ADAPTIVE BRIDGE REGRESSION MODELING AND SELECTION OF THE TUNING PARAMETERS

Kawano, Shuichi
Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University

https://doi.org/10.5109/1495409

出版情報：Bulletin of informatics and cybernetics. 44, pp.29-39, 2012-12. Research Association of Statistical Sciences
バージョン：
権利関係：

ADAPTIVE BRIDGE REGRESSION MODELING AND SELECTION OF THE TUNING PARAMETERS

by

Shuichi Kawano

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.44

FUKUOKA, JAPAN
2012

Bulletin of Informatics and Cybernetics, Vol. 44, 2012

ADAPTIVE BRIDGE REGRESSION MODELING AND SELECTION OF THE TUNING PARAMETERS

By

Shuichi Kawano*

Abstract

We consider the problem of constructing an adaptive bridge regression modeling, which is a penalized procedure where different weights are imposed on different coefficients in a bridge penalty term. A crucial issue in the modeling process is the choice of the adjusted parameters included in the models. Here, we treat the selection of the adjusted parameters as model selection and evaluation problems. In order to select the parameters, the model selection criteria are derived from information-theoretic and Bayesian approaches. We conduct some numerical studies to investigate the effectiveness of our proposed modeling strategy.

Key Words and Phrases: Adaptive penalty, Bayesian approach, Bridge regression, Information criterion, Penalized maximum likelihood method.

1. Introduction

Consider the linear regression model

$$\boldsymbol{y} = \beta_0 \mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an n-dimensional response vector, $\mathbf{1}_n$ is an n-dimensional vector whose elements are all one, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is an $n \times p$ design matrix, β_0 is an intercept parameter, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional coefficient parameter vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n-dimensional error vector distributed as $N(\mathbf{0}, \sigma^2 I_n)$, where σ (> 0) is an unknown parameter and I_n is an $n \times n$ identity matrix. Without loss of generality, we omit the intercept β_0 and the design matrix X is standardized; i.e., we assume that $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = n$ $(j = 1, \ldots, p)$.

We need to estimate the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$ in Equation (1) from data. Many procedures have been used to estimate the parameters; e.g., the maximum likelihood method, the ridge (Hoerl and Kennard, 1970), the lasso (Tibshirani, 1996), the bridge (Frank and Friedman, 1993) and the elastic net (Zou and Hastie, 2005). The bridge, or bridge regression, estimates the parameter vector $\boldsymbol{\theta}$ by maximizing the penalized log-likelihood function

$$\ell_{\lambda,q}(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) - \frac{n\lambda}{2} \sum_{j=1}^p |\beta_j|^q, \tag{2}$$

* Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, 1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan. skawano@ms.osakafu-u.ac.jp

where λ (> 0) is a regularization parameter, q (> 0) is a tuning parameter and $f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$ is a probability density function in the form

$$f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right], \quad i = 1, \ldots, n. \tag{3}$$

For any given q, the estimator $\hat{\boldsymbol{\theta}}$ that maximizes (2) is called a bridge estimator. Bridge regression involves two special cases: when $q = 1$, it becomes the lasso, and when $q = 2$, it becomes ridge regression. Also, when $0 < q \leq 1$, it is known that the estimator produces sparse solutions for the coefficient parameter $\boldsymbol{\beta}$, that is, some coefficients shrink to exactly zero. Hence, bridge regression for $0 < q \leq 1$ can simultaneously perform model selection and parameter estimation.

Bridge regression has been studied by many researchers. For example, Armagan (2009), Fu (1998) and Zou and Li (2008) presented efficient algorithms for estimating the parameters in bridge regression models. Kawano (2012) introduced a criterion from a Bayesian viewpoint for the selection of the tuning parameters for bridge regression models. Huang, Horowitz and Ma (2008) and Knight and Fu (2000) showed the asymptotic behavior of bridge estimators in the framework of linear regression models. Huang *et al.* (2009) and Park and Yoon (2011) extended bridge regression and the group lasso given by Yuan and Lin (2006) to propose a group bridge regression model. Although from the above research papers, we observe that bridge regression is useful, there is a remaining problem, in that bridge regression forces the coefficients to be equally penalized. This is not appropriate since we do not know, in general, if the covariates are equally relevant. Furthermore, it is known that the penalized regression models that impose equal penalties on the coefficients tend to produce less accurate predictions (e.g., see Grandvalet and Canu, 1999; Shimamura *et al.*, 2007, 2009; Zou, 2006; and Zou and Zhang, 2009).

In this paper, we present an adaptive bridge regression procedure that assigns different weights to different coefficients in the bridge penalty term. Thanks to these different weights, this adaptive model enables us to estimate the coefficients more flexibly than ordinary bridge regression models. A crucial issue in our modeling strategy is to choose values for the parameters that are included in the penalty term. In order to select the adjusted parameters objectively, we introduce model selection criteria from the viewpoint of both information-theoretic and Bayesian approaches. Some numerical examples are given to investigate the performance of our proposed modeling procedures.

The rest of this paper is organized as follows. Section 2 describes adaptive bridge regression models and an estimation algorithm for them. In Section 3, we derive model selection criteria from information-theoretic and Bayesian viewpoints to select some adjusted parameters in the models. Section 4 performs Monte Carlo simulations to examine the performance of our proposed strategy. Our concluding remarks are given in Section 5.

2. Adaptive bridge regression modeling

We propose estimating the parameter vector $\boldsymbol{\theta}$ based on the following penalized log-likelihood function:

$$\ell_{\lambda,q,\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) - \frac{n\lambda}{2} \sum_{j=1}^{p} \alpha_j |\beta_j|^q, \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ is a vector of positive weights. The estimator $\hat{\boldsymbol{\theta}}$ that maximizes (4) is the adaptive bridge estimator, and the estimation procedure is adaptive bridge regression. It is clear that the adaptive bridge regression model becomes the bridge regression model when all values of α_j $(j = 1, \ldots, p)$ are one.

Several adaptive versions of ordinary estimation methods have been proposed. For example, Grandvalet and Canu (1999) and Tipping (2001) proposed adaptive ridge regression models, and then they showed that the adaptive models yield sparse solutions for the coefficient vector $\boldsymbol{\beta}$. Huang, Ma, Xie and Zhang (2008), Shimamura *et al.* (2007) and Zou (2006) presented an adaptive lasso in the framework of linear regression models. Also, adaptive versions of elastic net were given by Shimamura *et al.* (2009) and Zou and Zhang (2009).

Since the penalty term in adaptive bridge regression models is a nonconvex function with respect to the parameter $\boldsymbol{\beta}$ when $0 < q < 1$, Equation (4) is a nonconcave optimization problem when $0 < q < 1$. Hence, the penalty is approximated with a convex function by using the local quadratic approximation (LQA) introduced by Fan and Li (2001). Under some conditions, we can approximate the penalty function at the initial values $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_p^{(0)})^T$ in the form

$$|\beta_j|^q \approx |\beta_j^{(0)}|^q + \frac{q}{2}\frac{|\beta_j^{(0)}|^{q-1}}{|\beta_j^{(0)}|}(\beta_j^2 - \beta_j^{(0)2}), \qquad j = 1, \ldots, p. \tag{5}$$

Then, Equation (4) can be expressed as

$$\ell^*_{\lambda,q,\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) - \frac{n\lambda q}{4}\sum_{j=1}^{p} \alpha_j |\beta_j^{(0)}|^{q-2}\beta_j^2. \tag{6}$$

For fixed weights $\boldsymbol{\alpha}$, it is easy to maximize Equation (6) with respect to the parameter vector $\boldsymbol{\theta}$. The estimator of the parameter vector $\boldsymbol{\theta}$ can be obtained by the following algorithm:

Step1 Set the values of the regularization parameter λ and the tuning parameter q.

Step2 Initialize $\boldsymbol{\alpha}^{(0)}$ by $\alpha_j^{(0)} = 1/|\hat{\beta}_j^{\mathrm{MLE}}|$ $(j = 1, \ldots, p)$, where $\hat{\boldsymbol{\beta}}^{\mathrm{MLE}} = (\hat{\beta}_1^{\mathrm{MLE}}, \ldots, \hat{\beta}_p^{\mathrm{MLE}})^T = (X^TX)^{-1}X^T\boldsymbol{y}$.

Step3 For $k = 0, 1, \ldots$, update the parameter vector $\boldsymbol{\theta}$ as follows:

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \mathrm{argmax}_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \log f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) - \frac{n\lambda q}{4}\sum_{j=1}^{p} \alpha_j^{(k)} |\beta_j^{(0)}|^{q-2}\beta_j^2 \right\}. \tag{7}$$

Step4 Update the weights $\boldsymbol{\alpha}$ in the form

$$\alpha_j^{(k+1)} = \frac{1}{|\hat{\beta}_j^{(k+1)}| + \delta}, \qquad j = 1, \ldots, p, \tag{8}$$

where $\hat{\boldsymbol{\theta}}^{(k+1)} = (\hat{\boldsymbol{\beta}}^{(k+1)T}, \hat{\sigma}^{(k+1)2})^T$ and δ is an arbitrary small number (e.g., $\delta = 10^{-5}$ in our simulations).

Step5 Repeat Step3 into Step4 until the condition

$$|\ell^*_{\lambda,q,\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}^{(k+1)}) - \ell^*_{\lambda,q,\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}^{(k)})| < \tau \tag{9}$$

is satisfied, where τ is an arbitrary small number (e.g., $\tau = 10^{-5}$ in our numerical examples), or until the number of iterations attains a predetermined M (e.g., $M = 100$ in our numerical studies).

From the above procedures, we obtain the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2)^T$ and the weights $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)^T$ corresponding to the estimator $\hat{\boldsymbol{\theta}}$. We then derive the following statistical model:

$$f(y_i|\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^2}\right], \quad i = 1, \ldots, n. \tag{10}$$

This model has adjusted parameters; the regularization parameter λ and the tuning parameter q. In order to select the parameters objectively, we introduce model selection criteria in terms of information-theoretic and Bayesian viewpoints in the next section.

Remarks

- Our adaptive bridge regression can be regarded as a special case of the adaptive group bridge regression presented by Park and Yoon (2011). However, our proposed procedure recursively updates the weights $\boldsymbol{\alpha}$ according to the recursive elastic net by Shimamura *et al.* (2009), while Park and Yoon (2011) simply provided weights as $\alpha_j = 1/|\hat{\beta}_j^{\mathrm{MLE}}|$ $(j = 1, \ldots, p)$. The effectiveness of updating the weights has not yet investigated by Shimamura *et al.* (2009) or any other researchers. We consider this to be a future topic for our research.

- As mentioned in Section 1, the theoretical properties of bridge regressions have been studied by Huang, Horowitz and Ma (2008) and Knight and Fu (1998). These papers showed that, under certain regularity conditions, bridge regressions enjoy the oracle property that was introduced by Fan and Li (2001), when $0 < q < 1$. In adaptive bridge regression models, Zou (2006) reported that the oracle property is satisfied when $q = 1$. It will be interesting to investigate if adaptive bridge regressions have the oracle or other theoretical properties when the tuning parameter q is not equal to one. We consider this to be a topic for future research.

3. Model selection criteria

3.1. Generalized Bayesian information criterion

The Bayesian approach to model selection or evaluation criteria is based on a maximization of the marginal likelihood

$$\mathrm{ML} = \int \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2)d\boldsymbol{\theta}, \tag{11}$$

where $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2)$ is the prior distribution of the parameter $\boldsymbol{\theta}$. Here, we consider the prior distribution $\pi(\sigma^2)$ to be the uniform distribution $U(0, a)$ for a large

positive value a (e.g., $a = 10^{10}$), which means a non-informative prior distribution, and the distribution $\pi(\boldsymbol{\beta}|\sigma^2) = \pi(\boldsymbol{\beta})$ to be

$$\pi(\boldsymbol{\beta}|\lambda, q) = \prod_{j=1}^{p} \pi(\beta_j|\lambda, q) = \prod_{j=1}^{p} \frac{q2^{-(1+1/q)}(n\lambda\alpha_j)^{1/q}}{\Gamma(1/q)} \exp\left\{-\frac{n\lambda\alpha_j}{2}|\beta_j|^q\right\}, \quad (12)$$

where $\Gamma(\cdot)$ is the Gamma function. In order to evaluate Equation (11), the Laplace approximation (Tierney and Kadane, 1986) is usually employed. However, in the cases that some components of the coefficients $\boldsymbol{\beta}$ are exactly zero by adaptive bridge approaches, the functional in the integral (11) is not differentiable at the origin, and then the approximation method cannot be directly applied.

Let $\mathcal{A} = \{j; \hat{\beta}_j \neq 0\}$ be an active set of the parameter $\boldsymbol{\beta}$. To overcome the above problems, we consider the partial marginal likelihood

$$\text{ML} \approx \text{PML} = \int \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\lambda, q)\pi(\sigma^2)d\boldsymbol{\theta}_{\mathcal{A}}, \quad (13)$$

where $\boldsymbol{\theta}_{\mathcal{A}} = (\boldsymbol{\beta}_{\mathcal{A}}^T, \sigma^2)^T$. Here $\boldsymbol{\beta}_{\mathcal{A}} = (\beta_{k_1}, \ldots, \beta_{k_r})^T$, where we set $\mathcal{A} = \{k_1, \ldots, k_r\}$ and $k_1 < \cdots < k_r$. The formula (13) can be calculated by integrating over the unknown parameter $\boldsymbol{\theta}_{\mathcal{A}}$. By applying the Laplace approximation for Equation (13), we derive

$$\int \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\lambda, q)\pi(\sigma^2)d\boldsymbol{\theta}_{\mathcal{A}} \approx \frac{(2\pi)^{|\mathcal{A}|+1}}{n^{|\mathcal{A}|+1}|V(\hat{\boldsymbol{\theta}}_{\mathcal{A}})|^{1/2}} \exp\{nv(\hat{\boldsymbol{\theta}}_{\mathcal{A}})\}, \quad (14)$$

where

$$v(\boldsymbol{\theta}) = \frac{1}{n} \log\left\{\prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\lambda, q)\pi(\sigma^2)\right\}, \quad V(\boldsymbol{\theta}) = -\frac{\partial^2 v(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_{\mathcal{A}}\partial\boldsymbol{\theta}_{\mathcal{A}}^T}$$

and $\hat{\boldsymbol{\theta}}_{\mathcal{A}} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^T, \hat{\sigma}^2)^T$, where $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ is the estimator of the coefficient $\boldsymbol{\beta}_{\mathcal{A}}$.

By taking the logarithm for the right-hand side of Equation (14), Konishi *et al.* (2004) proposed the generalized Bayesian information criterion (GBIC), which is an extension of the Bayesian information criterion by Schwarz (1978). By using the result of Konishi *et al.* (2004, p. 30), we obtain a model selection criterion

$$\begin{aligned}\text{GBIC} = {} & n\log(2\pi) + n\log\hat{\sigma}^2 + n - (|\mathcal{A}| + 1)\log\left(\frac{2\pi}{n}\right) \\ & + \log|R| - 2|\mathcal{A}|\log q + 2|\mathcal{A}|\left(1 + \frac{1}{q}\right)\log 2 - \frac{2|\mathcal{A}|}{q}\log(n\lambda) \\ & + 2\log a + 2|\mathcal{A}|\log\Gamma\left(\frac{1}{q}\right) - \frac{2}{q}\sum_{j\in\mathcal{A}}\log\hat{\alpha}_j + n\lambda\sum_{j\in\mathcal{A}}\hat{\alpha}_j|\hat{\beta}_j|^q, \end{aligned} \quad (15)$$

where R is a $(|\mathcal{A}| + 1) \times (|\mathcal{A}| + 1)$ matrix given by

$$R = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} X_{\mathcal{A}}^T X_{\mathcal{A}} + n\lambda\hat{\sigma}^2 q(q-1)K_1 & \frac{1}{\hat{\sigma}^2}X_{\mathcal{A}}^T\Lambda\mathbf{1}_n \\ \frac{1}{\hat{\sigma}^2}\mathbf{1}_n^T\Lambda X_{\mathcal{A}} & \frac{n}{2\hat{\sigma}^2} \end{pmatrix}. \quad (16)$$

Here $\Lambda = \mathrm{diag}(y_1 - \boldsymbol{x}_1^T\hat{\boldsymbol{\beta}}, \ldots, y_n - \boldsymbol{x}_n^T\hat{\boldsymbol{\beta}})$, $K_1 = \mathrm{diag}(\hat{\alpha}_{k_1}|\hat{\beta}_{k_1}|^{q-2}/2, \ldots, \hat{\alpha}_{k_r}|\hat{\beta}_{k_r}|^{q-2}/2)$ and

$$X_{\mathcal{A}} = [x_{ik}], \quad i = 1, \ldots, n; \quad k \in \mathcal{A}. \tag{17}$$

Note that our criterion, the GBIC in Equation (15), is equal to the GBIC proposed by Kawano (2012) if all α_js ($j = 1, \ldots, p$) are equal to one.

3.2. Generalized information criterion

It can be seen that the adaptive bridge estimator for the active set $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ is given as the solution of the implicit equation

$$\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{18}$$

where

$$\boldsymbol{\psi}(y_i, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}_{\mathcal{A}}} \left\{ \log f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) - \frac{\lambda}{2} \sum_{j=1}^{p} \alpha_j |\beta_j|^q \right\}. \tag{19}$$

This formula is included in the robust estimator (Huber, 2004). Thus, using the results of Konishi and Kitagawa (1996, p. 876), the generalized information criterion (GIC) is obtained by the form

$$\mathrm{GIC} = n\{\log(2\pi) + 1\} + n \log \hat{\sigma}^2 + 2\mathrm{tr}(R^{-1}Q), \tag{20}$$

where R is given by Equation (16) and Q is

$$Q = \frac{1}{n\hat{\sigma}^2} \begin{pmatrix} \frac{1}{\hat{\sigma}^2} X_{\mathcal{A}}^T \Lambda^2 X_{\mathcal{A}} - \lambda q K_2 \mathbf{1}_n^T \Lambda X_{\mathcal{A}} & \frac{1}{2\hat{\sigma}^4} X_{\mathcal{A}}^T \Lambda^3 \mathbf{1}_n - \frac{1}{2\hat{\sigma}^2} X_{\mathcal{A}}^T \Lambda \mathbf{1}_n \\ \frac{1}{2\hat{\sigma}^4} \mathbf{1}_n^T \Lambda^3 X_{\mathcal{A}} - \frac{1}{2\hat{\sigma}^2} \mathbf{1}_n^T \Lambda X_{\mathcal{A}} & \frac{1}{4\hat{\sigma}^6} \mathbf{1}_n^T \Lambda^4 \mathbf{1}_n - \frac{n}{4\hat{\sigma}^2} \end{pmatrix}. \tag{21}$$

Here, K_2 is a $|\mathcal{A}| \times |\mathcal{A}|$ matrix given by

$$K_2 = \mathrm{diag}\left(\frac{\hat{\alpha}_{k_1}|\hat{\beta}_{k_1}|^{q-1}\mathrm{sgn}(\hat{\beta}_{k_1})}{2}, \ldots, \frac{\hat{\alpha}_{k_r}|\hat{\beta}_{k_r}|^{q-1}\mathrm{sgn}(\hat{\beta}_{k_r})}{2} \right), \tag{22}$$

where sgn means the sign function. For more details of the derivation of the GIC, we refer the reader to Konishi and Kitagawa (2008) and to the references given therein.

We select the adjusted parameters, the regularization parameter λ and the tuning parameter q, from the minimizer of the GBIC in (15) or the GIC in (20).

Note that the GBIC and the GIC are useful and applicable only when the number of nonzero estimated parameters is smaller than the sample size. Hence, we must be careful about using these criteria in situations where $q > 1$ and the number of parameters is larger than the sample size, since not all the estimated coefficient parameters will shrink to zero.

4. Numerical studies

We conducted Monte Carlo simulations to investigate our proposed modeling procedure. The simulation has four settings. The design matrix X was generated from a multivariate normal distribution with a mean of zero and a variance of one, and the pairwise correlation structure between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ was $\mathrm{cor}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.5^{|i-j|}$. The response vector $\boldsymbol{y}$ was generated from the linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n). \tag{23}$$

The four simulation settings that we used are as follows:

- Setting 1 : The training data and the test data were given by 20 observations and 200 observations, respectively. The true parameters were $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, \underbrace{0, \ldots, 0}_{8})^T$ and $\sigma = 3$. This setting is a sparse case.
- Setting 2 : This setting is the same as Setting 1 except for $\boldsymbol{\beta} = (\underbrace{0.5, \ldots, 0.5}_{8})^T$. Setting 2 is a dense case.
- Setting 3 : This setting is also the same as Setting 1, except the true parameters were $\boldsymbol{\beta} = (1, \underbrace{0, \ldots, 0}_{7})^T$ and $\sigma = 2$. It is considered as a sparse case.
- Setting 4 : We generated 100 observations and 400 observations for the training data and the test data, respectively. We set

$$\boldsymbol{\beta} = (\underbrace{0, \ldots, 0}_{10}, \underbrace{1.85, \ldots, 1.85}_{10}, \underbrace{0, \ldots, 0}_{10})^T \tag{24}$$

and $\sigma = 3$. This setting also considers sparse.

We fitted adaptive bridge regression models to the generated data sets. The regularization parameter λ and the tuning parameter q in the adaptive bridge penalty were selected by the GBIC (ABGBIC) and the GIC (ABGIC) that were introduced in Section 3., where the candidate values of λ and q were set to $\{10^{-0.1i+3}; i = 1, \ldots, 100\}$ and $\{0.1, 0.4, 0.7, 1.0, 1.3, 1.7, 2.0, 2.3, 2.7\}$, respectively. We also constructed the ordinary bridge regression models (Bridge). The adjusted parameters involved in the bridge regression models were chosen by the model selection criterion given in Kawano (2012).

The mean squared error (MSE), defined by $\mathrm{MSE} = \sum_{i=1}^{n}(\hat{y}_i - y_i^*)^2/n$, was computed to validate the performance of our proposed procedures, where $y_1^*, \ldots, y_n^*$ indicate the test data for the response variable that were generated from the true model. We also computed the medians of the adjusted parameters λ and q. Table 1 presents the simulation results that were obtained by averaging over 100 Monte Carlo trials. The values in parentheses indicate the standard deviations for the means.

The simulation results are summarized as follows. In almost all settings, our proposed methods, the ABGBIC and the ABGIC, provided MSEs that were smaller than those from the Bridge. For the selection of the tuning parameter q, the ABGBIC and the ABGIC gave appropriate values of the tuning parameter in almost all situations except for Setting 2 for the ABGBIC and Setting 4 for the ABGIC; i.e., the value of the tuning

Table 1: Comparison of the mean squared errors (MSE), the values of the regularization parameter λ and the tuning parameter q. Figures in parentheses give the estimated standard deviations.

	ABGBIC	ABGIC	Bridge		ABGBIC	ABGIC	Bridge
Setting 1				Setting 2			
MSE	16.83	18.89	20.67	MSE	13.30	13.60	14.74
	(7.512)	(8.718)	(9.607)		(3.235)	(3.456)	(2.135)
$\log_{10}(\lambda)$	–1.00	–0.70	–0.70	$\log_{10}(\lambda)$	–1.00	–0.90	0.60
q	1.00	1.00	1.00	q	1.00	1.50	1.00
Setting 3				Setting 4			
MSE	5.152	5.176	5.230	MSE	10.98	12.54	10.99
	(1.399)	(1.256)	(0.713)		(1.387)	(2.767)	(1.001)
$\log_{10}(\lambda)$	–0.70	–0.50	0.90	$\log_{10}(\lambda)$	–1.70	–1.50	–1.50
q	1.00	2.00	1.00	q	1.00	1.30	0.70

parameter q was larger than one when the structure of the coefficient parameter $\boldsymbol{\beta}$ was dense, and $0 < q \leq 1$ is given when the coefficient parameter $\boldsymbol{\beta}$ had sparse patterns. Suitable values of the tuning parameter q were also given by the Bridge except for in Setting 2. From these results, we conclude that our proposed procedures outperform the Bridge.

We also compared the adaptive bridge regression models using the GBIC and the GIC with the ordinary least square (OLS), the ridge, the lasso and the elastic net (ENet). An adjusted parameter involved in ridge regressions was selected by the leave-one-out cross-validation, and the five-fold cross-validation was used to choose the adjusted parameters included in the lasso and the ENet. To investigate the performance of each method, we computed MSEs and constructed boxplots of the values for the 100 trials of the Monte Carlo experiments. Figure 1 shows the boxplots of the MSEs. We observe that, in almost all situations, our adaptive bridge regression model is superior to the Bridge, the OLS and the ridge from the aspect of minimizing MSEs, while our proposed methods are competitive with the lasso and the ENet.

5. Concluding remarks

We proposed an adaptive bridge regression modeling procedure that imposes different weights to different coefficients in the penalty term. We also provided an estimation algorithm for our models. In order to objectively select the adjusted parameters in the models, model selection criteria were derived in terms of information-theoretic and Bayesian approaches. Numerical examples showed that our proposed modeling strategy performed well in several situations from the viewpoint of yielding relatively lower prediction errors and selecting appropriate values of the tuning parameter q. It is important to extend the proposed modeling procedures into the framework of generalized linear models or nonlinear models, including generalized additive models. Also, it is interesting to analyze high-dimensional data by using the proposed methods. We consider this to be one of our future research topics.

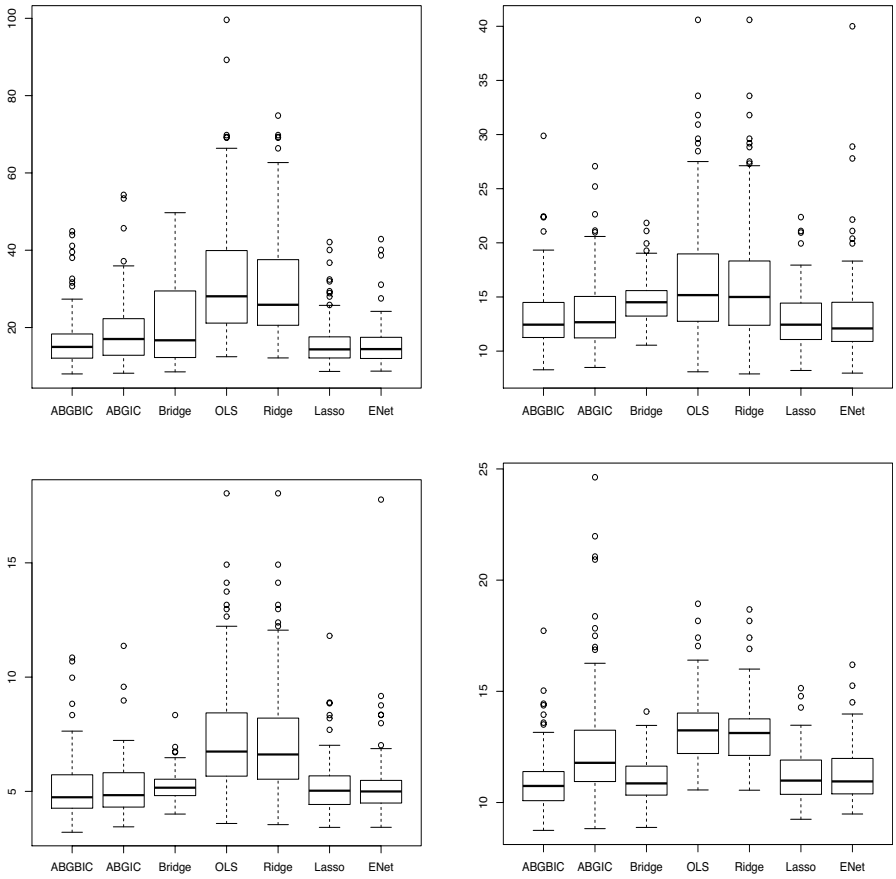

Figure 1: Boxplots of the MSEs. The left top panel shows the result for Setting 1, the right top panel that for Setting 2, the left bottom panel that for Setting 3 and the right bottom panel that for Setting 4.

Acknowledgement

The author would like to thank the anonymous reviewer for his helpful comments and suggestions. This work was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #24700280, 2012–2015.

References

Armagan, A. (2009). Variational bridge regression. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, **5**, 17–24.

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.

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

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

Grandvalet, Y. and Canu, S. (1999). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. *Advances in Neural Information Processing Systems*, **11**, 445–451.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, **36**, 587–613.

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistics Sinica*, **18**, 1603–1618.

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, **96**, 339–355.

Huber, P. (2004). *Robust Statistics*. New York: Wiley.

Kawano, S. (2012). Selection of tuning parameters in bridge regression models via Bayesian information criterion. Prerint, arXiv:1203.4326.

Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, **28**, 1356–1378.

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*. New York: Springer.

Park, C. and Yoon, Y. J. (2011). Bridge regression: adaptivity and group selection. *Journal of Statistical Planning and Inference*, **141**, 3506–3519.

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

Shimamura, T., Imoto, S., Yamaguchi, R., Fujita, A., Nagasaki, M. and Miyano, S. (2009). Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, **3**, 41.

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.

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

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, **1**, 211–244.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68**, 49–67.

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

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.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, **36**, 1509–1533.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, **37**, 1733–1751.

Received April 14, 2012
Revised August 28, 2012