Assistant Professor Hassan URAIBI, PhD  
Department of Statistics, University of Al-Qadisiyah, IRAQ  
E-mail: hssn.samil@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.
Hassan Uraibi, Habshah Midi

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 (see Xu, 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 (Hastic et al., 2001) and forward
stage wise regression (Freund and Schapire, 1997). The other direction concentrated
on suggesting new penalty functions, for instance, SCAD (Freund and Schapire, 1997),
adaptive LASSO (Zou, 2006), relaxed LASSO (Meinshausen, 2007) and the Dantzig
selector (Candes and Tao, 2005b).

Tibshirani (1996) used linear programming to find the optimal solution path for the
lasso. The Least Angle Regression (LARS) of Efron et 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 (Khan et al., 2007; Brink-Jensen and Thorn, 2014).

According to Tibshirani et 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, Meinshausen et al. (2009) proposed resampling by a single split procedure (Wasserman and 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 (Wu and 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 $\mathcal{MHL}$.

Hence in this paper, we propose a new estimator that we call $\mathcal{MHL}$. The rest of the paper is structured as follows. Section 2 & Section 3 present the $\mathcal{HLL}$ 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

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

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

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

where $t_i = r_i/s$.

Taking partial derivative with respect to $\beta$ 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 $\rho$, we would get

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

where

$$\psi(t)_i = \begin{cases} t_i, & \text{if } |t_i| \leq \delta \\ \delta \text{sign}(t_i), & \text{if } |t_i| > \delta \end{cases}$$

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 $\ell_1$ penalty to propose huberized Lasso algorithm which is typically tuned by $\lambda_\delta$ to achieve optimal prediction accuracy. They employed Least angle regression algorithm (Efron et 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,

\[
\hat{\beta}(\lambda_\delta) = \sum \mathcal{H}_\delta(t) + \lambda_\delta \sum_{j=1}^p |\beta(\lambda_\delta)|
\]

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

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

and then a piecewise constant vector can be computed from the following function,

\[
\frac{\partial \hat{\beta}(\lambda_\delta)}{\partial \lambda_\delta} \left/ \frac{\partial \hat{\beta}(\lambda_\delta)}{\partial \lambda_\delta} \right|.
\]

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

The optimal choice of \(\lambda_\delta\) 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 \(\lambda_1\) 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 \(\lambda_\delta\). This procedure involves a very cheap computation step and the \(\hat{\beta}(\lambda)\) solution is a piecewise linear and monotonically decreasing in \(\lambda_\delta\).

3. Random Multi-split Huberized LARS-Lasso (MHLL)

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 (HLL) and the MM-estimator (Yohai, 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 (Buhlmann et al., 2013; Zhang and Zhang, 2014; Lockhart et al., 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2013; Uraibi, 2019; Uraibiet al., 2017a; Uraibiet al., 2015; Uraibiet al., 2017). Consider linear regression equation,

\[
Y = X\beta + \epsilon,
\]

where \(Y\) is an \((n \times 1)\)response vector, \(X\) is \((n \times p)\) fixed design matrix of independent variables, \(\beta\) is an \((p \times 1)\) regression parameters vector and \(\epsilon\) is an \((n \times 1)\) random errors vector with 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
Robust Variable Selection Method Based on Huberized LARS-Lasso Regression

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,
   \[ z = \frac{\hat{\varepsilon} - \text{Med}(\hat{\varepsilon})}{\text{MAD}(\hat{\varepsilon})} \]
   where $\hat{\varepsilon}$ is the residuals vector of Least Median of Squares (LMS), and it is considered an outlier if $iz > 2.5$.
   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 $j$th row with $\text{RMD}_j > \chi^2_{0.975, p}$ 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 $\tilde{S}_{HLL}^{(b)}$ be the estimates of $\beta(\lambda)$ of the set of active covariates based on $D_{in}^{(b)}$ subsample data such that $\tilde{S}_{HLL}^{(b)} = \{ j; \tilde{\beta}_{HLL}^{(b)} \neq 0 \}$ and $N = \tilde{S}_{HLL}^{(c)} = \{ j; \tilde{\beta}_{HLL}^{(c)} = 0 \}$ (where $c$ refers to complement) is the set of inactive covariates. $\hat{\beta}(\lambda)$ is computed using $HLL$(Huberized LARS-Lasso).

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

\[ \tilde{P}_{\text{MMj}}^{(b)} = \begin{cases} \hat{P}_{\text{MMj}}^{(b)} & \text{if } j \in \tilde{S}_{HLL}^{(b)} \\ 1 & \text{if } j \notin \tilde{S}_{HLL}^{(b)} \end{cases} \] (7)

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

\[ \tilde{P}_{\text{MMj}}^{(b)} = \min \left( \tilde{P}_{\text{MMj}}^{(b)} \left| \tilde{S}_{HLL}^{(c)}, 1 \right. \right), j = 1, 2, \ldots, p \] (8)

**Step2.** The $B$ vectors of $\tilde{P}_{\text{MMj}}^{(b)}$ 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, (Meinshausen et al.,2009), defined $Q_j (\cdot)$ in Equation 9 as follows,

\[ Q_j (\gamma) = \min \left\{ 1, q\gamma \left( \left( \frac{p_{\text{MMj}}^{(b)}}{\gamma}; b = 1, \ldots, B \right) \right) \right\} \] (9)

where $q\gamma(\cdot)$ 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 $1 - \log(Y_{min})$ with upper bound equals 4. Subsequently, the robust adjusted p-values is given by,

$$P_{j}^{rob} = \min \left\{ 1, 1 - \log(Y_{min}) \right\} \inf_{\gamma \in (Y_{min}, 1)} Q_{j}(\gamma)$$

(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.

Simulation 1: 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 $\beta$ is set to be $\beta = (1,2,3,4,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$. The distribution of random errors $\epsilon$ is generated from the following contamination model,

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

where $\epsilon$ is the contamination ratio, $\sigma$ 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 ($\epsilon = \{0.05, 0.10, 0.15$ 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,

$$y_{(50 \times 1)} = x_{(50 \times 25)} \times \left(1,2,3,4,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\right)^{t} + F(\epsilon)_{(50 \times 1)}$$

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,

$$y_{(100 \times 1)} = x_{(150 \times 50)} \times \left(5,0,0,0,8,0,1,5,0,0,3,5,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\right)^{t} + F(\epsilon)_{(150 \times 1)}$$

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 $\beta = \left(4,0,0,0,3,5,0,0,0,2,5,0,0,5,0,0,0,4,5,0,0,3,5,0,0,0,0,0,0,0\right)$. Afterwards the dependent variables are computed as follows,

$$y_{(500 \times 1)} = x_{(500 \times 50)} \times (\beta)^{t} + F(\epsilon)_{(500 \times 1)}$$

The MHLL, LAD-Lasso and WLAD-Lasso were then applied to the simulated data. In each simulation runs, there were 5000 replications. Four criteria are
Robust Variable Selection Method Based on Huberized LARS-Lasso Regression

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 (\( \bar{mse} \)) 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 (\( \bar{mse} \)) 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 MHLL are fairly closed to 0% and 100%, respectively and having the smallest values of (\( \bar{mse} \)) and Med(mse), followed by, WLAD-Lasso and LAD-Lasso. For example, for simulation 3, with 5% contamination, the MHLL successfully select 100% of N. 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 \( \bar{mse} \) and Med(mse) greater than their counterparts of MHLL. The results are consistent for all percentages of contaminations. This indicates that the performance of MHLL 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

<table>
<thead>
<tr>
<th>ε</th>
<th>Model</th>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LAD-Lasso</td>
<td>WLAD-Lasso</td>
<td>MHLL</td>
</tr>
<tr>
<td>5%</td>
<td>Z. Coef</td>
<td>0.44</td>
<td>0.40</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>N. Z. Coef</td>
<td>0.56</td>
<td>0.60</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>mse</td>
<td>78.0</td>
<td>76.02</td>
<td>3.80</td>
</tr>
<tr>
<td></td>
<td>Med(mse)</td>
<td>18.43</td>
<td>18.11</td>
<td>3.08</td>
</tr>
<tr>
<td>10%</td>
<td>Z. Coef</td>
<td>0.44</td>
<td>0.43</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>N. Z. Coef</td>
<td>0.56</td>
<td>0.57</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>mse</td>
<td>997.52</td>
<td>730.96</td>
<td>6.68</td>
</tr>
<tr>
<td></td>
<td>Med(mse)</td>
<td>21.33</td>
<td>21.31</td>
<td>5.54</td>
</tr>
<tr>
<td>15%</td>
<td>Z. Coef</td>
<td>0.45</td>
<td>0.48</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>N. Z. Coef</td>
<td>0.55</td>
<td>0.52</td>
<td>0.83</td>
</tr>
</tbody>
</table>

DOI: 10.24818/18423264/54.3.20.09
5. Modified Hawkins BraduKass Data  
In order to evaluate the performance of M3LL, 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,

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

where $X_1 = [x_1 \ x_2 \ \ldots \ x_{120}]_{(10 \times 120)}$ and $X_2 = [x_1 \ x_2 \ \ldots \ x_{120}]_{(65 \times 120)}$.

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,

$$Y_1_{(10 \times 1)} = X_{1(10 \times 120)} \beta_{1(120 \times 1)},$$

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

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

$$Y_2_{(65 \times 1)} = X_{2(65 \times 120)} \beta_{2(120 \times 1)} + \varepsilon_{(65 \times 1)},$$

where, $\beta_{2(120 \times 1)} = (2, 2, 0, \ldots, 0_{120})^t$ and $\varepsilon_{(65 \times 1)}$ is distributed as Cauchy distribution.

---

<table>
<thead>
<tr>
<th></th>
<th>mse</th>
<th>Med(mse)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>334</td>
<td>23.0</td>
<td>2817.9</td>
<td>22.33</td>
<td>6.02</td>
<td>144.0</td>
<td>7.72</td>
<td>7.16</td>
<td>182.0</td>
</tr>
<tr>
<td></td>
<td>9.63</td>
<td>9.63</td>
<td>20</td>
<td>14.80</td>
<td>14.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>Z Coef</td>
<td>0.44</td>
<td>0.56</td>
<td>0.48</td>
<td>0.81</td>
<td>0.19</td>
<td>0.99</td>
<td>0.59</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>NZ Coef</td>
<td>0.56</td>
<td>0.81</td>
<td>0.22</td>
<td>0.19</td>
<td>0.99</td>
<td>0.59</td>
<td>0.45</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>mse</td>
<td>9</td>
<td>2255</td>
<td>7.75</td>
<td>5.16</td>
<td>205.5</td>
<td>485.5</td>
<td>31112</td>
<td>299457</td>
</tr>
<tr>
<td></td>
<td>Med(mse)</td>
<td>26.2</td>
<td>25.19</td>
<td>6.87</td>
<td>17.1</td>
<td>50.2</td>
<td>20.2</td>
<td>17.7</td>
<td>7.9</td>
</tr>
</tbody>
</table>
Consequentially, $Y_{(1×75)} = \begin{bmatrix} Y_{1(10×1)} \\ Y_{2(65×1)} \end{bmatrix}$.

The MHLL 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 MHLL 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.**

<table>
<thead>
<tr>
<th></th>
<th>LAD-Lasso</th>
<th>WLAD-Lasso</th>
<th>MHLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z. Coef</td>
<td>108</td>
<td>111</td>
<td>117</td>
</tr>
<tr>
<td>NZ. Coef</td>
<td>12</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>False variable</td>
<td>10 (0.083)</td>
<td>7 (0.058)</td>
<td>1 (0.008)</td>
</tr>
</tbody>
</table>

### 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.
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.

![Cook's Distance with 0.02 as threshold for Males Hand Grip Strength data](image)

**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.

![Identified leverage point and outliers in Males Hand Grip Strength data](image)

**Figure 2. Identified leverage point and outliers in Males Hand Grip Strength data**
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 $\text{M}^\text{HLL}$ select 2 non-zero coefficients (Height and right Upper arm circumference). The values of the Root of Mean Squares Error (RMSE) of $\text{M}^\text{HLL}$s is smaller than the LAD-Lasso and WLAD-lasso. Hence for this data set, the $\text{M}^\text{HLL}$s are 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>
<thead>
<tr>
<th></th>
<th>LAD-Lasso</th>
<th>WLAD-Lasso</th>
<th>$\text{M}^\text{HLL}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z.Coeff</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>NZ.Coeff</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>RMSE</td>
<td>10.87</td>
<td>9.60</td>
<td>9.38</td>
</tr>
</tbody>
</table>

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.
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

<table>
<thead>
<tr>
<th></th>
<th>LAD-Lasso</th>
<th>WLAD-Lasso</th>
<th>MHLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z. Coef</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>NZ. Coef</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>RMSE</td>
<td>9.48</td>
<td>8.68</td>
<td>7.85</td>
</tr>
</tbody>
</table>
Robust Variable Selection Method Based on Huberized LARS-Lasso Regression

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 method included all covariates. It is interesting to see that the M̃HLL 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 M̃HLL 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](image.png)

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 M̃HLL 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 M̃HLL offers a remarkable improvement over the LAD-Lasso and WLAD-Lasso.

The M̃HLL 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 M̃HLL 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.

DOI: 10.24818/18423264/54.3.20.09
REFERENCES


DOI: 10.24818/18423264/54.3.20.09
Robust Variable Selection Method Based on Huberized LARS-Lasso Regression

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

DOI: 10.24818/18423264/54.3.20.09
Hassan Uraibi, Habshah Midi

