Assistant Professor Hassan URAIBI, PhD
Department of Statistics, University of Al-Qadisiyah, IRAQ
E-mail: hssn.sami1@gmail.com , hassan.uraibi@qu.edu.iq
Professor Habshah MIDI,PhD
Faculty of Science and Institute for Mathematical Research
University Putra Malaysia
E-mail: habshah@upm.edu.my

ROBUST VARIABLE SELECTION METHOD BASED ON HUBERIZED LARS-LASSO REGRESSION

Abstract. *The combination of the least absolute deviation (LAD) and the least absolute shrinkage and selection operator (LASSO) (LAD-LASSO regression method) is a popular method to simultaneously perform robust parameter estimation and variable selection. The weighted version of LAD-LASSO is put forward to overcome the problem of LAD-LASSO which is only resistant to outliers in the response variable, but not resistant to outlying observations in the predictor variables (high leverage points (hlps)).Moreover, sometimes these methods lead to overfitting problems or increasing false selection rates with smaller changes in a data. The stability selection methods such as multi-split procedure is proposed to overcome these problems. Unfortunately, this procedure is not resistant to outliers. In this study, we developed a new variable selection method which is called Multi Split Huberized LARS-Lasso (MHLL) by combining the Huberized LARS-Lasso (HLL) and multi-split procedure, under certain conditions to improve stability and predictions. The performance of the proposed MHLL method is assessed extensively by real examples and simulation study. The results indicate that the MHLL is more efficient and reliable than the other two methods.*

Keywords*: Huberized Lasso, LAD-Lasso, WLAD-Lasso, Outliers, adjusted p-value.*

JEL Classification: C51, C52, C55

1. Introduction

The informatics revolution in the last century has led to huge developments in data collection technologies. Researchers in a wide variety of scientific areas have employed these technologies to aggregate large scale of data. Sometimes, one obtains a set of data with many predictor variables but do not know which one to use. By including unimportant explanatory variables may produce less accurate predictions and reduce the efficiency of the resulting estimation. On the contrary, deleting an important predictor may result biased estimates and inaccurate prediction. In this connection, a variable selection technique is very crucial

DOI: 10.24818/18423264/54.3.20.09

technique to be employed to choose the important variables to be included in a model. The aim of variable selection includes accurate prediction, interpretable models, stability and avoiding estimation biased. There are many commonly used variables selection methods in literatures such as all possible subsets, stepwise regression, forward selection and backward elimination. All of these traditional methods are not reliable as they fall short in one or more of these criteria due to using the Ordinary Least Squares (OLS) method which is easily affected by outliers. Moreover, those methods are difficult to apply to high dimensional data in which the number of candidate predictor variables (p) is larger than the number of observations (n).

Tibshirani (1996), proposed least absolute shrinkage and selection operator (Lasso) to tackle the traditional variable selection problem. The Lasso is a popular technique for simultaneous estimation and variable selection where these two procedures are combined in a single minimization problem. Nonetheless, the Lasso is not resistant to outliers because it is a special case of the penalized of the loss function of the OLS subject to L_1 penalty function. To remedy this problem, the Lad-Lasso regression method is developed by combining the Least Absolute Deviation (Lad) and the Lasso methods (seeXu,2005; Wang and Leng, 2007).

It is noted that the Lad-Lasso only resistant to outliers in the response variable. Weighted version of the Lad-Lasso method is then developed by combining the Wlad regression criterion and the adaptive Lasso penalty function which is proposed by Zou (2006), to make the resultant estimator resistant to hlps. Lacroix (2011) proposed a new estimator by combining Huber’s criterion and adaptive lasso penalty whereby this estimator is robust to heavy tailed errors or outliers in the response. The model that selected by this manner should be characterized by accuracy, stability, and interpretability. The respectable efforts have been paid in the statistical literature to develop the approach of penalizing the sum of residual squares. These efforts focus on two directions, first accelerating the stability and the prediction with L_1 penalty, such as boosting(Hastie et al., 2001) andforward stage wise regression (FreundandSchapire,1997). The other direction concentrated on suggesting new penalty functions, for instance, SCAD (FreundandSchapire,1997), adaptive LASSO(Zou,2006), relaxed LASSO(Meinshausen,2007)and the Dantzigselector (Candesand Tao,2005b).

Tibshirani (1996) used linear programming to find the optimal solution path for the lasso. The Least Angle Regression (LARS) of Efronet al. (2004) is a popular method for computing the lasso optimal solution path. Unfortunately, LARS has the ability to rank the most important variables but they do not need to be significant (Khanet al., 2007;Brink-Jensenand Thorn,2014).

According to Tibshiraniet al. (2012), Lasso does not have a unique minimum when the rank of covariates matrix is less than the number of covariates due to selecting some of covariates from a discrete probability distribution. Drastic changes may occur in lasso solution path as a result of small changes that may occur in the original data. Consequently, different lasso results are obtained when

DOI: 10.24818/18423264/54.3.20.09

subsamples are repeatedly drawn from the original data. To overcome the instability of lasso solution path, Meinshausenet al. (2009) proposed resampling by a single split procedure (Wassermanand Roeder,2009), which relies on subsampling technique of Politis and Romano (1994). This approach is capable of making asymptotically correct inference around Lasso coefficients and obtaining the optimal solution path based on the novel adjusted p-values.

Unfortunately, outliers may occur as a result of data collection from different subpopulations (Heterogeneity problem) and then the characterize approximate distribution of Lasso estimator becomes more complex (Wuand Ma,2014). Moreover, the random subsamples procedure gives each observation in a data set the same probability of being chosen to be in the specific subsample, even this observation is identified as outlier. One solution considered to overcome the instability is to propose combining random Multi-split procedure with Huberized LARS- Lasso (HLL) and call it Multi Split Huberized LARS Lasso denoted as $M\mathcal{H}LL$.

Hence in this paper, we propose a new estimator that we call $M\mathcal{H}LL$. The rest of the paper is structured as follows. Section 2 & Section 3 present the $\mathcal{H}LL$ regression and the proposed procedure, respectively. Section 4 illustrates the performance of the proposed method using simulation study.Real data sets are illustrated in section 5 and section 6. The conclusion of this paper is presented in Section 7.

2. The Huberized LARS-LASSO (HLL)

The robust regression parameter estimates are obtained by minimizing a Huber loss function $\mathcal{H}_\delta$ that can be expressed as

$$\min_{\beta} \sum_{i=1}^{n} \mathcal{H}_\delta\left(\frac{r_i}{s}\right) \tag{1}$$

where $s = 1.348\ \mathrm{median}|r_i - \mathrm{median}(r_i)|$, $r_i = y_i - x_i'\beta$ and $\mathcal{H}_\delta$ is a symmetric function with a unique minimum at zero,

$$\mathcal{H}_\delta(t)_i = \begin{cases} t_i^2/2, & \text{if } |t_i| \le \delta \\ \delta\,|t_i| - \delta^2/2, & \text{if } |t_i| > \delta \end{cases} \tag{2}$$

where $t_i = r_i/s$.

Taking partial derivative with respect to β and setting them equal to zero, producing a system of normal equations that can solve this minimization problem. Thus, by letting $\psi(t)$ as the derivative of ρ, we would get

$$\sum_{i=1}^{n} \psi(t)\,x_i = 0,$$

where

$$\psi(t)_i = \begin{cases} t_i, & \text{if } |t_i| \le \delta \\ \delta\,\mathrm{sign}(t_i), & \text{if } |t_i| > \delta \end{cases}. \tag{3}$$

Huber (1981)demonstrated that the asymptotic efficiency of 0.95 at normal errors can be achieve when $\delta = 1.345$. Rosset and Zhu(2007) combined Huber function with ℓ_1 penalty to propose huberized Lasso algorithm which is typically tuned by λ_δ to achieve optimal prediction accuracy. They employed Least angle regression algorithm (Efronet al., 2004), as a piecewise linear solution path of the

DOI: 10.24818/18423264/54.3.20.09

Huberized Lasso to improve the efficiency of the estimates. Consider the Huberized Lasso regression solution as follows,

$$\hat{\beta}(\lambda_\delta) = \sum \mathcal{H}_\delta(t) + \lambda_\delta \sum_{j=1}^{P} \left|\beta(\lambda_\delta)_j\right| \quad (4)$$

by taking the derivative of $\hat{\beta}(\lambda_\delta)$,

$$\frac{\partial\hat{\beta}(\lambda_\delta)}{\partial\lambda_\delta} = -\left[\nabla^2 \sum \mathcal{H}_\delta(t) + \nabla^2\lambda_\delta \sum_{j=1}^{P} \left|\beta_j\right|\right]^{-1} \nabla\lambda_\delta \sum_{j=1}^{P} \left|\beta(\lambda_\delta)_j\right| \quad (5)$$

and then a piecewise constant vector can be computed from the following function, $\frac{\partial\hat{\beta}(\lambda_\delta)}{\partial\lambda_\delta} \Big/ \left\|\frac{\partial\hat{\beta}(\lambda_\delta)}{\partial\lambda_\delta}\right\|$,

therefore λ_δ is a piecewise linear too. The $\hat{\beta}(\lambda_\delta)$ can have nonzero only when the generalize absolute correlation $\left|\nabla(\sum \mathcal{H}_\delta(t))_j\right| = \lambda_\delta$ which involves $\operatorname{sgn}\left(\nabla(\sum \mathcal{H}_\delta(t))_j\right) = -\operatorname{sgn}\left(\hat{\beta}(\lambda_\delta)_j\right)$

The optimal choice of λ_δ relies on the cross validation procedure to generate the solution paths of $\hat{\beta}(\lambda) = \hat{\beta}(\lambda_\delta) - (\lambda - \lambda_\delta)\gamma_\delta$. The procedure starts from one λ_1 and the solution is moving in a linear direction of LARS steps $\gamma_\delta$which is determined by the equiangular path until settle on the optimizing value λ_δ . This procedure involves a very cheap computation step and the $\hat{\beta}(\lambda)$ solution is a piecewise linear and monotonically decreasing in λ_δ.

3. Random Multi-split Huberized LARS-Lasso($\mathbf{M\mathcal{H}LL}$)

To improve the performance of the single split random selection technique, a stability selection or random multi-split procedure is proposed. This procedure repeatedly split randomly the data into two subsamples with equal size of n/2 for a certain number of times (at least 50 times) whereby the dimensional reduction of Huberized LARS- Lasso ($\mathcal{H}$LL) and the MM-estimatorYohai(1987) are applied to the first and the second subsamples, respectively and a set of p-values of regression coefficients are recorded.

This procedure combines all sets of p-values and only considers those variables with significant p values to be included in the final model. The details explanation of multi split procedures can be found in some literatures (Buhlmannet al., 2013;Zhang and Zhang,2014;Lockhartet al., 2014; Van de Geer et al., 2014;Javanmardand Montanari,2013;Uraibi,2019;Uraibiet al., 2017a;Uraibiet al, 2015;Uraibiet al., 2017). Consider linear regression equation,

$$Y = X\beta + \varepsilon, \quad (6)$$

where Y is an $(n \times 1)$response vector, X is $(n \times p)$fixed design matrix of independent variables, β is an $(p \times 1)$ regression parameters vector and ε is an $(n \times 1)$ random errors vector with $\text{iid.}\,N(0, \sigma^2)$.

Let B is the total number of times of the random splitting of original data such that $b = 1, \ldots, B$. The Multi-split algorithm of (Yohai,1987) is summarized as follows:

Step 1. For $b = 1, \ldots, B$

1. Let the full dataset denoted as $D_{\text{Full}} = [Y_{\text{Full}}, X_{\text{Full}}]$ be randomly splits into two disjoint groups of equal size $(n/2)$where the first and the second

DOI: 10.24818/18423264/54.3.20.09

groups are denoted as $D_{in}^{(b)} = (Y_{in}, X_{in})$ and $D_{out}^{(b)} = (Y_{out}, X_{out})$, respectively.

2. Identify leverage points and outliers individually for $D_{in}^{(b)}$ and $D_{out}^{(b)}$,as follows

   a. Identify the residuals outliers by using robust three sigma rule,

$$z = \frac{\hat{\varepsilon} - \mathrm{Med}(\hat{\varepsilon})}{\mathrm{MAD}(\hat{\varepsilon})}$$

   where $\hat{\varepsilon}$ is the residuals vector of Least Median of Squares (LMS), and it is considered an outlier if$z > 2.5$.

   b. The Robust Mahalanobis Distance (RMD) based on robust location and scatter matrix (such as MCD, MVE) is used with fixed design matrix of independent variables X to identify leverage points. Observations corresponding to the jth row with $\mathrm{RMD_j} > \chi^2_{(0.975,p)}$ is considered as leverage point. If the identified percentage of outliers or leverage points exceeds 50%, discards the whole subgroups and repeat step 1.

3. Let $\tilde{S}_{\mathcal{H}}^{(b)}$ be the estimates of $\beta(\lambda)$ of the set of active covariates based on $D_{in}^{(b)}$ subsample data such that $\tilde{S}_{\mathcal{H}LL}^{(b)} = \{j; \hat{\beta}_j^{\mathcal{H}LL} \neq 0\}$ and $N = \tilde{S}_{\mathcal{H}LL}^{c(b)} = \{j; \hat{\beta}_j^{\mathcal{H}LL} = 0\}$ (where c refers to complement) is the set of inactive covariates. $\hat{\beta}(\lambda)$is computed using $\mathcal{H}LL$(Huberized LARS-Lasso).

4. Employ the MM-estimator (Yohai,1987) to estimate the parameters of the set of active predictors in $\tilde{S}_{\mathcal{H}LL}^{(b)}$ for the $D_{out}^{(b)}$ subsample data and calculate the corresponding p-values as follows,

$$\tilde{P}_{MMj}^{(b)} = \begin{cases} \tilde{P}_{MMj}^{(b)} & \text{if} \quad j \in \tilde{S}_{\mathcal{H}LL}^{(b)} \\ 1 & \text{if} \quad j \notin \tilde{S}_{\mathcal{H}LL}^{(b)} \end{cases} \tag{7}$$

   and then without aggregated, adjusted $\tilde{P}_{MMj}^{(b)}$ values as

$$\hat{P}_{MMj}^{(b)} = \min\left(\tilde{P}_{MMj}^{(b)} \left|\tilde{S}_{\mathcal{H}LL}^{(b)}\right|, 1\right), j = 1,2, \dots, p \tag{8}$$

Step2. The B vectors of $\hat{P}_{MMj}^{(b)}$ for each predictor X_j are obtained from Step1. For any fixed $\gamma \in (0,1)$ with lower bound at least equals to 0.05, (Meinshausenet al.,2009), defined Qj () in Equation 9 as follows,

$$Q_j(\gamma) = \min\left\{1, q_\gamma\left(\left\{\frac{\hat{P}_{MMj}^{(b)}}{\gamma}; b = 1, \dots, B\right\}\right)\right\} \tag{9}$$

where $q_\gamma(.)$ is the empirical quantile function.

DOI: 10.24818/18423264/54.3.20.09

The choice of γ entails additional correction to control the Family-wise Error (FWER) rate at level α through the correction factor $1 - \log(\gamma_{\min})$ with upper bound equals 4. Subsequently, the robust adjusted p-values is given by,

$$P_j^{rob} = \min\left\{1, 1 - \log(\gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma)\right\} \tag{10}$$

The final model will include only those variables that possess P_j^{rob} which is not equal to one.

4. Simulation Study

In this section, we report a simulation study that is designed to investigate the performance of our proposed MHLL compared to LAD-Lasso and WLAD-Lasso. Here, we consider three different simulation scenarios.

Simulation1: The first simulation considers multiple linear regression with sample of size 50 $(n = 50)$ and 25 covariates $(p = 25)$ where each of the covariate is drawn from joint Gaussian marginal distribution with correlation structure $\rho = 0.5$. The true regression parameters β is set to be $\beta = \left(\underbrace{1,2,3,4,5}_{5}, \underbrace{0,0,0,0,0}_{20}\right)$. The distribution of random errors e is generated from the following contamination model,

$$F(e) = [(1-\varepsilon)N(0,1) + \varepsilon\, H(0,2)] \times \sigma$$

where ε is the contamination ratio, σ is a signal to noise which is chosen to be 3, N is standard normal distribution and H is Cauchy distribution to create heavy –tailed distribution. The variables are contaminated by certain ratio ($\varepsilon = \{0.05, 0.10, 0.15$ and $0.20\}$ of high leverage points. The high leverage points are created by replacing randomly some original observations with large values equals to 15. Subsequently, the respond variables are computed as follows,

$$y_{(50\times1)} = X_{(50\times25)} \times \left(\underbrace{1,2,3,4,5}_{5}, \underbrace{0,0,0,0,0}_{20}\right)^t + F(e)_{(50\times1)}$$

Simulation 2: The second simulation is similar to the first simulation process, except for different values of p and n $(p = 50\ , n = 150)$ and the respond variables are calculated as follows,

$$y_{(100\times1)} = X_{(150\times50)} \times \left(\underbrace{5,0,0,0,8}_{5}, \underbrace{0,1.5,0,0,3}_{5}, \underbrace{0,5,0,0,0}_{5}, \underbrace{0,\dots,0}_{35}\right)^t + F(e)_{(150\times1)}$$

Simulation 3: The third simulation scenario is similar to the second simulation $(p = 50)$ with a slight change where n is increased to 200 and $\beta = \left(\underbrace{4,0,0,0,3.5,0}_{6}, \underbrace{0,0,2.5,0,0}_{5} \underbrace{0,5,0,0,0,4.5}_{6}, \underbrace{0,\dots,0}_{33}\right)$. Afterwards the dependent variables are computed as follows

$$y_{(500\times1)} = X_{(500\times50)} \times (\beta)^t + F(e)_{(500\times1)}$$

The $M\mathcal{H}$LL, LAD-Lasso and WLAD-Lasso were then applied to the simulated data. In each simulation runs, there were 5000 replications. Four criteria are

DOI: 10.24818/18423264/54.3.20.09

considered to evaluate the performances of the three methods, namely: (1) the percentage of zero coefficients (Zero.coef), (2) the percentage of non-zero true coefficients (N.Zero.coef), (3) the average of mean squares of errors ($\overline{\text{mse}}$) and (4) the Median of mean squares of errors Med(mse). A good method is the one that possesses the lowest percentage of Zero.coef and the highest percentage of N.Zero.coef which is reasonably closed to 0% and the 100%, respectively, and having the least ($\overline{\text{mse}}$) and Med(mse)values.

Several interesting points appear from the results of Table 1. The results clearly show the merit of MHLL. It can be observed from Table 1 that the percentage of Zero.coef and N.Zero.coef of the $\mathrm{M}\mathcal{H}\mathrm{LL}$ are fairly closed to 0% and 100%, respectively and having the smallest values of ($\overline{\text{mse}}$) and Med(mse), followed by, WLAD-Lasso and LAD-Lasso. For example, for simulation 3, with 5% contamination, the $\mathrm{M}\mathcal{H}\mathrm{LL}$ successfully select 100% 0fN.Zero.coef and 0% zero.coef , while WLAD-Lasso suffers from underfitting problems (it selects 23% of N.Zero.coef and 77% Zero.coef) over 5000 replications with $\overline{\text{mse}}$ and Med(mse) greater than their counterparts of $\mathrm{M}\mathcal{H}\mathrm{LL}$. The results are consistent for all percentages of contaminations. This indicates that the performance of $\mathrm{M}\mathcal{H}\mathrm{LL}$ is more efficient than the other two methods.

Table 1: The percentage of Zero and non-zero coefficients, Average of MSE and median of MSE for three scenarios of Simulations

		Simulation 1			Simulation 2			Simulation 3		
ε	Model	LAD-Lasso	WLAD-Lasso	$\mathrm{M}\mathcal{H}\mathrm{LL}$	LAD-Lasso	WLAD-Lasso	$\mathrm{M}\mathcal{H}\mathrm{LL}$	LAD-Lasso	WLAD-Lasso	$\mathrm{M}\mathcal{H}\mathrm{LL}$
5%	Z.Coef	0.44	0.40	0.16	0.78	0.75	0.00	0.78	0.77	0.00
	NZ.Coef	0.56	0.60	0.84	0.22	0.25	1.00	0.22	0.23	1.00
	$\overline{\text{mse}}$	78.03	76.02	3.80	143.92	11.45	10.42	1686.1	1551.3	1505.4
	Med(mse)	18.40	18.11	3.08	148.15	2.87	2.62	174.28	4.90	4.74
10 %	Z.Coef	0.44	0.43	0.15	0.78	0.77	0.00	0.78	0.78	0.00
	NZ.Coef	0.56	0.57	0.85	0.22	0.23	1.00	0.22	0.22	1.00
	$\overline{\text{mse}}$	997.52	730.96	6.68	184.92	165.45	15.42	310.05	147.28	140.95
	Med(mse)	21.33	21.31	5.54	184.20	10.46	8.88	182.49	11.40	10.92
15 %	Z.Coef	0.45	0.48	0.17	0.78	0.79	0.00	0.78	0.76	0.00
	NZ.Coef	0.55	0.52	0.83	0.22	0.21	1.00	0.22	0.24	1.00

	$\overline{\text{mse}}$	3349.63	2817.97	6.72	287.20	178.82	14.60	372048	365874.6	348725
	Med(mse)	23.01	22.33	6.02	144.18	7.72	7.16	182.09	14.80	14.27
20 %	Z. Coef	0.44	0.52	0.19	0.78	0.81	0.01	0.41	0.55	0.00
	NZ. Coef	0.56	0.48	0.81	0.22	0.19	0.99	0.59	0.45	1.00
	$\overline{\text{mse}}$	2469	2255	7.75	516.8	205.5	485.5	311123	299457	25013
	Med(mse)	26.26	25.19	6.87	171.50	20.27	17.79	180.12	14.95	10.84

5. Modified Hawkins BraduKass Data

In order to evaluate the performance of $\mathrm{M}\mathcal{H}\mathrm{LL}$, an artificial data set constructed by Hawkins et al. [32,33] is used. This data set containing 75 observations of one response and three explanatory variables in which the first ten observations (cases 1-10) are constructed as bad leverage points and (cases 11-14) as good leverage points. Since the positions of outliers of this data set are exactly known, it has been enormously employed by many researchers to demonstrate the robustness of robust regression techniques. The idea of Artificial data is taken from Arslan (2012) who modified the Hawkins data set.

We modified the dimension of original data to have 120 covariates. The first three covariates are collected from the original dataset and the remaining covariates are generated from standard normal distribution. The new design matrix X is divided into two sub matrices X_1 and X_2 in which X_1 having the first 10 samples and X_2 involves the remaining 65 samples. Mathematically, this procedure can be written as follows,

$$\mathrm{X} = \begin{bmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \end{bmatrix},$$

where $\mathrm{X}_1 = [\mathrm{x}_1 \quad \mathrm{x}_2 \quad \cdots \quad \mathrm{x}_{120}]_{(10\times 120)}$ and $\mathrm{X}_2 = [\mathrm{x}_1 \quad \mathrm{x}_2 \quad \cdots \quad \mathrm{x}_{120}]_{(65\times 120)}$

To exclude the outliers from the first 10 samples and keeping only the bad leverage points, the response variable Y_1 is computed as follows,

$$\mathrm{Y}_{1(10\times 1)} = \mathrm{X}_{1(10\times 120)} \beta_{1.(120\times 1)},$$

where $\beta_{1.(120\times 1)} = \Big(\underbrace{-1,-1,0,\dots,0}_{120}\Big)^{t}$

On the other hand, the outliers are considered with the remaining 65 samples by using the following formula,

$\mathrm{Y}_{2(65\times 1)} = \mathrm{X}_{2(65\times 120)} \beta_{2.(120\times 1)} + \varepsilon_{(65\times 1)}$,

where, $\beta_{2.(120\times 1)} = \Big(\underbrace{2,2,0,\dots,0}_{120}\Big)^{t}$

and $\varepsilon_{(65\times 1)}$ is distributed as Cauchy distribution.

Consequentially, $Y_{(1\times 75)} = \begin{bmatrix} Y_{1(10\times 1)} \\ Y_{2(65\times 1)} \end{bmatrix}$.

The $M\mathcal{H}LL$ method along with LAD-Lasso and WLAD-Lasso were then applied to the modified Hawkins dataset to identify the zero, non-zero coefficients and the false selection variable. The best method is the one that includes non-zero coefficients of x_1 and x_2 covariates and having the least false selection rate. It can be observed from Table 2 that LAD-Lasso selects 108 and 12 zeros and non-zeros coefficients, respectively. The 12 covariates with non-zero coefficients included 10 false selection covariates. The WLAD-Lasso shows 111 covariates with zero coefficients, 9 covariates with non-zero coefficients are selected with 0.058 false selection rate. It is very interesting to observe that our proposed $M\mathcal{H}LL$ chooses three potential covariates with 117 non-zero coefficients, having the least probability of false selection (0.008).

Table 2: the number of Zero and Non-zero coefficients and number and percentage of False selection of covariates for four methods with modified Data.

	LAD-Lasso	WLAD-Lasso	$M\mathcal{H}LL$
Z. Coef	108	111	117
NZ. Coef	12	9	3
False selection variable	10 (0.083)	7 (0.058)	1 (0.008)

6. Hand Grip Strength data

Our second example is the Malaysian Hand Grip Data. Hand grip strength is a crucial measure employed to evaluate hand disorders and injuries and to monitor the progression of recovery and so on [35]. The original data of right hand grip strength is obtained from [36] where in our study we consider 304 (196 men and 108 women) healthy volunteers of staff, medical students and visitors of University of Malaya Medical Centre between January and April. It is noted that grip strength is influenced by a number of factors such as Age, Height, Weight, Right Upper arm circumference, and BMI of each subject. In this study we wish to investigate which of the five independent variables should be included in the final model for the prediction of grip strength of adult Malaysian. As such, we will apply our MHLL variable selection method to identify a few ‘best’ subsets X that will increase model predictive ability. Here, we separately analyses the data by gender.

DOI: 10.24818/18423264/54.3.20.09

6.1 Males Hand Grip Strength data

Firstly, we want to investigate whether the data has outliers (outlying observations in Y direction), leverage points (outlying observations in X direction), and influential observations.

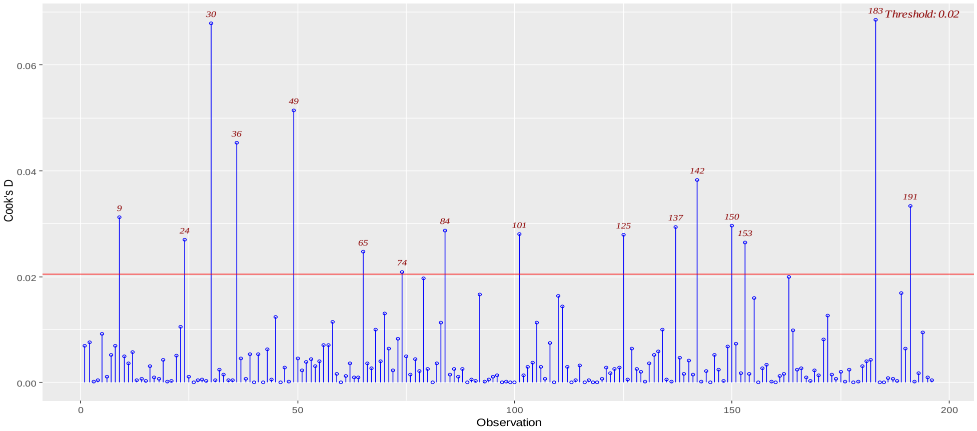

Figure 1. Cook's Distance with 0.02 as threshold for Males Hand Grip Strength data

Diagnostic methods namely the hat matrix (with threshold equals 0.061), the R-student (with cut-off point equals 2.5) and Cook's distance (with threshold equals 0.02) are employed for the identification of outliers, leverage points and influential observations. It can be observed from Figure 1 that 16 influential observation are detected while Figure 2 shows that this data has 19 leverage points and 4 outliers. Moreover, the plot of correlation matrix of the five covariates in Figure 3 indicates that Weight and right Upper arm circumference, are strongly correlated with BMI.

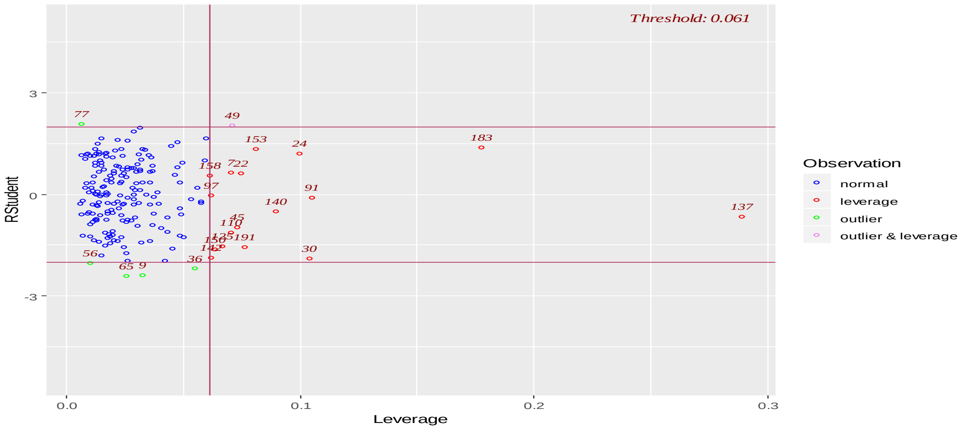

Figure 2. Identified leverage point and outliers in Males Hand Grip Strength data

DOI: 10.24818/18423264/54.3.20.09

Table 3 present the results of the three variable selection methods. It is obvious that LAD-Lasso and WLAD-Lasso select 5 non-zero coefficients with higher RMSE values, while $M\mathcal{H}LL$ select 2 non-zero coefficients (Height and right Upper arm circumference). The values of the Root of Mean Squares Error (RMSE) of $M\mathcal{H}LL$is smaller than the LAD-Lasso and WLAD-lasso. Hence for this data set, the $M\mathcal{H}LL$is more reliable and selects Height and right Upper arm circumference to be included in the model.

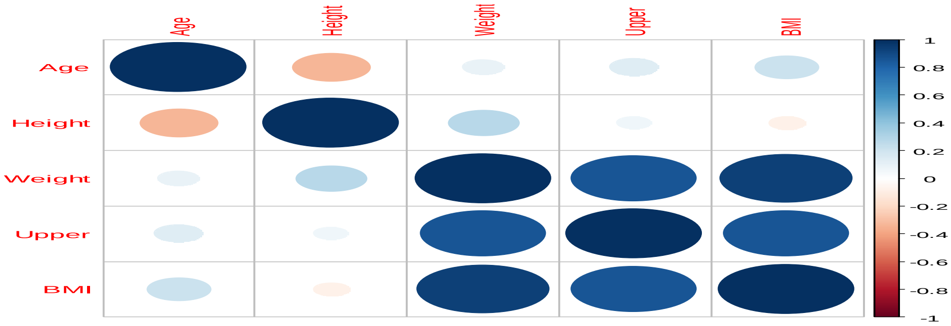

Figure 3.The correlation matrix of five covariates (Age, Height, Weight, Upper arm circumference right, and BMI) of Males Hand Grip Strength data

Table 3: The number of Zero and Non-zero coefficients and the root of mean of mean squares errors for four methods of Male Hand Grip Strength Data

	LAD-Lasso	WLAD-Lasso	$M\mathcal{H}LL$
Z. Coef	0	0	3
NZ. Coef	5	5	2
RMSE	10.87	9.60	9.38

6.2 Females Hand Grip Strength data

We employed the same diagnostic procedures as used in the previous data to the female hand grip data. Figure 4 shows that the women handgrip data has seven influential observations.

This data set also has 11 leverage points and 4 outliers as shown in Figure 5.

DOI: 10.24818/18423264/54.3.20.09

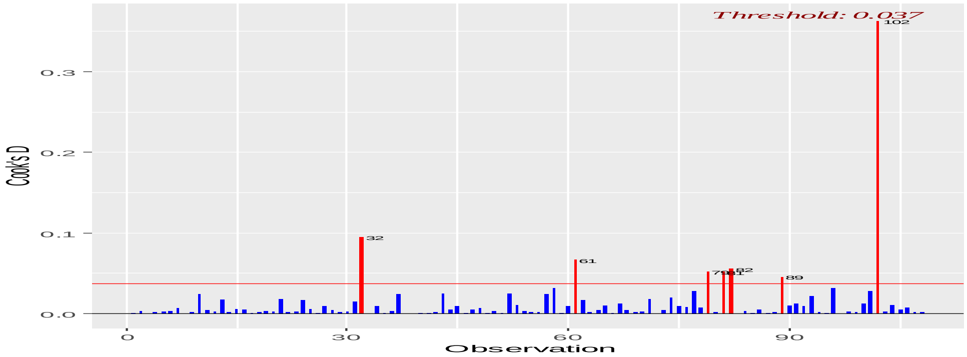

Figure 4. The results of Cook's Distance for Females Hand Grip Strength data

Similar to the Male Hand Grip Strength data, the Weight and the right Upper arm circumference of the women hand grip data, are strongly correlated with BMI as shown in Figure 6.

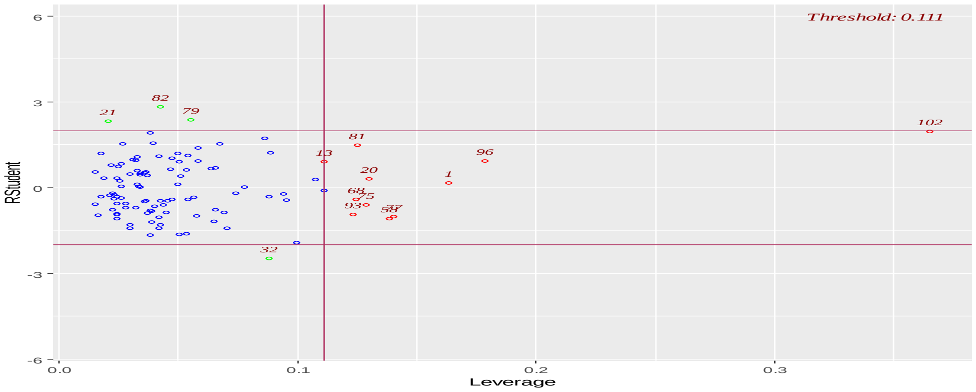

Figure 5. identified leverage point and outliers in Females Hand Grip Strength data

Table 4: the number of Zero and Non-zero coefficients and the root of mean of squares errors for four methods of Female Hand Grip Strength Data

	LAD-Lasso	WLAD-Lasso	$M\mathcal{H}LL$
Z. Coef	1	0	3
NZ. Coef	4	5	2
RMSE	9.48	8.68	7.85

DOI: 10.24818/18423264/54.3.20.09

The results of the three variable selection methods are exhibited in Table 4. The results signify that the LAD-Lasso method excluded only one covariate (Age), and the WLAD-Lasso included all covariates. It is interesting to see that the $M\mathcal{H}LL$ method selects Height and right Upper arm circumference with non-zero coefficients and having the least value of RMSE. The results show that the performance of $M\mathcal{H}LL$ is more efficient and more reliable than the other methods considered in this study.

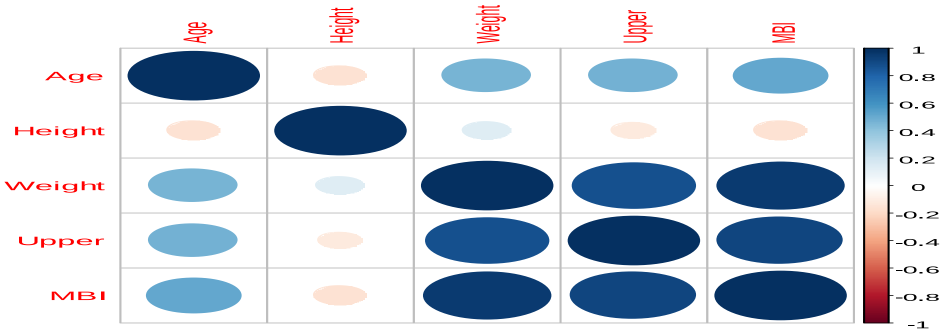

Figure 6: The correlation matrix of five covariates (Age, Height, Weight, Upper arm circumference right, and BMI) of Females Hand Grip Strength data

7. Conclusions

The main purpose of this paper was to develop a reliable alternative approach for improving lasso solution path. In this study, we proposed a multi split Huberized Lars-Lasso solution path by combining Huberized Lars-Lasso with multiple split procedure whereby overfitting problem is controlled using adjusted p-value. We have compared the $M\mathcal{H}LL$ with two other estimators namely the LAD-Lasso and WLAD-Lasso. The LAD-Lasso is not reliable at all and the WLAD-Lasso is not any better either. The simulation experiments and empirical studies signify that the $M\mathcal{H}LL$ offers a remarkable improvement over the LAD-Lasso and WLAD-Lasso.

The $M\mathcal{H}LL$ can significantly select the potential variables in the final model with the least value of RMSE and least probability of selecting false variables. The LAD-Lasso and WLAD-Lasso are not capable of selecting the correct variables in the final model having considerably large probability of false variable selection and large RMSE. Thus, we can contemplate that our proposed $M\mathcal{H}LL$ technique as a sound variable selection method and highly suggest employing this method particularly when outliers and high leverage points are present in a data.

REFERENCES

[1]**Arslan, Olcay (2012),*Weighted LAD-LASSO Method for Robust Parameter Estimation and Variable Selection in Regression.*** *Computational Statistics & Data Analysis,56(6);*
[2] **Bohannon R. W. (2001), *Dynamometer Measurements of Hand-Grip Strength Predict Multiple Outcomes.*** *Perceptual and Motor Skills, 93(2):323-328;*
[3] **Brink-Jensen, K., Thorn Ekstrm, C. (2014),*Inference for Feature Selection Using the Lasso with High-dimensional data .*** *eprintarXiv 1403.4296;*
[4] **Buhlmann, P. and others (2013),*Statistical Significance in High-dimensional Linear Models.*** *Bernoulli, 19(4);*
[5] **Candes, E. and Tao, T. (2005b),*The Dantzig Selector: Statistical Estimation when p is Much Larger than n.*** *Ann. Statist. 35 2313–2351;*
[6] **Efron, B., Hastie, T. & Johnstone, I. (2004), *Least Angle Regression.*** *The Annals of Statistics, 32(2), 407–499;*
[7] **Fan, J. & Li, R. (2001),*Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties.*** *Journal of the American Statistical Association, 96(456), 1348-1360;*
[8] **Freund Y. and Schapire R. (1997),*A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.*** *Journal of Computer and System Sciences,55(1), Pages 119-139;*
[9] **Hastie, T., Tibshirani, R. & Friedman, J. (2001),*The Elements of Statistical Learning; Data Mining, Inference and Prediction.*** *Springer Verlag, New York;*
[10] **Huber, P. J. (1981),*Robust Statistics.*** *New York: John Wiley and Sons, Inc;*
[11] **Javanmard, A. and Montanari, A. (2013),*Model Selection for High Dimensional Regression under the Generalized Irrepresentability Condition.*** *In Advances in Neural Information 2013. Processing Systems 26;*
[12] **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(480);*
[13] **Lacroix, S. (2011),*Robust Regression through the Hubers Criterion and Adaptive Lasso Penalty.*** *Electronic Journal of Statistics,5;*
[14] **Lockhart, R.,Taylor, J., Tibshirani, R. J. and Tibshirani, R.(2014),*A Significance Test for the Lasso.*** *Annals of Statistics,42(2);*
[15] **M. G. Hossain, R. Zyroul B. P. Pereira, T. Kamarul (2011),*Multiple Regression Analysis of Factors Influencing Dominant Hand Grip Strength in an Adult Malaysian Population.*** *Journal of Hand Surgery (European Volume), vol. 37, 1: pp. 65-70;*

DOI: 10.24818/18423264/54.3.20.09

[16] **Meinshausen, N. (2007),*Relaxed Lasso.***Comput. Statist. Data Anal. 52 374–393;*

[17] **Meinshausen, N., Meier, L. and Bu¨hlmann, P. (2009),*P-values for High-dimensional Regression.*** *Journal of the American Statistical Association, (104);*

[18] **Politis, D. N and Romano, J. P. (1994),*Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions.*** *The Annals of Statistics,2031–2050;*

[19] **Rosset, S. and Zhu, J. (2007),*Piecewise Linear Regularized Solution Paths.*** *The Annals of Statistics, 1012–1030;*

[20] **Rousseeuw, P. J. (1984),*Least Median of Squares Regression.*** *Journal of the American Statistical Association, 79, 871-880;*

[21] **Rousseeuw, P.J. and Yohai, V. J. (1984),*Robust Regression by means of S-Estimators Robust and Nonlinear Time Series Analysis.*** *Lecture Notes in Statistics. Heidelberg, Germany;*

[22] **Tibshirani, R. Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani, R J. (2012),*Strong Rules for Discarding Predictors in Lasso-Type Problems.*** *Journal of the Royal Statistical Society, Series B, 74(2);*

[23] **Tibshirani, R. (1996), *Regression Shrinkage and Selection via the Lasso.*** *Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288;*

[24] **Uraibi, H. S. (2019),*Roust Lasso Regression based on Weighted Subsamples p-values.*** *Electronic Journal of Applied Statistical Analysis, 12(1), in print;*

[25] **Uraibi, H. S., Midi, H. and Rana, S. (2015),*Robust Stability Best Subset Selection for Autocorrelated Data Based on Robust Location and Dispersion Estimator.*** *Journal of Probability and Statistics;*

[26] **Uraibi, H. S., Midi, H. and Rana, S. (2017a),*Selective Overview of Forward Selection in Terms of Robust Correlations.*** *Communications in Statistics-Simulation and Computation, 46(7);*

[27] **Uraibi, H. S., Midi, H. and Rana, S. (2017),*Robust multivariate least angle regression.***SCIENCEASIA, 43(1);*

[28] **Van de Geer, S., Bu¨hlmann, P., Ritov, Y. and Dezeure, R. and others, (2014),*On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models.*** *The Annals of Statistics,42(3);*

[29] **Wang, H., Leng, C. (2007),*Unified Lasso Estimation by Least Squares Approximation.*** *Journal of the American Statistical Association102,1039–1048;*

[30] **Wasserman, L. and Roeder, K. (2009),*High Dimensional Variable Selection.*** *Annals of Statistics, 37(5A);*

[31] **Wu, Cen and Ma, Shuangge (2014),*A Selective Review of Robust Variable Selection with Applications in Bioinformatics.*** *Briefings in bioinformatics, 16(5);*

[32] **Xu, J., Ying, Z. (2010),*Simultaneous Estimation and Variable Selection in Median Regression Using Lasso-Type Penalty.*** *Annals of the Institute of Statistical Mathematics62,487–514;*

[33] **Xu,J.(2005),*Parameter Estimation, Model Selection and Inference in L1 Based Linear Regression.*** *Unpublished PhD Thesis, Columbia University;*

DOI: 10.24818/18423264/54.3.20.09

[34] **Yohai, V. J. (1987),*High Breakdown-Point and High Efficiency Robust Estimates for Regression.*** *The Annals of Statistics, 642-656;*
[35] **Zhang, C. H. and Zhang, S. (2014),*Confidence Intervals for Low Dimensional Parameters in High-Dimensional Linear Models.*** *Journal of the Royal Statistical Society, Series B, 76;*
[36] **Zou, H. (2006),*The Adaptive Lasso and Its Oracle Properties.*** *Journal of the American Statistical Association, 101(476), 1418-1429,* doi:10.1198/016214506000000735.