Variable selection for functional regression model via the L_1 regularization

Matsui, Hidetoshi
Faculty of Mathematics, Kyushu University

Konishi, Sadanori
Faculty of Mathematics, Kyushu University

https://hdl.handle.net/2324/13205

出版情報：MI Preprint Series. 2009-3, 2011-12. Elsevier B.V.
バージョン：

MI Preprint Series

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

Variable selection for functional regression model via the L_1 regularization

H. Matsui & S. Konishi

MI 2009-3

(Received January 19, 2009)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Variable selection for functional regression model via the L_1 regularization

Hidetoshi Matsui and Sadanori Konishi
Kyushu University

Abstract

In regression analysis, the L_1 regularization such as the lasso or the SCAD provides sparse solutions, which leads to variable selection. We consider the variable selection problem where variables are given as functional forms, using the L_1 regularization. In order to select functional variables each of which is controlled by multiple parameters, we treat parameters as grouped parameters and then apply the group SCAD. A crucial issue in the regularization method is the choice of regularization parameters. We derive a model selection criterion for evaluating the model estimated by the regularization method via the group SCAD penalty. Results of simulation and real data analysis show the effectiveness of the proposed modeling strategy.

Key words: Functional data analysis, Group lasso, Information criterion, Model selection, Regularization, SCAD.

1 Introduction

Variable selection is one of the most important problems in regression analysis, and there have been considered several methods for selecting a set of necessary variables in linear regression models (see e.g., Burnham and Anderson, 2002; Miller, 1984). Tibshirani (1996) proposed applying L_1 regularization, called the lasso, and showed that the lasso penalty simultaneously shrinks parameters and selects variables, owing to the property that the lasso attempt to shrink some parameters toward exactly zero. Lasso-type regularization has received considerable attension in various fields of application such as medical science or bioinformatics (Tibshirani, 1997; Segal *et al.*, 2003). Furthermore, Fan and Li (2001) derived a new penalty function called the smoothly clipped absolute deviation (SCAD), which is considered to be an improved version of the lasso and the hard thresholding penalty (Donoho and Johnstone, 1994). The SCAD has been used in various models such as semiparametric models or proportional hazards models (Fan and Li, 2002; 2004; Cai *et al.*, 2005). We consider the problem of selecting variables using the SCAD penalty, where predictors are given as functions.

When the data are observed at possibly differing time points, traditional regression procedures cannot be directly applied. While more recently, functional data analysis (FDA) has been used in various fields of study, as the method for analyzing data sets which have functional forms (Ramsay and Silverman, 2002; 2005). The basic idea behind functional data analysis is to express discrete data as a smooth function and then draw information from the collection of functional data. The traditional regression models are easily extended to the framework of the functional data analysis, and we refer to it as functional regression models.

When the regularization method via the SCAD penalty is applied to the functional regression models based on basis expansions directly, variable selection of the functional predictor fails since multiple parameters exist for one predictor. On the other hand, Yuan and Lin (2006) considered selecting grouped variables rather than individual variables and proposed a group lasso. Furthermore, Wang *et al.* (2007) used a group SCAD estimation to varying-coefficient models.

We consider applying group SCAD regularization to the functional regression model with functional predictors and a scalar response, estimating and selecting models simultaneously. We also derive a model selection criterion for selecting regularization parameters involved in the maximum penalized likelihood method with the group SCAD penalty. The proposed modeling strategy is applied to a Monte Carlo simulation and real data analysis. Results show that our modeling strategy effectively estimate the model and select functional predictors.

The remainder of this paper is given as follows. In Section 2 we introduce a functional regression model with functional predictors and a scalar response. Section 3 briefly describes properties of group SCAD penalty, and then show how to estimate the model. In Section 4 we describe model selection criteria for evaluating models estimated by the method of regularization. Section 5 shows applications of the proposed modeling strategy to a simulation example and weather data. Some concluding remarks are included in Section 6.

2 Functional regression model

Suppose we have n observations $\{(y_\alpha,\, \boldsymbol{x}_\alpha(t));\, t \in \mathcal{T},\, \alpha = 1, \dots,\, n\}$, where y_α is a scalar response and $\boldsymbol{x}_\alpha(t) = (x_{\alpha 1}(t), \dots,\, x_{\alpha M}(t))^T$ are functional predictors with M variables, expressed via basis expansions as follows:

$$x_{\alpha m}(t) = \sum_{j=1}^{p_m} w_{\alpha m j} \phi_{mj}(t) = \boldsymbol{w}_{\alpha m}^T \boldsymbol{\phi}_m(t),$$

where $\boldsymbol{w}_{\alpha m} = (w_{\alpha m 1},\, \dots, w_{\alpha m p_m})^T$ are vectors of coefficients and $\boldsymbol{\phi}_m(t) = (\phi_{m1},\, \dots,\, \phi_{m p_m})^T$ are vectors of basis functions. Although Fourier series or B-splines are commonly used for basis functions, we apply Gaussian basis functions (Ando *et al.*, 2008) defined as

$$\phi_{mj}(t) = \exp\left\{ -\frac{(t - c_{mj})^2}{2\nu_m s_{mj}^2} \right\} \quad (j = 1, \dots, p), \tag{1}$$

where c_{mj} and s_{mj}^2 are centers and dispersions of basis functions respectively and ν_m is a hyperparameter. Advantages with respect to the use of Gaussian bases are that the resulting function is simply expressed and that it can easily be applied to surface fitting data. Coefficients $\boldsymbol{w}_\alpha$ and parameters involved in Gaussian basis functions are determined prior to the functional regression modeling procedure by smoothing methods, whose details are given in the Appendix. Then we consider a functional regression model (Ramsay and Silverman, 2005; Araki *et al.*, 2008) given by

$$y_\alpha = \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}} x_{\alpha m}(t) \beta_m(t) dt + \varepsilon_\alpha, \tag{2}$$

where β_0 is a constant term, the $\beta_m(t)$ are coefficient functions and ε_α is a Gaussian noise with mean 0 and variance σ^2. $\beta_m(t)$ are supposed to be expressed via basis expansions as follows:

$$\beta_m(t) = \sum_{j=1}^{p_m} b_{mj}^* \phi_{mj}(t) = \boldsymbol{b}_m^{*^T} \boldsymbol{\phi}_m(t),$$

where $\boldsymbol{b}_m^* = (b_{m1}^*, \dots, b_{m p_m}^*)^T$ are parameter vectors. Then the functional regression model (2) can be expressed in the following form:

$$\begin{aligned} y_\alpha &= \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}} \boldsymbol{w}_{\alpha m}^T \boldsymbol{\phi}_m(t) \boldsymbol{\phi}_m^T(t) \boldsymbol{b}_m^* dt + \varepsilon_\alpha \\ &= \boldsymbol{z}_\alpha^T \boldsymbol{b} + \varepsilon_\alpha, \end{aligned}$$

where $\boldsymbol{z}_\alpha = (1, \boldsymbol{w}_{\alpha 1}^T \boldsymbol{J}_{\phi_1}, \dots, \boldsymbol{w}_{\alpha M}^T \boldsymbol{J}_{\phi_M})^T$, $\boldsymbol{b} = (\beta_0, \boldsymbol{b}_1^{*^T}, \dots, \boldsymbol{b}_M^{*^T})^T$ and $\boldsymbol{J}_{\phi_m} = \int_{\mathcal{T}} \boldsymbol{\phi}_m(t) \boldsymbol{\phi}_m^T(t) dt$ are $p_m \times p_m$ cross product matrices. When we use Gaussian basis functions (1), the (j, k)-th element of J_{ϕ_m} has an analytical expression

$$J_{\phi_m}^{(j,k)} = \frac{\sqrt{2\pi\nu_m s_{mj}^2 s_{mk}^2}}{\sqrt{s_{mj}^2 + s_{mk}^2}} \exp\left\{ -\frac{(c_{mj} - c_{mk})^2}{2\nu_m (s_{mj}^2 + s_{mk}^2)} \right\}.$$

From these assumptions the functional regression model (2), given a functional predictor $\boldsymbol{x}_\alpha$, has a probability density function:

$$f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{b}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_\alpha - \boldsymbol{z}_\alpha^T \boldsymbol{b})^2}{2\sigma^2} \right\}. \tag{3}$$

3 Estimation via the group SCAD regularization

The parameters $\boldsymbol{\theta} = \{\boldsymbol{b}, \sigma^2\}$ in the functional regression model (3) is estimated by the regularization method with the group SCAD penalty. Consider maximizing a penalized log-likelihood function

$$l_\lambda(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - n \sum_{m=1}^{M} p_\lambda(\|\boldsymbol{b}_m^*\|_2), \tag{4}$$

where $l(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n} f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{\theta})$ is a log-likelihood function, $p_\lambda(\cdot)$ is a SCAD penalty function and $\|\boldsymbol{b}_m^*\|_2$ is given by

$$\|\boldsymbol{b}_m^*\|_2 = \sqrt{\boldsymbol{b}_m^{*^T} \boldsymbol{G}_m \boldsymbol{b}_m^*} \tag{5}$$

with $p_m \times p_m$ positive semi-definite matrix $\boldsymbol{G}_m$. The definition and the properties of the penalty are given in the following subsection.

3.1 Property of the group SCAD penalty

The first derivative of the SCAD penalty $p_\lambda(\cdot)$ is given by

$$p'_\lambda(|\theta|) = \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\} \tag{6}$$

with tuning parameters $\lambda (> 0)$ and $a (> 2)$. Fan and Li (2001) reported that $a \approx 3.7$ results in minimal Bayes risk, and thus we use it. We easily determine that $p_\lambda(\cdot)$ is given by

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta| & (|\theta| \leq \lambda), \\ -\dfrac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)} & (\lambda < |\theta| \leq a\lambda), \\ \dfrac{(a+1)\lambda^2}{2} & (a\lambda < |\theta|). \end{cases}$$

The SCAD penalty is derived in order to remove drawbacks of the lasso and the hard thresholding penalty and takes the following three properties into consideration simultaneously:

1. Sparsity: The estimator shrinks small parameters towards exactly zero.

2. Continuity: The estimator is continuous at threshold levels to avoid instability in model prediction.

3. Unbiasedness: Sufficiently large estimators are unbiased to avoid unnecessary bias.

Yuan and Lin (2006) considered substituting the form given in (5) into the lasso penalty function $p_\lambda(\theta) = |\theta|$. It enables all p_m elements of the parameter vector $\boldsymbol{b}_m^*$ involving the m-th variable to shrink towards zero when the value of $||\boldsymbol{b}_m^*||_2$ is sufficiently small, which provides an appropriate variable selection. It is the basic concept behind the group lasso. Furthermore, Wang *et al.* (2007) proposed to use a group SCAD penalty for estimating varying-coefficient models with scalar predictors and a functional response. They also applied it to the analysis of gene expression data, selecting transcriptional factors that are relevant to gene expression.

3.2 Estimation

It is difficult to derive the SCAD estimator analytically because of the inclusion of the L_1 penalty. Therefore, we need to approximate (4) analytically or numerically. Fan and Li (2001) locally approximated the SCAD penalty by the quadratic function and then derived the estimator via an iterative procedure, in the framework of generalized linear models. Consider initial values $\boldsymbol{b}^{(0)} = (\beta_0^{(0)}, \boldsymbol{b}_1^{*(0)^T}, \ldots, \boldsymbol{b}_M^{*(0)^T})^T$ and $\sigma^{(0)2}$. For $\boldsymbol{b}^{(0)}$, for example, ridge estimator or maximum likelihood estimator with generalized inverse can be considered. Then we approximate the SCAD penalty as follows:

$$p_\lambda(\|\boldsymbol{b}_m^*\|_2) \approx p_\lambda(\|\boldsymbol{b}_m^{*(0)}\|_2) + \frac{1}{2}\frac{p'_\lambda(\|\boldsymbol{b}_m^{*(0)}\|_2)}{\|\boldsymbol{b}_m^{*(0)}\|_2}(\boldsymbol{b}_m^{*T}\boldsymbol{b}_m^* - \boldsymbol{b}_m^{*(0)T}\boldsymbol{b}_m^{*(0)}) \qquad \text{for } \boldsymbol{b}_m^* \approx \boldsymbol{b}_m^{*(0)}.$$

Therefore the penalized log-likelihood function (4) can be approximated by

$$\begin{aligned} l(\boldsymbol{b}) \approx{}& l(\boldsymbol{b}^{(0)}) + \nabla l(\boldsymbol{b}^{(0)})^T(\boldsymbol{b} - \boldsymbol{b}^{(0)}) \\ &+ \frac{1}{2}(\boldsymbol{b} - \boldsymbol{b}^{(0)})^T \nabla^2 l(\boldsymbol{b}^{(0)})(\boldsymbol{b} - \boldsymbol{b}^{(0)}) - \frac{n}{2}\boldsymbol{b}^T\Sigma(\boldsymbol{b}^{(0)})\boldsymbol{b}, \end{aligned} \tag{7}$$

where $\Sigma(\boldsymbol{b}) = \mathrm{diag}\{0, p'_\lambda(\|\boldsymbol{b}_1^*\|_2)/\|\boldsymbol{b}_1^*\|_2\mathbf{1}_{p_1}, \ldots, p'_\lambda(\|\boldsymbol{b}_M^*\|_2)/\|\boldsymbol{b}_M^*\|_2\mathbf{1}_{p_M}\}$ and the constant term with respect to $\boldsymbol{b}$ is omitted. By maximizing (7), the $(k+1)$-th updated value of $\boldsymbol{b}$ is given by

$$\boldsymbol{b}^{(k+1)} = \boldsymbol{b}^{(k)} - \{\nabla^2 l(\boldsymbol{b}^{(k)}) - n\Sigma(\boldsymbol{b}^{(k)})\}^{-1}\{\nabla l(\boldsymbol{b}^{(k)}) - n\Sigma(\boldsymbol{b}^{(k)})\boldsymbol{b}^{(k)}\}. \tag{8}$$

In particular, for the Gaussian model, (8) can be rewritten as

$$\boldsymbol{b}^{(k+1)} = \left(Z^TZ + n\sigma^{(k)2}\Sigma(\boldsymbol{b}^{(k)})\right)^{-1} Z^T\boldsymbol{y},$$

where $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. The parameter $\boldsymbol{b}$ is updated until the stopping criterion is satisfied. Furthermore, the variance σ^2 is updated by

$$\sigma^{(k+1)2} = \frac{1}{n}(\boldsymbol{y} - Z\boldsymbol{b}^{(k+1)})^T(\boldsymbol{y} - Z\boldsymbol{b}^{(k+1)}). \tag{9}$$

Since the updated values of $\boldsymbol{b}$ and σ^2 depend on each other, they are updated until convergence. Then we have regularized estimator $\hat{\boldsymbol{b}}$ and $\hat{\sigma}^2$.

4 Model selection criteria

The statistical model estimated by the SCAD regularization depends on the regularization parameter λ and thus it is very important to determine its value. We introduce some model selection criteria for objectively selecting λ.

Fan and Li (2001) used a GCV to evaluate models estimated by the SCAD regularization, given by

$$\mathrm{GCV} = \frac{1}{n}\frac{\|\boldsymbol{y} - Z\hat{\boldsymbol{b}}\|^2}{(1 - \hat{df}/n)^2}, \tag{10}$$

where $\hat{df} = \mathrm{tr}\{Z(Z^TZ + n\hat{\sigma}^2\Sigma(\hat{\boldsymbol{b}}))^{-1}Z^T\}$ is a effective degrees of freedom. While on the other hand, Wang *et al.* (2007) proved that the GCV does not provide consistent models and proposed using a BIC that is consistent with respect to model selection. It is given by

$$\mathrm{BIC} = -2\sum_{\alpha=1}^{n} f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) + \hat{df}\log n.$$

However, the BIC is originally derived for evaluating models estimated by the maximum likelihood method, not by the maximum penalized likelihood method including the group SCAD regularization. We consider using a model selection criterion GIC (Konishi and Kitagawa, 1996) which is derived for evaluating the model based on the framework of the M-estimator, involving the maximum penalized likelihood estimator. Using the result of Konishi and Kitagawa, we derive a GIC for evaluating the functional regression model estimated by the group SCAD regularization, given by

$$\mathrm{GIC} = -2\sum_{\alpha=1}^{n} f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) + 2\mathrm{tr}\left\{R(\hat{\boldsymbol{\theta}})^{-1}Q(\hat{\boldsymbol{\theta}})\right\},$$

where $R(\boldsymbol{\theta})$ and $Q(\boldsymbol{\theta})$ are defined by

$$\begin{aligned}
R(\boldsymbol{\theta}) &= -\frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial^2\{\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) - \boldsymbol{b}^T\Sigma(\boldsymbol{b})\boldsymbol{b}/2\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \\
&= \frac{1}{n\sigma^2}\begin{pmatrix} Z^TZ + n\sigma^2\Sigma(\boldsymbol{b}) & \dfrac{1}{\sigma^2}Z^T\boldsymbol{\Lambda}\mathbf{1}_n \\ \dfrac{1}{\sigma^2}\mathbf{1}_n^T\boldsymbol{\Lambda}Z & \dfrac{n}{2\sigma^2}\end{pmatrix}, \\
Q(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial\{\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) - \boldsymbol{b}^T\Sigma(\boldsymbol{b})\boldsymbol{b}/2\}}{\partial\boldsymbol{\theta}}\frac{\partial\{\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}^T} \\
&= \frac{1}{n\sigma^2}\begin{pmatrix} \dfrac{1}{\sigma^2}Z^T\boldsymbol{\Lambda}^2Z - \Sigma(\boldsymbol{b})\boldsymbol{b}\mathbf{1}_n^T\boldsymbol{\Lambda}Z & \dfrac{1}{\sigma^4}Z^T\boldsymbol{\Lambda}^3\mathbf{1}_n - \dfrac{1}{\sigma^2}Z^T\boldsymbol{\Lambda}\mathbf{1}_N \\ \dfrac{1}{\sigma^4}\mathbf{1}_n^T\boldsymbol{\Lambda}^3Z - \dfrac{1}{\sigma^2}\mathbf{1}_n^T\boldsymbol{\Lambda}Z & \dfrac{1}{2\sigma^6}\mathbf{1}_n^T\boldsymbol{\Lambda}^4\mathbf{1}_n - \dfrac{n}{4\sigma^2}\end{pmatrix}
\end{aligned}$$

respectively, where $\boldsymbol{\Lambda} = \mathrm{diag}\left\{y_1 - \boldsymbol{z}_1^T\boldsymbol{b}, \ldots, y_n - \boldsymbol{z}_n^T\boldsymbol{b}\right\}$. We select λ minimizing these criteria, and then select the corresponding model as the optimal one.

5 Examples

The functional regression modeling is applied to Monte Carlo simulations and analysis of real data. We examined the effectiveness of the proposed modeling strategy by investigating whether our method appropriately select functional predictors.

5.1 Numerical example

In the Monte Carlo simulation, we simulated n sets of 3–variate functional predictors and a scalar response $\{(x_{\alpha m}(t), y_\alpha); t \in \mathcal{T}_m, \alpha = 1, \ldots, n, m = 1, 2, 3\}$, then selected functional predictors using the functional regression model and group SCAD regularization. First, we generated predictors $x_{\alpha mi}$ $(i = 1, \ldots, 50)$ corresponding to the m-th predictor X_m according to the following rule:

$$x_{\alpha mi} = u_{\alpha m}(t_{mi}) + \varepsilon_{\alpha mi}, \qquad \varepsilon_{\alpha mi} \sim N(0, 0.025 r_{x\alpha m}^2), \tag{11}$$

where $r_{x\alpha m} = \max_i(u_{\alpha m}(t_{mi})) - \min_i(u_{\alpha m}(t_{mi}))$ and we assume $u_{\alpha m}(t)$ as follows:

$$\begin{array}{lll} X_1: & u_{\alpha 1}(t) = \cos(2\pi(t - a_1)) + a_2 t, \ \ \mathcal{T}_1 = [0, 1], & a_1 \sim N(-5, 3^2), \ \ a_2 \sim N(7, 1), \\ X_2: & u_{\alpha 2}(t) = b_1 \sin(2t) + b_2, \ \ \mathcal{T}_2 = [0, \pi/3], & b_1 \sim U(3, 7), \ \ b_2 \sim N(0, 1), \\ X_3: & u_{\alpha 3}(t) = c_1 t^3 + c_2 t^2 + c_3 t + c_4, \ \ \mathcal{T}_3 = [-1, 1], & c_1 \sim N(-3, 1.2^2), \ \ c_2 \sim N(2, 0.5^2), \\ & & c_3 \sim N(-2, 1), \ \ c_4 \sim N(2, 1.5^2). \end{array}$$

Next, scalar response Y is generated as follows:

$$\begin{aligned} y_\alpha &= g(\boldsymbol{u}_\alpha) + \varepsilon_\alpha, \\ g(\boldsymbol{u}_\alpha) &= \sum_{m=1}^{3} \int_{\mathcal{T}_m} u_{\alpha m}(t)\beta_m(t)dt, \ \ \varepsilon_\alpha \sim N(0, (c r_y)^2). \end{aligned}$$

where $r_y = \max(g(\boldsymbol{u}_\alpha)) - \min(g(\boldsymbol{u}_\alpha))$ and coefficient functions $\beta_m(t)$ are given by

$$\beta_1(t) = \sin(2\pi t), \quad \beta_2(t) = \sin(\pi t), \quad \beta_3(t) = 0.$$

In other words, only X_1 and X_2 relate to Y; X_3 does not. The aim of this analysis is to select functional predictors correctly.

As a first step of the analysis, we converted the data $x_{\alpha mi}$ into functional data $x_{\alpha m}(t)$ by the smoothing method based on the basis expansion. For simplicity we restricted the number of basis functions of the 3 predictors to be the same. It was selected to be 6 by the model selection criterion GBIC (Konishi *et al.*, 2004). For these data, we assumed the functional regression model

$$y_\alpha = \sum_{m=1}^{3} \int_{\mathcal{T}_m} x_{\alpha m}(t)\beta_m(t)dt + \varepsilon_\alpha, \ \ \varepsilon_\alpha \sim N(0, \sigma^2),$$

Table 1: Results from the simulation example.

	$n = 50$				$n = 100$			
	GCV	BIC	GIC	glasso	GCV	BIC	GIC	glasso
(c=0.05)								
AMSE	1.98	1.98	1.98	1.70	2.76	2.76	2.77	2.79
Correct	80	80	80	78	90	91	90	85
(c=0.1)								
AMSE	7.58	7.58	7.42	5.96	7.87	7.88	7.88	8.06
Correct	51	51	53	48	80	80	80	72

and estimated it by group SCAD regularization. Furthermore, we evaluated the model using the three model selection criteria described in the previous section. We also compared group SCAD regularization with group lasso regularization. We repeated the above procedure for 100 times, then computed the averages of the mean squared errors and the number of correctly selected models.

Table 1 shows results of the average mean squared error and the number of correctly selected models. From the table we observe that when the sample size n is small the AMSE of the group SCAD is larger than that of the group lasso. On the other hand, when n is relatively large, differences become small, and the group SCAD is slightly preferable to the group lasso. The results also show that the group SCAD procedure is better than the group lasso at selecting the correct model, and that differences increase as the sample size becomes large. For model selection criteria in the group SCAD regularization, the proposed GIC performs as well as the ordinary criteria.

5.2 Analysis of weather data

We applied functional regression modeling via group SCAD regularization to the analysis of weather data, available on Chronological Scientific Tables 2005, selecting variables concerning weather information. We used weather data observed at 79 stations in Japan. The data includes monthly and annual total observations averaged from 1971 to 2000: monthly observed average temperatures (TEMP), average atmospheric pressure (PRESS), time of daylight (LIGHT), average humidity (HUMID), maximum temperature (MAX.TEMP), minimum temperature (MIN.TEMP) and annual total precipitation. The aim of the analysis is to select monthly observed weather data that have a relationship with annual total

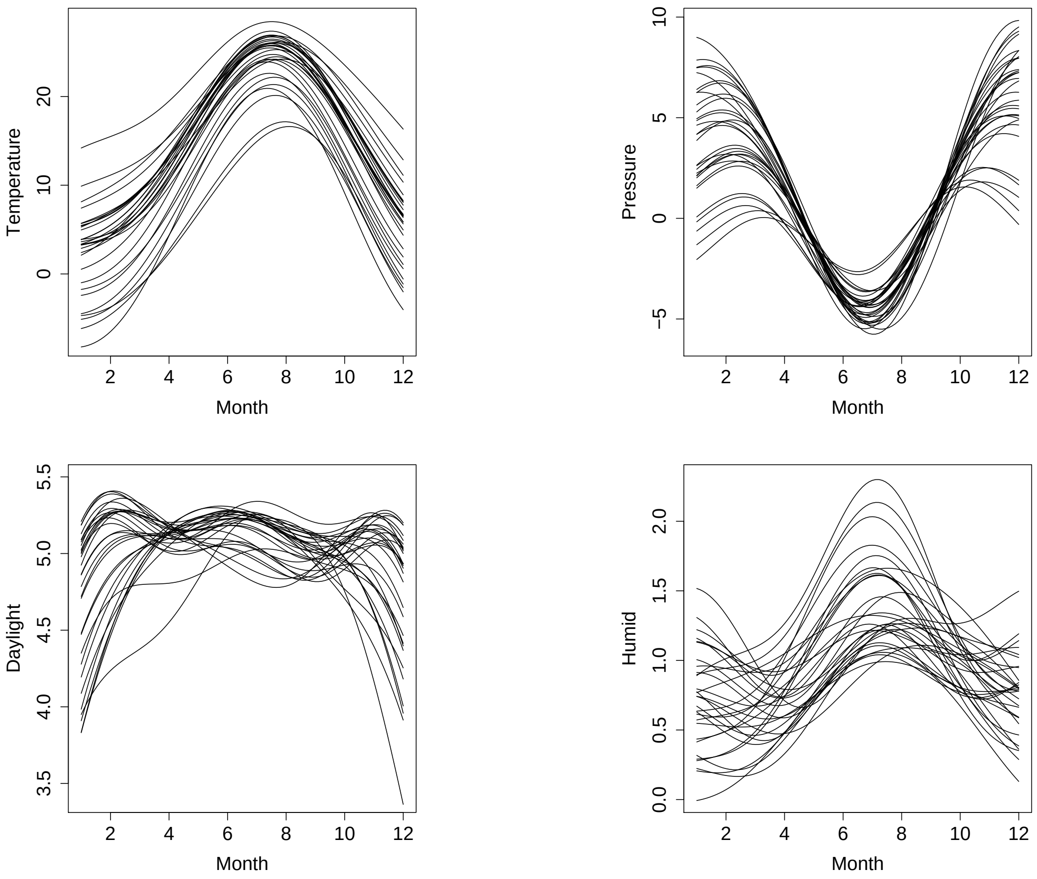

Figure 1: Examples of monthly observed data converted into functions.

precipitation. Since the dimension of monthly data elements may become large we used them as functional data, and then applied functional regression modeling with the group SCAD penalty to weather data.

Monthly data, observed at 12 points, were firstly converted into functions using Gaussian basis expansions with regularization. Since the selected number of Gaussian basis functions is 6 for all kinds of functional variables, the number of parameters for each variable in functional regression model was taken to be 6. Then we estimated it via the regularization method with group SCAD penalty. The estimated model was evaluated using the model selection criterion GIC.

Estimated parameters for each variable of the functional regression model $\hat{\boldsymbol{b}}_m^*$ are shown in Table 2. From this result, parameters concerning average temperature, average atmospheric pressure and minimum temperature are estimated to be zero. It indicates that there is no relationship between these variables and the precipitation and that the remain-

Table 2: Estimators of parameters.

Variable	TEMP	PRESS	LIGHT	HUMID	MAX.TEMP	MIN.TEMP
Estimated coefficients	0	0	4.12	−5.77	−0.52	0
	0	0	−0.23	1.05	−0.43	0
	0	0	−4.55	5.99	−0.32	0
	0	0	0.01	−1.14	−0.20	0
	0	0	1.89	−3.57	1.10	0
	0	0	−1.42	3.58	0.35	0

Table 3: Number of correctly selected models.

Variable	TEMP	PRESS	LIGHT	HUMID	MAX.TEMP	MIN.TEMP
Select	39	31	45	72	52	30

ing variables, namely the time of daylight, average humidity and maximum temperature, is considered to relate to the precipitation. Furthermore, we generated 100 bootstrap samples from the weather data. For each bootstrap sample functional regression modeling was performed, then we examined how many times each variable was selected. The results are shown in Table 3. The mean humidity was selected most frequently among the 6 variables, followed by the maximum temperature and the time of daylight. It reveals the relationships of these variables to the precipitation. On the other hand, the average atmospheric pressure and the minimum temperature are less selected. From the results, there seems to be less of a relationship between these variables and the precipitation.

6 Concluding remarks

We considered the problem of selecting functional variables using the L_1 regularization. Time-course observations are converted into functional forms using Gaussian basis function expansions and regularization, then we constructed functional regression models. Since there are multiple parameters in each functional predictors we treated them as grouped parameters, then applied the group SCAD regularization. In order to select regularization parameters we derived a model selection criterion for evaluating models estimated by the maximum pnelized likelihood method. The proposed modeling strategy is applied to the analysis of a simulation example and real data, selecting functional predictors effectively.

The proposed modeling procedure may be extended to the framework of generalized linear models. Then we can select functional predictors which is relevant for classifying data into some distinct groups, using functional logistic modelings. Furthermore, future works reminds on applying the group SCAD regularization to functional regression models with functional response and predictors. We believe that the proposed method may be efficient solution for analyzing high dimensional data, especially when the dimension of predictors are much greater than the number of observations.

Appendix: Converting discrete data to functional data

Since data are generally obtained discretely we need to express these data as functions. We apply a smoothing method via the regularized basis expansion for converting raw data into functional data. Here we omit the suffix of the index of the functional predictor m for simplicity.

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

$$x_{\alpha i} = u_\alpha(t_{\alpha i}) + \varepsilon_{\alpha i}, \quad i = 1, \ldots, N_\alpha, \tag{12}$$

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

We assume that $u_\alpha(t)$ are represented by the basis function expansion such as

$$u_\alpha(t) = \sum_{j=1}^{p} w_{\alpha j} \phi_j(t) = \boldsymbol{w}_\alpha' \boldsymbol{\phi}(t), \tag{13}$$

where $\boldsymbol{w}_\alpha = (w_{\alpha 1}, \ldots, w_{\alpha p})'$ are vectors of coefficient parameters and $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_p(t))'$ are vectors of Gaussian basis functions

$$\phi_j(t) = \exp\left\{ -\frac{(t - c_j)^2}{2\nu s_j^2} \right\} \quad (j = 1, \ldots, p), \tag{14}$$

where c_j and s_j^2 are center and dispersion parameters of basis functions respectively and ν is a hyperparameter. We need to estimate c_j and s_j^2 in addition to coefficients $w_{\alpha j}$. Although we can estimate their values simultaneously, this method may cause unstable estimates or may be computationally expensive. A useful technique is to determine these values prior to estimating $w_{\alpha j}$ by applying clustering algorithms such as k-means method to observational points (Moody and Darken, 1989). First, $\sum_\alpha N_\alpha$ observational points $\{t_{\alpha i};\ \alpha = 1, \dots, n,\ i = 1, \dots, N_\alpha\}$ are divided into p clusters $\{C_1, \dots, C_p\}$, and then c_j and s_j^2 are determined by

$$\hat{c}_j = \frac{1}{d_j} \sum_{t_{\alpha i} \in C_j} t_{\alpha i}, \quad \hat{s}_j^2 = \frac{1}{d_j} \sum_{t_{\alpha i} \in C_j} (t_{\alpha i} - \hat{c}_j)^2$$

respectively, where $d_j = \#\{t_{\alpha i} \in C_j\}$. The number of clusters p becomes smaller than $\min_\alpha N_\alpha$. From these results the regression model (12) has a probability density function

$$f(x_{\alpha i} | t_{\alpha i}; \boldsymbol{w}_\alpha, \sigma_{x\alpha}^2) = \frac{1}{\sqrt{2\pi\sigma_{x\alpha}^2}} \exp\left\{ -\frac{(x_{\alpha i} - \boldsymbol{w}'_\alpha \boldsymbol{\phi}(t_{\alpha i}))^2}{2\sigma_{x\alpha}^2} \right\}. \tag{15}$$

The parameters $\boldsymbol{w}_\alpha$ and $\sigma_{x\alpha}^2$ are secondly estimated by using regularization method, which maximizes a penalized log-likelihood function

$$l_{\zeta_\alpha}(\boldsymbol{w}_\alpha, \sigma_{x\alpha}^2) = \sum_{i=1}^{N_\alpha} \log f(x_{\alpha i} | t_{\alpha i}; \boldsymbol{w}_\alpha, \sigma_{x\alpha}^2) - \frac{N_\alpha \zeta_\alpha}{2} \boldsymbol{w}'_\alpha \Omega \boldsymbol{w}_\alpha, \tag{16}$$

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

$$\hat{\boldsymbol{w}}_\alpha = (\Phi'_\alpha \Phi_\alpha + N_\alpha \zeta_\alpha \hat{\sigma}_{x\alpha}^2 \Omega)^{-1} \Phi'_\alpha \boldsymbol{x}_{(\alpha)}, \quad \hat{\sigma}_{x\alpha}^2 = \frac{1}{N_\alpha} (\boldsymbol{x}_{(\alpha)} - \Phi_\alpha \hat{\boldsymbol{w}}_\alpha)' (\boldsymbol{x}_{(\alpha)} - \Phi_\alpha \hat{\boldsymbol{w}}_\alpha), \tag{17}$$

respectively, where $\Phi_\alpha = (\boldsymbol{\phi}(t_{\alpha 1}), \dots, \boldsymbol{\phi}(t_{\alpha N_\alpha}))'$ and $\boldsymbol{x}_{(\alpha)} = (x_{\alpha 1}, \dots, x_{\alpha N_\alpha})'$.

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

$$x_\alpha(t) \equiv \hat{u}_\alpha(t) = \hat{\boldsymbol{w}}'_\alpha \boldsymbol{\phi}(t). \tag{18}$$

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

References

Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks. *J. Stat. Plan. Infer.* **138**, 3617–3633.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2008). Functional Regression Modeling via Regularized Gaussian Basis Expansions. To appear in *Ann. Inst. Statist. Math.*

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information–theoretical Approach (2nd ed.)* Springer–Verlag, New York.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303–316.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J. and Li, R. (2002), Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710–723.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function network, *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, New York.

Miller, A. J. (1990). *Subset Selection in Regression.* Chapman & Hall, London.

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

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

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis (2nd ed.)* Springer–

Verlag, New York.

Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003). Regression approaches for microarray data analysis. *J. Comp. Biol.* **10**, 961–980.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Stat. Med.* **16**, 385–395.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486–1494.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**(3), 553–568.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49–67.

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