Selection of tuning parameters in bridge regression models via Bayesian information criterion

Shuichi Kawano

Received: 15 December 2012 / Revised: 29 July 2013 / Published online: 15 October 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract We consider bridge regression models, which can produce a sparse or non-sparse model by controlling a tuning parameter in the penalty term. A crucial part of a model building strategy is the selection of the values for adjusted parameters, such as regularization and tuning parameters. Indeed, this can be viewed as a problem in selecting and evaluating the model. We propose a Bayesian selection criterion for evaluating bridge regression models. This criterion enables us to objectively select the values of the adjusted parameters. We investigate the effectiveness of our proposed modeling strategy with some numerical examples.

Keywords Bridge penalty · Model selection · Penalized maximum likelihood method · Sparse regression

Mathematics Subject Classification 62J05 · 62G05 · 62F15

1 Introduction

With the appearance of high-throughput data of unprecedented size and complexity, statistical methods have become increasingly important. In particular, linear regression has become a widely used and fundamental tool. The parameters in regression models are usually estimated by using the ordinary least squares (OLS) or maximum likelihood method. However, the models produced by these methods often have unstable estimators and yield large prediction errors, especially when there is a multicollinearity problem.
In order to overcome this problem, various penalties and penalized regression methods have been proposed: e.g., ridge regression (Hoerl and Kennard 1970), the lasso (Tibshirani 1996), bridge regression (Frank and Friedman 1993), the elastic net (Zou and Hastie 2005), the smoothly clipped absolute deviation (SCAD; Fan and Li 2001), and the minimax concave penalty (MCP; Zhang 2010). We will focus on bridge linear regression, which is a linear regression model in which the parameters are estimated with the bridge penalty. An advantage of bridge regression is to be able to produce a sparse model, which has received considerable attention in high-dimensional data analysis that has extensively studied in recent machine learning and statistical literature (see, e.g., Bühlmann and van de Geer 2011), or a non-sparse model by controlling a tuning parameter included in the penalty term. Additionally, many researchers have shown that bridge regression models are useful from practical and theoretical perspectives (e.g., see Armagan 2009; Fu 1998; Huang et al. 2008; Knight and Fu 2000). Although bridge regression is useful, there remains a problem of evaluating bridge regression models, which leads to the selection of values for adjusted parameters involved in the constructed bridge regression models. Cross-validation (CV) is often used to evaluate the resulting model, although it tends to have a very high computational time. In addition, CV’s high variability and tendency to undersmooth are not negligible, since the selectors are repeatedly applied.

In this paper, we introduce a model selection criterion for evaluating the models estimated by the penalized maximum likelihood method with the bridge penalty from the viewpoint of Bayesian approach. The proposed criterion enables us to select objectively the appropriate values of the adjusted parameters in a bridge regression model. Through some numerical studies, we investigate the performance of our proposed methodology.

This paper is organized as follows. Section 2 describes the bridge linear regression models with the estimation algorithm. In Sect. 3, we introduce a criterion derived from the Bayesian viewpoint for selecting the values of some adjusted parameters in the models. Section 4 conducts Monte Carlo simulations and a real data analysis to examine the performance of our proposed strategy and to compare several types of criteria and methods. Some concluding remarks are given in Sect. 5.

2 Bridge regression modeling

2.1 Preliminary

Suppose that we have a data set \{((y_i, x_i); i = 1, \ldots, n}\}, where \(y_i \in \mathbb{R}\) is a response variable and \(x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p\) denotes a \(p\)-dimensional covariate vector. Without loss of generality, it can be assumed that the response is centered and the covariate is standardized, that is,

\[
\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad j = 1, \ldots, p.
\]
In order to capture a relationship between the response $y_i$ and the covariate vector $x_i$, we consider the linear regression model

$$y = X \beta + \varepsilon,$$

where $y = (y_1, \ldots, y_n)^T$ is an $n$-dimensional response vector, $X$ is an $n \times p$ design matrix, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional coefficient vector and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n$-dimensional error vector. In addition, we assume that the error $\varepsilon_i$ ($i = 1, \ldots, n$) is independently distributed as a normal distribution with a mean zero and a variance $\sigma^2$.

From the above some assumptions, the probability density function for the response $y$ is

$$f(y_i|x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-x_i^T\beta)^2}{2\sigma^2}\right), \quad i = 1, \ldots, n,$$

where $\theta = (\beta^T, \sigma^2)^T$ is the parameter vector to be estimated. This leads to a log-likelihood function given by

$$\ell(\theta) = \sum_{i=1}^n \log f(y_i|x_i; \theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n(y_i-x_i^T\beta)^2.$$

2.2 Estimation via the bridge penalty

The unknown parameter $\theta$ is estimated by the penalized maximum likelihood method, that is, by maximizing the penalized log-likelihood function given by

$$\ell_\lambda(\theta) = \ell(\theta) - n \sum_{j=1}^p p_\lambda(\beta_j),$$

where $p_\lambda(\cdot)$ is a penalty function and $\lambda$ ($> 0$) is a regularization parameter. There have been many penalty functions proposed: e.g., the $L_2$ penalty or ridge penalty $p_\lambda(\beta) = \lambda \beta^2 / 2$ (Hoerl and Kennard 1970), the $L_1$ penalty or lasso penalty $p_\lambda(\beta) = \lambda \|\beta\|$ (Tibshirani 1996), the elastic net penalty $p_\lambda(\beta) = \lambda (\alpha \beta^2 / 2 + (1-\alpha) \|\beta\|)$, where $\alpha$ ($0 \leq \alpha \leq 1$) is a tuning parameter (Zou and Hastie 2005), the SCAD penalty $p_\lambda'(\beta) = \lambda I(|\beta| \leq \lambda) + (a\lambda - |\beta|)I(|\beta| > \lambda)/(a(\lambda - 1)|\beta|)$, where $a$ ($> 2$) is a tuning parameter and $(x)_+ = \max(0, x)$ (Fan and Li 2001), and the MCP $p_\lambda'(\beta) = (a\lambda - |\beta|)_+ / a$, where $a$ ($> 0$) is a tuning parameter (Zhang 2010). Note that the ridge, lasso and elastic net penalties are convex functions, while the SCAD and MCP penalties are non-convex. Also, the lasso, elastic net, SCAD and MCP penalties can produce sparse solutions for coefficient parameters, while the ridge penalty cannot. For more penalty functions and the variants, we refer to Antoniadis et al. (2011, 2013), Fan and Lv (2008), Lv and Fan (2009), and Neykov et al. (2013).
In this paper, since we consider regression models with the bridge penalty $p_\lambda(\beta) = \lambda|\beta|^q/2$ (Frank and Friedman 1993). From this, we obtain

$$\ell_{\lambda,q}(\theta) = \ell(\theta) - \frac{n\lambda}{2} \sum_{j=1}^{p} |\beta_j|^q,$$

(1)

where $q (> 0)$ is a tuning parameter. It is obvious that the bridge penalty is the $L_1$ penalty when $q = 1$ and the $L_2$ penalty when $q = 2$. In addition, it is known that the bridge penalty yields sparse models if $0 < q \leq 1$, but non-sparse models if $q > 1$.

Bridge regression has been studied from various points of view. For example, Arman-gan (2009), Fu (1998) and Zou and Li (2008) proposed efficient algorithms for estimating the coefficient parameters. Huang et al. (2008) and Knight and Fu (2000) showed the asymptotic properties for linear regression model with the bridge penalty. Huang et al. (2009) and Park and Yoon (2011) extended the bridge penalty into the group bridge penalty, which is an extension of the group lasso penalty presented by Yuan and Lin (2006).

Since the bridge penalty is a convex function when $q \geq 1$, Eq. (1) is a concave optimization problem. Hence, in order to obtain the estimators of the coefficients, we can use the usual optimization algorithms; e.g., the shooting algorithm (Fu 1998). However, since the bridge penalty is non-convex when $0 < q < 1$, in this case, Eq. (1) becomes a non-concave optimization problem. Thus, we need to approximate the bridge penalty with a convex function. In this paper, we apply the local quadratic approximation (LQA) introduced by Fan and Li (2001) for the bridge penalty.

For the LQA, under some conditions, the penalty function can be approximated at initial values $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_p^{(0)})^T$ in the form

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

Then, Eq. (1) can be expressed as

$$\ell_{\lambda,q}(\theta) \approx \ell(\theta) - \frac{n\lambda q}{4} \sum_{j=1}^{p} |\beta_j^{(0)}|^{q-2} \beta_j^2 - \frac{n\lambda}{2} \sum_{j=1}^{p} \left(1 - \frac{q}{2}\right) |\beta_j^{(0)}|^q. \quad (2)$$

By omitting the constant terms in Eq. (2), we obtain

$$\ell^\star_{\lambda,q}(\theta) = \ell(\theta) - \frac{n\lambda q}{4} \sum_{j=1}^{p} |\beta_j^{(0)}|^{q-2} \beta_j^2.$$

This formulation leads to a concave optimization problem, since the bridge penalty is replaced with the quadratic function with respect to the coefficient parameters.
\( \beta_j \ (j = 1, \ldots, p) \). It is thus easy to estimate the parameter \( \theta \), which can be derived according to the following algorithm (Park and Yoon 2011):

Step 1 Set the values of the regularization parameter \( \lambda \) and tuning parameter \( q \), respectively.

Step 2 Initialize \( \beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_p^{(0)})^T \) and \( \sigma^{(0)2} \). In our numerical studies, we set

\[
\beta^{(0)} = (X^T X + n \gamma I_p)^{-1} X^T y, \quad \sigma^{(0)2} = 1,
\]

where \( \gamma = 10^{-5} \) and \( I_p \) is a \( p \times p \) identity matrix.

Step 3 Update the coefficient vector \( \beta \) as follows:

\[
\hat{\beta}^{(k+1)} = (X^T X + \Sigma_{\lambda,q} \hat{\beta}^{(k)})^{-1} X^T y, \quad k = 0, 1, 2, \ldots,
\]

where \( \Sigma_{\lambda,q} \hat{\beta}^{(k)} = \text{diag}(n \lambda \hat{\sigma}^{(k)2} | \hat{\beta}_1^{(k)}|^{q-2}/4, \ldots, n \lambda \hat{\sigma}^{(k)2} | \hat{\beta}_p^{(k)}|^{q-2}/4) \).

Step 4 Update the parameter \( \sigma^2 \) in the form

\[
\hat{\sigma}^{(k+1)2} = \frac{1}{n} (y - X \hat{\beta}^{(k+1)})^T (y - X \hat{\beta}^{(k+1)}).
\]

Step 5 Iterate Step 3 and Step 4 until the following condition is satisfied:

\[
|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}| < \delta,
\]

where \( \delta (> 0) \) is an arbitrary small number (e.g., \( 10^{-5} \) in our numerical examples).

From the procedure, we obtain the estimator \( \hat{\theta} = (\hat{\beta}^T, \hat{\sigma}^2)^T \), and we then derive the statistical model

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

The statistical model includes some adjusted parameters, i.e., the regularization parameter \( \lambda \) and the tuning parameter \( q \). In order to choose these values objectively, we introduce a model selection criterion in terms of Bayesian approach.

### 3 Model selection criteria

#### 3.1 Proposed criterion

Schwarz (1978) used Bayesian theory to propose the Bayesian information criterion (BIC). The BIC, however, is only applicable to models estimated by the maximum likelihood method. Konishi et al. (2004) extended it such that it could be used for evaluating statistical models estimated by the penalized maximum likelihood method.
The Bayesian approach is to select values of the regularization parameter \( \lambda \) and tuning parameter \( q \) that maximize the marginal likelihood. The marginal likelihood is calculated by integrating over the unknown parameter \( \theta \) and is defined by

\[
\text{ML} = \int \prod_{i=1}^{n} f(y_i|x_i; \theta)\pi(\theta)d\theta = \int \prod_{i=1}^{n} f(y_i|x_i; \theta)\pi(\beta|\sigma^2)\pi(\sigma^2)d\theta, \tag{3}
\]

where \( \pi(\theta) = \pi(\beta|\sigma^2)\pi(\sigma^2) \) is the prior distribution of the parameter \( \theta \). In our bridge regression models, the prior distribution \( \pi(\sigma^2) \) is assumed to be any non-informative prior distribution and the prior distribution \( \pi(\beta|\sigma^2) = \pi(\beta) \) is from Fu (1998) as follows:

\[
\pi(\beta|\lambda, q) = \prod_{j=1}^{p} \pi(\beta_j|\lambda, q) = \prod_{j=1}^{p} q 2^{-(1+1/q)}(n\lambda)^{1/q} \Gamma(1/q) \exp\left\{ -\frac{n\lambda}{2}|\beta_j|^q \right\}, \tag{3}
\]

where \( \Gamma(\cdot) \) is the Gamma function.

In general, it is difficult to evaluate Eq. (3), since it requires the calculation of a high-dimensional integral. Hence, some approximation methods are usually applied for the integral, for example, the Laplace approximation (Tierney and Kadane 1986). However, in some bridge methods, some of the components of \( \beta \) are exactly zero, and the functional in the integral (3) is not differentiable at the origin. In this case, the approximation methods cannot be applied directly.

Let \( \mathcal{A} = \{j; \hat{\beta}_j \neq 0\} \) be the active set of the parameter \( \beta \). In order to overcome the above problem, we consider the partial marginal likelihood given by

\[
\text{ML} \approx \text{PML} = \int \prod_{i=1}^{n} f(y_i|x_i; \theta_\mathcal{A})\pi(\beta|\lambda, q)d\theta_\mathcal{A}, \tag{4}
\]

where \( \theta_\mathcal{A} = (\beta_\mathcal{A}, \sigma^2)^T \). Here \( \beta_\mathcal{A} = (\beta_{k_1}, \ldots, \beta_{k_r})^T \), where we set \( \mathcal{A} = \{k_1, \ldots, k_r\} \) and \( k_1 < \cdots < k_r \). The partial marginal likelihood was originally introduced by Shimamura et al. (2007) for evaluating graphical Gaussian models estimated with the weighted lasso. The quantity, Eq. (4), is calculated by integrating over the unknown parameter \( \theta_\mathcal{A} \) included with the active set \( \mathcal{A} \). Applying the Laplace approximation for Eq. (4), we obtain

\[
\int \prod_{i=1}^{n} f(y_i|x_i; \theta_\mathcal{A})\pi(\beta|\lambda, q)d\theta_\mathcal{A} = \frac{(2\pi)^{|\mathcal{A}|+1}}{n^{|\mathcal{A}|+1}|V(\hat{\theta}_\mathcal{A})|^{1/2}} \exp\{nv(\hat{\theta}_\mathcal{A})\} \left\{ 1 + O_p(n^{-1}) \right\},
\]

where

\[
v(\theta) = \frac{1}{n} \log \left\{ \prod_{i=1}^{n} f(y_i|x_i; \theta)\pi(\beta|\lambda, q) \right\}, \quad V(\theta) = -\frac{\partial^2 v(\theta)}{\partial \theta_\mathcal{A}^T \theta_\mathcal{A}}
\]

and \( \hat{\theta}_\mathcal{A} = (\hat{\beta}_\mathcal{A}, \hat{\sigma}^2)^T \), where \( \hat{\beta}_\mathcal{A} \) is the estimator of the coefficient \( \beta_\mathcal{A} \).
By taking the logarithm of the formula calculated by the Laplace approximation, Konishi et al. (2004) presented the generalized Bayesian information criterion (GBIC) to evaluate models estimated by the penalized maximum likelihood method. Using the result of Konishi et al. (2004, p. 30), we derive a model selection criterion

\[
\text{GBIC} = n \log(2\pi) + n \log \hat{\sigma}^2 + n - (|A| + 1) \log \left( \frac{2\pi}{n} \right) \\
+ \log |J| - 2|A| \log q + 2|A| \left( 1 + \frac{1}{q} \right) \log 2 - \frac{2|A|}{q} \log(n\lambda) \\
+ 2|A| \log \Gamma \left( \frac{1}{q} \right) + n\lambda \sum_{j \in A} |\hat{\beta}_j|^q, \tag{5}
\]

where \( J \) is a \((|A| + 1) \times (|A| + 1) \) matrix given by

\[
J = \frac{1}{n\hat{\sigma}^2} \left( \frac{X^T_A X_A + n\lambda\hat{\sigma}^2 q(q - 1) K}{\hat{\sigma}^2} \right) \Lambda_n \frac{1}{n\hat{\sigma}^2} X^T_A \Lambda_n \frac{1}{n\hat{\sigma}^2} X_A n \hat{\sigma}^2 \frac{1}{n\hat{\sigma}^2}.
\]

Here \( \Lambda_n = (1, \ldots, 1)^T \) is an \( n \)-dimensional vector, \( K = \text{diag}(|\hat{\beta}_1|^q - 2/2, \ldots, |\hat{\beta}_r|^q - 2/2) \) and

\[
X_A = [x_{ik}], \quad i = 1, \ldots, n; \quad k \in A.
\]

We choose the values of the adjusted parameters, including the regularization parameter \( \lambda \) and the tuning parameter \( q \), from the minimizer of GBIC in Eq. (5).

3.2 Other criteria

This section describes other selection criteria for choosing the values of the adjusted parameters included in bridge regression models.

3.2.1 Modified AIC and modified BIC

As an approximation of the effective degrees of freedom in a model, Hastie and Tibshirani (1990) proposed using the trace of the hat matrix. In bridge regression models, the hat matrix is given by \( S = X_A (X^T_A X_A + \Sigma_{\lambda,q} (\hat{\beta}_A))^{-1} X^T_A \), where \( \Sigma_{\lambda,q} (\hat{\beta}_A) = \text{diag}(n\lambda\hat{\sigma}^2 q|\hat{\beta}_k|^q/2, \ldots, n\lambda\hat{\sigma}^2 q|\hat{\beta}_r|^q/2) \). By replacing the number of parameters in AIC (Akaike 1974) and BIC (Schwarz 1978) with the trace of the hat matrix \( S \), we obtain the modified AIC, which was originally proposed by Eilers and Marx (1996) for selecting the values of the tuning parameters in \( P \)-splines, and modified BIC, respectively, given by
\[
\text{mAIC} = -2 \sum_{i=1}^{n} \log f (y_i | x_i; \hat{\theta}) + 2 \text{tr} S,
\]
\[
\text{mBIC} = -2 \sum_{i=1}^{n} \log f (y_i | x_i; \hat{\theta}) + (\text{tr} S) \log n.
\]

There may be a problem with the theoretical justification for the use of the bias-correction term, since AIC and BIC only cover statistical models estimated by the maximum likelihood method.

3.2.2 Bias-corrected AIC

Hurvich and Tsai (1989) and Sugiura (1978) proposed an improved version of AIC in the context of linear regression models and autoregressive time series models estimated by the maximum likelihood method. Hurvich et al. (1998) proposed replacing the number of parameters in the improved version of AIC with the trace of the hat matrix, and introduced the criterion

\[
\text{AICc} = -2 \sum_{i=1}^{n} \log f (y_i | x_i; \hat{\theta}) + 2 \frac{(\text{tr} S + 1)}{n - \text{tr} S - 2}.
\]

3.2.3 Cross-validation and generalized cross-validation

Cross-validation is a technique that evaluates how the statistical model for each observation fits the remaining data. Let \(\hat{y}^{(-i)}\) be a regression response value that is estimated after removing \((y_i, x_i)\) from the observed data. The cross-validation criterion is then

\[
\text{CV} = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \hat{y}^{(-i)} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{1 - s_{ii}} \right)^2,
\]

where \(\hat{y}_1, \ldots, \hat{y}_n\) are fitted values and \(s_{ii}\) is an \(i\)th diagonal element of the hat matrix \(S\).

Craven and Wahba (1979) proposed a generalized cross-validation method in which \(s_{ii}\) in Eq. (6) is replaced with the trace of the hat matrix as follows:

\[
\text{GCV} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{1 - \text{tr} S/n} \right)^2.
\]

3.2.4 Extended information criterion

Let \(\{(y_i^{(b)}, x_i^{(b)}); i = 1, \ldots, n\} (b = 1, \ldots, B)\) be the \(b\)th bootstrap sample by resampling, and let \(\hat{\theta}^{(b)}\) be the bridge estimator based on the \(b\)th bootstrap sample. The extended information criterion proposed by Ishiguro et al. (1997) is then defined by
EIC = −2 \sum_{i=1}^{n} \log f(y_i|x_i; \hat{\theta}) + \frac{2}{B} \sum_{b=1}^{B} \left\{ \log f(y_i^{(b)}|x_i^{(b)}; \hat{\theta}^{(b)}) - \log f(y_i|x_i; \hat{\theta}) \right\}.

In our numerical experiments, \( B \) was set to 100.

4 Numerical results

In order to show the efficiency of our proposed modeling strategy, we conducted some numerical examples. Monte Carlo simulations and analysis of real data are given to illustrate the proposed bridge modeling procedure.

4.1 Simulated examples

We performed a simulation study to validate our proposed modeling procedure. The simulation has five settings, and the design matrix \( X \) was generated from a multivariate normal distribution with mean zero and variance one for Settings 1, 2, 3 and 4, and then the correlation structure was given for each setting. The response vector \( y \) was generated from the true regression model

\[ y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n), \]

where \( I_n \) is an \( n \times n \) identity matrix. Our five settings, which are very similar to those used in Tibshirani (1996) and Zou and Hastie (2005), are as follows:

- Setting 1: The training and test data consisted of 20 and 200 observations, respectively. The true parameter was \( \beta = (3, 15, 7.5, 5, 2, 0, 0, 0, 0)^T \) and \( \sigma = 3 \). The pairwise correlation between \( x_i \) and \( x_j \) was \( \text{cor}(x_i, x_j) = 0.5^{|i-j|} \). This model is sparse.
- Setting 2: This setting is the same as Setting 1 except for \( \beta_j = 10 \) (\( j = 1, \ldots, 10 \)). This model is dense.
- Setting 3: This setting is also the same as Setting 1, except that the true parameter was \( \beta = (5, 0, 0, 0, 0, 0, 0)^T \) and \( \sigma = 2 \). This model is sparse.
- Setting 4: The training and test data consisted of 100 and 400 observations, respectively. We set

\[ \beta = (0, \ldots, 0, 5, \ldots, 5, 0, \ldots, 0, 3, \ldots, 3)^T \]

and \( \sigma = 3 \). The pairwise correlation between \( x_i \) and \( x_j \) was \( \text{cor}(x_i, x_j) = 0.95^{|i-j|} \). This model is sparse.
- Setting 5: The generating procedure for the training and test data was the same as for Setting 4. The true parameter was

\[ \beta = (10, \ldots, 10, 0, \ldots, 0)^T \]

and \( \sigma = 35 \). The pairwise correlation between \( x_i \) and \( x_j \) was \( \text{cor}(x_i, x_j) = 0.95^{|i-j|} \). This model is sparse.
and $\sigma = 3$. The design matrix $X$ was generated as follows:

$$x_{ij} = Z_k + \varepsilon_j, \quad Z_k \sim N(0, 1),$$

$$j = 5k - 4, \ldots, 5k, \ k = 1, \ldots, 7 \text{ for all } i,$$

$$x_{ij} \sim N(0, 1), \quad j = 36, \ldots, 40 \text{ for all } i,$$

where the $\varepsilon_j$ were identically distributed as $N(0, 0.1^2)$ for $j = 1, \ldots, 35$. This model is sparse.

We fitted the bridge regression models to the simulated data. The regularization parameter $\lambda$ and tuning parameter $q$ in the bridge penalty were selected by GBIC, mAIC, mBIC, AICc, CV, GCV, and EIC, where we set the candidate values of $\lambda$ and $q$ as $\{10^{-0.1i+3}; i = 1, \ldots, 100\}$ and $\{0.1, 0.4, 0.7, 1.0, 1.3, 1.7, 2.0, 2.3, 2.7\}$, respectively.

We computed the mean squared error (MSE) defined by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y^*_i)^2 / n,$$

where $y^*_1, \ldots, y^*_n$ denote the test data for the response variable generated from the true model. The simulation results were obtained by averaging over 100 Monte Carlo trials, which are shown in Tables 1, 2, 3, 4 and 5. The values in parentheses indicate the standard deviations. For the 100 repetitions, we computed the means and standard deviations of the selected values for the tuning parameters $\lambda$ and $q$. We also computed the true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP), defined as

$$\text{TP} = \#\{j \in [1, \ldots, p] | \hat{\beta}_j \neq 0 \land \beta_j \neq 0\},$$

$$\text{FN} = \#\{j \in [1, \ldots, p] | \hat{\beta}_j = 0 \land \beta_j \neq 0\},$$

$$\text{TN} = \#\{j \in [1, \ldots, p] | \hat{\beta}_j = 0 \land \beta_j = 0\},$$

$$\text{FP} = \#\{j \in [1, \ldots, p] | \hat{\beta}_j \neq 0 \land \beta_j = 0\}.$$

Table 1  Comparison of the averages and standard deviations of the MSEs, two adjusted parameters ($\lambda$ and $q$), TP, FN, TN, and FP for Setting 1

|                | GBIC | mAIC | mBIC | AICc | CV  | GCV | EIC  |
|----------------|------|------|------|------|-----|-----|------|
| MSE            | 15.67| 17.13| 16.18| 15.80| 16.27| 16.02| 17.89|
| log_{10}(\lambda) | -1.137| -0.803| -0.544| -0.412| -0.599| -0.593| -0.625|
| $q$            | 0.598| 0.946| 0.805| 0.745| 0.832| 0.841| 0.890|
| TP             | 4.79 | 4.88 | 4.76 | 4.7 | 4.79 | 4.83 | 4.67 |
| FN             | 0.21 | 0.12 | 0.24 | 0.3 | 0.21 | 0.17 | 0.33 |
| TN             | 4.05 | 2.34 | 3.56 | 4.27 | 3.01 | 3.24 | 3.81 |
| FP             | 0.95 | 2.66 | 1.44 | 0.73 | 1.99 | 1.76 | 1.19 |

Figures in parentheses give the estimated standard deviations.
### Table 2: Comparison of the averages and standard deviations of the MSEs, two adjusted parameters ($\lambda$ and $q$), TP, FN, TN, and FP for Setting 2

|               | GBIC  | mAIC  | mBIC  | AICc  | CV   | GCV  | EIC   |
|---------------|-------|-------|-------|-------|------|------|-------|
| MSE           | 20.72 | 21.23 | 21.42 | 24.02 | 21.82| 21.63| 81.06 |
| $(\lambda)$   | $-4.148$ | $-1.106$ | $-1.07$ | $-0.932$ | $-0.983$ | $-0.963$ | $-2.518$ |
| $q$           | 2.700 | 1.140 | 1.176 | 1.372 | 1.289 | 1.201 | 2.560 |
| TP            | 10    | 10    | 10    | 10    | 10   | 10   | 10    |
| FN            | 0     | 0     | 0     | 0     | 0    | 0    | 0     |
| TN            | NA    | NA    | NA    | NA    | NA   | NA   | NA    |
| FP            | NA    | NA    | NA    | NA    | NA   | NA   | NA    |

Figures in parentheses give the estimated standard deviations.

### Table 3: Comparison of the averages and standard deviations of the MSEs, two adjusted parameters ($\lambda$ and $q$), TP, FN, TN, and FP for Setting 3

|               | GBIC  | mAIC  | mBIC  | AICc  | CV   | GCV  | EIC   |
|---------------|-------|-------|-------|-------|------|------|-------|
| MSE           | 4.986 | 5.836 | 5.213 | 5.068 | 5.470| 5.551| 5.500 |
| $(\lambda)$   | $-0.741$ | $-0.550$ | $-0.158$ | $-0.352$ | $-0.498$ | $-0.477$ | $-0.618$ |
| $q$           | 0.466 | 0.844 | 0.556 | 0.565 | 0.652 | 0.754 | 0.778 |
| TP            | 1     | 1     | 1     | 1     | 1    | 1    | 1     |
| FN            | 0     | 0     | 0     | 0     | 0    | 0    | 0     |
| TN            | 6.04  | 3.12  | 4.87  | 5.31  | 4.03 | 3.85 | 4.44  |
| FP            | 0.96  | 3.88  | 2.13  | 1.69  | 2.97 | 3.15 | 2.56  |

Figures in parentheses give the estimated standard deviations.

$$TN = \# \{ j \in \{1, \ldots, p\} \mid \hat{\beta}_j = 0 \land \beta_j = 0 \},$$
$$FP = \# \{ j \in \{1, \ldots, p\} \mid \hat{\beta}_j \neq 0 \land \beta_j = 0 \}.$$

The simulation results are summarized as follows.
Table 4  Comparison of the averages and standard deviations of the MSEs, two adjusted parameters ($\lambda$ and $q$), TP, FN, TN, and FP for Setting 4

|          | GBIC | mAIC | mBIC | AICc | CV   | GCV  | EIC |
|----------|------|------|------|------|------|------|-----|
| MSE      | 11.76| 11.93| 12.21| 11.92| 11.87| 11.87| 1.54|
| log$_{10}(\lambda)$ | -2.094 | -0.788 | -0.548 | -0.664 | -0.710 | -0.699 | -1.834 |
| $q$      | 0.874 | 1.030 | 1.000 | 1.003 | .009 | 1.006 | 1.890 |
| TP       | 19.4 | 19.9 | 19.95 | 19.92 | 19.9 | 19.91 | 20 |
| FN       | 0.6 | 0.1 | 0.05 | 0.08 | 0.1 | 0.09 | 0 |
| TN       | 16.98 | 12.75 | 13.94 | 14.02 | 13.81 | 13.93 | 0 |
| FP       | 3.02 | 7.25 | 6.06 | 5.98 | 6.19 | 6.07 | 20 |

Figures in parentheses give the estimated standard deviations.

Table 5  Comparison of the averages and standard deviations of the MSEs, two adjusted parameters ($\lambda$ and $q$), TP, FN, TN, and FP for Setting 5

|          | GBIC | mAIC | mBIC | AICc | CV   | GCV  | EIC |
|----------|------|------|------|------|------|------|-----|
| MSE      | 14.39| 14.62| 15.41| 14.91| 4.72 | 14.73| 10.95 |
| log$_{10}(\lambda)$ | -3.768 | -0.945 | -0.779 | -0.858 | -0.910 | -0.901 | -2.925 |
| $q$      | 1.827 | 1.009 | 1.000 | 1.000 | 1.006 | 1.006 | 2.34 |
| TP       | 34.93 | 35 | 35 | 35 | 35 | 35 | 35 |
| FN       | 0.07 | 0 | 0 | 0 | 0 | 0 | 0 |
| TN       | 1.8 | 3.59 | 4.19 | 3.94 | 3.72 | 3.74 | 0 |
| FP       | 3.2 | 1.41 | 0.81 | 1.06 | 1.28 | 1.26 | 5 |

Figures in parentheses give the estimated standard deviations.

- For Settings 1, 2, and 3, all the criteria provide appropriate values of the tuning parameter $q$: i.e., the tuning parameter $q > 1$ when the true structure of the coefficient parameter $\beta$ is dense, and $0 < q \leq 1$ when the true structure of the coefficient parameter $\beta$ is sparse. For Setting 4, GBIC and mBIC yield sparse solutions for the
coefficient vectors $\beta$, while the other criteria produce dense solutions. For Setting 5, mBIC and AICc select appropriate values of the tuning parameter $q$, whereas the other criteria (including GBIC, which is the criterion proposed in this paper), do not.

- Setting 4 assumes multicollinearity problems. In this situation, we observe that EIC, where all coefficient estimates are nonzero, will have more accurate predictions than those for which some estimates are zero. For Setting 5, we also considered the case in which some predictors were grouped. This setting also produces lower prediction errors when the number of estimated nonzero coefficients is large; i.e., GBIC and EIC have the smallest MSEs. Therefore, for Setting 5, the values of the FPs for both these methods are larger than the other methods.

- Except for EIC for Settings 4 and 5, GBIC has the smallest MSE. However, EIC appears to be unstable, since it has the worst MSE and larger standard deviations for Settings 1 and 2. We also note that EIC requires a heavy computational load.

From the above discussion, we believe that our proposed criterion, GBIC, seems to be useful in terms of computation time and minimizing the MSE.

We also compared our bridge regression modeling procedure with OLS, ridge, lasso, elastic net (ENet), SCAD, and MCP. The value of the adjusted parameter included in the ridge regression was selected by the leave-one-out cross-validation, and the adjusted parameters involved in the lasso and ENet were selected by fivefold cross-validation. For SCAD and MCP, we selected the values of the adjusted parameter $\lambda$ by using fivefold cross-validation. The value of the parameter $a$ in the penalties was set to 3.0 for all simulations. In order to evaluate the performance of each model, we computed the MSE, and drew boxplots for 100 Monte Carlo simulations. Figure 1 shows boxplots of the MSEs. In almost situations, our proposed method may perform well; i.e., it produces a relatively small median with small variance.

We can also observe some interesting points from Figure 1. For example, the performance of the ridge estimator for Setting 4, which assumes multicollinearity problems, is poor, while the ridge regression is generally robust to multicollinearity. For Setting 5, the prediction accuracy of ENet is poor, although Setting 5 is suitable for ENet. At this time, we do not have intuitively explanations and theoretical considerations for these problems. We leave as an area of future research.

4.2 Real data analysis

We applied our method to a pollution data set that was analyzed by McDonald and Schwing (1973), Liu et al. (2007) and Park and Yoon (2011). The data set consists of 60 observations and 15 covariates. The response variable is the total age-adjusted mortality rate obtained for the years 1959–1961 for 201 Standard Metropolitan Statistical Areas. It is available from the SMPracticals package in the software R.

In order to compute the prediction errors, we randomly divided the data set into 40 training and 20 test observations. Using the training data set, we constructed the bridge regression models. The values of the regularization parameter $\lambda$ and tuning parameter $q$ were chosen by using GBIC. Here, we set the candidate values of $\lambda$ and $q$
Fig. 1  Boxplots of the MSE. a the result for Setting 1, b that for Setting 2, c that for Setting 3, d that for Setting 4 and e that for Setting 5
as \( \{10^{-0.1i+3}; i = 1, \ldots, 100\} \) and \( \{0.1, 0.25, 0.4, 0.55, 0.7, 0.85, 1.0, 1.3, 1.7, 2.0\} \), respectively. The selected values of the adjusted parameters were \( \lambda = 0.007943 \) and \( q = 0.7 \).

We compared the performance of our modeling procedure with that of OLS, ridge, lasso, ENet, SCAD, and MCP. The values of the adjusted parameters in the ridge, lasso, ENet, SCAD, and MCP were selected as in Sect. 4.1. Table 6 summarizes the prediction errors (PE) from these methods. We observe that the bridge regression model outperforms the other methods from the viewpoint of minimizing the PE. Table 7 presents all of the selected variables using the entire pollution data set. Lasso and ENet choose most of the same variables as were found by McDonald and Schwing (1973) and our proposed method selected the smallest model among them. From this, we conclude that the variables 1, 8, 9, and 14 are likely to be relevant to the response variable, since they were selected by all the methods.

We also randomly split the data set five separate times and computed the PE for each of these in order to avoid any bias related to a particular division of the data. Table 8 shows the medians and standard deviations of the PE for each method. We note that the median of PE for our proposed method is the smallest, but the standard deviation of our method seems to be relatively large. We keep the problem of the instability of our proposed method as a future research.

### Table 6 Prediction errors (PE) for the pollution data set

| Method       | Bridge | OLS   | Ridge | Lasso | ENet | SCAD  | MCP   |
|--------------|--------|-------|-------|-------|------|-------|-------|
| PE           | 1,663.51 | 1,822.31 | 1,817.68 | 1,735.71 | 1,720.65 | 1,794.14 | 1,950.36 |

### Table 7 Selected variables for the pollution data set

| Method                        | Selected variables |
|-------------------------------|--------------------|
| McDonald and Schwing          | (1, 2, 6, 8, 9, 14) |
| Luo et al.                    | (1, 2, 6, 9, 14)   |
| Park and Yoon (LQA)           | (1, 2, 3, 6, 8, 9, 14) |
| Park and Yoon (LLA)           | (1, 2, 3, 6, 7, 8, 9, 14, 15) |
| Lasso                         | (1, 2, 6, 7, 8, 9, 14) |
| ENet                          | (1, 2, 6, 7, 8, 9, 14) |
| SCAD                          | (1, 2, 3, 5, 6, 8, 9, 14) |
| MCP                           | (1, 2, 3, 5, 6, 8, 9, 14) |
| Bridge with GBIC              | (1, 8, 9, 14)      |

### Table 8 Medians of PE for randomly split data sets

| Method       | Bridge | OLS   | Ridge | Lasso | ENet | SCAD  | MCP   |
|--------------|--------|-------|-------|-------|------|-------|-------|
| PE           | 1,133.40 | 1,800.40 | 1,292.27 | 1,165.49 | 1,323.73 | 1,218.70 | 1,950.36 |
|              | (488.70) | (866.12) | (557.07) | (441.72) | (472.30) | (430.47) | (413.26) |

Figures in parentheses give the estimated standard deviations.
5 Concluding remarks

In this paper, we considered the problem of evaluating linear regression models estimated by the penalized maximum likelihood method with the bridge penalty. In order to select the optimal values of the adjusted parameters including the regularization parameter in the penalized maximum likelihood function and the tuning parameter in the bridge penalty, we introduced a model selection criterion in terms of Bayesian theory. Monte Carlo simulations and analyzing a real data showed that our proposed modeling procedure performs well in various situations from the viewpoint of yielding relatively lower prediction errors than previously developed criteria and methods.

An area of future work is to apply the proposed procedure to high-dimensional data sets and to extend our procedure into the framework of generalized linear models or varying coefficient models. We note that Polson et al. (2011) proposed a full Bayesian approach for bridge regression models, and it will be an interesting topic for future research to compare this with our proposed method.

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

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Contr AC-19:716–723
Antoniadis A, Gijbels I, Nikolova M (2011) Penalized likelihood regression for generalized linear models with non-quadratic penalties. Ann Inst Stat Math 63:585–615
Antoniadis A, Gijbels I, Lambert-Lacroix S (2013) Penalized estimation in additive varying coefficient models using grouped regularization. Stat Pap. doi:10.1007/s00362-013-0522-1
Armagan A (2009) Variational bridge regression. In: Proceedings of the 12th international conference artificial intelligence statistics, vol 5, pp 17–24
Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data. Springer, New York
Craven P, Wahba G (1979) Smoothing noisy data with spline functions. Numer Math 31:377–403
Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11:89–121
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B 70:849–911
Frank JE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35:109–148
Fu WJ (1998) Penalized regression: the bridge versus the lasso. J Comput Graph Stat 7:397–416
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman & Hall, London
Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67
Huang J, Horowitz JL, Ma S (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann Stat 36:587–613
Huang J, Ma S, Xie H, Zhang C-H (2009) A group bridge approach for variable selection. Biometrika 96:339–355
Hurvich CM, Tsai C-H (1989) Regression and time series model selection in small samples. Biometrika 76:297–307
Hurvich CM, Simonoff JS, Tsai C-H (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Stat Soc Ser B 60:271–293
Ishiguro M, Sakamoto Y, Kitagawa G (1997) Bootstrapping log likelihood and EIC, an extension of AIC. Ann Inst Stat Math 49:411–434
Tuning parameter selection in bridge regression models

Knight K, Fu WJ (2000) Asymptotics for lasso-type estimators. Ann Stat 28:1356–1378
Konishi S, Ando T, Imoto S (2004) Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika 91:27–43
Liu Y, Zhang H, Park C, Ahn J (2007) Support vector machines with adaptive $L_q$ penalty. Comput Stat Data Anal 51:6380–6394
Lv J, Fan Y (2009) A unified approach to model selection and sparse recovery using regularized least squares. Ann Stat 37:3498–3528
McDonald G, Schwing R (1973) Instabilities of regression estimates relating air pollution to mortality. Technometrics 15:463–482
Neykov NM, Filzmoser P, Neytchev PN (2013) Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. Stat Pap. doi: 10.1007/s00362-013-0516-z
Park C, Yoon YJ (2011) Bridge regression: adaptivity and group selection. J Stat Plan Inference 141:3506–3519
Polson NG, Scott JG, Windle J (2011) The Bayesian bridge. arXiv:1109.2279
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Shimamura T, Imoto S, Yamaguchi R, Miyano S (2007) Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. Genome Inform 19:142–153
Sugiura N (1978) Further analysis of the data by Akaike’s information criterion and finite corrections. Comm Stat Theory Methods A7:13–26
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288
Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc 81:82–86
Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B 68:49–67
Zhang C-H (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38:894–942
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67:301–320
Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models. Ann Stat 36:1509–1533