NOVEL RESAMPLING METHODS FOR TUNING PARAMETER SELECTION IN ROBUST SPARSE REGRESSION MODELING

Park, Heewon
Graduate School of Science and Engineering, Chuo University

https://doi.org/10.5109/1495411

出版情報：Bulletin of informatics and cybernetics. 44, pp.49-64, 2012-12. Research Association of Statistical Sciences
バージョン：
権利関係：

NOVEL RESAMPLING METHODS FOR TUNING PARAMETER SELECTION IN ROBUST SPARSE REGRESSION MODELING

by

Heewon Park

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

NOVEL RESAMPLING METHODS FOR TUNING PARAMETER SELECTION IN ROBUST SPARSE REGRESSION MODELING

By

Heewon Park*

Abstract

The robust lasso-type regularized regression is a useful tool for simultaneous estimation and variable selection even in the presence of outliers. Crucial issues in the robust modeling procedure include the selection of regularization parameters and also a tuning constant in outlier detection. Although the performance of the robust sparse regression strongly depends on the proper choice of these tuning parameters, little attention was paid for this issue, particularly in the presence of outliers. We consider the problem of choosing the tuning parameters and propose an information-theoretic criterion based on the bootstrap. Although the bootstrap information criterion has several advantages on its flexibility and weak assumptions, a bootstrap sample may contain more outliers compared with those included in the original sample, since the bootstrap sample is drawn randomly. This implies that the bootstrap information criterion may be obtained from the highly contaminated bootstrap sample by outliers, so the resulting criterion may produce biased results. In order to overcome the drawback, we propose a robust bootstrap information criterion via winsorizing technique (Srivastava et al., 2010) in line with the efficient bootstrap information criterion (Konishi and Kitagawa, 1996) for choosing an optimal set of tuning parameters. Monte Carlo simulations and real data analysis are conducted to investigate the effectiveness of the proposed method. We observe that the proposed robust efficient bootstrap information criterion produces reliable model estimates and performs well in the presence of outliers.

Key Words and Phrases: Efficient bootstrap information criterion, Robust sparse regression modeling, Tuning parameter selector, Winsorization technique.

1. Introduction

Sparse regression models are constructed by optimizing the penalized least squares loss function with various L_1-type of norms (see, e.g., Hastie et al., 2009; Kawano et al., 2010). By replacing the least squares loss function with the robust loss function, robust lasso-type regularization can effectively perform simultaneous parameter estimation and variable selection even in the presence of outliers. Although the robust sparse regression modeling heavily depends on an appropriate choice of the regularization parameters and tuning constant in outlier detection, little attention was paid for this issue. In fact, existing studies on the M-lasso and M-adaptive lasso (Zhang et al., 2009; Lambert-Lacroix and Zwald, 2011) selected only the regularization parameters controlling the

* Graduate School of Science and Engineering, Chuo University, 1–13–27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. heewonn@gug.math.chuo-u.ac.jp

model complexity without regard for selection of the tuning constant (i.e., under the fixed tuning constant). The issue, however, should be considered as a selection of an optimal set of the regularization parameters and tuning constant at once. Also, we should consider the effect of outliers in the model evaluation, since the outliers may have considerable effect on the evaluation process, and thus they yield distort model selection results. Ronchetti et al. (1997) proposed a robust version of the cross-validation using a robust loss function. Tharmaratnam (2010) also proposed a robust version of the AIC (Akaike, 1973) using the weighted Kullback-Leibler distance.

We consider the use of the bootstrap information criterion for choosing the tuning parameters. Although the bootstrap technique is a practical method, it has a demerit in the presence of outliers that a bootstrap sample may contain more outliers than those in the original sample, since the bootstrap sample is drawn randomly. This implies that the bootstrap information criterion may be obtained from the contaminated bootstrap sample by outliers, and thus the resulting criterion may produce biased results. In order to overcome the drawback, we propose a robust bootstrap information criterion via winsorization technique (Srivastava et al., 2010) in line with the efficient bootstrap information criterion (Konishi and Kitagawa, 1996) for choosing an optimal set of the regularization parameters and a tuning constant robustly. We observe that the variance due to the ordinal bootstrap resampling can be reduced significantly, and thus the number of bootstrap replications may be greatly reduced. Furthermore, we also observe that the proposed robust bootstrap information criterion produces stable results even in the presence of outliers by using the winsorization technique. In short, using the proposed method, we can perform effective and robust sparse regression modeling. Although we focus on the proposed method as a robust tuning parameter selector, it is a useful tool for robust evaluation of various modeling techniques.

The rest of this paper is organized as follows. We present the methodology of the robust sparse regression modeling in Section 2. In Section 3, we propose the robust efficient bootstrap information criterion via winsorization technique for the robust sparse regression modeling. We briefly introduce the robust lasso-type regularization methods in Section 4. Monte Carlo simulations are conducted to investigate the performance of proposed technique in Section 5. The real world example is shown in Section 6. Some concluding remarks are given in Section 7.

2. Methodology

Suppose we have n independent observations $\{(y_i, \boldsymbol{x}_i); i = 1, ..., n\}$, where y_i are random response variables and $\boldsymbol{x}_i$ are p-dimensional vectors of the predictor variables.

Consider the linear regression model,

$$y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, ..., n, \tag{1}$$

where β_0 is an intercept, $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and ε_i are the random errors which are assumed to be independently, identically distributed with mean 0 and variance σ^2. We assume that the y_i are centered and x_{ij} are standardized by their mean and standard deviation: $\sum_i^n y_i/n = 0$, $\sum_i^n x_{ij}/n = 0$ and $\sum_i^n x_{ij}^2/n = 1$.

To construct an outlier-resistant regression model for (1), we estimate the regression

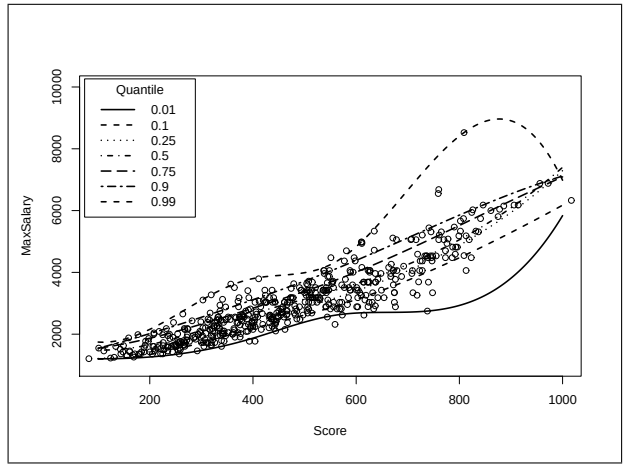

Figure 1: Quantile regression with various quantiles

coefficient vector by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \Big[\sum_{i=1}^{n} \rho(r_i; k) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \Big], \tag{2}$$

where $\rho(r_i; k)$ is a robust loss function (e.g., least absolute deviation, least trimmed squares loss function, M-function, and etc.) with tuning constant k for outlier detection, $r_i = y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\sum_{j=1}^{p} p_\lambda(|\beta_j|)$ is a lasso-type penalty (e.g., the penalty of lasso, adaptive lasso, elastic net, and etc.) with regularization parameter λ (see Section 4). The robust sparse regression modeling procedure performs variable selection and estimation simultaneously by an appropriate choice of the regularization parameter λ, and outliers are controlled by the tuning constant k in $\rho(r_i; k)$.

The ordinary lasso-type approaches consist of the least squares loss function with the L_1-type penalty term. Although the lasso-type regularization effectively performs simultaneous variable selection and estimation by imposing the L_1-type penalty, its performance takes a sudden turn for the worst in the presence of outliers, since it is based on the least squares loss function. To overcome the problem, numerous studies have attempted to achieve the robustness of lasso-type regularization by replacing the least squares loss function with the robust loss function (see Section 4).

It is well known that the choice of the regularization parameter is a vital matter, since it controls the sparsity of a constructed model. Furthermore, in the robust regression modeling, the tuning constant plays a key role for outlier-resistant modeling procedure by controlling the outliers. Figure 1 shows the quantile regression for salary data (Weisberg, 2005), which is one of the robust regression modeling, with various quantiles. The x-axis is the score of job difficulty for job classes as a predictor variable and the y-axis is the maximum salary as a response variable. In the quantile regression, quantile can be seen as a tuning constant for controlling outliers. As shown in Figure 1, the regression fitting line is significantly changed as increasing the quantile. This implies that choosing the tuning constant is crucial in the robust regression modeling.

Although not only the regularization parameters, but also the tuning constant

plays a key role for robust sparse regression modeling, existing studies on the robust L_1-type regularization, such as the M-lasso and M-adaptive lasso, were conducted by the choice of only the regularization parameters without considering a tuning constant (i.e., under the fixed tuning constant $k = 1.34$ (Lambert-Lacroix and Zwald, 2011)). This issue, however, should be considered as a selection of optimal set of regularization parameters and tuning constant, since the selection of the regularization parameters is also influenced by outliers.

3. Novel resampling method for tuning parameter selection

We propose the robust method for choosing an optimal set of regularization parameters and tuning constant via the bootstrap information criterion which showed effective performance for robust sparse regression modeling in Park et al. (2012). Although the bootstrap information criterion is effective for choosing the tuning parameters, we should point out the demerit of the bootstrap technique in the presence of outliers that the bootstrap sample may be more contaminated by outliers than original sample, because of random drawing procedure of the bootstrap resampling. This implies that the bootstrap information criterion may be obtained from the highly contaminated bootstrap sample by outliers, and thus the resulting criterion may produce biased results. In order to overcome the drawback, we propose a robust bootstrap information criterion via winsorization technique in line with the efficient bootstrap information criterion for robust sparse regression modeling.

We first introduce the efficient bootstrap information criterion (Konishi and Kitagawa, 1996) in the next section.

3.1. Efficient bootstrap information criterion

Consider the case in which a model is given in the form of a probability distribution $\{f(y|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$ having p-dimensional parameters. We assume that the data $\mathbf{y}_n = \{y_1, ..., y_n\}$ are generated from the true distribution function $G(y)$. Our task is to evaluate the performance of the estimated model $f(z|\hat{\boldsymbol{\theta}})$ when it is used to predict the independent future data $Z = z$ generated from the unknown true distribution $G(z)$.

The general form of an information criterion is constructed as follows;

$$IC(\mathbf{y}_n; \hat{G}) = -2 \sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{\text{estimator for } b(G)\}, \tag{3}$$

where $b(G)$ is a bias of the log-likelihood as an estimator of the expected log-likelihood depending on the unknown probability distribution G. That is, the bias $b(G)$ is given by

$$b(G) = E_{G(\mathbf{y}_n)} \left[\log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}(\mathbf{y}_n)) - n E_{G(z)} \left[\log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{y}_n)) \right] \right], \tag{4}$$

where $\log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}(\mathbf{y}_n)) = \sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}(\mathbf{y}_n))$ and the expectation $E_{G(\mathbf{y}_n)}$ is taken with respect to the joint distribution, $\prod_{i=1}^{n} G(y_i) = G(\mathbf{y}_n)$ of the sample $\mathbf{y}_n$ (Konishi and Kitagawa, 2008). Note that $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y}_n)$ depends on the sample $\mathbf{y}_n$.

In order to improve the model evaluation accuracy, numerous studies on estimation of the bias in (4) have been conducted. Konishi and Kitagawa (1996) showed that the

difference between the log-likelihood of the model and n times the expected log-likelihood

$$D(\mathbf{y}_n; G) = \log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}) - n \int \log f(z|\hat{\boldsymbol{\theta}}) dG(z), \tag{5}$$

can be decomposed into three terms

$$D(\mathbf{y}_n; G) = D_1(\mathbf{y}_n; G) + D_2(\mathbf{y}_n; G) + D_3(\mathbf{y}_n; G), \tag{6}$$

where

$$\begin{aligned} D_1(\mathbf{y}_n; G) &= \log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}) - \log f(\mathbf{y}_n|\boldsymbol{\theta}), \\ D_2(\mathbf{y}_n; G) &= \log f(\mathbf{y}_n|\boldsymbol{\theta}) - n \int \log f(z|\boldsymbol{\theta}) dG(z), \\ D_3(\mathbf{y}_n; G) &= n \int \log f(z|\boldsymbol{\theta}) dG(z) - n \int \log f(z|\hat{\boldsymbol{\theta}}) dG(z). \end{aligned} \tag{7}$$

By taking the expectation term by term on (6), the second term is

$$\begin{aligned} E_G[D_2(\mathbf{y}_n; G)] &= E_G \left[\log f(\mathbf{y}_n|\boldsymbol{\theta}) - n \int \log f(z|\boldsymbol{\theta}) dG(z) \right] \\ &= \sum_{i=1}^{n} E_G \left[\log f(y_i|\boldsymbol{\theta}) - n E_G[\log f(Z|\boldsymbol{\theta})] \right] \\ &= 0. \end{aligned} \tag{8}$$

Thus, the expectation of (5) can be expressed without $D_2(\mathbf{y}_n; G)$ term as follows;

$$E_G[D(\mathbf{y}_n; G)] = E_G[D_1(\mathbf{y}_n; G) + D_3(\mathbf{y}_n; G)]. \tag{9}$$

In the bootstrap information criteria, the true distribution $G(y)$ is replaced with an empirical distribution function $\hat{G}(y)$. With this replacement, the random variable and estimator in (4) are substituted as follows;

$$\begin{gathered} G(y) \longrightarrow \hat{G}(y), \\ y_i \sim G(y) \longrightarrow y_i^* \sim \hat{G}(y), \\ Z \sim G(z) \longrightarrow Z^* \sim \hat{G}(z), \\ E_{G(\mathbf{y})}, E_{G(z)} \longrightarrow E_{\hat{G}(\mathbf{y}^*)}, E_{\hat{G}(z^*)}, \\ \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y}) \longrightarrow \hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}(\mathbf{y}^*). \end{gathered}$$

Therefore, the bootstrap bias estimate of (4) is given by

$$b^*(\hat{G}) = E_{\hat{G}(\mathbf{y}^*)} \left[\sum_{i=1}^{n} \log f(y_i^*|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^*)) - n E_{\hat{G}(z^*)} \left[\log f(Z^*|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^*)) \right] \right]. \tag{10}$$

Let us draw B sets of bootstrap samples of size n from the observed data and write the b^{th} bootstrap sample as $\mathbf{y}_n^*(b) = \{y_1^*(b), ..., y_n^*(b)\}$. In the bootstrap estimate, (9) is replaced by

$$E_{\hat{G}}[D(\mathbf{y}_n^*; \hat{G})] = E_{\hat{G}}[D_1(\mathbf{y}_n^*; \hat{G}) + D_3(\mathbf{y}_n^*; \hat{G})]. \tag{11}$$

Therefore, we can use

$$b_B(\hat{G}) = \frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*(b);\hat{G}) + D_3(\mathbf{y}_n^*(b);\hat{G})\} \tag{12}$$

as a bootstrap bias estimate.

Conditional on the observed data, Konishi and Kitagawa (1996) showed that the orders of asymptotic conditional variances of two bootstrap estimates are

$$\begin{aligned}
&\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\{D(\mathbf{y}_n^*;\hat{G})\}\right] = \frac{1}{B}O(n), \\
&\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*;\hat{G}) + D_3(\mathbf{y}_n^*;\hat{G})\}\right] = \frac{1}{B}O(1).
\end{aligned} \tag{13}$$

This implies that the variance due to the bootstrap resampling can be reduced significantly, and thus we can expect it to produce an efficient modeling.

Consequently, the efficient bootstrap information criterion based on variance reduction method is defined as follows;

$$\begin{aligned}
\mathrm{EIC}_{\mathrm{eff}} &= -2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{b_B(\hat{G})\} \\
&= -2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}}) + 2\left[\frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*(b);\hat{G}) + D_3(\mathbf{y}_n^*(b);\hat{G})\}\right].
\end{aligned} \tag{14}$$

For details on the theoretical justification for bootstrap variance reduction technique, see Konishi and Kitagawa (1996; 2008).

3.2. Novel resampling method for tuning parameter selection

The bootstrap information criterion is a useful tool for evaluating models constructed by the robust lasso-type regularization with L_1-type penalty, since it is a flexible technique and can be applied to complex problems employing very weak assumptions. Although the bootstrap information criterion has several advantages, it has a considerable demerit that the bootstrap sample may include more outliers than those in the original sample, since the bootstrap sample is drawn randomly.

Table 1 shows the seriousness of the problem that bootstrap sample contains more outliers than those in the original sample over 100,000 simulated datasets. As shown in Table 1, overall, more than 35% of bootstrap samples contain more outliers than those in the original samples. This implies that the resulting criterion from the bootstrap sample may produce biased results in the presence of outliers, and hence the bootstrap information criterion does not perform well as a tuning parameter selector. To overcome the demerit, we use a winsorization technique to the efficient bootstrap information criterion.

A winsorization is a statistical technique that aims to reduce the effect of outliers in the sample (Yale and Forsythe, 1976). First, we introduce a winsorization bootstrap method (Singh, 1998; Srivastava et al., 2010). Suppose that the order statistics of the

Table 1: Percentage that bootstrap sample is more contaminated than original sample

	Proportion (%) of outliers in the original sample			
n	1%	5%	10%	15%
100	0.26	0.39	0.42	0.43
500	0.38	0.45	0.46	0.47
1000	0.42	0.46	0.48	0.48

original data be denoted by $y_{[1]}, y_{[2]}, ..., y_{[n]}$. For some δ between 0 and 1/2, δ-winsorized sample for $\{y_i\}$ is given by

$$\begin{aligned} y_i^* &= y_{[l+1]}, && \text{if } y_i \le y_{[l]}, \\ &= y_{[n-l]}, && \text{if } y_i \ge y_{[n-l+1]}, \\ &= y_i, && \text{otherwise}, \end{aligned} \tag{15}$$

where $\delta = l/n$, $0 \le \delta \le 1/2$ represents the winsorizing proportion. The winsorized bootstrap sample $\{y_i^{**}\}$ are randomly drawn from the δ-winsorized sample $\{y_i^*\}$. This implies that the winsorized bootstrap sample may not be affected by outliers which are greater than $y_{[l]}$ or smaller than $y_{[n-l+1]}$. Thus, we can reduce the effect of outliers in the bootstrap technique.

For outlier-resistant model evaluation, we propose a robust efficient bootstrap information criterion via winsorized bootstrap sample. By using the winsorized bootstrap sample, the bootstrap bias estimate of (4) is given by

$$b^{**}(\hat{G}) = E_{\hat{G}(\mathbf{y}^{**})}\left[\sum_{i=1}^{n} \log f(y_i^{**}|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^{**})) - nE_{\hat{G}(z^{**})}\left[\log f(Z^{**}|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^{**}))\right]\right]. \tag{16}$$

Let us draw B sets of winsorized bootstrap samples of size n and write the b^{th} winsorized bootstrap sample as $\mathbf{y}_n^{**}(b) = \{y_1^{**}(b), ..., y_n^{**}(b)\}$. In the winsorized bootstrap estimate, (9) and (11) are replaced by

$$E_{\hat{G}}[D(\mathbf{y}_n^{**}; \hat{G})] = E_{\hat{G}}[D_1(\mathbf{y}_n^{**}; \hat{G}) + D_3(\mathbf{y}_n^{**}; \hat{G})]. \tag{17}$$

Therefore, the bootstrap bias estimate of (4) is substituted by

$$b_B^w(\hat{G}) = \frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^{**}(b); \hat{G}) + D_3(\mathbf{y}_n^{**}(b); \hat{G})\}. \tag{18}$$

Consequently, the proposed robust efficient bootstrap information criterion is given by

$$\begin{aligned} \text{R.EIC}_{\text{eff}} &= -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{b_B^w(\hat{G})\} \\ &= -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{B}\sum_{i=1}^{B}\{D_1(\mathbf{y}_n^{**}(b); \hat{G}) + D_3(\mathbf{y}_n^{**}(b); \hat{G})\}. \end{aligned} \tag{19}$$

By using the R.EIC$_{\text{eff}}$, the variance of the bootstrap estimates caused by simulation can be reduced extensively and then the number of bootstrap replications may be greatly reduced. Furthermore, we can perform accurate and stable model evaluation even in the presence of outliers.

Using the R.EIC$_{\text{eff}}$, we choose an optimal set of the regularization parameters and a tuning constant in the robust lasso-type regularization based on the grid search. Under the assumption that ε_i in (1) are the random errors from $N(0, \sigma^2)$, the linear regression model is expressed as

$$f(y_i|\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\{y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}\}^2}{2\sigma^2}\right]. \tag{20}$$

To calculate the R.EIC$_{\text{eff}}$ for the robust sparse regression model, the winsorized bootstrap samples denoted as $\mathbf{y}_n^{**} = \{y_1^{**}, ..., y_n^{**}\}$ are generated using a x-fixing method. In the x-fixing method, predictor variables $\boldsymbol{x}_n$ are considered as not random variables and $\mathbf{y}_n^{**} = \hat{\beta}_0 + \boldsymbol{x}_n^T\hat{\boldsymbol{\beta}} + \mathbf{e}_n^{**}$, where $\mathbf{e}_n^{**}$ are randomly drawn from winsorized sample $\mathbf{e}_n^*$ of $\mathbf{e}_n(= \mathbf{y}_n - \hat{\beta}_0 - \boldsymbol{x}_n^T\hat{\boldsymbol{\beta}})$,

$$\begin{aligned} e_i^* &= e_{[l+1]}, && \text{if } \ e_i \le e_{[l]}, \\ &= e_{[n-l]}, && \text{if } \ e_i \ge e_{[n-l+1]}, \\ &= e_i, && \text{otherwise.} \end{aligned} \tag{21}$$

Afterwards, we calculate the R.EIC$_{\text{eff}}$ based on $\hat{\boldsymbol{\beta}}$ estimated by the robust lasso-type approaches at each set of the regularization parameters and a tuning constant. Finally, we perform model selection and estimation by choosing an optimal set of these tuning parameters that minimizes R.EIC$_{\text{eff}}$.

4. Examples: Robust lasso-type regularization methods

Several robust lasso-type regularization methods were proposed by replacing the least squares loss function with the robust loss function for robust sparse regression modeling.

- Least trimmed squares lasso (Mateos and Giannakis, 2012):

$$\hat{\boldsymbol{\beta}}^{\text{LTS-lasso}} = \arg\min_{\boldsymbol{\beta}} \Big[\sum_{i=1}^{s} r_{[i]}^2 + \lambda\sum_{j=1}^{p}|\beta_j|\Big], \tag{22}$$

  where s is a tuning constant, $r_{[i]}^2$ is the i-th order statistic of squared residuals.

- M-lasso (Zhang et al., 2009):

$$\hat{\boldsymbol{\beta}}^{\text{M-lasso}} = \arg\min_{\boldsymbol{\beta}} \Big[\sum_{i=1}^{n} \rho(r_i) + \lambda\sum_{j=1}^{p}|\beta_j|\Big], \tag{23}$$

  where $\rho(\cdot)$ is the M-estimation function,

-Huber function:

$$\rho(r) = \begin{cases} r^2/2, & \text{if } |r| < k, \\ k(|r| - k/2), & \text{if } |r| \geq k. \end{cases}$$

-Tukey function:

$$\rho(r) = \begin{cases} (k^2/6)(1 - [1 - (r/k)^2]^3), & \text{if } |r| < k, \\ k^2/6, & \text{if } |r| \geq k, \end{cases}$$

where k is a tuning constant.

To improve the performance of robust sparse regression modeling, we also consider the combination of Huber M-function and smoothly clipped absolute deviation (SCAD) penalty having good properties: unbiasedness, sparsity and continuity (Fan and Li, 2001),

$$\hat{\boldsymbol{\beta}}^{\text{M-SCAD}} = \arg\min_{\boldsymbol{\beta}} \Big[\sum_{i=1}^{n} \rho(r_i) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \Big], \tag{24}$$

where

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda, \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right), & \text{if } \lambda < |\beta_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta_j| > a\lambda. \end{cases}$$

Figure 2 (a) and (b) show the estimator of the lasso and SCAD, respectively. The x-axis is the least square estimator, and the y-axis of Figure 2 (a) and (b) is the estimator of lasso and SCAD, respectively. As shown in Figure 2, the SCAD produces unbiased estimation results in large $|\beta|$ unlike to the lasso, and thus we can expect better performance for modeling by using the M-SCAD than by using the M-lasso.

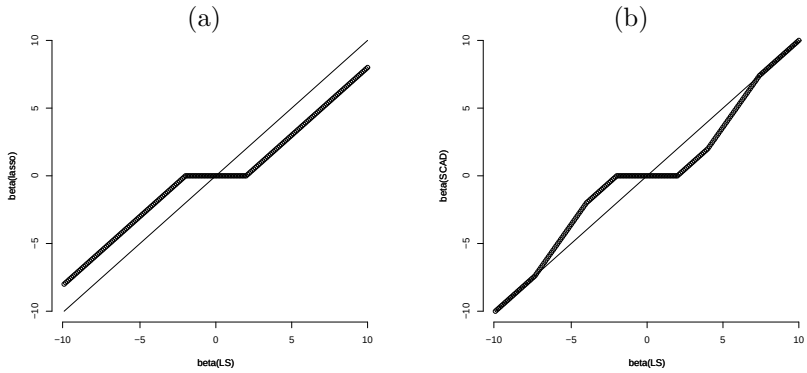

Figure 2: Thresholding function with $\lambda = 2$ for (a) the lasso and (b) the SCAD (a=3.7)

5. Simulation studies

We examine, through Monte Carlo experiments, the effectiveness of the proposed modeling strategy as a robust tuning parameter selector by comparing with the ordinary bootstrap information criterion and cross-validation. In the winsorization technique, choosing the winsorizing proportion δ is crucial in practice. The simplest way to choose the δ is to specify them in advance (Chen et al., 2001). The δ was determined adaptively from the data in literatures (Welsh, 1987; Dodge and Jurecková, 1997). Chen et al. (2001) mentioned that this issue is largely a philosophical question as to which approaches individual users prefer. Srivastara et al. (2010) showed that the winsorization technique with $\delta \approx$ "proportion of outliers in the original sample" outperforms in the bootstrap regression. Therefore, we use the winsorizing proportion $\delta =$ "proportion of outliers in the original sample" in simulation studies.

First, we show the stability of the proposed technique. Figure 3 shows bar plots of the standard deviation of bootstrap estimates D, D_1+D_3, D_1, D_2 and D_3, for the sample size $n = 100$. From the Figure 3, it can be clearly seen that the bootstrap estimates D, $D_1 + D_3$, D_1, D_2, and D_3 in the proposed robust bootstrap information criterion (black bar plots) show smaller standard deviation compared with those in the existing one (white bar plots). It implies that the proposed robust bootstrap information criterion is more effective and stable against outliers than the existing one, and thus we can expect efficient and robust sparse regression modeling by using the proposed method.

We evaluate the proposed method as a tuning parameter selector for robust sparse regression modeling. We simulated 50 datasets consisting of n observations from the

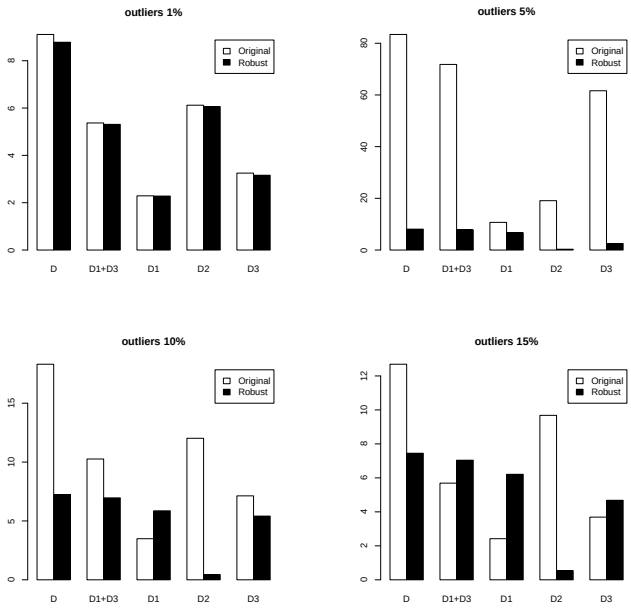

Figure 3: Standard deviation of bootstrap estimate of the D, $D_1 + D_3$, D_1, D_2 and D_3

model

$$y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, ..., n, \tag{25}$$

where ε_i are standard normal. In the numerical study, we assume that $\beta_0 = 0$. The correlation between x_l and x_m is $\rho^{|l-m|}$ with ρ=0.5. Simulations are conducted in the presence of 5%, 10%, and 15% outliers for $\varepsilon_i \sim N(30, 3)$. To evaluate the proposed method, we choose the regularization parameters and tuning constant by the robust efficient bootstrap information criterion and the ordinary bootstrap information criterion. We also compare results by the 10-fold cross validation. For model estimation by robust L_1-type regularization, we used an iterative reweighted least square (IRLS) algorithm (Zhang et al., 2009).

Two Monte Carlo simulations are conducted for the robust sparse regression modeling by the LTS-lasso, M-lasso and M-SCAD based on the Huber function,

Simulation 1: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $n = 80$.

Simulation 2: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0, 3, 1.5, 0, 0, 2, 0, 0, 0, 3, 1.5, 0, 0)^T$ and $n = 50$.

Table 2, Table 3 and Table 4 compare the simulation results for variable selection and forecasting accuracy of robust sparse regression modeling by the LTS-lasso, M-lasso and M-SCAD, respectively, where the bold numbers indicate the best performance among the three criteria. The values of "T.N" indicate the average proportion of five and twelve true zero coefficients that were correctly set to zero, called true negative, and the values of "F.N" indicate average proportion of the three and eight truly non-zero coefficients that incorrectly set to zero, called false negative. The forecasting root mean square errors (RMSE) over 50 simulated datasets are summarized in last column

- LTS-lasso

Table 2: Simulation results: LTS-lasso

Outlier	Method	Simulation 1			Simulation 2		
		T.N	F.N	RMSE	T.N	F.N	RMSE
5%	CV	0.036	0.000	1.82	0.117	0.005	**2.01**
	Eff.Boot.IC	0.036	0.000	1.99	0.345	0.017	2.61
	Robust.Eff.Boot.IC	**0.068**	0.000	**1.67**	**0.488**	0.025	2.21
10%	CV	0.016	0.000	3.17	0.033	0.007	4.70
	Eff.Boot.IC	**0.020**	0.000	3.33	0.242	0.037	4.56
	Robust.Eff.Boot.IC	0.016	0.000	**2.83**	**0.273**	0.015	**3.74**
15%	CV	0.004	0.000	4.56	0.042	0.010	6.53
	Eff.Boot.IC	0.008	0.000	5.00	0.225	0.066	5.92
	Robust.Eff.Boot.IC	**0.020**	0.000	**4.07**	**0.303**	0.037	**5.00**

- M-lasso

Table 3: Simulation results: M-lasso

Outlier	Method	Simulation 1			Simulation 2		
		T.N	F.N	RMSE	T.N	F.N	RMSE
5%	CV	0.072	0.000	1.75	0.148	0.005	1.94
	Eff.Boot.IC	0.072	0.000	1.75	**0.353**	0.000	2.05
	Robust.Eff.Boot.IC	**0.136**	0.000	1.75	0.290	0.010	**1.76**
10%	CV	0.020	0.000	3.54	0.098	0.030	**4.53**
	Eff.Boot.IC	0.032	0.000	3.53	0.088	0.020	4.87
	Robust.Eff.Boot.IC	**0.056**	0.013	**3.48**	**0.110**	0.012	4.71
15%	CV	0.012	0.000	5.36	0.092	0.040	6.50
	Eff.Boot.IC	0.056	0.020	5.08	0.093	0.050	6.68
	Robust.Eff.Boot.IC	**0.068**	0.007	**5.02**	**0.130**	0.037	**6.10**

- M-SCAD

Table 4: Simulation results: M-SCAD

Outlier	Method	Simulation 1			Simulation 2		
		T.N	F.N	RMSE	T.N	F.N	RMSE
5%	CV	0.076	0.000	1.75	0.075	0.005	1.92
	Eff.Boot.IC	0.084	0.000	1.73	0.202	0.000	1.88
	Robust.Eff.Boot.IC	**0.152**	0.000	**1.72**	**0.245**	0.000	**1.83**
10%	CV	0.036	0.000	**3.45**	**0.160**	0.000	**4.58**
	Eff.Boot.IC	0.036	0.000	3.58	0.072	0.007	4.81
	Robust.Eff.Boot.IC	**0.064**	0.020	3.58	0.090	0.007	4.60
15%	CV	0.008	0.070	5.39	0.082	0.037	6.86
	Eff.Boot.IC	0.052	0.040	5.29	0.095	0.035	6.71
	Robust.Eff.Boot.IC	**0.084**	0.070	**5.27**	**0.097**	0.025	**6.47**

in each Table. From the columns "T.N" in all Tables, it can be seen that the proposed robust efficient bootstrap information criterion is a useful tool for choosing the tuning

parameters in the presence of outliers in the viewpoint of the "sparsity" (i.e., some coefficients in the estimated model are exactly zero (Tibshirani et al., 2005)), which is a crucial property of the lasso-type approaches. It can be also seen that proposed method is superior to the existing ones for the forecasting accuracy in overall (see columns "RMSE"). In short, the proposed technique is an efficient tool for roust sparse regression modeling via LTS-lasso, M-lasso and M-SCAD in the viewpoint of the "sparsity" and forecasting accuracy.

6. Real-world example

We illustrate the proposed procedure through the analysis of a crime dataset (Agresti and Finlay, 1997) to evaluate the practicality. The dataset consists of p=9 variables for $n = 51$ observations as follows,

- crime: violent crimes per 100,000 people
- sid: state id
- state: state name
- murder: murders per 1,000,000 people
- pctmetro: the percent of the population living in metropolitan areas
- pctwhite: the percent of the population that is white
- pcths: percent of population with a high school education or above
- poverty: percent of population living under poverty line
- single: percent of population that are single parents

The variable "crime" is considered as a response variable and the variables "murder, pctmetro, pctwhite, pcths, poverty" and "single" are considered as predictor variables (i.e., p=6). The robust sparse regression modeling is conducted via the LTS-lasso, M-lasso and M-SCAD based on the Huber function. We also compare results of the ordinary lasso. The regularization parameters and tuning constant are selected by the proposed robust efficient bootstrap information criterion. The regression model is estimated using observations 1 to 40, and then we calculate forecasting RMSE based on observations 41

Table 5: Robust sparse regression modeling for Crime data

	Cross validation	Eff.Boot.IC	Robust.eff.Boot.IC
lasso	130.56	130.56	**130.55**
LTS-lasso	128.01	128.03	**127.99**
M-lasso	130.73	129.02	**127.45**
M-SCAD	130.77	130.20	**127.89**

to 51 in Table 5, where the bold numbers indicate the best performance among the three criteria. It can be first seen from Table 5 that the robust lasso-type approaches outperform the ordinary lasso in the presence of outliers. We can also see that the proposed robust efficient bootstrap information criterion is outstanding for the forecasting accuracy with all examples of the L_1-type regularization, lasso, LTS-lasso, M-lasso and M-SCAD. This implies that the proposed technique is also useful for analysis of contaminated real world data.

7. Concluding remarks

We have proposed the robust bootstrap information criterion via winsorizing technique in line with the efficient bootstrap information criterion for the robust sparse regression modeling. In order to robustly select the tuning parameters, we use the winsorization bootstrap technique. We observed through Monte Carlo experiments that the proposed robust efficient bootstrap information criterion is more stable against outliers than existing one. In addition, our simulation studies also showed the efficiency of the proposed technique for choosing an optimal set of regularization parameters and a tuning constant in the viewpoint of the sparsity and forecasting accuracy. The results of the real-world example also showed the superiority of the proposed method. Future work remains to be done towards considering various robust bootstrap techniques (e.g., trimming method).

Acknowledgement

The author would like to thank Professors Sadanori Konishi and Fumitake Sakaori for their valuable advice, and the anonymous reviewer for helpful comments and suggestions that improved the quality of the paper considerably.

References

Agresti, A. and Finlay, B. (1997). *Statistical Methods for the Social Sciences*. Prentice Hall, New Jersey.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *in Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademiai Kiado, 267-281.

Chen, L.-A., Welsh, A.H. and Chan, W. (2001). Estimators for the linear regression model based on winsorized observations. *Statistica Sinica* **11**(1), 147-172.

Dodge, Y. and Jurecková, J. (1997). Adaptive choice of trimming proportion in trimmed least-squares estimation. *Statistics & Probability Letters* **33**(2), 167-176.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association* **96**(456), 1348-1360.

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

Kawano, S., Hirose, K., Taeishi, S. and Konishi, S. (2010). Recent development in regression modeling and L_1 type regularization. *Journal of the Japan Statistical Society* **39**(2), 211-242.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**(4), 875-890.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Hubers criterion and adaptive lasso penalty. *Electronic Journal of Statistics* **5**, 1015-1053.

Mateos, G. and Giannakis, G.B. (2012). Robust nonparametric regression via sparsity control with application to load curve data cleansing. *IEEE Transactions on Signal Processing* **60**(612), 1571-1584.

Park, H., Sakaori, F. and Konishi, S. (2012). Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria. *in preperation*.

Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of American Statistical Association* **60**(439), 1017-1023.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics* **26**(5), 1719-1732.

Srivastava, D.K., Pan, J.M., Sarkar, I. and Mudholkar, G.S. (2010). Robust winsorized regression using bootstrap approach. *Communications in Statistics - Simulation and Computation* **39**(1), 45-67.

Tharmaratnam, K. and Claeskens, G. (2010). A comparison of robust versions of the AIC based on M, S and MM-estimators. *Technical report*, Katholieke Universiteit Leuven.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, **67**, 91-108.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley, New York.

Welsh, A.H. (1987). The trimmed mean in the linear model. *The Annals of Statistics* **15**(1), 20-36.

Yale, C. and Forsythe, A. B. (1976). Winsorized regression. *Technometrics* **18**(3), 291-300.

Zhang, Z. G., Chan, S. C., Zhou, Y. and Hu, Y. (2009). Robust linear estimation using M-estimation and weighted L_1 regularization: model selection and recursive implementation. *Proceedings of the 2009 International Symposium on Circuits and Systems*, 1193-1196.

Received July 14, 2012
Revised October 23, 2012