Journal of Data Science **9**(2011), 549-564

Selection of Smoothing Parameter for One-Step Sparse Estimates with L_q Penalty

Masaru Kanba and Kanta Naito*
Shimane University

Abstract: This paper discusses the selection of the smoothing parameter necessary to implement a penalized regression using a nonconcave penalty function. The proposed method can be derived from a Bayesian viewpoint, and the resultant smoothing parameter is guaranteed to satisfy the sufficient conditions for the oracle properties of a one-step estimator. The results of simulation and application to some real data sets reveal that our proposal works efficiently, especially for discrete outputs.

Key words: One-step estimator, oracle properties, penalized likelihood, smoothing parameter, variable selection.

1. Introduction

A crucial problem in statistical modeling is variable selection, which affects the accuracy of inferences. This is especially important for the selection of explanatory variables of regression analysis of recent huge data sets. Many methods for variable selection in regression have been developed and summarized in standard textbooks, such as Chatterjee, Hadi and Price (2000). Cross validation (CV), AIC and BIC are known to be convenient techniques, not only for variable selection in regression but also for general model selection. However, when the set of variables is large, these techniques have a high computational cost, since all combinations of the variables must be calculated. The LASSO type method can overcome this difficulty by its simultaneous implementation of estimation and variable selection (see Tibshirani, 1996). Progress on using this simultaneous implementation includes Fan and Li (2001), in which nonconcave penalty functions play an important role. Fan and Li (2001) show such estimators possess the oracle properties. Zou and Li (2008) discuss using local linear and quadratic approximations to avoid the singularities of nonconcave penalty functions and show that the obtained estimator possesses the oracle properties as well.

*Corresponding author.

On the other hand, penalized regression approaches generally include a smoothing parameter. Because the estimates obviously depend on the actual value of the smoothing parameter, the selection of the smoothing parameter ultimately affects variable selection. This paper presents a simple selection method for the smoothing parameter used in nonconcave penalized likelihood models. Zou and Li (2008) utilized a conventional 5-fold CV to determine the smoothing parameter included in the model. However it is not clear that one-step estimates obtained through the model with a smoothing parameter determined by 5-fold CV preserve the oracle properties. We propose a simple and effective method for the selection of the smoothing parameter and show that the obtained smoothing parameter satisfies the sufficient conditions for the oracle properties of one-step estimates. The method is developed by using an appropriate prior of the parameter in the model, the idea of which has been discussed in the field of model selection in mixed models (Ruppert, Wand and Carroll, 2003; Yoshida, Kanba and Naito, 2010).

The paper is organized as follows. Section 2 gives a summary of the one-step estimator discussed in Zou and Li (2008). In Section 3, our method for selection of the smoothing parameter is proposed. Practical examples of regression models and functions to be optimized are shown in Section 4. Section 5 reports the results of applying the proposal to real data sets as well as simulation studies. It will be seen that our proposed method is especially efficient for models with discrete output, such as binary and Poisson models. Final discussions are contained in Section 6.

2. One-Step Estimator

Let $\{(y_i, \boldsymbol{x}_i) \mid i = 1, \cdots, n\}$ be a data set, where $\boldsymbol{x}_i$ is the p-dimensional explanatory variable and y_i is the scalar response variable. Assume that y_i depends on $\boldsymbol{x}_i$ through a linear combination $\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ and has the density $f(y_i)$. The conditional log-likelihood given $\boldsymbol{x}_i$ is assumed to be expressed as $\ell_i(\boldsymbol{\beta}, \phi)$, where ϕ is a dispersion parameter. In the setting of the generalized linear model, ϕ is the variance of the error for the linear model, and we can take $\phi \equiv 1$ for the logistic and Poisson models. For simplicity, we use $\ell(\boldsymbol{\beta})$ to stand for the log-likelihood $\ell(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}, \phi)$. Zou and Li (2008) considered the variable selection method of maximizing the penalized log-likelihood function

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|), \tag{2.1}$$

where p_{λ_n} is a nonconcave penalty function, such as SCAD and L_q, and λ_n is the smoothing parameter. In the special case of a separable penalty function

$p_{\lambda_n}(\cdot) = \lambda_n p(\cdot)$, for smoothing parameter λ_n and some function $p(\cdot)$, we can obtain estimates of $\boldsymbol{\beta}$ by maximizing

$$\ell(\boldsymbol{\beta}) - n\lambda_n \sum_{j=1}^{p} p(|\beta_j|), \tag{2.2}$$

with respect to $\boldsymbol{\beta}$. In fact, since maximizing (2.1) and (2.2) is challenging by its singularity involved in $p(\cdot)$, we aim to utilize an approximation of the penalized log-likelihood. That is, using the Taylor expansion of the log-likelihood around MLE $\boldsymbol{\beta}^{(0)}$

$$\ell(\boldsymbol{\beta}) \approx \ell\big(\boldsymbol{\beta}^{(0)}\big) + \frac{1}{2}\big(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\big)^{\mathsf{T}} \nabla^2 \ell\big(\boldsymbol{\beta}^{(0)}\big)\big(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\big),$$

and the local linear approximation (LLA) of the penalty function

$$p_{\lambda_n}(|\beta_j|) \approx p_{\lambda_n}\big(|\beta_j^{(0)}|\big) + p'_{\lambda_n}\big(|\beta_j^{(0)}|\big)\big(|\beta_j| - |\beta_j^{(0)}|\big),$$

for $\beta_j \approx \beta_j^{(0)}$, we optimize

$$\begin{aligned}\ell(\boldsymbol{\beta}) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \approx{}& \ell\big(\boldsymbol{\beta}^{(0)}\big) + \frac{1}{2}\big(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\big)^{\mathsf{T}} \nabla^2 \ell\big(\boldsymbol{\beta}^{(0)}\big)\big(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\big) \\ & - p_{\lambda_n}\big(|\beta_j^{(0)}|\big) - p'_{\lambda_n}\big(|\beta_j^{(0)}|\big)\big(|\beta_j| - |\beta_j^{(0)}|\big),\end{aligned}$$

with respect to $\boldsymbol{\beta}$. Zou and Li (2008) then defined OSE as follows:

$$\widehat{\boldsymbol{\beta}}(\text{ose}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2}\big(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\big)^{\mathsf{T}} \big[- \bigtriangledown^2 \ell\big(\boldsymbol{\beta}^{(0)}\big)\big]\big(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\big) + n\sum_{j=1}^{p} p'_{\lambda_n}\big(|\beta_j^{(0)}|\big)|\beta_j| \right\}. \tag{2.3}$$

Here, let $\boldsymbol{\beta}_0 = \big[\boldsymbol{\beta}_{10}^{\mathsf{T}}\ \boldsymbol{\beta}_{20}^{\mathsf{T}}\big]^{\mathsf{T}}$ be the true parameter and let $\boldsymbol{\beta}_{20} = \mathbf{0}$. Similarly, let $\widehat{\boldsymbol{\beta}}(\text{ose}) = \big[\widehat{\boldsymbol{\beta}}(\text{ose})_1^{\mathsf{T}}\ \widehat{\boldsymbol{\beta}}(\text{ose})_2^{\mathsf{T}}\big]^{\mathsf{T}}$, corresponding to division of true parameter $\boldsymbol{\beta}_0 = \big[\boldsymbol{\beta}_{10}^{\mathsf{T}}\ \boldsymbol{\beta}_{20}^{\mathsf{T}}\big]^{\mathsf{T}}$. Also, let $I(\boldsymbol{\beta}_0)$ be the Fisher information matrix and let

$$I(\boldsymbol{\beta}_0) = \begin{bmatrix} I_1(\boldsymbol{\beta}_{10}) & * \\ * & * \end{bmatrix},$$

where $I_1(\boldsymbol{\beta}_{10})$ is a submatrix of $I(\boldsymbol{\beta}_0)$ corresponding to $\boldsymbol{\beta}_{10}$.

In the special case of a penalty function which separates into a smoothing parameter λ_n and a function $p(\cdot)$ satisfying:

Condition 1 $p'(\cdot)$ is continuous on $(0, \infty)$,

Condition 2 there is some $s > 0$ such that $p'(t) = O(t^{-s})$ as $t \to 0+$,

Condition 3 $n^{(1+s)/2}\lambda_n \to \infty$ and $\sqrt{n}\lambda_n \to 0$.

OSE has oracle properties (Zou and Li, 2008):

(a) Sparsity: $P(\widehat{\boldsymbol{\beta}}(\text{ose})_2 = \mathbf{0}) \to 1$, as $n \to \infty$.

(b) Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\beta}}(\text{ose})_1 - \boldsymbol{\beta}_{10}) \to_D N(0, I_1(\boldsymbol{\beta}_{10})^{-1})$, as $n \to \infty$.

Zou and Li (2008) used 5-fold CV to determine λ_n. However, the λ_n determined by 5-fold CV does not necessarily always satisfy Condition 3, and the computational cost of implementing 5-fold CV itself is not cheap. A simple and reliable selection of smoothing parameter λ_n is proposed in the subsequent discussion.

3. Selection of Smoothing Parameter

In this section, when $p_{\lambda_n}(\cdot) = \lambda_n p(\cdot)$, we show that λ_n is expressed in terms of the parameter of a prior distribution for $\boldsymbol{\beta}$.

3.1 Decision of Smoothing Parameter

Assume that the prior distribution of $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_p]^{\mathsf{T}}$ belongs to the exponential family. That is, it has density

$$g(\boldsymbol{\beta}; \theta) = d(\theta) \exp\left(\frac{1}{\theta}\sum_{j=1}^{p} T(\beta_j)\right),$$

where θ is a scalar parameter, $d(\theta)$ is the normalizing constant and $T(\cdot)$ is a measurable function. Then the log of the joint density of $(y_1, \cdots, y_n)$ is

$$\log\left[\left\{\prod_{i=1}^{n} f(y_i)\right\} g(\boldsymbol{\beta}; \theta)\right] = \ell(\boldsymbol{\beta}) + \frac{1}{\theta}\sum_{j=1}^{p} T(\beta_j) + \log d(\theta),$$

where $f(y_i)$ is the density function of y_i given $\boldsymbol{\beta}$. We note that maximizing this formula with $\theta = 1/(n\lambda_n)$ and $T(\beta_j) = -p(|\beta_j|)$ is the same as maximizing the penalized likelihood (2.2). Hence, using the parameter θ, the smoothing parameter λ_n can be expressed as

$$\lambda_n = \frac{1}{n\theta}. \tag{3.1}$$

3.2 Conditions for Smoothing Parameter and Penalty Function

Here, we show that the OSE with the smoothing parameter (3.1) satisfies the oracle properties. If Conditions 1-3 hold, then the OSE has the oracle properties. Conditions 1 and 2 are concerned with the penalty function, and Condition 3 is for the smoothing parameter. Hence, we verify whether the smoothing parameter (3.1) satisfies Condition 3.

First, suppose θ is a constant. In this case $\sqrt{n}\lambda_n = (\sqrt{n}\theta)^{-1} \to 0$. However, since

$$n^{(1+s)/2}\lambda_n = n^{(1+s)/2}\frac{1}{n\theta} = n^{(s-1)/2}\frac{1}{\theta} \to 0,$$

for $0 < s < 1$, the smoothing parameter (3.1) with a constant θ does not satisfy Condition 3 for $0 < s < 1$.

To satisfy Condition 3, suppose θ is a function of n. For example, we take

$$\theta = \theta_1(n) = \frac{\log(1+n)}{\sqrt{n}}.$$

Then we can see that $\lambda_n = 1/(n\theta_1(n))$ satisfies Condition 3. Also, if we take

$$\theta = \theta_2(n, \alpha) = \frac{1}{n^\alpha}, \quad \alpha \in \mathbb{R},$$

then we have $n^{(1+s)/2}\lambda_n = n^{(s-1)/2+\alpha}$. Therefore, if we take α such that

$$\frac{1-s}{2} < \alpha < \frac{1}{2}, \tag{3.2}$$

for $0 < s < 1$, (3.1) with $\theta = n^{-\alpha}$ satisfies Condition 3. In particular, if we consider an L_q penalty $p(|\beta|) = |\beta|^q$, $0 < q < 1$, then this penalty function satisfies Conditions 1 and 2, because

$$p'(t) = qt^{q-1} = O(t^{q-1})$$

is continuous on $(0, \infty)$ and s of Condition 2 is given by $s = 1 - q$. Hence, from (3.2) α should satisfy

$$\frac{q}{2} < \alpha < \frac{1}{2}.$$

For example, we can take $\alpha = \alpha_1(q) = q$ for $q < 1/2$. On the other hand, we can take $\alpha = \alpha_2(q) = q(1 - q/2)$ for $q \in (0, 1)$. This is a point divided $q/2$ and $1/2$ into $1 - q : q$.

4. Examples

In this section, we show some models which are included in our framework.

4.1 Linear Regression Model

For linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the Hessian matrix is

$$\nabla^2 \ell(\boldsymbol{\beta}^{(0)}) = -\frac{1}{\phi}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}.$$

Hence, we can get the OSE by optimizing

$$\frac{1}{2\phi}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^{(0)}\|^2 + n\lambda_n \sum_{j=1}^{p} q|\beta_j^{(0)}|^{q-1}|\beta_j|,$$

from (2.3). We use MLE for ϕ.

4.2 Logistic Model

For a data set $\{(y_i, x_{i1}, \cdots, x_{ip}) | i = 1, \cdots, n\}$, the logistic model has joint density

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \pi(\boldsymbol{x}_i)^{y_i}\{1 - \pi(\boldsymbol{x}_i)\}^{1-y_i}, \quad i = 1, \cdots, n,$$

where

$$\pi(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})}.$$

The Hessian matrix is

$$\nabla^2 \ell(\boldsymbol{\beta}^{(0)}) = -\boldsymbol{X}^{\mathsf{T}} \operatorname{diag}\left(\frac{\exp(\boldsymbol{X}\boldsymbol{\beta}^{(0)})}{\{\boldsymbol{1} + \exp(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\}^2}\right)\boldsymbol{X} \equiv -H_1,$$

where $\boldsymbol{1} = [1, \cdots, 1]^{\mathsf{T}}$, $\operatorname{diag}(\boldsymbol{x})$ is a diagonal matrix with diagonal components $x_1, \cdots, x_n$ for $\boldsymbol{x} = [x_1, \cdots, x_n]^{\mathsf{T}}$ and

$$\frac{\exp(\boldsymbol{X}\boldsymbol{\beta}^{(0)})}{\{\boldsymbol{1} + \exp(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\}^2} = \left[\frac{\exp(\boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{\beta}^{(0)})}{\{1 + \exp(\boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{\beta}^{(0)})\}^2}, \cdots, \frac{\exp(\boldsymbol{x}_n^{\mathsf{T}}\boldsymbol{\beta}^{(0)})}{\{1 + \exp(\boldsymbol{x}_n^{\mathsf{T}}\boldsymbol{\beta}^{(0)})\}^2}\right]^{\mathsf{T}}.$$

Hence, we can get the OSE by optimizing

$$\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^{\mathsf{T}} H_1 (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + n\lambda_n \sum_{j=1}^{p} q|\beta_j^{(0)}|^{q-1}|\beta_j|,$$

from (2.3).

4.3 Poisson Model

For a data set $\{(y_i, x_{i1}, \cdots, x_{ip})|i = 1, \cdots, n\}$, the Poisson model has joint density

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\mu(\boldsymbol{x}_i)^{y_i} \exp(-\mu(\boldsymbol{x}_i))}{y_i!}, \quad i = 1, \cdots, n,$$

where $\pi(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})$. Since the Hessian matrix is

$$\nabla^2 \ell(\boldsymbol{\beta}^{(0)}) = -\boldsymbol{X}^{\mathsf{T}} \operatorname{diag}\left(\exp(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\right) \boldsymbol{X} \equiv -H_2,$$

we can get the OSE by optimizing

$$\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^{\mathsf{T}} H_2 (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + n\lambda_n \sum_{j=1}^{p} q|\beta_j^{(0)}|^{q-1}|\beta_j|,$$

from (2.3), where $\exp(\boldsymbol{x}) = [\exp(x_1), \cdots, \exp(x_n)]^{\mathsf{T}}$ for $\boldsymbol{x} = [x_1, \cdots, x_n]^{\mathsf{T}}$.

5. Implementation

In this section, we report the results of applying the proposed OSE to real data sets. Simulation results are addressed as well. Also, we compare the models obtained by AIC, BIC, LASSO, OSE(CV) and our proposed OSE, where OSE(CV) refers to an OSE with λ_n, minimized with respect to 5-fold CV with $q = 0.01$.

5.1 Real Data Sets

Here, we performed a comparison of estimation methods using Ozone, Diabetes, Parkinson, Glass and Wine data sets. We used the 5-fold CV value to compare the prediction ability (PA) of each method. The variables selected by each method are also compared. All results are tabulated in Table 1.

In Table 1, EST1 stands for OSE with $\lambda_n = (n\theta_1(n))^{-1}$ and $q = 0.01$ and EST2 stands for OSE with either $\lambda_n = (n\theta_2(n, \alpha_1(q)))^{-1}$ or $\lambda_n = (n\theta_2(n, \alpha_2(q)))^{-1}$ minimizing the 5-fold CV value on an appropriate grid of q.

For the Ozone data set, we applied a linear regression model with quadratic interaction

$$\begin{aligned} y_i = {} & \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \beta_6 x_{i3}^2 \\ & + \beta_7 x_{i1} x_{i2} + \beta_8 x_{i1} x_{i2} + \beta_9 x_{i2} x_{i3} + \varepsilon_i, \end{aligned}$$

Table 1: Data sets and results of comparison

No	1	2	3	4	5
Data[1]	Ozone	Diabetes	Parkinson	Glass	Wine
Sample size	111	442	195	214	178
Number of predictor	9	10	22	9	13
Model	Linear	Linear	Logistic	Poisson	Poisson
PA					
EST1	0.104	0.211	1.39	1.86	1.01
EST2	0.101	0.193	1.39	1.43	0.36
OSE(CV)	0.102	0.081	1.39	1.92	0.44
LASSO	0.096	0.231	1.39	1.57	0.39
AIC	0.043	0.051	1.14	1.88	0.34
BIC	0.043	0.052	1.10	1.46	0.33
Selected variables					
EST1	2,5-9	2-4,6,7,9,10	-	1,2,5	1,4
EST2	5,7-9,	2-7,9,10	-	1,2,5	1,4,7,8,13
OSE(CV)	5-9	3,9	1-7, 10, 12-15,17-22	1	1-13
LASSO	5,7,9	3,9	1-7, 9,10, 12-22	1	1-13
AIC	2,3,6-8	2-7,	1,4,6,7,11,13,17,20,22	2,3,4,7	1,4,7,13
BIC	2,3,6-8	2-7,	13,16,19,21	2,3,4,7	4,7

[1]Reference: No.1 from Hastie, Tibshirani and Friedman (2009); No.2 from Efron, Hastie, Johnstone and Tibshirani (2004); No.3,4,5 from the data warehouse of University of California Irvine (http://archive.ics.uci.edu/ml)

where $(x_{i1}, x_{i2}, x_{i3}) = (\text{solar radiation}_i, \text{daily maximum temperature}_i, \text{wind speed}_i)$ for $i = 1, \cdots, 111$. We can observe from Table 1 that PAs of EST1, EST2, OSE(CV) and LASSO are larger than those of AIC and BIC. The 7th variable solar radiation $\times$ daily maximum temperature, should be the most important in this model since all methods selected it. All methods here except AIC and BIC selected the 5th variable, squared daily maximum temperature. This inclusion of the 5th variable might be the reason their PAs are worse.

For the Diabetes data set, the PAs of AIC, BIC and OSE(CV) are smaller than those of the other methods. EST1, EST2 and LASSO performed similarly, with respect to PA. It is interesting that there is a significant difference between the PA of OSE(CV) and that of LASSO although they selected the same two variables. This might be the effect of the choice of q: for LASSO, $q = 1$, while for OSE(CV), $q = 0.01$.

For the Parkinson data set, all methods have similar PAs. The proposed OSE selected no variables although other methods selected several variables. This means that EST1 and EST2 both suggest "*Random Guess*" in this model. This

might be an extreme solution, but it motivates us to look at the set of explanatory variables carefully.

For the Glass data set, the PA of EST2 is the smallest and EST2 selected only 3 variables from 9 explanatory variables. This means that EST2 achieves a precise prediction using only 3 variables.

For the Wine data set, there is little difference among PAs, except for EST1. BIC selected just two variables and has the best PA. EST1 and EST2 selected only a few variables although OSE(CV) and LASSO selected all variables. From the sets of selected variables and the PAs, one can observe that the behavior of EST2 is similar to that of AIC in the case of the Poisson model.

For the Parkinson, Glass and Wine data sets, there is little difference among the PAs of the methods. On the other hand, the PAs of AIC and BIC are better than other methods for the Ozone and Diabetes data sets using a linear model. As expected, AIC and BIC performed well with linear models. Note that the sets of variables selected by EST1 and EST2 were similar to those selected by AIC and BIC. For logistic and Poisson regression models with discrete output, there was little difference between the PAs of the methods, but EST1 and EST2 selected only a relatively small number of variables. Hence it can be claimed that the proposed OSE is efficient for models with discrete output.

The proposed method does not require a grid search for the smoothing parameter and it can select variables at the same time as it estimates $\boldsymbol{\beta}$. Hence, the proposed method can be executed at little computational cost compared to AIC and BIC, which are computationally costly because they compute for every combination of the variables. This is an attractive point for the use of the proposed OSE.

5.2 Simulation

Here, to see the behavior of proposed OSE, we carried out simulations. First we will explain the structures of the simulations.

We utilized the following regression models:

$$\text{Model 1 \& Model 2}: \ y_i = \mu(\boldsymbol{x}_i) + \varepsilon_i, \ \mu(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}_j \ (j = 1, 2),$$

$$\boldsymbol{\beta}_1 = [3, -5, \underbrace{0, \cdots, 0}_{10-\text{fold}}]^{\mathsf{T}} \in \mathbb{R}^{12}, \ \boldsymbol{\beta}_2 = [3, 1.5, 0, 0, 2, \underbrace{0, \cdots, 0}_{7-\text{fold}}]^{\mathsf{T}} \in \mathbb{R}^{12},$$

$$\text{Model 3 \& Model 4}: \ y_i \sim \mathcal{B}er\left(\mu(\boldsymbol{x}_i)\right), \ \mu(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}_j)}{1 + \exp(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}_j)} \ (j = 3, 4),$$

$$\boldsymbol{\beta}_3 = [0.7, -1.2, \underbrace{0, \cdots, 0}_{10-\text{fold}}]^{\mathsf{T}} \in \mathbb{R}^{12}, \ \boldsymbol{\beta}_4 = [1, -0.4, 0, 0, -0.6, \underbrace{0, \cdots, 0}_{7-\text{fold}}]^{\mathsf{T}} \in \mathbb{R}^{12},$$

$$\text{Model 5 \& Model 6 : } y_i \sim \mathcal{P}[\mu(\boldsymbol{x}_i)],\ \mu(\boldsymbol{x}) = \exp(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}_j)\ (j = 5, 6),$$
$$\boldsymbol{\beta}_5 = [0.7, -0.5, \underbrace{0, \cdots, 0}_{10-\text{fold}}]^{\mathsf{T}} \in \mathbb{R}^{12},\ \boldsymbol{\beta}_6 = [1.2, 0.6, 0, 0, 0.8, \underbrace{0, \cdots, 0}_{7-\text{fold}}]^{\mathsf{T}} \in \mathbb{R}^{12},$$

for $i = 1, \cdots, n$. Covariates and errors are generated according to $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \Sigma)$ and $\varepsilon_1, \cdots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, where the (i, j)-component of Σ is $0.5^{|i-j|}$.

Here we will use the model error

$$\text{ME}[\hat{\mu}] = E[(\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}))^2],$$

for the constructed prediction $\hat{\mu}(\cdot)$ based on the variable selection, where the expectation is taken with respect to a new observation $\boldsymbol{x}$. The simulation algorithm is as follows:

1. Generate $\{(y_i, \boldsymbol{x}_i) \mid i = 1, \cdots, n\}$ from the model.
2. For MLE $\boldsymbol{\beta}^{(0)}$, calculate $\hat{\mu}_{\text{MLE}}(\cdot)$.
3. For the estimates of $\boldsymbol{\beta}$ obtained by each method, calculate $\hat{\mu}(\cdot)$.
4. Evaluate the prediction ability (PA) by minimizing the 5-fold CV value and calculate the ratio of model error:
$$\text{RME}[\hat{\mu}] = \frac{\text{ME}[\hat{\mu}]}{\text{ME}[\hat{\mu}_{\text{MLE}}]}.$$
5. Calculate Correct, the number of nonzero coefficients correctly estimated to be nonzero and Incorrect, the number of zero coefficients incorrectly estimated to be nonzero.
6. Repeat steps 1 through 5 a thousand times, and then calculate the MRME (the median of RME), PA (the average of the 5-fold CV), C (the average of Correct) and IC (the average of Incorrect).

The above algorithm was implemented for $n = 50, 100, 200$. Methods for variable selection compared throughout these simulation studies are LASSO, AIC, BIC, the proposed OSE with $\theta_1(n)$ for $q = 0.01$, $\theta_2(n, \alpha_1(q))$ for $q = 0.01, 0.25, 0.49$, $\theta_2(n, \alpha_2(q))$ for $q = 0.01, 0.5, 0.99$, and OSE(CV) for $q = 0.01$ and λ_n minimizing the 5-fold CV value.

For Model 1, C = 2 means the method can select two important nonzero variables and IC = 0 means the method can delete all unnecessary variables. Tables 2 and 3 show MRME, PA, C and IC for the linear regression model. Similarly, Table 4 shows these values for the logistic regression model, and Tables

5 and 6 show these values for the Poisson regression model. We compare these values for the different methods.

In Tables 2 and 3, both the MRME and PA of BIC are the smallest of all the methods for both Models 1 and 2. The MRMEs of the proposed OSEs are larger than those of other methods and their PAs are also somewhat large. We can observe from the C column of Tables 2 and 3 that all methods could select the important nonzero variables for both Models 1 and 2, because C for all methods was near to 2 for Model 1 and equal to 3 for Model 2. The values of IC in Tables 2 and 3 show that the proposed OSEs could frequently delete the unnecessary variables, since the ICs of the proposed OSEs are near to 0, but the ICs of the other methods are slightly larger than those of the proposed OSEs. The overall impression from considering Tables 2 and 3 is that the proposed OSE could select important variables as well as other methods, and could delete unnecessary variables more often than other methods. Hence it might be claimed that the proposed OSE is accurate in the sense of avoiding false positives, see Bühlmann and Meier (2008). Though its ability to predict and to fit data is not satisfactory, the proposed OSEs performed well at variable selection.

Table 2: Linear regression model ($n = 50$)

Model	Model 1				Model 2			
Method	**MRME**	**PA**	**C**	**IC**	**MRME**	**PA**	**C**	**IC**
$\theta_1(n)$	0.746	1.312	2	1.76	0.552	1.237	3	1.69
$\theta_2(n, \alpha_1(0.25))$	0.597	1.523	2	2.39	0.848	1.339	3	0.93
$\theta_2(n, \alpha_1(0.49))$	1.348	3.507	2	0.88	4.278	2.390	3	0.17
$\theta_2(n, \alpha_2(0.5))$	6.740	2.145	1.996	0.11	1.873	1.660	3	0.40
$\theta_2(n, \alpha_2(0.99))$	8.543	3.680	1.984	0.09	4.617	2.487	3	0.16
OSE(CV)	0.544	1.254	2	4.05	0.554	1.250	3	2.99
LASSO	0.534	1.248	2	4.66	0.555	1.243	3	3.55
AIC	0.619	1.289	2	2.04	0.670	1.054	3	1.83
BIC	0.297	1.175	2	0.68	0.396	1.028	3	0.59

As shown in Table 4, there is little difference in MRME among the methods, for both Model 3 and Model 4. There was also little difference in PA. Hence, the fit and PA of methods were nearly the same. According to C and IC in Table 4, the proposed OSE with $\theta_2(n, \alpha_1(0.49))$ and $\theta_2(n, \alpha_2(0.99))$ selected no variables. However according to C of Model 3, the proposed OSE with $\theta_2(n, \alpha_1(0.01))$ and $\theta_2(n, \alpha_2(0.01))$ could select two important variables, similar to AIC and BIC, but BIC would appear to attain the best balance of C and IC values. For Model 4, only the C of LASSO is near 3, so these methods could not select three important variables. The values of C of the proposed OSE with $\theta_2(n, \alpha_1(0.01))$ and

Table 3: Linear regression model ($n = 200$)

Model	Model 1				Model 2			
Method	**MRME**	**PA**	**C**	**IC**	**MRME**	**PA**	**C**	**IC**
$\theta_1(n)$	2.30	1.15	2	0.25	1.33	1.094	3	0.513
$\theta_2(n, \alpha_1(0.25))$	4.43	1.30	2	0.05	2.39	1.169	3	0.203
$\theta_2(n, \alpha_1(0.49))$	55.22	4.28	2	0.00	27.73	2.740	3	0.001
$\theta_2(n, \alpha_2(0.5))$	16.41	2.03	2	0.00	8.43	1.554	3	0.018
$\theta_2(n, \alpha_2(0.99))$	61.44	4.63	1.999	0.00	30.71	2.923	3	0.001
OSE(CV)	0.55	1.05	2	3.63	0.57	1.048	3	2.934
LASSO	0.84	1.04	2	8.62	0.56	1.046	3	3.483
AIC	0.62	1.05	2	1.64	0.67	1.054	3	1.541
BIC	0.18	1.02	2	0.20	0.28	1.028	3	0.197

$\theta_2(n, \alpha_2(0.01))$, OSE(CV), AIC and BIC are all similar, BIC having the smallest IC. For Model 4, the values of C of the proposed OSE are similar to those of OSE(CV), and the values of IC for the proposed OSE are smaller than the value of IC for OSE(CV). Hence we can conclude that the proposed OSE perform better than OSE(CV) in the models simulated.

Table 4: Logistic regression model ($n = 200$)

Model	Model 3				Model 4			
Method	**MRME**	**PA**	**C**	**IC**	**MRME**	**PA**	**C**	**IC**
$\theta_1(n)$	0.9991	1.334	0.95	0.01	1.0027	1.352	1.32	0.03
$\theta_2(n, \alpha_1(0.01))$	0.9988	1.246	1.96	1.65	1.0007	1.246	2.49	1.75
$\theta_2(n, \alpha_1(0.49))$	0.9998	1.386	0.00	0.00	1.0027	1.386	0.00	0.00
$\theta_2(n, \alpha_2(0.01))$	0.9988	1.246	1.96	1.65	1.0007	1.246	2.49	1.76
$\theta_2(n, \alpha_2(0.99))$	0.9998	1.386	0.00	0.00	1.0027	1.386	0.00	0.00
OSE(CV)	0.9996	1.277	1.96	3.43	1.0004	1.236	2.67	3.31
LASSO	1.0000	1.280	2.00	10.00	1.0000	1.278	3.00	9.00
AIC	1.0001	1.245	2.00	1.62	0.9997	1.253	2.70	1.60
BIC	0.9993	1.217	1.94	0.20	0.9989	1.240	2.33	0.28

According to Tables 5 and 6, the MRMEs of the proposed OSE are comparatively small for both Model 5 and Model 6. The MRME of the proposed OSE with $\theta_2(n, \alpha_1(0.25))$ is the smallest for Model 5, and the MRME of LASSO is the smallest for Model 6. Also, there is little difference in the PAs of the methods for Model 5. For Model 6, the PA of the proposed OSE and OSE(CV) are relatively small. The PAs of LASSO, AIC and BIC are larger than the proposed OSE, which means that the proposed OSE has a good PA. For Model 5, the Cs of the proposed OSE, OSE(CV), AIC and BIC are near to 2 and the ICs of the proposed

OSE and BIC are near to 0. The proposed OSE with $\theta_2(n, \alpha_1(0.25))$ seems to be the best for both $n = 100$ and $n = 200$ in that it has the best balance of C and IC. We can see that AIC and BIC could not select the important variables well, because for Model 6, while the ICs of AIC and BIC are also near to 0, so are the Cs. For the Poisson regression model, the proposed OSE has a good fit and PA, and it can select the set of important variables well.

Table 5: Poisson regression model ($n = 100$)

Model	Model 5				Model 6			
Method	**MRME**	**PA**	**C**	**IC**	**MRME**	**PA**	**C**	**IC**
$\theta_1(n)$	0.704	1.206	2	0.643	0.863	1.33	3	0.34
$\theta_2(n, \alpha_1(0.25))$	0.631	0.625	1.99	0.146	0.784	1.56	2.99	0.32
$\theta_2(n, \alpha_1(0.49))$	0.459	0.717	0.90	0	0.546	3.75	2.79	0.03
$\theta_2(n, \alpha_2(0.5))$	0.521	0.661	1.79	0.004	0.672	2.33	2.97	0.08
$\theta_2(n, \alpha_2(0.99))$	0.457	0.722	0.81	0	0.535	3.91	2.76	0.03
OSE(CV)	0.771	0.665	1.927	3.55	0.941	1.19	3	3.18
LASSO	0.488	1.639	1.559	0.033	0.31	8.51	1.57	0.01
AIC	0.977	0.622	1.967	1.448	1.03	4.72	1.11	0.06
BIC	0.966	0.613	1.967	0.295	1.03	4.71	1.11	0.01

Table 6: Poisson regression model ($n = 200$)

Model	Model 5				Model 6			
Method	**MRME**	**PA**	**C**	**IC**	**MRME**	**PA**	**C**	**IC**
$\theta_1(n)$	0.732	1.149	2	0.147	0.86	1.12	3	0.337
$\theta_2(n, \alpha_1(0.25))$	0.685	0.620	2	0.037	0.82	1.24	3	0.096
$\theta_2(n, \alpha_1(0.49))$	0.466	0.724	0.956	0.000	0.57	3.13	2.936	0.005
$\theta_2(n, \alpha_2(0.5))$	0.544	0.653	1.94	0.000	0.70	1.80	2.999	0.009
$\theta_2(n, \alpha_2(0.99))$	0.463	0.730	0.822	0.000	0.56	3.30	2.907	0.003
OSE(CV)	0.837	0.722	1.995	3.213	0.94	1.04	3	3.186
LASSO	0.486	1.729	1.525	0.001	0.27	8.76	1.486	0.000
AIC	0.988	0.618	1.921	1.380	1.04	4.88	1.002	0.002
BIC	0.981	0.613	1.921	0.199	1.04	4.88	1.002	0.000

Results of the simulations look similar to those from applying the methods to real data sets. For linear regression models, the PAs of AIC and BIC are better than other methods. For logistic and Poisson regression models, the PAs of all methods are almost the same, but the proposed OSE performed variable selection better than other methods.

6. Discussion

We consider the penalized log-likelihood with an L_q penalty as the MAP using the prior of $\boldsymbol{\beta}$ in Section 3.1. We consider how the form of prior of $\boldsymbol{\beta}$ affects the selection of the smoothing parameter? To determine this, we focus on the variance matrix of the prior of $\boldsymbol{\beta}$, calculated by

$$V[\boldsymbol{\beta}] = C_q(\theta)\mathbf{I}_p, \quad C_q(\theta) = \frac{\theta^{2/q}\Gamma(3/q)}{\Gamma(1/q)},$$

where $\Gamma(\cdot)$ is the gamma function and $\mathbf{I}_p$ is the $p \times p$ identity matrix. If the variance component $C_q(\theta)$ is small, then this means that the prior of $\boldsymbol{\beta}$ is tightly distributed around the mean $\mathbf{0}$; hence $\boldsymbol{\beta} \approx \mathbf{0}$, and the probability that many variables will be deleted from the model can be expected to be high. In this sense, the proposed method may tend to be strict with respect to selecting variables.

In the case of using either $\theta = \theta_1(n)$ or $\theta = \theta_2(n, \alpha)$, it is clear that the variance of $\boldsymbol{\beta}$ decreases for increasing n. The smoothing parameter $\lambda_n = (n\theta)^{-1}$ also decreases as n grows for both $\theta = \theta_1(n)$ and $\theta = \theta_2(n, \alpha)$.

We discuss here the effect of the variance of the prior and the smoothing parameter on variable selection. Roughly speaking, the sum of values of C and IC in Tables 2-6 is the number of variables selected by each method. In the simulation results of the Poisson regression for Model 5 tabulated in Tables 5 and 6, the number of variables (the sum of C and IC) selected by OSE with either $\theta_1(200)(q = 0.01)$ or $\theta_2(200, \alpha_1(0.25))$ is fewer than those with $n = 100$. On the other hand, the number of variables selected by OSE with either $\theta_2(200, \alpha_1(0.49))$, $\theta_2(200, \alpha_2(0.5))$ or $\theta_2(200, \alpha_2(0.99))$ is slightly larger than those with $n = 100$.

We can understand these different results by considering the variance component of the prior distribution. For $\theta_1(n)(q = 0.01)$ and $\theta_2(n, \alpha_1(0.25))$, the variance components decrease drastically as n grows from $n = 100$ to $n = 200$; this change of variance has a more serious affect than the decrease of the smoothing parameter, hence the number of variables selected is decreasing.

In the cases of $\theta_2(n, \alpha_1(0.49))$, $\theta_2(n, \alpha_2(0.5))$ and $\theta_2(n, \alpha_2(0.99))$, the change in variance is not large, hence the decrease in the smoothing parameter from $n = 100$ to $n = 200$ only affects variable selection: the number of variables selected increases.

In this sense, variable selection is affected not only by the value of the smoothing parameter but also by the variance of prior distribution of $\boldsymbol{\beta}$.

In the case of discrete output (binary and Poisson), variable selection by ordinary methods, such as AIC, BIC and LASSO, were not satisfactory according to our simulation. This is because discrete output can generally be explained by a smaller number of explanatory variables than continuous output. AIC type methods are aimed at prediction, which means that these methods tend to include many variables in the model. A set of variables selected by AIC might include

redundant variables for prediction, which explains the results seen in Tables 5 and 6.

Bühlmann and Meier (2008) and Zhang (2008) proposed new methods of variable selection corresponding to high-dimension low sample size (HDLSS) problems where $p \gg n$, which are called the MSA-LASSO method and the MC+ method, respectively. MSA-LASSO can be regarded as an adaptive way to search for the best initial estimator in the one-step paradigm. These authors pointed out that the number of false positives is perhaps more important than reducing prediction errors in high-dimensional data analysis. As a method for having a low probability of false positives, the proposed OSE in this paper would be worth applying to the HDLSS setting.

Acknowledgements

The research of the second author is supported by KAKENHI 20500257.

References

Bühlmann, P. and Meier, L. (2008). Discussion on the paper by Zou and Li. *Annals of Statistics* **36**, 1534-1541.

Chatterjee, S., Hadi, A. S. and Price, B. (2000). *Regression Analysis by Example*, 3rd ed. Wiley, New York.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407-499.

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Springer, New York.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

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

Yoshida, T., Kanba, M. and Naito, K. (2010). A computationally efficient model selection in the generalized linear mixed model. *Computational Statistics* **25**, 463-484.

Zhang, C. (2008). Discussion on the paper by Zou and Li. *Annals of Statistics* **36**, 1553-1560.

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

Received June 16, 2010; accepted December 20, 2010.

Masaru Kanba
Graduate school of Science and Engineering
Shimane University
Matsue 690-8504, Japan
masaru-kanba@hotmail.co.jp

Kanta Naito
Department of Mathematics
Shimane University
Matsue 690-8504, Japan
naito@riko.shimane-u.ac.jp