Robust regression modeling via L_1 regularization

Mathematics Course, Heewon Park

***Abstract*—There is currently much discussion about lasso-type regularized regression which is a useful tool for simultaneous estimation and variable selection. Although the L_1-type regularization has several advantages in regression modeling, it suffers from outliers because their procedures are constructed by non-robust manners. To overcome the problem, we propose robust modeling strategies via L_1-type regularization in the various perspectives: methodology, model estimation and evaluation. We observe through simulation studies and real data analyses that our robust modeling strategies outperform for regression modeling in the presence of outliers.**

I. Introduction

The lasso-type regularization consisting of the least squares loss function with L_1-type penalty has drawn a large amount of attention in recent years. By imposing an L_1-type penalty to a least squares loss function, the L_1-type regularization methods can perform not only variable selection and estimation simultaneously, but also stable regression modeling by preventing high variances of estimates. Although the modeling procedure shows the remarkable performance in regression modeling, its performance takes a sudden turn for the worst in the presence of outliers, since the modeling procedures are based on non-robust manners.

We introduce the robust modeling strategies via the L_1-type regularization in various perspectives. We first discuss a robust L_1-type regularization based on the robust loss function not least squares loss function. Then, robust algorithms are proposed to estimate L_1-type regularized regression model. We also introduce the methods for tuning parameter selection, which is crucial in the robust sparse regression modeling, in the viewpoint of information-theoretic approaches.

The efficiency of the proposed robust modeling procedure is investigated through Monte Carlo experiments and real data analyses.

II. Robust regression modeling via L_1 regularization

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

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

where $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and ε_i are the random errors which are assumed to be independently and identically distributed with mean 0 and variance σ^2. For estimating the parameters in (1), there is currently much discussion about L_1-type regularized regression (e.g., lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001)). Although the modeling procedure is a useful tool for regression modeling, owing to its sparsity, it suffers from outliers since their procedures are based on non-robust manners. To overcome this drawback, we propose robust modeling strategies via the L_1-type regularization in the aspects of methodology, model estimation and evaluation.

A. Robust L_1-type approaches

The robust L_1-type regularization was proposed to overcome the demerit of the L_1-type regularization by replacing the least squares loss function with robust loss function as follows:

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

where $\rho(r_i; k)$ is a robust loss function (e.g., least absolute deviation, M-function) with tuning constant k for outlier detection, $r_i = y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\sum_{j=1}^{p} p_\lambda(|\beta_j|)$ is a lasso-type penalty with regularization parameter λ. The existing robust L_1-type regularization, however, do not have high breakdown point, since M-function and LAD loss function have a low breakdown point.

For the robust lasso-type approaches having a high breakdown point, we consider the least trimmed squares (LTS) estimator having the maximum breakdown point $\{[(n-p)/2]+1\}/n$, which is asymptotically equal to 50%, for $s = [(n+p+1)/2]$ (see Rousseeuw and Leroy (1987), Theorem 6). In the LTS procedure, the sample size used to estimate is decreased from n to s. In other words, there is a possibility that sample size s, is smaller than the number of predictor variables p. Hence, although the LTS has a high breakdown point, it is unsuitable for using with the lasso because of the limitation of the lasso as variable selection method in $p > n$ situation. Thus, we consider an elastic net, which was proposed to settle the problem of the lasso in $p > n$ situation (Zou and Hastie, 2005), and propose a robust elastic net (Park et al., 2012a), called a least trimmed squares-elastic net (LTS-Ela),

$$\hat{\boldsymbol{\beta}}^{\text{LTS-Ela}} = \arg\min_{\boldsymbol{\beta}} \{ \sum_{i=1}^{s} r_{[i]}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \}, \tag{3}$$

where s is the tuning constant and $r_{[i]}^2$ is the i^{th} order statistic of squared residuals. The proposed LTS-Ela performs well regression modeling in the presence of outliers, and is useful for $p \gg n$ data.

B. Estimation of sparse regression model

The lasso-type approaches provide a useful tool for the sparse regression modeling. Their estimates, however, cannot be analytically derived due to indifferentiability of the L_1-type penalty term. To settle on this issue, several algorithms have been proposed, such as local quadratic approximation (Fan and Li, 2001), LARS (Efron et al., 2004) and coordinate descent algorithm (Friedman et al., 2007). We focus on the coordinate descent algorithm which is competitive with the well known LARS. Although the coordinate descent algorithm effectively performs sparse regression modeling, it suffers from outliers, since the solution path is delivered based on standardized data by mean, standard deviation, and inner product of predictor x_{ij} and partial residual $(y_i - \tilde{y}_i)$ obtained by a non-robust manner. To overcome this drawback, we first standardize data by median and median absolute deviation instead of mean and standard deviation like Khan et al. (2007), and propose robust coordinate descent algorithm based on an outlier-resistant inner product.

B.1. W.coordinate descent algorithm

In order to control the outliers in j^{th} predictor and partial residual, we proposed a robust bivariate winsorization via the robust Mahalanobis distance based on the minimum covariance determinant,

$$\text{RD.W.ob}_j = \min(\sqrt{k/RD(\mathbf{x}_j)}, 1)\ \mathbf{x}_j, \tag{4}$$

where $\mathbf{x}_j = (x_{ij}, (y_i - \tilde{y}_i))^T$, $RD(\mathbf{x}_j)$ is robust Mahalanobis distance and $k = 5.99$ as the 95% quantile of the $\chi^2(\text{df} = 2)$ distribution like Khan et al. (2007). The coordinate descent procedure is conducted based on a robust inner product calculated by the winsorized data RD.W.ob$_j$ as follows:

$$\tilde{\beta}_j \leftarrow S\Big(\sum_{i=1}^{n} x_{ij}^{W}(y_i - \tilde{y}_i)^{W}, \gamma\Big), \tag{5}$$

where x_j^W and $(y_i - \tilde{y}_i)^W$ are winsorized data by (5).

B.2. T.coordinate descent algorithm

We also propose the robust bivariate trimming technique, which controls an effect of outliers by eliminating extreme observations at each tails, similar to robust bivariate winsorization,

$$\text{RD.T.ob}_j = \mathbf{x}_j\{I(\sqrt{k/RD(\mathbf{x}_j)} \geq 1)\}. \tag{6}$$

RD.T.ob$_j$ is updated in each iterations step, and then the variable selection and estimation are conducted base on the robust inner product by the trimmed data RD.T.ob$_j$ in the coordinate descent procedure as follows:

$$\tilde{\beta}_j \leftarrow S\Big(\sum_{i=1} x_{ij}^{TR}(y_i - \tilde{y}_i)^{TR}, \gamma\Big), \tag{7}$$

where x_j^{TR} and $(y_i - \tilde{y}_i)^{TR}$ are trimmed data by (7).

C. Model evaluation criteria for tuning parameter selection

Crucial issues in the robust modeling procedure include the selection of regularization parameters and also a tuning constant in outlier detection. We introduce novel methods for choosing the tuning parameters in an information-theoretic viewpoint.

C.1. Efficient bootstrap information criterion

We introduce to use the efficient bootstrap information criteria (Konishi and Kitagawa, 1996) for choosing the optimal set of the tuning parameters of the robust lasso-type approaches. 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 $G(y)$. The model is estimated by some estimator $\hat{\boldsymbol{\theta}}$. Our task is to evaluate the expected goodness or badness 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.

The general form of an information criterion can be given as follows:

$$IC(\mathbf{y}_n; \hat{G}) = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{\text{estimator for } b(G)\}, \tag{8}$$

where $b(G)$ is the bias of the log-likelihood as an estimator of the expected log-likelihood defined by

$$b(G) = E_G\Big[\log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}(\mathbf{y}_n)) - nE_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{y}_n))\Big]\Big]. \tag{9}$$

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

$$D(\mathbf{y}_n; G) = \log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}) - n\int \log f(z|\hat{\boldsymbol{\theta}})dG(z), \tag{10}$$

can be decomposed into sum of the following three terms

$$\begin{aligned} D_1(\mathbf{y}_n; G) &= \log f(\mathbf{y}_n|\hat{\boldsymbol{\theta}}) - \log f(\mathbf{y}_n|\boldsymbol{\theta}), \qquad (11)\\ D_2(\mathbf{y}_n; G) &= \log f(\mathbf{y}_n|\boldsymbol{\theta}) - n\int \log f(z|\boldsymbol{\theta})dG(z),\\ D_3(\mathbf{y}_n; G) &= n\int \log f(z|\boldsymbol{\theta})dG(z) - n\int \log f(z|\hat{\boldsymbol{\theta}})dG(z). \end{aligned}$$

In the information criterion, the bias represents as the expected value of $D(\mathbf{y}_n; G)$. By taking the expectation term by term in (11), we obtain the second term: $E_G[D_2(\mathbf{y}_n; G)] = 0$. Thus, the expectation of (10) can be expressed as follows;

$$E_G[D(\mathbf{y}_n; G)] = E_G[D_1(\mathbf{y}_n; G) + D_3(\mathbf{y}_n; G)]. \tag{12}$$

In constructing the bootstrap information criterion, the true distribution $G(y)$ is replaced with an empirical distribution function $\hat{G}(y)$. Let us extract B sets of bootstrap samples of size n and write the b^{th} bootstrap sample as $\mathbf{y}_n^*(b) = \{y_1^*(b), ..., y_n^*(b)\}$. In the bootstrap estimate, (12) is replaced by

$$E_{\hat{G}}[D(\mathbf{y}_n^*; \hat{G})] = E_{\hat{G}}[D_1(\mathbf{y}_n^*; \hat{G}) + D_3(\mathbf{y}_n^*; \hat{G})]. \tag{13}$$

Conditional on the observed data, it can be shown that the orders of asymptotic conditional variances of two bootstrap estimates are Var $\left[\frac{1}{B}\sum_{b=1}^{B}\{D(\mathbf{y}_n^*; \hat{G})\}\right] = \frac{1}{B}O(n)$ and Var $\left[\frac{1}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*; \hat{G}) + D_3(\mathbf{y}_n^*; \hat{G})\}\right] = \frac{1}{B}O(1)$. This implies that the variance due to the bootstrap resampling can be reduced significantly, and thus we can expect to efficient modeling. Consequently, the efficient bootstrap information criterion based on variance reduction method is defined as follows

$$EIC_{\text{eff}} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{B}\sum_{b=1}^{B}\{D_1(\mathbf{y}_n^*(b); \hat{G}) + D_3(\mathbf{y}_n^*(b); \hat{G})\}. \tag{14}$$

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

C.2. Generalized information criterion for M-lasso

We derive an information criterion to evaluate robust regularized regression model in line with the generalized information criterion (GIC) (Konishi and Kitagawa, 1996). The GIC, which is constructed with asymptotic bias for a statistical model as a function of the influence function and score function, can evaluate models estimated by various techniques not only maximum likelihood method. In case of the lasso-type approaches, however, we cannot calculate an influence function, which plays a key role in deriving the GIC, because of L_1-norm type penalty. To settle on this issue, we use the local quadratic approximation (Fan and Li, 2001), and then derive the information criterion (Park et al., 2012b).

For the linear regression model, we use robust lasso-type estimates (e.g., with Huber function) of the regression coefficients $\boldsymbol{\beta}$ given as the following solution. If $\mathbf{T}_{R.la,0}(G)$ is very close to 0, then set $\mathbf{T}_{R.la}(G) = 0$. Otherwise it is assumed that the functional $\mathbf{T}_{R.la}(G)$ is given as a solution of the implicit equation: $\int\{\psi(y-\mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x} - p_\lambda^{'}(|\boldsymbol{T}_{R.la,0}|)\text{sign}(\boldsymbol{T}_{R.la})\}dG(y) = \mathbf{0}$. By using the local quadratic approximations (Fan and Li, 2001), we can rewrite the implicit equation as

$$\int \Big(\psi(y-\mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x} \qquad (15)$$
$$-\{p_\lambda^{'}(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}\boldsymbol{T}_{R.la}(G)\Big)dG(y) = \mathbf{0}.$$

To derive the influence function $\mathbf{T}^{(1)}_{R.la}(G)$, G in (15) is substituted with $(1-\varepsilon)G + \varepsilon\delta_y$, and then differentiating both sides of the equation with respect to ε and setting ε=0. Consequently, we calculate the influence function, $\mathbf{T}^{(1)}_{R.la}(G)$, of the functional that defines the robust lasso-type estimator, and then derive the asymptotic bias of the log-likelihood of $f(y|\mathbf{x}, \hat{\boldsymbol{\beta}}_{R.la}(\lambda))$ is given by

$$b^{(1)}_{R.la} =$$
$$\text{tr}\Big(\Big[\int \psi^{'}(y-\mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x}\mathbf{x}^T + \{p_\lambda^{'}(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|\}dG\Big]^{-1}$$
$$\times[\int \Big[\psi(y-\mathbf{x}^T\mathbf{T}_{R.la}(G))\mathbf{x} - [p_\lambda^{'}(|\boldsymbol{T}_{R.la,0}|)/|\boldsymbol{T}_{R.la,0}|]\boldsymbol{T}_{R.la}(G)\Big]$$
$$\cdot\frac{\partial \log f(y|\mathbf{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\mathbf{T}_{R.la}(G)}dG]\Big) + O(n^{-1}). \qquad (16)$$

By replacing the unknown distribution G by the empirical distribution $\hat{G}$, and subtracting the asymptotic bias estimate from the log-likelihood, we have an information criterion for the statistical model $f(y|\mathbf{x}, \hat{\boldsymbol{\beta}}_{R.la}(\lambda))$ with the functional estimator, $\hat{\boldsymbol{\beta}}_{R.la}(\lambda) = \mathbf{T}(\hat{G})$, in the following

$$GIC_{R.la} = \qquad (17)$$
$$-2\Big[\sum_{i=1}^{n}\log f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_{R.la}(\lambda)) - \{R(\psi,\hat{G})^{-1}Q(\psi,\hat{G})\}\Big],$$

where $R(\psi,\hat{G})$ is

$$-\frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial\beta_j}\psi(y_i-\mathbf{x}_i^T\mathbf{T}_{R.la})x_{i,j}^2 + \{p_\lambda^{'}(|T_{R.la,j0}|)/|T_{R.la,j0}|\},$$

and $Q(\psi,\hat{G})$ is

$$\frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}\Big[\Big(\psi(y_i-\mathbf{x}_i^T\mathbf{T}_{R.la})x_{i,j} - \{p_\lambda^{'}(|T_{R.la,j0}|)/|T_{R.la,j0}|\}T_{R.la,j}\Big)$$
$$\cdot\frac{\partial \log f(y_i|\mathbf{x}_i,\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\mathbf{T}_{R.la}}\Big].$$

C.3. Robust efficient bootstrap information criterion

We introduce the robust tuning parameter selection, which is critical in the robust sparse regression modeling. Although the bootstrap information criterion has several advantages on its flexibility and weak assumptions, a bootstrap sample may contain more outliers compared with those in the original sample, since the bootstrap sample is drawn randomly. This implies that the resulting criterion from the highly contaminated bootstrap sample may produce biased results. 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; Park, 2012).

First, we introduce a winsorization bootstrap method (Srivastava et al., 2010). Suppose that the order statistics of the original data be denoted by $y_{[1]}, y_{[2]}, ..., y_{[n]}$. For winsorizing proportion $\delta = l/n$, $0 \le \delta \le 1/2$, a δ-winsorized sample $\{y_i^*\}$ is $y_{[l+1]}$ if $y_i \le y_{[l]}$ and $y_{[n-l]}$ if $y_i \ge y_{[n-l+1]}$, otherwise $\{y_i^*\}$ is y_i. The winsorized bootstrap sample $\{y_i^{**}\}$ are randomly drawn from the δ-winsorized sample $\{y_i^*\}$.

By using the winsorized bootstrap sample, the bootstrap bias estimate of (9), $b^{**}(\hat{G})$, is given by $E_{\hat{G}}\Big[\sum_{i=1}^{n}\log f(y_i^{**}|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^{**})) - nE_{\hat{G}(z^{**})}\Big[\log f(Z^{**}|\hat{\boldsymbol{\theta}}(\mathbf{y}_n^{**}))\Big]\Big]$.

Let us extract B sets of winsorized bootstrap samples of size n and write the b^{th} winsorized bootstrap sample as $\mathbf{y}_n^{**}(b) = \{y_1^{**}(b), ..., y_n^{**}(b)\}$. The winsorized bootstrap bias estimate of (9) is substituted by

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

Consequently, we propose a robust efficient bootstrap information criterion as follows;

$$R.EIC_{\text{eff}} = -2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}}) + 2\{b_B^w(\hat{G})\}. \qquad (19)$$

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.

III. Simulation studies

We simulated N datasets consisting of n observations from the following model

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma\varepsilon_i, \qquad i = 1, ..., n, \qquad (20)$$

where $\boldsymbol{\beta}$ is a p-dimensional vector and ε_i are distributed as a standard normal. The correlation between x_l and x_m is $\rho^{|l-m|}$ with ρ=0.5. In the simulation studies, the results of variable selection are shown as true negative 'T.N" (i.e., the average percentage of true zero coefficients) and false negative "F.N" (i.e., average number of the true non-zero coefficients, incorrectly set to zero). And, forecasting accuracy is measured forecasting root mean squares error (RMSE) by N simulated datasets.

Part 1: LTS-Ela and EIC$_{\text{eff}}$.

In Part 1, we evaluate the proposed LTS-Ela compared with LTS-lasso, and efficient bootstrap information criterion as a tuning parameter selector in Table 1. We simulated $N = 50$ datasets in $\sigma = 1$. The simulation is conducted in the presence of 10% and 20% outliers for $\varepsilon_i \sim N(10, 3)$ and $\boldsymbol{\beta} = (2, 3, 0, 0, 1.5, 0, 0, 1, 0, 0)^T$. From the column "T.N"

TABLE I

			T.N (%)		F.N (%)	
			LTS-lasso	LTS-Ela	LTS-lasso	LTS-Ela
Outliers	10%	CV	0.43	0.31	0.01	0.03
		BIC	0.05	0.38	0.00	0.11
		EIC_{eff}	0.50	**0.61**	0.11	0.15
	20%	CV	0.44	0.37	0.05	0.07
		BIC	0.05	0.25	0.02	0.05
		EIC_{eff}	0.42	**0.60**	0.11	0.19

in Table 1, it can be seen that the proposed LTS-Ela based on the efficient bootstrap information criterion is outstanding in the viewpoint of the "sparsity", which is a crucial property of the lasso-type approaches.

Part 2: Robust coordinate descent procedure.

In part 2, we evaluate the proposed robust coordinate descent algorithm for lasso. We simulated $N = 100$ datasets consisting of $n = 80$ with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 1$. We consider the two situations: (a) 10% outliers for $N(30, 1)$ in only y_i and (b) 5% outliers for $N(30, 1)$ in y_i, and 5% outliers for $N(0, 5)$ in x_1 and x_5. We evaluate the forecasting accuracy and variable selection results. From the results, it is observed that the proposed robust procedures produce reliable regression modeling results even in the presence of outliers, especially the W.coordinate descent algorithm shows the best performance.

Part 3: $\text{GIC}_{R.la}$ via local quadratic approximation.

In part 3, we evaluate the proposed $\text{GIC}_{R.la}$ as a tuning parameter selector for M-lasso and M-SCAD with Huber function. In order to evaluate the $\text{GIC}_{R.la}$, we compare with results by the BIC and cross-validation. The simulation was conducted under the cases ε_i are standard normal with 1% outliers for $\varepsilon_i \sim N(10, 1)$ in $\sigma = 1$, and sensible outliers for $\varepsilon_i \sim D/\sqrt{\text{var}(D)}, \sigma = 9.67$, where D is a standard double exponential distribution with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ for $n = 80$. For model estimation of the M-lasso and M-SCAD, we used an iterative reweighted least square (IRLS) algorithm. We observed that the proposed $\text{GIC}_{R.la}$ is a useful tool for robust L_1-type regression modeling in the viewpoint of the variable selection.

Part 4: Robust efficient bootstrap information criterion.

In part 4, we evaluate the robust efficient bootstrap information criterion as a tuning parameter selector compared with the ordinary bootstrap information criterion and 10-fold cross-validation. We simulated $N = 50$ datasets consisting of $n = 80$ with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 1$. Simulations are conducted in the presence of 5%, and 10% outliers for $\varepsilon_i \sim N(30, 3)$. To find the solution of the M-lasso with Huber function, we use the IRLS algorithm. From the column "T.N" and "RMSE" in Table 2, it can be seen that the proposed robust efficient bootstrap information criterion is effective for robust sparse regression modeling.

TABLE II

Outlier	Method	T.N	F.N	RMSE
5%	CV	0.076	0.000	1.75
	Eff.Boot.IC	0.076	0.000	1.75
	Robust.Eff.Boot.IC	**0.086**	0.003	**1.74**
10%	CV	0.076	0.007	3.59
	Eff.Boot.IC	0.054	0.000	3.47
	Robust.Eff.Boot.IC	**0.086**	0.003	**3.44**

IV. Concluding remarks

We have discussed the robust regression modeling via L_1 regularization. The robust modeling strategies have been proposed in the various perspectives: methodology, model estimation, evaluation and numerical aspect. We have observed through Monte Carlo experiments that the proposed robust modeling strategies are superior to existing ones, in the presence of outliers. For the future studies, the present studies may be extended to robust non-linear regression modeling. Furthermore, a work remains to be done for constructing a model for high dimensional data via robust coordinate descent algorithm.

References

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.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1, 302-332.

Khan, J.A., Van Aelst, S. and Zamar, R.H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102, 1289-1299.

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

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

Park, H.W. (2012). Novel resampling methods for tuning parameter selection in robust sparse regression modeling. *To appear in Bulletin of Informatics and Cybernetics*.

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

Park, H.W., Sakaori, F. and Konishi, S. (2012b). Selection of tuning parameters in robust sparse regression modeling. *Proceedings of the COMPSTAT'2012*, pp 713-723.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

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, 45-67.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67, 301-320.