each basis function.

It is a crucial issue to determine the tuning parameters, including the number of basis functions, a smoothing parameter and a hyperparameter associated with Gaussian basis functions. Unlike ridge regression, the lasso is able to automatically select the number of basis functions. We consider selecting the remaining two tuning parameters using model selection criteria derived from information-theoretic and Bayesian viewpoints (Spiegelhalter *et al.*, 2002). The proposed nonlinear modeling procedure is investigated by analyzing Monte Carlo simulations including curve fitting and surface fitting data. The results demonstrate the effectiveness of the proposed method in terms of prediction

accuracy.

This paper is organized as follows. Section 2 describes the framework of basis expansions and Gaussian basis function models. In section 3 we present a new lasso penalty for nonlinear regression models. Section 4 provides model selection criteria for evaluating statistical models estimated by the regularization method with the proposed lasso penalty. In section 5 we investigate the performance of proposed nonlinear regression modeling techniques through Monte Carlo simulations. Some concluding remarks are presented in Section 6.

2 Gaussian basis function model

Suppose that we have n independent observations $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, 2, \cdots, n\}$, where y_α are random response variables and $\boldsymbol{x}_\alpha$ are vectors of p-dimensional explanatory variables. We consider the regression model

$$y_\alpha = u(\boldsymbol{x}_\alpha) + \epsilon_\alpha, \quad \alpha = 1, 2, \cdots, n, \tag{1}$$

where $u(\cdot)$ is an unknown smooth function and ϵ_α are independently, normally distributed with mean zero and variance σ^2. It is assumed that the function $u(\cdot)$ can be expressed as a linear combination of basis functions $\phi_j(\boldsymbol{x})$ $(j = 1, 2, \cdots, m)$ in the form

$$u(\boldsymbol{x}; \boldsymbol{w}) = w_0 + \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \ , \tag{2}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = (1, \phi_1(\boldsymbol{x}), \cdots, \phi_m(\boldsymbol{x}))^T$ is a vector of basis functions and $\boldsymbol{w} = (\ w_0, \ w_1, \cdots, \ w_m \ )^T$ 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

$$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_j||^2}{2h^2}\right), \quad j = 1, 2, \cdots, m, \tag{3}$$

where $\boldsymbol{\mu}_j$ is a p-dimensional vector determining the center of the basis function, h_j^2 is a parameter that determines the dispersion and $||\cdot||$ is the Euclidian norm. Unknown parameters in the regression model (1) include the coefficient parameters w_j $(j = 1, \cdots, m)$, and the centers $\boldsymbol{\mu}_j$ and dispersion parameters h_j^2 required for Gaussian basis functions.

These parameters are generally determined in a two-stage procedure in order to avoid local minimum and identification problems (Moody and Darken, 1989). In the first stage,

the centers $\boldsymbol{\mu}_j$ and dispersion h_j^2 are determined by using the k-means clustering algorithm. The data set of observations of the explanatory variables $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ are divided into m clusters $\{C_1, \cdots, C_m\}$; then centers $\boldsymbol{\mu}_j$ and dispersions h_j^2 are determined by

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n_j} \sum_{\boldsymbol{x}_\alpha \in C_j} \boldsymbol{x}_\alpha, \quad \hat{h}_j^2 = \frac{1}{n_j} \sum_{\boldsymbol{x}_\alpha \in C_j} ||\boldsymbol{x}_\alpha - \boldsymbol{\mu}_j||^2, \tag{4}$$

where n_j is the number of observations included in the the j-th cluster C_j. Replacing $\boldsymbol{\mu}_j$ and h_j^2 in equation (3) by $\hat{\boldsymbol{\mu}}_j$ and $\hat{h}_j^2$ respectively, we obtain a set of m basis functions

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

In the second stage, the coefficient parameters w_j $(j = 0, 1, \cdots, m)$ are estimated by the maximum penalized likelihood method. Details are described in section 3.

However, basis functions (5) often yield inadequate results because of the lack of overlapping among basis functions. In order to overcome this problem, Ando *et al.* (2008) proposed the use of Gaussian basis functions with a hyperparameter, i.e. functions of the form

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

where ν is a hyperparameter that adjusts the dispersion of basis functions. Ando *et al.* (2008) showed that nonlinear models with these basis functions were effective in capturing the information from the data.

3 Estimation

For n indipendent observations $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, \cdots, n\}$, the nonlinear regression model based on Gaussian basis functions $\phi_j(\boldsymbol{x})$ $(j = 1, \cdots, m)$ given in Section 2 is expressed as

$$y_\alpha = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha) + \epsilon_\alpha, \ \alpha = 1, \cdots, n, \tag{7}$$

where $\boldsymbol{\phi}(\boldsymbol{x}_\alpha) = (1, \phi_1(\boldsymbol{x}_\alpha), \cdots, \phi_m(\boldsymbol{x}_\alpha))^T$, $\boldsymbol{w} = (w_0, w_1, \cdots, w_m)^T$ and ϵ_α are error terms. If the error terms ϵ_α are independently and normally distributed with mean 0 and variance σ^2, the nonlinear regression model (7) has a probability density function

$$f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\{y_\alpha - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}^2}{2\sigma^2}\right], \ \alpha = 1, \cdots, n. \tag{8}$$

Then the maximum likelihood estimates of the coefficient vectors $\boldsymbol{w}$ and σ^2 are respectively given by

$$\hat{\boldsymbol{w}} = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}})^T(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}}),$$

where $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_n))^T$ and $\boldsymbol{y} = (y_1, \cdots, y_n)$. However, when fitting a nonlinear model to data with a complex structure the maximum likelihood method often yields unstable estimates and leads to overfitting. We therefore estimate $\boldsymbol{w}$ and σ^2 by the method of regularization. Instead of using the log-likelihood function, we consider maximizing the penalized log-likelihood function

$$l_\lambda(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n} \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w}, \sigma^2) - n\lambda H(\boldsymbol{w}), \tag{9}$$

where $\boldsymbol{\theta} = \{\boldsymbol{w}, \sigma^2\}$ and λ is a smoothing parameter that controls the smoothness of the fitted model and $H(\boldsymbol{w})$ is a penalty function for $\boldsymbol{w}$.

One of the typical forms for the penalty function is the ridge penalty (Hoerl and Kennard, 1970), imposing an L_2 norm on the parameter $\boldsymbol{w}$, that is,

$$H(\boldsymbol{w}) = \frac{1}{2}\sum_{j=1}^{m} w_j^2 = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}. \tag{10}$$

Then, the maximum penalized likelihood estimates of $\boldsymbol{w}$ and σ^2 are respectively given by

$$\hat{\boldsymbol{w}} = (\Phi^T\Phi + n\lambda\hat{\sigma}^2 I_{m+1})^{-1}\Phi^T\boldsymbol{y}, \quad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}})^T(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}}). \tag{11}$$

Note that these estimators depend on each other. Therefore, we provide an initial value for the variance $\sigma^2_{x(0)}$ first, then $\hat{\boldsymbol{w}}$ and $\hat{\sigma}^2_x$ are updated until convergence. The ridge estimators continuously shrink the coefficients as λ increases.

Another type of penalty function is the lasso penalty (Tibshirani, 1996) given by

$$H(\boldsymbol{w}) = \sum_{j=1}^{m} |w_j|. \tag{12}$$

When using the lasso penalty, if λ is sufficiently large some coefficients are shrunk to exactly zero. Hence the lasso can be used as a regularization technique for variable selection, in particular for a linear regression model (Tibshirani, 1996).

The maximum penalized likelihood estimators utilizing basis expansions depend on the number of basis functions m, the smoothing parameter λ and the hyperparameter

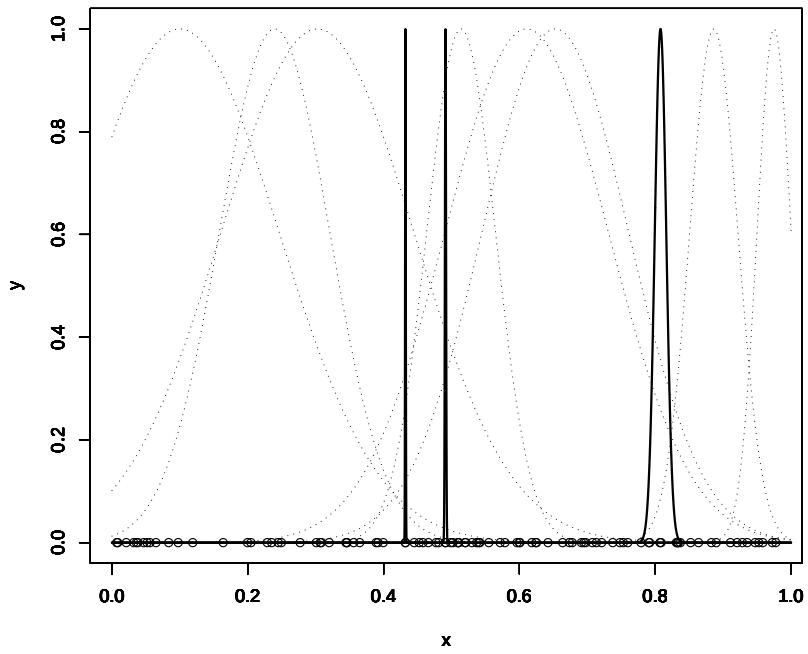

Fig. 1: Gaussian basis functions constructed for uniformly distributed predictors on $[0, 1]$ (points). Solid lines are functions with comparatively too little dispersion; this is not the case for the dotted lines.

ν of Gaussian basis functions, and it is a crucial issue to select suitable values of these tuning parameters. When we apply the ridge penalty for $H(\boldsymbol{w})$ we have to select all three parameters using model selection criteria described in the next section, which may be computationally expensive. On the other hand, if we use the lasso penalty, it is expected that the number of basis functions m can be selected automatically and appropriately at the same time as estimating the coefficient $\boldsymbol{w}$, utilizing the property that some parameters are shrunk towards exactly zero.

Figure 1 shows Gaussian basis functions constructed for unequally spaced predictors. We observe find that there are some bases with very small dispersions. In this case the ridge estimator is affected by these narrow basis functions directly, yielding unsmooth or unstable results. In contrast, the lasso estimates may effectively omit these basis functions and obtain smooth curves.

However, when we apply the ordinary lasso penalty (12) to nonlinear regression models, unfavorable results are obtained. Figure 2 plots estimated curves for $f(x) = \exp(-2x)$

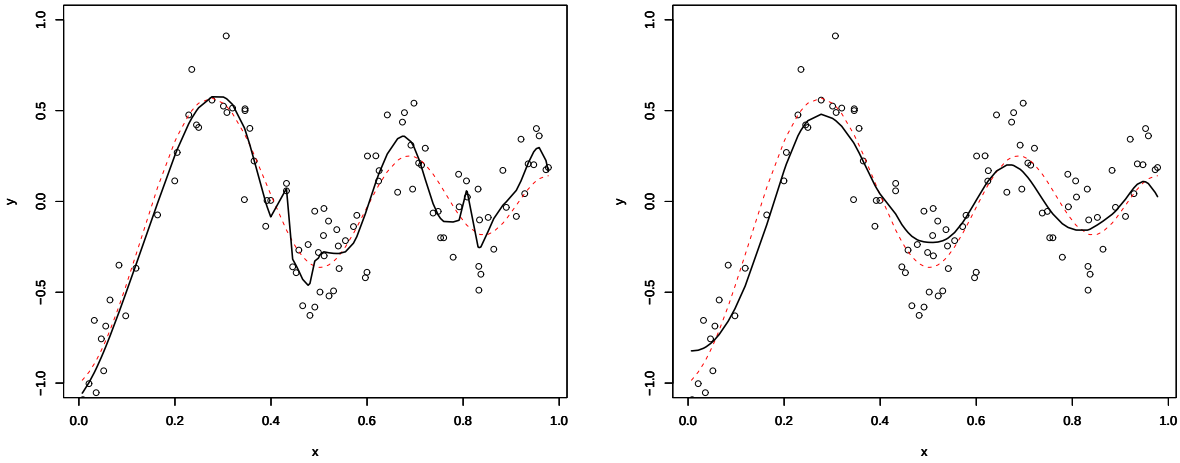

Fig. 2: Estimated curves obtained by lasso regularization. Dotted curves depict the true functions, points depict observed values and solid curves depict the estimated functions with $\lambda = 0.01$ (left) and $\lambda = 0.19$ (right).

$\cos(3\pi \exp(x))$ given by the regularization method with the lasso penalty, when smoothing parameters are small and large. The result depicted in the left panel shows that, globally, the appropriate regularization is achieved; however, the resulting function is not smooth locally due to some narrow basis functions. On the other hand, whereas the estimated function depicted in the right panel is smoothed becaused the estimated coefficients for narrow basis functions are zero, excessive regularization has been imposed, and therefore in this situation a function that is too smooth is obtained.

In order to overcome this problem and obtain appropriately smooth functions, we propose a new lasso-type penalty which imposes a weighted L_1 penalty on each coefficient, given by

$$H(\boldsymbol{w}) = \sum_{j=1}^{m} c_j |w_j|, \tag{13}$$

$$c_j = \begin{cases} 1 & (\bar{h}^2 \leq h_j^2) \\ \bar{h}^2 / h_j^2 & (\bar{h}^2 > h_j^2) \end{cases}, \bar{h}^2 = \frac{1}{m} \sum_{k=1}^{m} h_k^2.$$

The weights c_j $(j = 1, \cdots, m)$ encourage coefficients of narrow basis functions to shrink towards exactly zero so that the estimated regression function captures the structure of the data. Similarly, Zou (2006) proposed the adaptive lasso, which imposes weights

on the penalty, giving an "oracle property" to estimates of the linear regression model. Shimamura *et al.* (2007) also considered another weighted lasso and applied it to graphical Gaussian models for estimating large gene networks from DNA microarray data.

The maximum penalized likelihood estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{w}}, \hat{\sigma}^2)^T$ with the penalty (13) is given by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{w},\sigma^2} \left(\sum_{\alpha=1}^{n} \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w}, \sigma^2) - n\lambda \sum_{j=0}^{m} c_j |w_j| \right). \tag{14}$$

Since the lasso estimate is non-differentiable in $\boldsymbol{w}$, it is difficult to obtain its analytical form. Here we apply the shooting algorithm (Fu, 1998) to the maximum penalized likelihood method using the weighted lasso penalty (13). The detailed algorithm is given in Appendix A.1. We obtain a statistical model by replacing the unknown parameter $\boldsymbol{\theta} = (\boldsymbol{w}, \sigma^2)^T$ with its corresponding estimator $\hat{\boldsymbol{\theta}}$, that is,

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

4 Information criterion for model selection

The statistical model (15) estimated by the regularization method depends upon the number of basis functions m, the smoothing parameter λ and the hyperparameter ν of the Gaussian basis functions, and it is a crucial issue to determine these values appropriately. Although there are several model selection criteria for evaluating statistical models estimated by the regularization method (Konishi and Kitagawa, 1996; Konishi *et al.*, 2004), they cannot be directly applied to the lasso estimates since they need derivatives of penalized log-likelihood functions. On the other hand, Spiegelhalter *et al.* (2002) introduced a model evaluation criterion DIC for statistical models estimated by Bayesian estimation procedures such as posterior mean, mode and median. Since the lasso estimator is also considered to be a Bayesian estimators (Tibshirani, 1996), we can apply the DIC for the lasso estimates. Here, we describe the relationship between the lasso and Bayesian estimates.

Suppose that $\boldsymbol{w}$ has a Laplace prior density

$$\pi(w_j) = \frac{nc_j\lambda}{2} \exp\left(-nc_j\lambda|w_j|\right) \quad (j = 0, 1, \cdots, m) \tag{16}$$

and that σ^2 has a flat prior; then the posterior distribution of the parameter $\boldsymbol{\theta}$ is obtained by

$$\log \pi(\boldsymbol{w}, \sigma^2|\boldsymbol{y}) = \log f(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) - n\lambda \sum_{j=1}^{m} c_j |w_j| + C,$$

where C is a constant term with respect to $\boldsymbol{w}$ and σ. Therefore, the maximum a posteriori (MAP) estimator corresponds to maximum penalized likelihood estimators (14). Similarly, when we use a normal prior instead of the Laplace density (16), we have ridge estimates (11) as a MAP estimator.

The DIC for evaluating the nonlinear regression model based on Gaussian basis functions estimated by the maximum penalized likelihood method with a lasso penalty (13) is given by

$$\text{DIC} = -2 \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2p_D \ , \tag{17}$$

where p_D is the effective number of parameters. Using the posterior density (17), Spiegelhalter *et al.* (2002) defined p_D as follows:

$$p_D = E_{\boldsymbol{\theta}|\boldsymbol{y}} \left[-2 \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{\theta}) \right] + 2 \log f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{\theta}}). \tag{18}$$

Although it is generally difficult to derive p_D analytically, we can apply the Gibbs sampling algorithm (Geman and Geman, 1984) which is a simple and widely applicable Markov Chain Monte Carlo method. Details of the algorithm are given in Appendix A.2. We select values of the smoothing parameter and the hyperparameter which minimize the DIC, then take the corresponding model as the optimal model to be the optimal one.

5 Numerical examples

In this section, Monte Carlo simulations were conducted to investigate the effectiveness of our proposed nonlinear regression modeling. Here, two simulation examples are considered: curve fitting and surface fitting.

5.1 Curve fitting

For the first study, repeated random samples $\{(x_\alpha, y_\alpha); \alpha = 1, \cdots, n\}$ with $n = 130$ were generated from a true regression model $y_\alpha = u(x_\alpha) + \epsilon_\alpha$. The design points x_α are

Table 1: Comparison of results for curve fitting.

function		mean λ	SD λ	AMSE	SD MSE
(a)	w-Lasso	6.24×10^{-3}	4.95×10^{-3}	2.39×10^{-3}	7.24×10^{-4}
	Lasso	1.29×10^{-2}	6.36×10^{-3}	2.48×10^{-3}	7.15×10^{-4}
	Ridge	1.29×10^{-2}	6.35×10^{-3}	2.51×10^{-3}	7.18×10^{-4}
(b)	w-Lasso	1.62×10^{-3}	1.42×10^{-3}	2.06×10^{-2}	7.35×10^{-3}
	Lasso	3.75×10^{-3}	3.90×10^{-3}	2.17×10^{-2}	7.68×10^{-3}
	Ridge	3.75×10^{-3}	3.89×10^{-3}	2.18×10^{-2}	7.52×10^{-3}
(c)	w-Lasso	4.40×10^{-3}	4.73×10^{-3}	6.73×10^{-3}	2.54×10^{-3}
	Lasso	6.57×10^{-3}	4.20×10^{-3}	7.02×10^{-3}	2.63×10^{-3}
	Ridge	6.56×10^{-3}	4.20×10^{-3}	7.15×10^{-3}	2.66×10^{-3}
(d)	w-Lasso	1.15×10^{-2}	1.34×10^{-2}	7.12×10^{-4}	2.30×10^{-4}
	Lasso	2.77×10^{-2}	1.92×10^{-2}	7.59×10^{-4}	2.55×10^{-4}
	Ridge	2.77×10^{-2}	1.90×10^{-2}	7.78×10^{-4}	2.56×10^{-4}

uniformly distributed in $[0, 1]$ and the errors ϵ_α are independently, normally distributed with mean 0 and standard deviation $\tau = 0.1R_u$ with R_u being the range of $u(x)$ over $x \in [0, 1]$. We considered the following four cases for the true regression model:

$$
\begin{aligned}
&\text{(a)} \quad u(x) &&= \exp(-2x)\cos(3\pi\exp(x)),\\
&\text{(b)} \quad u(x) &&= 1 - 39x + 201x^2 - 318x^3 + 158x^4,\\
&\text{(c)} \quad u(x) &&= \cos(3\pi x^4),\\
&\text{(d)} \quad u(x) &&= 0.3\exp\{-50(x-0.3)^2\} + 0.7\exp\{-250(x-0.7)^2\}.
\end{aligned}
$$

We compared the performance of nonlinear regression models based on Gaussian basis functions estimated by the weighted lasso (13) (w-Lasso) with that of models estimated by the ordinary lasso (Lasso) and the ridge (Ridge). The smoothing parameter λ and the hyperparameter ν for the Gaussian basis functions were selected using the model selection criterion DIC. Although the number of basis functions was set at $m = 29$, the weighted lasso and the ordinary lasso can automatically reduce it to the appropriate number, whereas the ridge does not. We performed 50 repetitions, then calculated averages of the mean square errors (AMSE) defined by MSE $= \sum_{\alpha=1}^{n}\{u(x_\alpha) - \hat{y}_\alpha\}^2/n$ and standard deviations in order to assess the goodness of fit. Furthermore, we examined averages and standard deviations of the selected smoothing parameters mean λ and SD λ respectively.

Table 1 shows summaries of the simulation results (a) to (d). In almost all situations,

our proposed modeling procedure minimized the AMSE, thus improving the accuracy of prediction. Furthermore, the proposed method tends to select smaller values of smoothing parameters; that is, it employs more complex models than other methods.

5.2 Surface fitting

Next, we applied the modeling strategy to two-dimensional or surface data. We generated random samples $\{(y_\alpha, x_{1\alpha}, x_{2\alpha}); \alpha = 1, \cdots, n\}$ with $n = 150$ from a true model $y_\alpha = u(x_{1\alpha}, x_{2\alpha}) + \epsilon_\alpha$, for three cases, where

$$\begin{aligned}
\text{(e)} \quad u(x_1, x_2) &= \sin(5x_1x_2) + \cos(3(x_1 + x_2)), \\
\text{(f)} \quad u(x_1, x_2) &= \sin(\pi x_1) + \cos(\pi x_2), \\
\text{(g)} \quad u(x_1, x_2) &= \frac{\sin(10\sqrt{x_1^2 + x_2^2})}{10\sqrt{x_1^2 + x_2^2}}.
\end{aligned}$$

and the design points were uniformly distributed in $[0, 1] \times [0, 1]$. The errors ϵ_α were independently, normally distributed with mean 0 and standard deviation τ, where the standard deviation was taken as $\tau = 0.1R_u$ with R_u being the range of $u(x_1, x_2)$ over $(x_1, x_2) \in [0, 1] \times [0, 1]$.

We fitted the proposed nonlinear regression model, comparing the result of the ordinary lasso and the ridge estimator. Tuning parameters λ and ν were selected by the model selection criterion DIC, and the number of basis functions m was taken as 39. Figure 3 compares true surfaces with fitted ones. We observe that our modeling strategy is effective in capturing the true data structures well. We also compared the mean squared errors MSE $= \sum_{\alpha=1}^{n} \{u(x_{1\alpha}, x_{2\alpha}) - \hat{y}_\alpha\}^2/n$.

Table 2 presents summaries of the simulation results form the surface fitting. These results were obtained by averaging over 50 Monte Carlo simulations. This table demonstrates that our method is superior to others in almost all cases in termes of the prediction accuracy. Similar to the results for curve fitting, the proposed method tends to select smaller smoothing parameters.

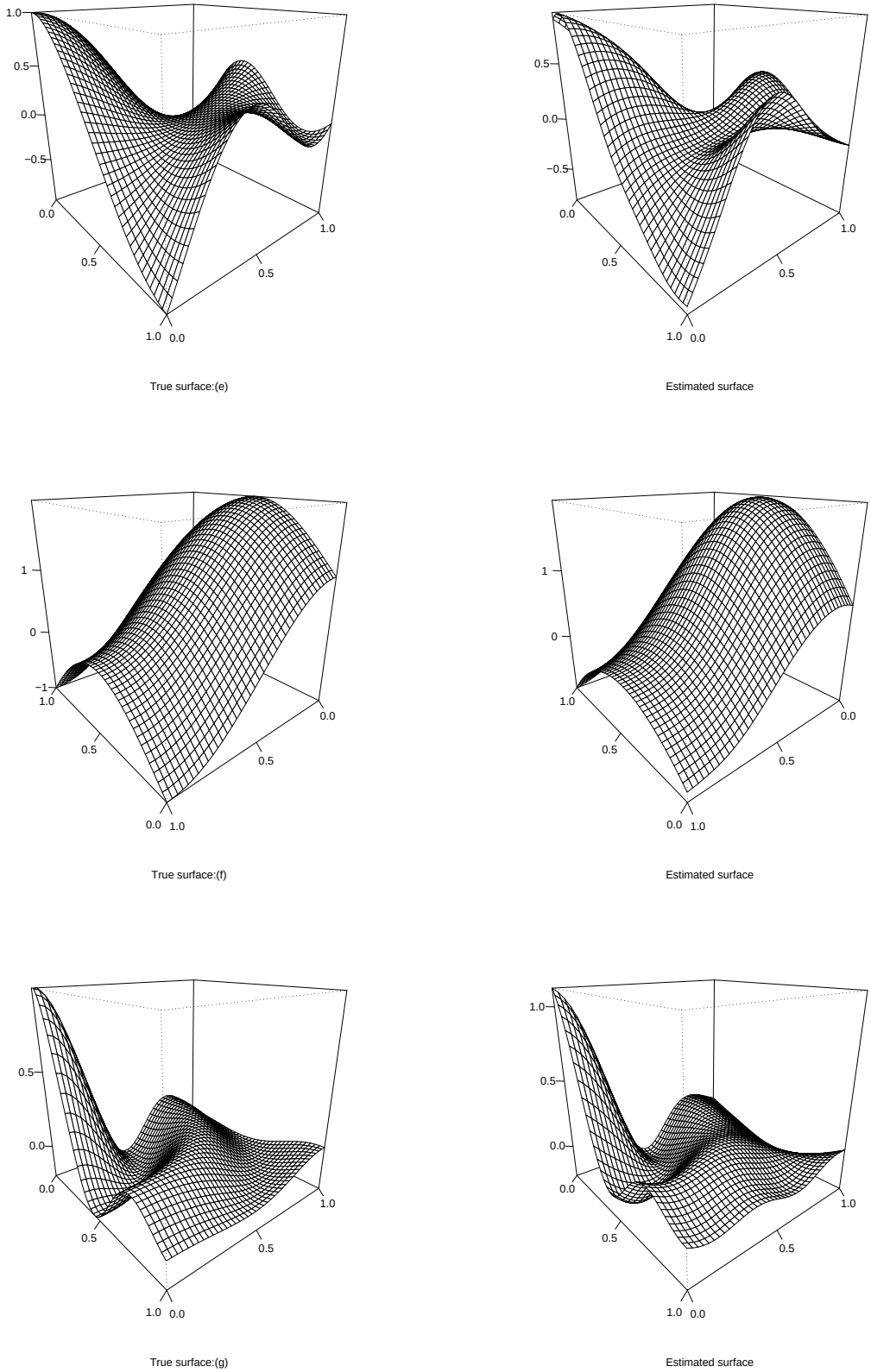

Fig. 3: Comparison of the true surfaces with estimated surfaces for simulation (e), (f) and (g).

Table 2: comparison of the mean squared errors for surface fitting.

function		mean λ	SD λ	MSE	SD MSE
(e)	w-Lasso	2.48×10^{-3}	1.63×10^{-3}	5.15×10^{-3}	1.51×10^{-3}
	Lasso	4.48×10^{-3}	3.12×10^{-3}	5.29×10^{-3}	1.52×10^{-3}
	Ridge	4.41×10^{-3}	3.09×10^{-3}	5.30×10^{-3}	1.51×10^{-3}
(f)	w-Lasso	3.53×10^{-3}	2.33×10^{-3}	7.43×10^{-3}	2.88×10^{-3}
	Lasso	6.30×10^{-3}	4.13×10^{-3}	8.18×10^{-3}	3.25×10^{-3}
	Ridge	6.76×10^{-3}	4.78×10^{-3}	8.21×10^{-3}	3.21×10^{-3}
(g)	w-Lasso	6.23×10^{-3}	4.20×10^{-3}	2.17×10^{-3}	7.03×10^{-3}
	Lasso	6.56×10^{-3}	4.08×10^{-3}	2.21×10^{-3}	7.25×10^{-3}
	Ridge	6.55×10^{-3}	4.07×10^{-3}	2.22×10^{-3}	7.24×10^{-3}

6 Concluding remarks

We have proposed a nonlinear regression modeling procedure that makes use of regularization. When we apply it to Gaussian basis expansions, we need to select multiple tuning parameters, which can be computationally expensive. Lasso regularization was applied in order to conduct proper smoothing and selection of basis functions. For the choice of tuning parameters, except for the number of basis functions, a model selection criterion DIC was derived using a Markov Chain Monte Carlo method. The ordinary lasso penalty has been extensively used in the framework of linear regression models; however, sufficient results have not been obtained for nonlinear regression models with Gaussian basis functions. The simulation results reported here demonstrate the effectiveness of the proposed modeling strategy in terms of prediction of the response variables.

A Appendix

A.1 Algorithm for deriving the lasso estimator

It is difficult to derive the lasso estimator analytically because of the inclusion of L_1 norm of parameters. Here we describe the details of the algorithm for obtaining the weighted lasso estimator. The algorithm is given in terms of the following steps:

Step 1. Start with initial values $\boldsymbol{w}_{old} = (w_{0,old}, \cdots, w_{m,old})^T$ and σ^2_{old}.

Step 2. For $j = 1, \cdots, m$, if $w_{j,old}$ is 0 then $w_{j,new}$ is 0; otherwise update the coefficient

parameters using the following rules:

$$w_{0,new} = \frac{\sigma^2_{old} S_0}{\phi_0(\boldsymbol{x})^T \phi_0(\boldsymbol{x})},$$

$$w_{j,new} = \begin{cases} \dfrac{\sigma^2_{old}(S_0 - n\lambda)}{\phi_j(\boldsymbol{x})^T \phi_j(\boldsymbol{x})} & \text{if} \quad S_0 > n\lambda \\ \dfrac{\sigma^2_{old}(S_0 + n\lambda)}{\phi_j(\boldsymbol{x})^T \phi_j(\boldsymbol{x})} & \text{if} \quad S_0 < -n\lambda \\ 0 & \text{if} \quad |S_0| \leq n\lambda, \end{cases}$$

$$S_0 = \frac{\partial}{\partial w_i} \log f(\boldsymbol{y}|w_i, \boldsymbol{w}_{-i,old}, \sigma^2_{old})|_{w_j=0}$$

where $\boldsymbol{w}_{-j,old} = (w_{0,new}, w_{1,new}, \cdots, w_{j-1,new}, w_{j+1,old}, \cdots, w_{m,old})^T$, $\phi_j(\boldsymbol{x}) = (\ \phi_j(x_1), \cdots, \phi_j(x_n)\ )^T$ and, in particular, $\phi_0(\boldsymbol{x}) = (1, \cdots, 1)^T$. Then obtain updated values $\boldsymbol{w}_{new} = (w_{0,new}, \cdots, w_{m,new})^T$.

Step 3. Update the variance parameter by

$$\sigma^2_{new} = \frac{1}{n}(\boldsymbol{y} - \Phi \boldsymbol{w}_{new})^T (\boldsymbol{y} - \Phi \boldsymbol{w}_{new}).$$

Step 4. Repeat **Step 2** and **Step 3** until $\boldsymbol{w}$ and σ^2 converge.

A.2 Random number generation from posterior distribution

It is necessary to sample from the posterior distribution in order to evaluating DIC numerically. We use Gibbs sampling similar to that in Park and Casella (2005) who used a Bayesian lasso. The Gibbs sampler algorithm is given by the following:

Step 1. Start with initial values $\boldsymbol{w}^{(0)} = (w_0^{(0)}, \cdots, w_m^{(0)})^T$, $\sigma^{2(0)}$ and let $t = 1$.

Step 2. For each $i = 0, 1, \cdots, m$, generate random variable s from the exponential distribution with mean $(nc_i\lambda)^2/2$; then generate $w_i^{(t)}$ from a normal distribution with mean μ_i and variance τ_i where

$$\mu_i = \frac{s\sum_{\alpha=1}^n \{\phi_i(x_\alpha)\}(y_\alpha - \boldsymbol{w}_{-i}^T \boldsymbol{\phi}(x_\alpha)_{-i})}{s\sum_{\alpha=1}^n \{\phi_i(x_\alpha)\}^2 + \sigma^{2(t-1)}}, \quad \tau_i = \frac{s\sigma^{2(t-1)}}{s\sum_{\alpha=1}^n \{\phi_i(x_\alpha)\}^2 + \sigma^{2(t-1)}},$$

$$\begin{aligned} \boldsymbol{\phi}(x_\alpha)_{-i} &= (1, \phi_1(x_\alpha), \cdots, \phi_{i-1}(x_\alpha), \phi_{i+1}(x_\alpha), \cdots, \phi_m(x_\alpha))^T, \\ \boldsymbol{w}_{-i} &= (w_0^{(t)}, \cdots, w_{i-1}^{(t)}, w_{i+1}^{(t-1)}, \cdots, w_m^{(t-1)})^T. \end{aligned}$$

Step 3. Generate $\sigma^{2(t)}$ from the inverse gamma distribution with shape parameter $n/2-1$, and scale parameter $(\boldsymbol{y} - \Phi\boldsymbol{w}^{(t)})^T(\boldsymbol{y} - \Phi\boldsymbol{w}^{(t)})/2$. Replace t with $t+1$, and return to **Step 2**.

References

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

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

[3] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[4] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 2313–2351.

[5] de Boor, C. (2001). *A practical Guide to Splines*. Springer.

[6] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

[7] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

[8] Fu, W. J. (1998). Penalized Regression: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

[9] Geman, S. and Geman, D. (1986). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[10] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Leaning*. Springer–Verlag, New York.

[11] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* **12**, 55–67.

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

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

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

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

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

[17] Park, T. and Casella, G. (2005). The Bayesian lasso. *Journal of the American Statistical Association*. **103**, 681–686.

[18] Shimamura, T., Imoto, S., Yamaguchi, R. and Miyano, S. (2007). Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics* **19**, 142–153.

[19] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* **64**, 583–616.

[20] Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

[21] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

[22] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67**, 301–320.

List of MI Preprint Series, Kyushu University

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

MI

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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