Communications for Statistical Applications and Methods
2017, Vol. 24, No. 6, 673–683

https://doi.org/10.29220/CSAM.2017.24.6.673
Print ISSN 2287-7843 / Online ISSN 2383-4757

Penalized rank regression estimator with the smoothly clipped absolute deviation function

Jong-Tae Park[a], Kang-Mo Jung[1,b]

[a]Department of Data Information, Pyeongtaek University, Korea;
[b]Department of Statistics and Computer Science, Kunsan National University, Korea

Abstract

The least absolute shrinkage and selection operator (LASSO) has been a popular regression estimator with simultaneous variable selection. However, LASSO does not have the oracle property and its robust version is needed in the case of heavy-tailed errors or serious outliers. We propose a robust penalized regression estimator which provide a simultaneous variable selection and estimator. It is based on the rank regression and the non-convex penalty function, the smoothly clipped absolute deviation (SCAD) function which has the oracle property. The proposed method combines the robustness of the rank regression and the oracle property of the SCAD penalty. We develop an efficient algorithm to compute the proposed estimator that includes a SCAD estimate based on the local linear approximation and the tuning parameter of the penalty function. Our estimate can be obtained by the least absolute deviation method. We used an optimal tuning parameter based on the Bayesian information criterion and the cross validation method. Numerical simulation shows that the proposed estimator is robust and effective to analyze contaminated data.

Keywords: local linear approximation, rank regression, robust methods, smoothly clipped absolute deviation, variable selection

1. Introduction

Classical variable selection methods such as stepwise selection needed two sequential steps for estimation and variable selection. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO), which can estimate regression parameters with simultaneous variable selection. However, LASSO can be biased for coefficients with large absolute values. This problem was addressed by Fan and Li (2001) who proposed a non-convex penalty function to make up for the deficiencies of LASSO, the smoothly clipped absolute deviation (SCAD) penalty. The SCAD penalty has an oracle property: the asymptotic bias, variance and distribution of the resulting estimate are the same and as if the correct subset were known in advance. Variable selection methods for regression models are covered by Jung and Park (2015) and Lee (2015).

The least squares estimate (LSE) is sensitive to even a single outlier. One alternative is the least absolute deviation (LAD) estimate. Jung (2011, 2013) proposed robust estimators and outlier detection methods in regression models and support vector machine. There are several robust versions of LASSO. For example, Wang *et al.* (2007) used the LAD loss function with the L_1 penalty function (LAD-LASSO), Alfons *et al.* (2013) proposed a least trimmed squares regression, and Chen *et al.*

[1] Corresponding author: Department of Statistics and Computer Science, Kunsan National University, 558 Daehakro, Kunsan 54150, Korea. E-mail: kmjung@kunsan.ac.kr

Published 30 November 2017 / journal homepage: http://csam.or.kr

(2010) considered a penalized M-estimator based on Huber's loss function. In this paper, we consider the SCAD penalty function and the error function proposed by Jaeckel (1972), which is a loss function from the point of non-parametric view. Kim *et al.* (2015) proposed a robust LASSO method based on ranks (RANK-LASSO).

In this paper we develop a robust rank regression method with a pairwise difference of residuals and the SCAD penalty function (RANK-SCAD) to improve the performance of a variable selection for RANK-LASSO. The proposed method combines the robustness of the rank regression and the oracle property of the SCAD penalty. Like RANK-LASSO, the proposed estimator RANK-SCAD has an outlier-resistant loss function. We conjecture that RANK-SCAD has the oracle property in the same manner as SCAD. RANK-SCAD may be a robust model selector having the oracle property.

This paper is organized as follows. In Section 2 we propose RANK-SCAD and describe its statistical properties. The sum of pairwise difference of residuals and the SCAD penalty functions are briefly reviewed. Section 3 gives an efficient algorithm to implement RANK-SCAD. We use a local linear approximation (LLA) algorithm which can be a linear system to minimize a non-differentiable and non-convex objective function. Zou and Li (2008) proposed LLA for nonconcave penalized likelihood models. Section 4 provides the simulation results and it shows that the performance of RANK-SCAD is superior to LSE, LASSO, LAD-LASSO, and RANK-LASSO in the view of selecting the correct model and the prediction error. Furthermore, it shows that RANK-SCAD is more robust than LASSO and RANK-LASSO. Section 5 provides some discussion and concluding remarks.

2. Rank-smoothly clipped absolute deviation estimator

We consider the classic linear regression model

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where α is the unknown intercept term, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the $n \times d$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ is the unknown vector of regression coefficients, d is the number of predictors and n is the number of observations. There are two cumbersome points. In case $d > n$, the problem becomes underdetermined and LSE cannot be computed. Another problem is that LSE can be distorted when the error has a heavy tailed or the data has even single outlier. See Rousseeuw and Leroy (1987) for details.

The underdetermined problem can be treated by penalized regression models. Hoerl and Kennard (1970) proposed ridge regression with squares errors and the L_2 norm penalty function. Tibshirani (1996) considered a penalty function of the L_1 norm $\sum_{j=1}^{d} |\beta_j|$ instead of the L_2 norm $\sum_{j=1}^{d} \beta_j^2$ for discarding small coefficients. The LASSO criterion is a penalized regression estimator with the square errors and the L_1 penalty:

$$\sum_{i=1}^{n} \left\{ y_i - \left(\alpha + \mathbf{x}_i^T \boldsymbol{\beta} \right) \right\}^2 + \lambda \sum_{j=1}^{d} |\beta_j|, \tag{2.2}$$

where λ is the tuning parameter. However, Leng *et al.* (2006) showed that LASSO is not asymptotically consistent. LASSO can be biased due to only one tuning parameter λ for large the number of predictors. Fan and Li (2001) dealt with this problem which can be resolved by the SCAD penalty function.

Fan and Li (2001) described the conditions of a good penalty function "(1) Unbiasedness: The resulting estimator is nearly unbiased when the true unknown parameter is large. (2) Sparsity: The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero

to reduce model complexity. (3) Continuity: The resulting estimator is continuous in the data to avoid instability in model prediction". The SCAD penalty is defined as $\sum_{j=1}^{d} p_\lambda(|\beta_j|)$ such that

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if } 0 \le |\beta_j| < \lambda, \\ \dfrac{(a^2-1)\lambda^2 - (|\beta_j| - a\lambda)^2}{2(a-1)}, & \text{if } \lambda \le |\beta_j| < a\lambda, \\ \dfrac{(a+1)^2\lambda^2}{2}, & \text{if } |\beta_j| \ge a\lambda, \end{cases}$$

and so its derivative becomes

$$p'_\lambda(|\beta_j|) = \begin{cases} \lambda, & \text{if } 0 \le |\beta_j| < \lambda, \\ \dfrac{a\lambda - |\beta_j|}{a-1}, & \text{if } \lambda \le |\beta_j| < a\lambda, \\ 0, & \text{if } |\beta_j| \ge a\lambda, \end{cases}$$

where a can be chosen using cross-validation (CV) or generalized CV. However, Fan and Li (2001) recommended $a = 3.7$, because simulation results say that the value is approximately optimal. In this paper we set $a = 3.7$. With the SCAD penalty function instead of the L_1 penalty function, the objective function (2.2) can be written by

$$\sum_{i=1}^{n} \left\{ y_i - \left(\alpha + \mathbf{x}_i^T \boldsymbol{\beta} \right) \right\}^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.3}$$

Nonetheless, the least squares criterion used in (2.3) is still very sensitive to even a single outlier (Rousseeuw and Leroy, 1987). One of alternatives to LSE is LAD estimate which uses the objective function that is the sum of the absolute deviations of the errors instead of the sum of the squares of the errors. One major advantage of LAD lies in its robustness with the comparison of LSE. LAD estimates are less affected by the presence of a few outliers or influential observations. Wang *et al.* (2007) proposed LAD-LASSO with the LAD loss function with the L_1 penalty

$$\sum_{i=1}^{n} \left| y_i - \left(\alpha + \mathbf{x}_i^T \boldsymbol{\beta} \right) \right| + n \sum_{j=1}^{d} \lambda_j |\beta_j|. \tag{2.4}$$

Jung (2012) also considered the problem with the LAD loss function with the SCAD penalty function by replacing the L_1 penalized term in the above equation by the SCAD function as

$$\sum_{i=1}^{n} \left| y_i - \left(\alpha + \mathbf{x}_i^T \boldsymbol{\beta} \right) \right| + \sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.5}$$

There are several robust loss function such as trimmed squares criterion, Huber's loss function. However, these stem from the parametric method and not the nonparametric method. The property of the nonparametric methods has a free distribution of errors and is robust regardless of error distributions. Jaeckel (1972) proposed a rank regression that minimizes the dispersion of the residuals. That can be written by minimizing the objective function

$$\sum_{i<j} \left| e_i - e_j \right|, \tag{2.6}$$

where $e_i = y_i - x_i^T\boldsymbol{\beta}$ denotes the i^{th} residual and the summation ranges over all pairwise difference. Equation (2.6) reduces the rank dispersion function for Wilcoxon scores $\sum_{i<j}|e_i - e_j| = 2\sum_{i=1}^{n} R(e_i)e_i$, where $R(e_i)$ denotes the rank of e_i among $e_1, \ldots, e_n$. Kim *et al.* (2015) used the objective function

$$\sum_{i<j}|e_i - e_j| + \lambda\sum_{j=1}^{d}|\beta_j|$$

which is called the RANK-LASSO estimator.

For the purpose of the robustness of a shrinkage estimator, in the objective function (2.5) we adopt the model error as (2.6). We consider the following objective function that consists of the dispersion of residuals and the SCAD penalty function

$$C(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i<j}|e_i - e_j| + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.7}$$

The pairwise differences of the residuals offers 95% improved efficiency for the normal distribution relative to the sum of squared residuals. So does the objective function (2.7). It preserves the goodness of the SCAD penalty function such as unbiasedness, sparsity and continuity. Unlike LASSO, the objective function (2.7) does not require the preprocessing of the data, for example the centering of the data. We call the estimator minimizing the objective function (2.7) the RANK-SCAD estimator.

The SCAD penalty in (2.7) is affected by the penalty parameter λ which controls the trade-off between model fitting and model complexity. Larger λ will simplify the system because it overfits the regression coefficients. Smaller λ yields an exact system adjusted to the data. The latter has the same results of the rank regression with little sparsity. Therefore, we should set a selection method of the tuning parameter λ (Jung, 2012). Data driven methods such as CV or generalized CV are often used to determine the penalty parameter λ. Wang *et al.* (2007) showed that under the random error with mean 0 and constant variance the Bayesian information criterion (BIC) can identify the true model consistently in the least squares penalized regression with the SCAD penalty. Since BIC = $n\ln(\hat{\sigma}(\lambda)^2)+\mathrm{df}(\lambda)\cdot\ln n$ in linear regression settings, in this paper we select an optimal tuning parameter by minimizing the quantity

$$n\ln\left(\frac{\hat{\sigma}(\lambda)}{n}\right) + \mathrm{df}(\lambda)\cdot\ln n, \tag{2.8}$$

where $\hat{\sigma}(\lambda)$ is the scale estimator in the model and $df(\lambda)$ denotes the number of the degrees of freedom of the model. We use the quantity $\hat{\sigma}(\lambda) = \sum_{i<j}|e_i - e_j|/n$ for the fixed λ and $\mathrm{df}(\lambda)$ equals the number of non-zero regression coefficients.

In this paper we use two choice methods of the penalty parameter such as the 5-folds CV method and the BIC criterion in Section 4.

3. Computation algorithm

When the objective function is differentiable at every point, we can use a standard optimization package to find the argument of the minimum of the objective function $C(\boldsymbol{\beta})$ in (2.7). However, the SCAD penalty function is not convex in $\boldsymbol{\beta}$ and it is not differentiable at zero. By the approximation of the SCAD function in the absolute terms of the regression coefficients, we can obtain an approximate objective function of (2.7).

Jung (2012) used the local quadratic approximation of the SCAD function (Fan and Li, 2001) which has a drawback of backward stepwise variable selection. Instead, we use the following approximation of the SCAD function

$$p_\lambda(|\beta_j|) \approx p_\lambda\left(\left|\beta_j^{(0)}\right|\right) + p'_\lambda\left(\left|\beta_j^{(0)}\right|\right)\left(|\beta_j| - \left|\beta_j^{(0)}\right|\right) \quad \text{for } \beta_j \approx \beta_j^{(0)}. \tag{3.1}$$

Approximation of the SCAD function transforms the objective function into linear equations that allows a simple way to find the solution of the RANK-SCAD estimator. Putting the approximation (3.1) into the objective function (2.7) gives

$$C(\boldsymbol{\beta}) \approx \frac{1}{n}\sum_{i<j}|e_i - e_j| + n\sum_{j=1}^{d}\left\{p_\lambda\left(\left|\beta_j^{(0)}\right|\right) + p'_\lambda\left(\left|\beta_j^{(0)}\right|\right)\left(|\beta_j| - \left|\beta_j^{(0)}\right|\right)\right\} \quad \text{for } \beta_j \approx \beta_j^{(0)}.$$

Then up to an additive constant we obtain the objective function such as the sum of absolute function

$$\tilde{C}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i<j}|e_i - e_j| + n\sum_{j=1}^{d} p'_\lambda\left(\left|\beta_j^{(0)}\right|\right)|\beta_j|. \tag{3.2}$$

The pairwise differences of the residuals (Kim *et al.*, 2015) can be written as

$$\sum_{i<j}|e_i - e_j| = \sum_{i<j}\left|(y_i - y_j) - (\mathbf{x}_i - \mathbf{x}_j)^T\boldsymbol{\beta}\right| = \sum_{k=1}^{N}\left|\tilde{y}_k - \tilde{\mathbf{x}}_k^T\boldsymbol{\beta}\right|,$$

where $\tilde{y}_k = y_i - y_j$ and $\tilde{\mathbf{x}}_k = \mathbf{x}_i - \mathbf{x}_j$ for an appropriate index k, $N = n(n-1)/2$. Let us define the matrices

$$\tilde{\mathbf{y}} = \begin{pmatrix} y_1 - y_2 \\ \vdots \\ y_{n-1} - y_n \end{pmatrix}, \qquad \tilde{\mathbf{X}} = \begin{pmatrix} (\mathbf{x}_1 - \mathbf{x}_2)^T \\ \vdots \\ (\mathbf{x}_{n-1} - \mathbf{x}_n)^T \end{pmatrix}.$$

And we define the augmented matrices such as

$$\tilde{\mathbf{y}}_a = \begin{pmatrix} \tilde{\mathbf{y}}/n \\ \mathbf{0}_d \end{pmatrix}, \qquad \tilde{\mathbf{X}}_a = \begin{pmatrix} \tilde{\mathbf{X}}/n \\ nP_\lambda^1 \end{pmatrix}.$$

where $\mathbf{0}_d$ is the $d \times 1$ vector consisting of zeros, P_λ^1 is the $d \times d$ diagonal matrix whose element is $p'_\lambda(|\beta_j^{(0)}|)$, $\tilde{\mathbf{y}}_a$ is the $(N+d)^{th}$ vector and $\tilde{\mathbf{X}}_a$ is the $(N+d) \times d$ matrix. In this paper we use the initial value $\hat{\boldsymbol{\beta}}^{(0)}$ as the unpenalized rank regression estimate. Then the solution of (3.2) can be obtained by

$$\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{RS}} = \arg\min_{\boldsymbol{\beta}} \tilde{C}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \left\|\tilde{\mathbf{y}}_a - \tilde{\mathbf{X}}_a\boldsymbol{\beta}\right\|_1, \tag{3.3}$$

where $\|\cdot\|$ is the l_1 norm on vectors, e.g., $\|\mathbf{a}\|_1 = \sum_i |a_i|$. Here the superscript means the Rank-SCAD estimator. Equation (3.3) is a traditional LAD criterion found in any standard unpenalized LAD procedure (e.g., l1fit function in the L1PACK package of R).

The proposed estimator (3.3) can be considered as a LAD estimator for the augmented data $\tilde{\mathbf{X}}_a$ and $\tilde{\mathbf{y}}_a$. Therefore, the proposed estimator satisfies the properties of the LAD estimator in the linear

regression model. Wang *et al.* (2007) proved the $\sqrt{n}$ consistency, sparsity and the oracle property of the LAD-LASSO estimator. We conjecture that the proposed estimator preserves the properties of the LAD-LASSO estimator, because the model in (3.3) can be considered as the linear regression model (2.1) without the intercept term.

Next, we find the estimator of the intercept α in linear regression model (2.1). From the estimating equation (3.3) we cannot find the intercept term, because equation (3.3) does not have the intercept term. It also means that the preprocessing as centering for LASSO is not required. Kim *et al.* (2015) recommended the Hodges-Lehmann median estimator for the intercept, which can be found in a nonparametric estimation. We adopt the estimator for the intercept term as

$$\hat{\alpha}_\lambda^{\mathrm{RS}} = \arg\min_\alpha \sum_{i=1}^{n} |\hat{e}_i(\lambda) - \alpha| = \mathrm{median}_i\left(\hat{e}_i(\lambda)\right), \tag{3.4}$$

where $\hat{e}_i(\lambda) = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda^{\mathrm{RS}}$.

Finally, we focus on finding an optimal λ. Since the solution of (3.3) does not have a closed form, it is usually to compute $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{RS}}$ in the grid of λ values, $[\lambda_{\min}, \ldots, \lambda_{\max}]$. Here $\lambda_{\min}$ is the minimum value of λ close to zero, which converts into unpenalized model and the value $\lambda_{\max}$ is the smallest value of λ which shrinks all regression coefficients to zero. Here we set $\lambda_{\min} = 2^{-5}$. The objective function (3.2) can be rewritten by $\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_1 + \|P_\lambda^1 \boldsymbol{\beta}\|_1$ and the normal equation becomes $-\tilde{\mathbf{X}}^T \mathrm{sign}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) = P_\lambda^1 \mathrm{sign}(\boldsymbol{\beta})$, where $\mathrm{sign}(x) = -1, 0, 1$ if $x < 0, = 0, > 0$. Since $\lim_{x \to 0} p'_\lambda(x) = \lambda$, to be zeros of $\boldsymbol{\beta}$ we have $\lambda_{\max} = \|\tilde{\mathbf{X}}^T \mathrm{sign}(\tilde{\mathbf{y}})\|_\infty$, where $\|\cdot\|_\infty$ denotes the l_∞-norm (e.g., for a vector $\mathbf{a}$, $\|\mathbf{a}\|_\infty = \max_i |a_i|$).

The algorithm can be summarized as the following iterative steps:

Step 1. Obtain the unpenalized rank estimator satisfying (2.6) for an initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$.

Step 2. For given $\lambda \in [\lambda_{min}, \lambda_{max}]$ we obtain the RANK-SCAD estimator $\hat{\boldsymbol{\beta}}^{\mathrm{RS}}$ from (3.3) for the augmented matrices $\tilde{\mathbf{X}}_a, \tilde{\mathbf{y}}_a$ and $\hat{\boldsymbol{\beta}}^{(0)}$. Set $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_\lambda^{\mathrm{RS}}$ and obtain $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{RS}}$ from (3.3).

Step 3. Compute the quantity (2.8) for given λ. Go to Step 2 for all grids λ.

Step 4. Find the optimal λ minimizing (2.8). Compute $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{RS}}$ and $\hat{\alpha}^{\mathrm{RS}}$ from (3.3) and (3.4), respectively.

4. Numerical comparisons

4.1. Simulation

In this section we conducted simulation in various situations to show the robustness of the RANK-SCAD estimate which is resistant to heavy-tailed errors and outliers. We compare the proposed method RANK-SCAD estimate with LSE, LASSO from (2.2), LAD-LASSO from (2.4), the RANK-LASSO estimate and the oracle estimate. Here we computed the oracle estimate using LSE on the predictors having nonzero coefficients. All simulations are performed on R program.

We simulated the data set consisting of observations from the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma\epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and ϵ_i is drawn from the standard normal distribution. The predictor $\mathbf{x}_i$ is multivariate normally distributed with the correlation between the $(j,k)^{th}$ columns of $\mathbf{x}_i$ is $0.5^{|j-k|}$ for $j, k = 1, \ldots, 8$. This setting is similar to Tibshirani (1996) and Fan and Li (2001).

Table 1: Simulation results for normal errors without outliers

σ	n	Method	Correct	Incorrect	AMAD
1	30	LSE	0.00 (0.000)	0.00 (0.000)	0.458 (0.136)
		LASSO	4.28 (0.188)	0.00 (0.000)	0.484 (0.188)
		LAD-LASSO	3.62 (1.117)	0.00 (0.000)	0.362 (0.168)
		RANK-LASSO	3.34 (1.125)	0.00 (0.000)	0.436 (0.164)
		RANK-SCAD	4.96 (0.197)	0.02 (0.141)	0.292 (0.173)
		oracle	5.00 (0.000)	0.00 (0.000)	0.254 (0.105)
	60	LSE	0.00 (0.000)	0.00 (0.000)	0.318 (0.081)
		LASSO	4.32 (0.863)	0.00 (0.000)	0.376 (0.118)
		LAD-LASSO	3.71 (1.038)	0.00 (0.000)	0.244 (0.103)
		RANK-LASSO	3.18 (1.395)	0.00 (0.000)	0.325 (0.118)
		RANK-SCAD	4.99 (0.100)	0.00 (0.000)	0.197 (0.086)
		oracle	5.00 (0.000)	0.00 (0.000)	0.185 (0.080)
$\sqrt{3}$	30	LSE	0.00 (0.000)	0.00 (0.000)	0.808 (0.235)
		LASSO	4.25 (0.947)	0.01 (0.100)	0.954 (0.338)
		LAD-LASSO	3.07 (1.139)	0.03 (0.171)	0.740 (0.301)
		RANK-LASSO	3.48 (1.259)	0.01 (0.100)	0.857 (0.333)
		RANK-SCAD	4.85 (0.458)	0.29 (0.498)	0.709 (0.393)
		oracle	5.00 (0.000)	0.00 (0.000)	0.448 (0.206)
	60	LSE	0.00 (0.000)	0.00 (0.000)	0.529 (0.139)
		LASSO	4.40 (0.725)	0.00 (0.000)	0.666 (0.192)
		LAD-LASSO	2.94 (1.238)	0.00 (0.000)	0.454 (0.217)
		RANK-LASSO	3.53 (1.159)	0.00 (0.000)	0.530 (0.189)
		RANK-SCAD	4.97 (0.223)	0.06 (0.239)	0.376 (0.230)
		oracle	5.00 (0.000)	0.00 (0.000)	0.312 (0.123)

AMAD = average of the mean absolute deviations; LSE = least squares estimate; LASSO = least absolute shrinkage and selection operator; LAD = least absolute deviation.

We chose the sample sizes $n = 30, 60$, and two different values $\sigma = 1, \sqrt{3}$ for the standard deviation of errors. For heavy-tailed error we took into account the t distribution with degrees of freedom 3 instead of the normal distribution. For checking the resistibility to outliers we made a contaminated data with 20% regression outliers.

The simulation data consists of training data and test data. After obtaining the regression coefficients based on the training data, the performance of the estimate is evaluated on the test data of the sample size 1,000 generating from the normal distribution with the covariance matrix described above. We conducted 100 simulation iterations to evaluate the error of the estimates in the average of the mean absolute deviations on the test data. The evaluation of the sparseness of an estimate can be measured by the average number of correctly estimated zero coefficients which is the column labeled "correct" in the tables. Analogously the column labeled "incorrect" denotes the number of zero estimates which are not zero coefficients, and it can be a measure of inaccuracy for discarding erroneously important variables. The simulation results are summarized in Tables 1–3. The number in the parenthesis is the sample standard deviation.

Tables 1 and 2 show that the proposed method RANK-SCAD outperforms the other methods. Furthermore, the performance of RANK-SCAD is quite comparable to that of the oracle, especially in the view of sparseness and the model errors. RANK-SCAD is considered to be better than other estimates such as LASSO and RANK-LASSO in terms of variable selection and prediction. The term "incorrect" of RANK-SCAD is negligible. Especially, under all distributions of errors the model error of RANK-SCAD is comparable to the oracle because it is evident from the property of nonparametric methods. All methods have smaller standard deviations as the sample size grows that also imply the consistency of the method.

Table 2: Simulation results for $t(3)$ errors without outliers

σ	n	Method	Correct	Incorrect	AMAD
1	30	LSE	0.00 (0.000)	0.00 (0.000)	0.818 (0.545)
		LASSO	4.18 (1.067)	0.10 (0.362)	0.935 (0.560)
		LAD-LASSO	3.38 (1.023)	0.03 (0.171)	0.485 (0.263)
		RANK-LASSO	3.23 (1.434)	0.01 (0.100)	0.626 (0.278)
		RANK-SCAD	4.93 (0.383)	0.09 (0.321)	0.429 (0.543)
		oracle	5.00 (0.000)	0.00 (0.000)	0.446 (0.543)
	60	LSE	0.00 (0.000)	0.00 (0.000)	0.472 (0.176)
		LASSO	4.60 (0.603)	0.01 (0.100)	0.777 (0.332)
		LAD-LASSO	3.77 (1.014)	0.00 (0.000)	0.283 (0.133)
		RANK-LASSO	3.62 (1.179)	0.00 (0.000)	0.404 (0.155)
		RANK-SCAD	4.98 (0.141)	0.02 (0.141)	0.234 (0.156)
		oracle	5.00 (0.000)	0.00 (0.000)	0.279 (0.153)
$\sqrt{3}$	30	LSE	0.00 (0.000)	0.00 (0.000)	1.330 (0.538)
		LASSO	4.19 (1.032)	0.35 (0.642)	1.536 (0.732)
		LAD-LASSO	2.68 (1.154)	0.08 (0.273)	0.947 (0.418)
		RANK-LASSO	3.25 (1.493)	0.13 (0.418)	1.103 (0.464)
		RANK-SCAD	4.83 (0.451)	0.69 (0.677)	1.173 (0.673)
		oracle	5.00 (0.000)	0.00 (0.000)	0.706 (0.312)
	60	LSE	0.00 (0.000)	0.00 (0.000)	0.905 (0.366)
		LASSO	4.46 (0.702)	0.18 (0.626)	1.367 (0.682)
		LAD-LASSO	2.88 (1.174)	0.01 (0.100)	0.525 (0.265)
		RANK-LASSO	3.66 (1.174)	0.01 (0.100)	0.705 (0.265)
		RANK-SCAD	4.97 (0.171)	0.20 (0.402)	0.498 (0.346)
		oracle	5.00 (0.000)	0.00 (0.000)	0.483 (0.317)

AMAD = average of the mean absolute deviations; LSE = least squares estimate; LASSO = least absolute shrinkage and selection operator; LAD = least absolute deviation.

Now we conducted the simulation on the data with outliers, which is consisted of adding the 20% regression outliers to the data by changing $y = y+10$. In this case Table 3 summarizes the performance of the methods when the errors are generated from the t distribution with degrees of freedom 3. We had similar results for the normal errors. Under the contaminated data RANK-SCAD outperforms LASSO and RANK-LASSO in view of the model parsimony. However, in terms of the model error RANK-LASSO is a little larger than RANK-LASSO and LAD-LASSO, but it is smaller than LASSO. Even though RANK-SCAD gives the best sparse model and also the worst in the "incorrect" term. In the future we will improve a variable selection method to diminish the "incorrect" value of RANK-SCAD.

4.2. Real data analysis

In this section we adopted the proposed method to the prostate cancer data with larger explanatory variables and sample size, which is analyzed in Tibshirani (1996). The data consists of 97 observations with the response variable *lpsa* (log prostate specific antigen) and eight explanatory variables such as *lcavol* (log cancer volume), *lweight* (log prostate weight), *age*, *lbph* (log benign prostatic hyperplasia amount), *svi* (seminal vesicle invasion), *lcp* (log capsular penetration), *gleason* (Gleason score), and *pgg45* (percentage Gleason scores 4 or 5).

We used RANK-SCAD methods with two tuning parameters by CV (RS-CV) and BIC criterion (RS-BIC). We compared the RANK-SCAD estimator with LASSO, LAD-LASSO, RANK, and RANK-LASSO. We use 67 randomly chosen observations as the training data and the remaining as the testing data for 100 replications. The prediction accuracies of these method are measured by the

Table 3: Simulation results for $t(3)$ errors with 20% outliers

σ	n	Method	Correct	Incorrect	AMAD
1	30	LSE	0.00 (0.000)	0.00 (0.000)	2.359 (0.705)
		LASSO	4.43 (0.902)	1.07 (0.832)	2.222 (0.700)
		LAD-LASSO	2.86 (1.271)	0.13 (0.338)	0.875 (0.567)
		RANK-LASSO	4.15 (0.999)	0.47 (0.688)	1.244 (0.698)
		RANK-SCAD	4.91 (0.288)	1.24 (0.726)	1.632 (0.796)
		oracle	5.00 (0.000)	0.00 (0.000)	1.294 (0.537)
	60	LSE	0.00 (0.000)	0.00 (0.000)	1.481 (0.377)
		LASSO	4.46 (0.797)	0.49 (0.703)	1.827 (0.629)
		LAD-LASSO	3.41 (1.065)	0.00 (0.000)	0.370 (0.172)
		RANK-LASSO	3.97 (0.979)	0.03 (0.223)	0.713 (0.313)
		RANK-SCAD	4.98 (0.141)	0.41 (0.552)	0.630 (0.492)
		oracle	5.00 (0.000)	0.00 (0.000)	0.758 (0.320)
$\sqrt{3}$	30	LSE	0.00 (0.000)	0.00 (0.000)	2.464 (0.745)
		LASSO	4.62 (0.663)	1.34 (0.855)	2.528 (0.758)
		LAD-LASSO	2.25 (1.266)	0.26 (0.463)	1.604 (0.791)
		RANK-LASSO	4.29 (1.066)	1.03 (0.926)	1.959 (0.833)
		RANK-SCAD	4.91 (0.321)	1.71 (0.782)	2.176 (0.861)
		oracle	5.00 (0.000)	0.00 (0.000)	1.351 (0.547)
	60	LSE	0.00 (0.000)	0.00 (0.000)	1.615 (0.444)
		LASSO	4.44 (0.935)	0.77 (0.874)	2.101 (0.748)
		LAD-LASSO	2.57 (1.208)	0.03 (0.171)	0.754 (0.341)
		RANK-LASSO	3.91 (1.045)	0.17 (0.451)	1.177 (0.486)
		RANK-SCAD	4.96 (0.197)	1.01 (0.703)	1.304 (0.598)
		oracle	5.00 (0.000)	0.00 (0.000)	0.900 (0.396)

AMAD = average of the mean absolute deviations; LSE = least squares estimate; LASSO = least absolute shrinkage and selection operator; LAD = least absolute deviation.

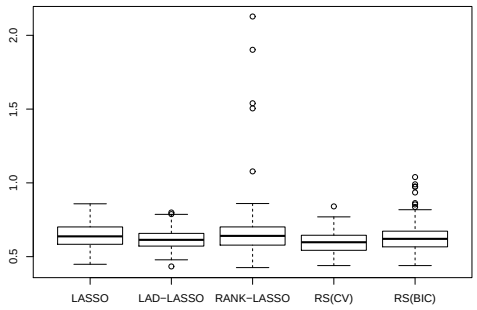

Figure 1: *Boxplots of the mean absolute prediction errors from 100 replications. LASSO = least absolute shrinkage and selection operator; LAD = least absolute deviation; RS = rank smoothly clipped absolute deviation; CV = cross validation; BIC = Bayesian information criterion.*

mean absolute prediction error (MAPE) on the testing data. The distribution of these 100 MAPE values is depicted graphically in Figure 1. Our proposed estimator with CV method yields the smallest median of MAPE. The prediction accuracy of the RANK-SCAD remains satisfactory. The standard error of the MAPE estimate for LASSO, LAD-LASSO, RANK-LASSO, and RANK-SCAD(CV), RANK-SCAD(BIC) are also reported as 0.085, 0.071, 0.246, 0.073, 0.113, respectively.

Similarly to Kim *et al.* (2015), we tested the effect of outliers by changing the first two response variable to $10\max(|y_i|)$, $5\max(|y_i|)$. Then we had the following estimates for the full data and the

Table 4: Estimates for the prostate cancer data and the modified data

	LASSO	LAD-LASSO	RANK-LASSO	RS(CV)	RS(BIC)
lcavol	0.472 (0.000)	0.524 (0.450)	0.562 (0.244)	0.552 (0.000)	0.519 (0.419)
lweight	0.187 (0.000)	0.480 (0.516)	0.350 (0.000)	0.496 (0.000)	0.483 (0.698)
age	0.000 (0.000)	0.000 (0.000)	−0.018 (0.000)	−0.018 (0.000)	−0.009 (−0.017)
lbph	0.000 (0.000)	0.000 (0.000)	0.121 (0.000)	0.102 (0.000)	0.017 (0.000)
svi	0.368 (0.000)	0.249 (0.319)	0.529 (0.000)	0.691 (0.000)	0.567 (0.610)
lcp	0.000 (0.000)	0.000 (0.000)	−0.054 (0.000)	−0.074 (0.000)	0.000 (0.000)
gleason	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
pgg45	0.000 (0.000)	0.000 (0.000)	0.006 (0.010)	0.005 (0.000)	0.004 (0.004)

The numbers in the parenthesis are the estimates on the modified data.
LASSO = least absolute shrinkage and selection operator; LAD = least absolute deviation; RS = rank smoothly clipped absolute deviation; CV = cross validation; BIC = Bayesian information criterion; lcavol = log cancer volume; lweight = log prostate weight; lbph = log benign prostatic hyperplasia amount; svi = seminal vesicle invasion; lcp = log capsular penetration; gleason = Gleason score; pgg45 = percentage Gleason scores 4 or 5.

modified data in Table 4. It shows that the estimates of LASSO, RANK-LASSO, RANK-SCAD(CV) gave a constant regression model. Both LAD-LASSO and RANK-SCAD(BIC) are significant and also they are robust to regression outliers. Consequently, it seems that the proposed RANK-SCAD estimator is an alternative to robust penalized regression estimators.

5. Concluding remarks

We proposed a robust variable selection method regardless of error distributions in the regression model using the rank regression estimator with the SCAD penalty function. The method preserves the advantages of both robustness from the rank regression and simultaneous variable selection from the SCAD penalty function, which will especially provide a conjecture that the proposed estimator has an oracle property. The proposed algorithm is much faster and effective. Numerical simulation shows that the proposed method RANK-SCAD performs well as the oracle estimate for variable selection.

In future we will study penalized regressions to reduce the influence of leverage points based on the rank regression simply by considering the weight to the loss function. We only have studied the linear regression model; however, the proposed method can be extended to other regression models such as generalized linear models.

Acknowledgements

Jung's research was supported by Basic Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01060528).

References

Alfons A, Croux C, and Gelper S (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets, *The Annals of Applied Statistics*, **7**, 226–248.

Chen X, Wang J, and McKeown MJ (2010). Asymptotic analysis of robust LASSOs in the presence of noise with large variance, *IEEE Transactions on Information Theory*, **56**, 5131–5149.

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.

Hoerl AE and Kennard RW (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Jaeckel LA (1972). Estimating regression coefficients by minimizing the dispersion of the residuals, *The Annals of Mathematical Statistics*, **43**, 1449–1458.

Jung KM (2011). Weighted least absolute deviation LASSO estimator, *Communications for Statistical Applications and Methods*, **18**, 733–739.

Jung KM (2012). Weighted least absolute deviation regression estimator with the SCAD function, *Journal of the Korean Data Analysis Society*, **14**, 2305–2312.

Jung KM (2013). Weighted support vector machines with the SCAD penalty, *Communications for Statistical Applications and Methods*, **20**, 481–490.

Jung SY and Park C (2015). Variable selection with nonconcave penalty function on reduced-rank regression, *Communications for Statistical Applications and Methods*, **22**, 41–54.

Kim HJ, Ollila E, and Koivunen V (2015). New robust LASSO method based on ranks. In *Proceedings of the 23rd European Signal Processing Conference*, Nice, France, 704–708.

Lee S (2015). An additive sparse penalty for variable selection in high-dimensional linear regression model, *Communications for Statistical Applications and Methods*, **22**, 147–157.

Leng C, Lin Y, and Wahba G (2006). A note on the LASSO and related procedures in model selection, *Statistica Sinica*, **16**, 1273–1284.

Rousseeuw PJ and Leroy AM (1987). *Robust Regression and Outlier Detection*, John Wiley, New York.

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

Wang H, Li G, and Jiang G (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso, *Journal of Business & Economic Statistics*, **25**, 347–355.

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

Received September 5, 2017; Revised October 12, 2017; Accepted October 17, 2017