https://doi.org/10.26682/sjuod.2022.25.2.53
Journal of University of Duhok., Vol. 25, No.2(Pure Engineering Sciences), Pp 601-613, 2022

ACOMPARATIVE STUDY OF SOME VARIABLES SELECTION METHODS IN HIGH DIMENSIONAL MULTIPLE LINER REGRESSION VIA SIMULATION

MEDIA **S**HAMSADDIN **B**ARI[*] and **H**USSEIN **A**BDULRAHMAN **H**ASHEM[**]
[*]Dept. of Mathmetic ,College of Basic Education,University of Duhok,Kurdistan Region,Iraq
[**]Dept. of Mathematics, College of Sciences, University of Duhok, Kurdistan Region, Iraq

(Received: October 9, 2022; Accepted for Publication: November 15, 2022)

ABSSTRACT

In this study, we surveyed many strategies for picking relevant variables in high-dimensional MLR analyses. Parameters in linear regression are often estimated using traditional approaches like the Ordinary Least Squares (OLS) methodology. However, OLS estimates do not fare well when the dataset contains outliers or when the assumption of normality is broken, as in the case of heavy-tailed errors. Huber Lasso (Rosset and Zhu, 2007) and quantile regression (Koenker and Bassett, 1978) are two examples of resilient regularized regression techniques presented as solutions to this issue. This study examines the differences between the Whitening Lasso (WLasso) estimates, adaptable Huber Lasso (HLasso) estimates, adaptive LAD Lasso (ALasso) estimates, genLasso (generative least squares) estimates, gamma (gam) estimates, and Split Regularized Regression (SRR) estimates.

KEYWORDS: High-dimensional regression; Lasso; Split regularization.

1. INTRODUCTION

Many fields of study, including biology, collaborative filtering, and signal processing, variable choice is crucial. When doing a microarray experiment, for instance, one may test hundreds of variables all at once (genes, proteins). The data sets generated by these trials are often vast in terms of the number of predictors (p), but typically limited in terms of the number of biological samples themselves (n). Commonly referred to as the "big p and small n problem" ($p \gg n$), this issue is a significant barrier to conventional statistical approaches in regression analysis. Over the course of many decades, several statistical approaches have been created in response to the growing database volumes caused by the proliferation of computers and other data gathering technology. Parameter estimates, model choice, and variable identification present especially difficult problems. A variety of strong regression techniques are presented with the goal of fitting different regression models, particularly in the scenario where $\geq n$. According to Lasso (Least Absolute Shrinkage and Selection Operator), proposed by Tibshirani (1996), this makes the sum of squares conform to a L 1-norm requirement. When using the Lasso penalty, certain coefficients are estimated as zero, which serves as a proxy for both estimation and variable selection. Different extensions of the Lasso, such as the adaptive Lasso (Zou, 2006), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001), etc., were created after Tibshirani's (1996) foundational article. Koenker and Bassett (1978) introduced quantile regression, which may be used to estimate several quantiles of the conditional distribution, including the median. This allows us to see and contrast how different quantiles of the response variable are affected by different predictor factors.

In order to carry out variable selection in high-dimensional data containing outliers, several of the approaches use a combination of regularized and robust regression techniques. The Huber Lasso technique, proposed by Rosset and Zhu (2007), combines the Lasso penalty with Huber's criteria loss, to provide just one example. Combining Least Absolute Deviance (LAD) with the concept of adaptive Lasso, Wang et al. (2007) introduced the LAD-adaptive Lasso approach. Huber's loss function was combined with an adaptable Lasso penalty, and Lambert-Lacroix and Zwald (2011) produced

Huber's Criterion with an adjustable Lasso. The gamma divergence for regression was first proposed by Fujisawa and Eguchi (2008). It quantifies the dissimilarity between two conditional probability density functions. With the R package genLasso, Arnold and Tibshirani's (2016) dual algorithm may be used. The gamma Lasso (GL) technique, proposed by Taddy (2016), is a multi-convex relaxation of optimal variable selection that is computationally more appealing. To improve the speed and scalability of computing the solution routes of penalized quantile regression, Yi and Huang (2016) devised a method called Semismooth Newton Coordinate Descent (SNCD). Maximum Tangent Likelihood Estimation was proposed by Qin et al. (2017). (MTE). The Split Regularized Regression (SRR) technique, developed by Christidis et al. (2020), is a multi-convex relaxation of optimal split selection that is computationally more appealing. After applying a whitening modification to the data, the extended Lasso criteria developed by Tibshirani and Taylor may be used to get rid of the correlations, as proposed by Zhu et al. (2021). (2011). This study will summarize many regularized and resilient approaches to variable selection in linear regression in the next part.

2.MATERIAL AND METHOD

We start from the multiple linear regression standard model to define the methods of regression regularization. Let the data $(x_1, y_1),. \ . \ .,(x_n, y_n)$, and the design matrix denoted by $\boldsymbol{X} = (x_1^T, \ . \ \ . \ , x_n^T)^T$, the general linear model is usually represented as

$$y = \boldsymbol{X}\beta + \epsilon \qquad (1)$$

Here $\beta = (\beta_1,. \ . \ . \ ,\beta_p)^T$ are the regression coefficients $\epsilon = (\epsilon_1,. \ . \ .,\epsilon_n)^T \sim N(0, \sigma^2 I_n)$ are the random errors, x_i are the regressors for observation $i, i = 1,. \ . \ .,n$ and $y = (y_1 ,. \ . \ . \ ,y_n)^T$. The ordinary least squares (OLS) method estimates β by minimizing the residual squared error, i.e. $\hat{\beta}_{OLS} = \min_{\beta}\{(y - \boldsymbol{X}\beta)^T(y - \boldsymbol{X}\beta)\}$.

OLS estimates often have modest biases, but big variances and improved prediction accuracy are sometimes gained by lowering the variance with a little higher bias.

2.1 Lasso Regression

Tibshirani (1996) proposed the Lasso penalty, a regularization technique for simultaneous estimation and variable selection for large data sets. The Lasso estimate $\widehat{\boldsymbol{\beta}}$ is well-defined by:

$$\hat{\beta}_{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2\right\}, \ \text{s.t.} \ \ \sum_{j=1}^{p}|\beta_j| \le t, \ t \ge 0. \qquad (2)$$

An equivalent form of the Lasso is,

$$\hat{\beta}_{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda\sum_j|\beta_j|\right\}, \qquad (3)$$

or

$$\hat{\beta}_{lasso} = \min_{\beta}\|y - x\beta\|_2^2 + \lambda\|\beta\|_1 . \qquad (4)$$

The relative importance of the penalty term to the total absolute value of the coefficients is determined by a parameter called lambda.

As the absolute value of the coefficients makes up the punishment term, lambda is the parameter that determines the relative importance of lowering the RSS and the penalty term.

As long as the absolute value of the coefficients is smaller than a constant, the Lasso algorithm will minimize the sum of squares of the residuals. For models with many variables but few data points, Lasso is a regression shrinkage approach often used. Lasso's primary function is to carry out variable selection when a regression line is being fitted to the data. We do this by reducing the values of certain coefficients and also by setting others to zero. In order to achieve a L_1 regularization, Lasso imposes a penalty on the target aim. Which coefficients are reduced and by how much is determined by this penalty, which is the total of the absolute values of the coefficients.

2.2 Adaptive Lasso

The adaptive Lasso, as proposed by Zou (2006), is an improved version of the original Lasso. With the adaptive Lasso, the penalized least squares are precisely defined as

$$\hat{\beta}_{\text{adaptive Lasso}} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\widehat{w}|\beta_j|\right\} \qquad (5)$$

Adaptive weights are introduced to penalize distinct coefficients in a variety of ways, rather

than relying just on the absolute value of the parameters. The use of approximated weights, as recommended by Zou (2006), $\widehat{w}_j = \frac{1}{|\widehat{\beta}_j|^{\gamma}}$, where $\hat{\beta}_j$ comes from reducing the OLS or Lasso and γ is a user-chosen constant. The choice of $\widehat{w}_j$ is very significant and Zou (2006) proposed using OLS while γ can be chosen by K -fold cross-validation. The adaptive Lasso chooses the accurate set of nonzero coefficients with possibility tending to one.

2.3 Huber Lasso

Lasso's effectiveness decreases when outliers are present in the regression answer. Rosset and Zhu (2007) discuss the Huber loss function as an alternative to the least-squares loss function of traditional Lasso.

$$\hat{\beta}_{Huber\ lasso} = \min_{\beta} \sum_{i=1}^{n} \rho\left(y_i - x_i^T \beta\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| , \quad (6)$$

where $\rho(t) = \begin{cases} t^2 & if \quad |t| \le M \\ 2M|t| - M^2 & if \quad |t| > M \end{cases}$.

The robustness of the predictions produced by the model is controlled by the tuning constant M, with smaller M values yielding more accurate predictions.

2.4 Adaptive Huber Lasso

Although the LAD loss function is quite stable, Lambert-Lacroix and Zwald (2011) noted that it is not as effective when used with data that follows a normal distribution. What they suggested was employing As the loss function, Huber's Criterion with a concurrent scale

$$\hat{\beta}_{Hadl} = \min_{\beta} \mathcal{L}_{\rho}(\beta, s) + \lambda \sum_{j=1}^{p} \widehat{w}_j^{Hadl} \left|\beta_j\right|$$

where $\widehat{w}_j^{ladl} = \left(\widehat{w}_1^{Hladl}, \ldots, \widehat{w}_p^{Hladl}\right)$ are a known weights vector and Huber's criterion is defined by

$$\mathcal{L}_{\rho}(\beta, s) = \begin{cases} ns + \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=1}^{p} \beta_j x_{ij}}{s}\right) s & if \quad s > 0, \\ 2M \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right| & if \quad s = 0, \\ +\infty & if \quad s < 0, \end{cases}$$

In this context, $\rho(t)$ is definite as (6), $s > 0$ is a scale limit for the distribution. The $\rho(t)$ is a function controlled by M that uses a combination of absolute errors for relatively big mistakes and squared errors for lesser errors. The rank loss functions are resistant to responses from influential locations in the same way that the check loss function and its variations are. The value of M, a constant, is often determined by the amount of noise and outliers present in the data $M = 1.345$.

2.5 LAD -Lasso

By combining the popular Lasso approach for shrinkage estimation and variable selection with the Least Absolute Deviation (LAD) regression technique, Wang et al. (2007) created a new method known as LAD-Lasso that is especially useful for robust regression. Wording options for the LAD-Lasso include (Wang et al., 2007).

$$\hat{\beta}_{Lad\ lasso} = \min_{\beta} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right| + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \quad (7)$$

Realizing that the LAD- criteria combines the Lasso penalty with the LAD criterion, it follows that the resulting estimator is likely to be both sparse and robust against outliers.

2.6 Adaptive LAD-LASSO

For datasets that are prone to heavy-tailed errors or outliers, the LAD estimator is preferred over the OLS due to its greater stability. The shrinkage estimation method Lasso is commonly used. Adaptive LAD-Lasso proposes a robust detection approach for estimating change points in the mean-shift model by connecting the two classical notions. The fundamental concept is to reformat the change point estimation issue as a penalized variable selection problem. The notation for the Adaptive LAD-Lasso looks like this: (Lambert-Lacroix and Zwald, 2011).

$$\hat{\beta}_{ladl} = \min_{\beta} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right| + \lambda \sum_{j=1}^{p} \widehat{w}_j^{ladl} \left|\beta_j\right| \quad (8)$$

where $\widehat{w}_j^{ladl} = \left(\widehat{w}_1^{ladl}, \ldots, \widehat{w}_p^{ladl}\right)$ is a recognized weights vector. In this current model, the estimator is robust to outliers since the squared loss is replaced by the l_1-loss.

2.7 Genlasoo Method:

Arnold and Tibshirani (2016) modified the original goal function by including a tiny ridge penalty for the increasingly common high-dimensional scenario where $n < p$.

$$minimize \frac{1}{2}\|y - X\beta\|_2^2 + \rho \quad (9)$$

Subject to $A\beta = b$ and $C\beta \leq d$

Where $y \epsilon R^n$ is the response vector, $X \epsilon R^{n \times p}$ is the design matrix of predictors/covariates, $\beta \epsilon R^p$ is the vector of unknown regression coefficients, and $\rho \geq 0$ is a tuning parameter that determines the level of regularization. The constraint matrices, A, and C are assumed to have full row rank. When this occurs, the issue is

$$minimize \frac{1}{2}\|y - X\beta\|_2^2 + \rho\|\beta\|_1 + \frac{\varepsilon}{2}\|\beta\|_2^2 \quad (10)$$

Subject to $A\beta = b$ and $C\beta \leq d$

where ε is some small constant. Note that objective (10) can be re-arranged into standard constrained Lasso form (9)

$$minimize \frac{1}{2}\|y^* - (X^*)\beta\|_2^2 + \rho\|\beta\|_1 \quad (11)$$

Subject to $A\beta = b$ and $C\beta \leq d$

using the augmented data $y^* = \begin{pmatrix} y \\ 0_p \end{pmatrix}$ and $X^* = \begin{pmatrix} X \\ \sqrt{\varepsilon} I_p \end{pmatrix}$

The column rank of the improved design matrix is complete. The dual algorithm was implemented by Arnold and Tibshirani in 2016 and may be found in the genLasso R package. To conduct our analysis, we make use of the genLasso utility found within the package.

2.8 The Gamma Lasso

Recent work by Taddy (2016) developed the gamma Lasso (GL) approach, which can be thought of as a multi-convex relaxation of optimal variable selection that is more appealing from a computational standpoint. In order to accommodate nonconvex cost functions in the L_0 and L_1 norms, the gamma Lasso algorithm generates regularization pathways that map to these values. Similar to the glmnet package (which performs the same function for penalization between L_1 and L_2 norms), this package's usage is as close as possible to that of the $glmnet$. The dual algorithm was implemented by Taddy (2016), and their code can be found in the $gaml$ r R package. To perform the evaluation, we make use of the $gaml$r built-in to the package.

2.9 Split Regularized Regression (SRR)

The Split Regularized Regression (SRR) method, recently published by Christidis et al. (2020), can be viewed as a computationally more appealing, multi-convex relaxation of optimum split selection. The suggested method for high-dimensional regression creates an ensemble of models by partitioning the set of covariates into distinct, though sometimes overlapping, classes. Model stacking is utilized to make reliable predictions, and a penalty term is incorporated to promote variation between groups.

A variable is considered to be part of a particular cluster if its coefficient in the related coefficient vector is nonzero, even if SRR does not perform an explicit search for variable clusters. This method can detect overlapping clusters, and it does not need that the coefficients of variables inside the same cluster converge on the same value. SRR's end goal is

$$J(b_1, \ldots, b_K) = \sum_{k=1}^{K}\{\frac{1}{2n}\|y - Xb_k\|_2^2 + \delta\left[\alpha\sum_{j=1}^{p}|b_{jk}| + (1-\alpha)\sum_{j=1}^{p} b_{jk}^2\right] + \lambda\sum_{g\neq k}\sum_{j=1}^{p}|b_{jk}||b_{jg}|\}$$

The matrix $B = [b_1, \ldots, b_K]$ is essentially a cluster membership matrix, where variable j belongs to cluster k if $b_{jk} \neq 0$. Variables can belong to multiple clusters, but hard clusters $C_1, \ldots, C_K$ can also be defined such that $C_k: \{j | k = armax_l |b_{jl}|\}$. Maximal diversity is achieved when the rows of B contain only one nonzero element and thus each variable belongs to a single cluster (i.e. $|b_{jk}||b_{jg}| = 0 \; \forall \, j, k, g$). The final vector of regression coefficients used for prediction is an average across all vectors b_k: $\bar{b} = \frac{1}{K}\sum_{k=1}^{K} b_k$.

2.10 Whitening Lasso(WLasso) Method

By first performing a whitening treatment to the data, Whitening Lasso (WLasso), suggested by Zhu et al. (2021) removes correlations before running the analysis through the generalized Lasso criterion developed by Tibshirani and Taylor (2011). By include the covariance matrix in the penalty function, the WLasso mitigates the impact of high correlation on variable selection. Since the WLasso needs a decomposition of covariance matrices, it can be computationally time-consuming. Additionally, the predicted covariance matrix $\sum$ is provided with two blocks based on the assumption of a robust connection

between the active and inactive variables (one block including mainly active variables and the other one including mainly inactive variables). However, the block structure's partitioning may lead to erratic variable selection and other problems.

3. SIMULATION STUDY

The author examines the performance of several standardized regression techniques in low-dimensional with sparse and non-sparse coefficients ($p = 15$, $n = 100$) and high-dimensional with sparse coefficients ($p = 100$, $n = 50$) environments. Researchers often employ a traditional simulation environment when working with sparse data. For a non-sparse environment, we use _j=0.2 for every j, as in Bradic and Fan (2011), where $y = \beta_0 + x\beta + u$, with $\beta_0 = 0$ and $\beta = (3, 1.5, 0, 0, 2, 0, \dots, 0)$. In this study, x is selected at random from a multivariate normal distribution $N(0, \Sigma_x)$. By assigning x_i and x_j is set to be $(\Sigma_x)_{ij} = r^{|i-j|}$ as the pairwise covariance, we may study their relationship in greater detail. We choose non-normal distributions for the error u to test the robustness of the techniques. We focus on the following special cases: $: u \sim N(0, 1)$, Double Exponential (DE), Gamma distribution $G(3,1)$, t-distribution (t_3) with 3 degrees of freedom, and Chi-squared distribution $(\chi^2_{(3)})$. The adaptive LAD Lasso, Huber adaptive Lasso (Xu and Ying, 2010; Lambert-Lacroix and Zwald, 2011), genlasoo method (Arnold and Tibshirani, 2016), gamma Lasso (Taddy, 2016), Split Regularized Regression (SRR) (Christidis et al., 2020), and Whitening Lasso (WLasso) method are all defined in the previous (Zhu et .al, 2021). To implement the $genlasoo$ approach, we pull in the genlasoo R package, the gamma Lasso from the $gaml$ R package, and the adaptive LAD Lasso and adaptive Huber Lasso from the $parcor$ R package, modifying a few of its functions. $SplitReg$ is a R package used for the SRR method, and WLasso is a R package used for the WLasso method.

3.1 Example A.Low –dimesional with sparse coefficients

The researcher takes into account $p = 15$ and $n = 100$ data. The results of the simulation are presented in Table 1A, Table 1B, and Figure 1. We analyse scenarios where the predictors have a low $r = 0.5$ and a high $r = 0.95$ degree of correlation. The model error is computed as $(\hat{\beta} - \beta)^T S_x (\hat{\beta} - \beta)$, where $\hat{\beta}$ is the expected parameter and S_x is the sample covariance, and the median model error after 500 iterations is reported in the top panels (same results for the mean error). The true positives, or the number of non-zero coefficients that were accurately identified, are reported in the bottom panels. In this situation, three indicates that the detection of all non-zero coefficients was successful.

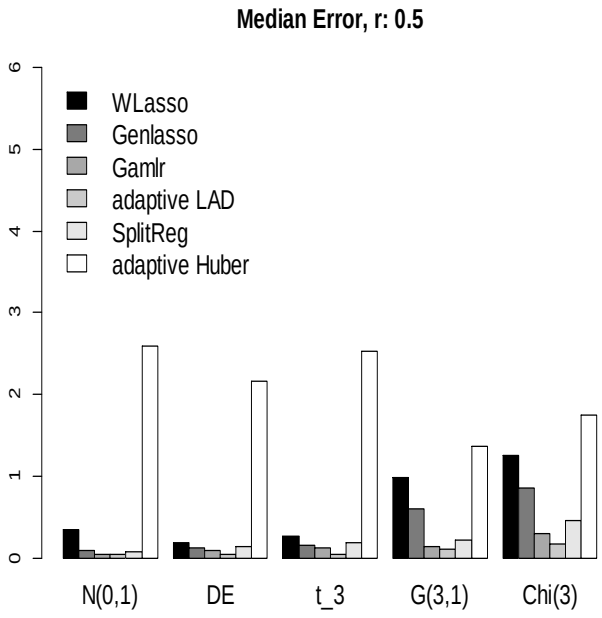
We find that the adaptive Huber Lasso technique underperforms when the predictors are complex, while the adaptive LAD and gamma Lasso ($gamlr$) methods perform better for the vast majority of error distributions.

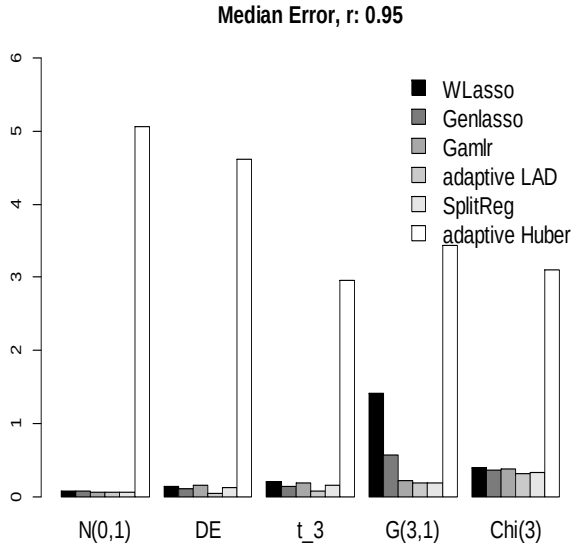
Table (1A): Average Median Model Error over 500 replications for the case: $p = 15, n = 100, r = 0.5$, and β values as in example 1, Best method indicated in bold.

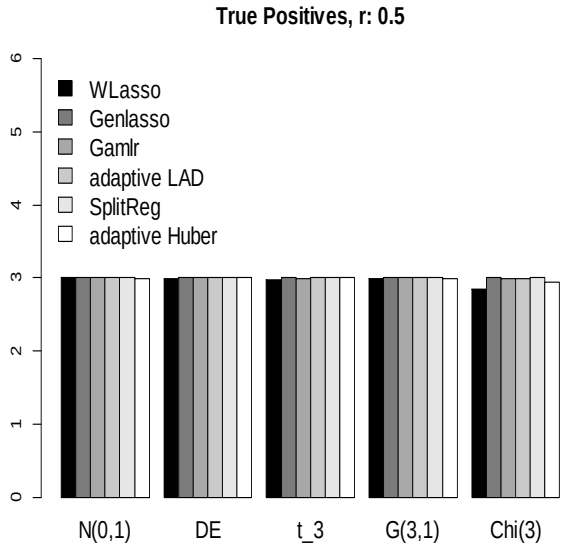
	WLasso	Genlasso	Gamlr	adaptive LAD	SplitReg	adaptive Huber
$N(0,1)$	0.345	0.088	0.042	**0.041**	0.069	2.594
DE	0.191	0.124	0.087	**0.034**	0.133	2.166
t_3	0.268	0.155	0.120	**0.049**	0.185	2.531
$G(3,1)$	0.976	0.596	0.137	**0.107**	0.214	1.359
$Chi(3)$	1.246	0.858	0.292	**0.174**	0.453	1.744

Table (1B): Average Median Model Error over 500 replications for the case: $p = 15, n = 100, r = 0.95$, and β values as in example 1, Best method indicated in bold.

	WLasso	Genlasso	Gamlr	adaptive LAD	SplitReg	adaptive Huber
$N(0,1)$	0.068	0.077	0.050	**0.056**	0.061	5.061
DE	0.138	0.104	0.147	**0.047**	0.123	4.611
t_3	0.197	0.142	0.183	**0.080**	0.149	2.959
$G(3,1)$	1.414	0.564	0.217	**0.187**	0.186	3.433
$Chi(3)$	0.400	0.368	0.381	**0.317**	0.331	3.107




Median Error, r: 0.95
WLasso
Genlasso
Gamlr
adaptive LAD
SplitReg
adaptive Huber
0
1
2
3
4
5
6
N(0,1)
DE
t_3
G(3,1)
Chi(3)


True Positives, r: 0.5
WLasso
Genlasso
Gamlr
adaptive LAD
SplitReg
adaptive Huber
0
1
2
3
4
5
6
N(0,1)
DE
t_3
G(3,1)
Chi(3)

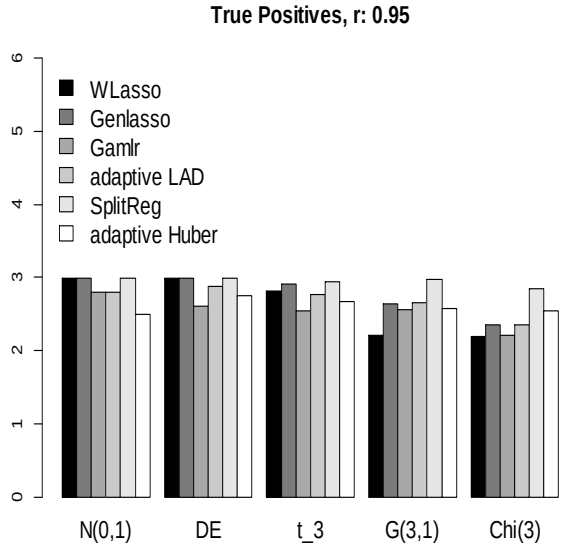


Fig. (1): Comparison of variable selection methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for instance 1 and the bottom panels the average true positives when $p = 15$ and $n = 100$.

Example B. high-dimensional with sparse coefficients

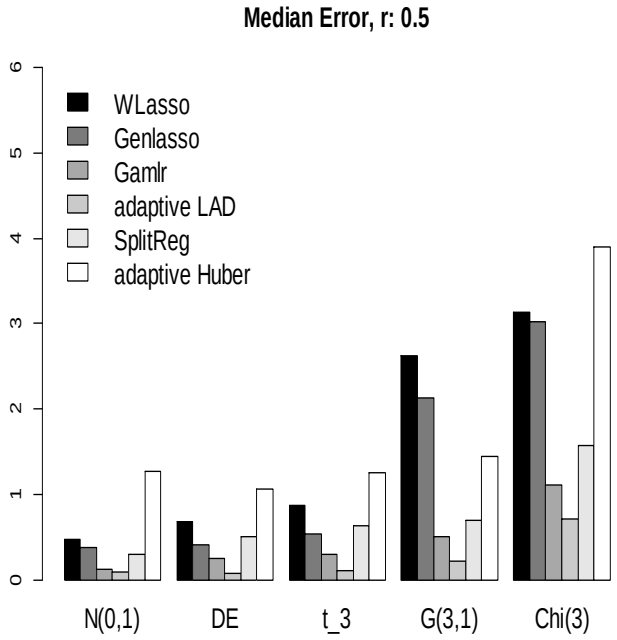
A scenario like simulation 3.1 is considered, but with different parameters and a smaller sample size. A high-dimensional example with sparse coefficients $p = 100$ and $n = 50$ is given special attention by the researcher. With the current simulation setup, this is a somewhat sparse problem. one when very many coefficients are zero. The outcomes of the simulations are displayed in Tables 2A and 2B, as well as in Figure 2. The model error was estimated with over 500 repetitions as an Example 1, with the results displayed in the upper panels. The true plus, which is the count of non-zero coefficients that were correctly labelled, is displayed in the bottom panels.

Table (2A): Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.5$, and β values as in example 2, Best method indicated in bold.

	WLasso	**Genlasso**	**Gamlr**	**adaptive LAD**	**SplitReg**	**adaptive Huber**
$N(0,1)$	0.479	0.379	0.123	**0.097**	0.295	1.261
DE	0.681	0.401	0.247	**0.081**	0.502	1.054
t_3	0.866	0.536	0.295	**0.103**	0.630	1.249
$G(3,1)$	2.619	2.129	0.510	**0.218**	0.687	1.451
$Chi(3)$	3.137	3.022	1.116	**0.706**	1.577	3.900

Table (2B): Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.95$, and β values as in example 2, Best method indicated in bold.

	WLasso	**Genlasso**	**Gamlr**	**adaptive LAD**	**SplitReg**	**adaptive Huber**
$N(0,1)$	0.347	0.164	0.252	**0.124**	0.177	3.475
DE	0.725	0.313	0.498	**0.146**	0.362	2.621
t_3	0.661	0.376	0.829	**0.346**	0.438	2.431
$G(3,1)$	2.056	2.148	1.081	**0.461**	0.509	2.293
$Chi(3)$	2.255	1.815	1.678	**0.602**	0.909	1.907



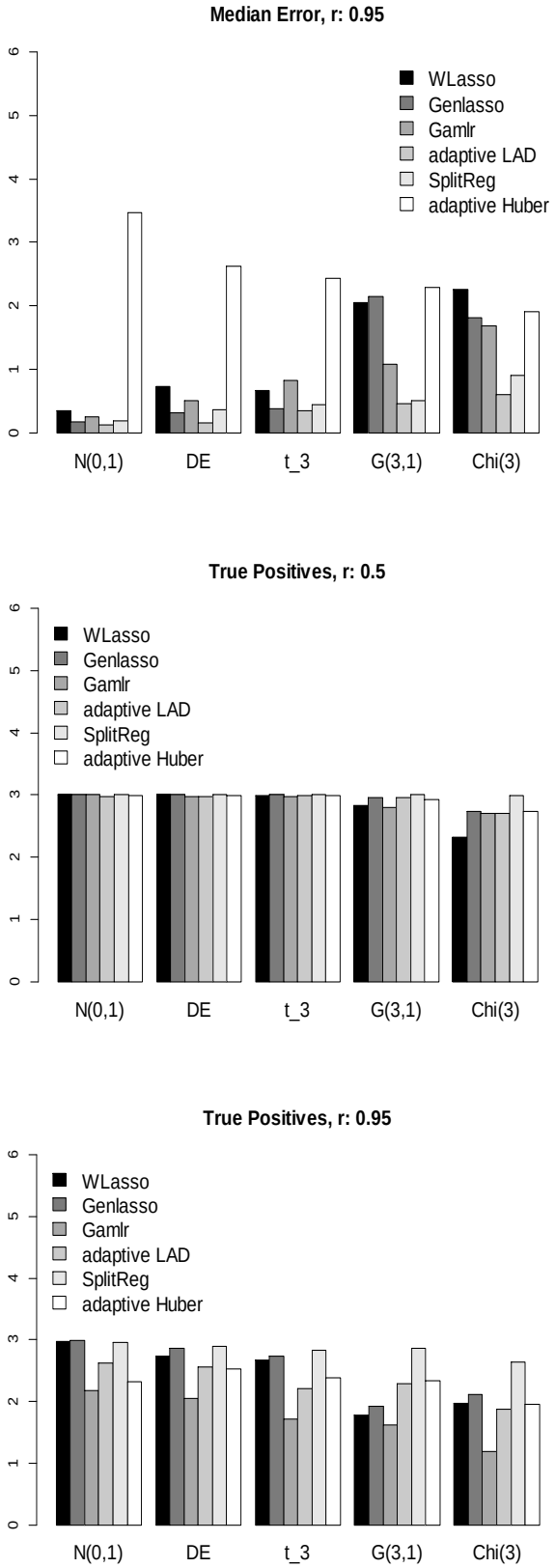


Fig. (2): Comparison of variable selection methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for example 2 and the bottom panels the average true positives when $p = 100$ and $n = 50$.

The results back up the efficacy of the approaches: adaptive Huber Lasso and WLasso don't work well when dealing with highly correlated predictors, while adaptive LAD and the Split Regularized Regression (SRR) methods perform best of all when dealing with increasing deviations from normality. This is especially clear in the case of highly correlated predictors and the $G(3,1)$ and $\chi^2_{(3)}$ simulation, both of which display a significant departure from normalcy.

3.3 Example C. low- dimensional with non-sparse coefficients

To consider the variable selection methods performance in example 1, the researcher conducts a new simulation where $\beta_j = 0.2$ for all j, which is a non-sparse situation. Table 3A, Table 3B, and Figure 3, all report the median model error over 500 replications for the case $p = 15$ and $n = 100$

Table (3A): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 15, \mathrm{n} = 100, \mathrm{r} = 0.5$, and β values as in example 3, Best method indicated in bold.

	WLasso	Genlasso	Gamlr	adaptive LAD	SplitReg	adaptive Huber
$N(0,1)$	0.969	0.247	0.142	0.807	**0.099**	1.257
DE	0.794	0.311	0.251	0.863	**0.180**	1.260
t_3	0.644	0.345	0.333	0.775	**0.240**	1.053
$G(3,1)$	1.029	0.915	0.363	0.872	**0.259**	1.281
$Chi(3)$	0.885	1.189	0.647	0.972	**0.457**	1.359

Table (3B): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 15, \mathrm{n} = 100, \mathrm{r} = 0.95$, and β values as in example 3, Best method indicated in bold.

	WLasso	Genlasso	Gamlr	adaptive LAD	SplitReg	adaptive Huber
$N(0,1)$	0.078	0.085	0.090	0.262	**0.054**	3.526
DE	0.222	0.120	0.150	0.311	**0.090**	2.473
t_3	0.426	0.159	0.193	0.300	**0.120**	2.300
$G(3,1)$	0.869	0.352	0.231	0.302	**0.137**	2.024
$Chi(3)$	0.425	0.344	0.362	0.373	**0.234**	2.219

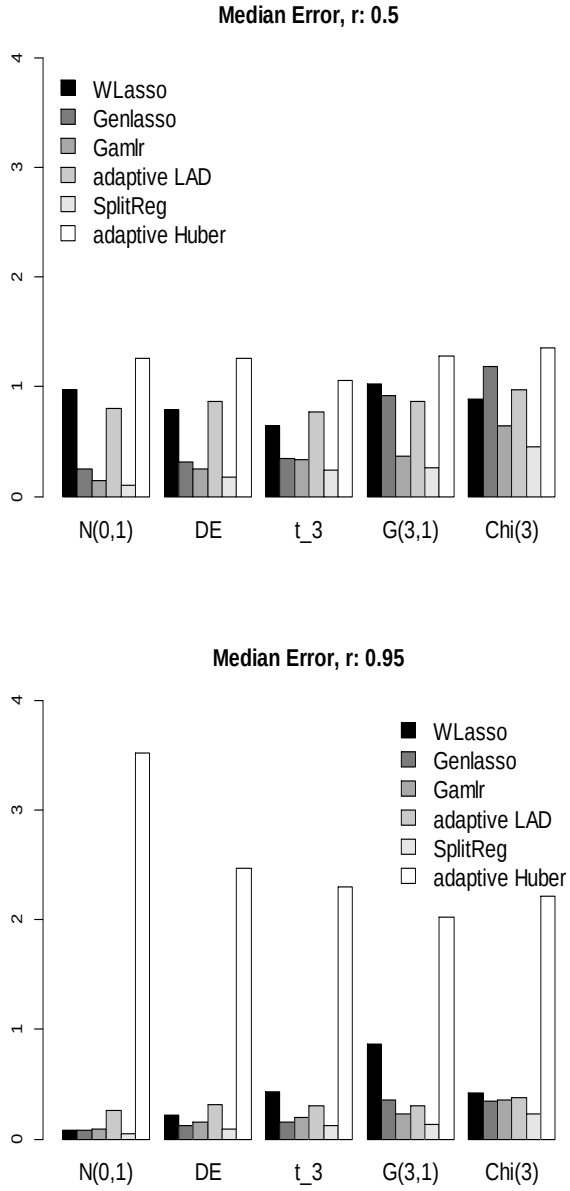


Fig. (3): Comparison of variable selection methods under different error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 500 replications for example 3 when $p = 15$ and $n = 100$.

Our simulation analysis shows that the Split Regularized Regression (SRR) technique performs best when deviation from normality increases (see Table 3A, Table 3B, and Figure 3). When the predictors are highly associated, this fact becomes especially clear.

4. CONCLUDING REMARKS

Many established statistical methods rely on the normalcy assumption. These methods are not well suited for data that exhibits substantial non-normality. This is a common result of working with tainted data, which might cause unexpected results. Recent advances in robust regularized regression techniques, such as the LAD approaches and the Split Regularized Regression, are taken into account in this study (SRR). When dealing with large numbers of dimensions $p \geq n$. In a simulation analysis, we demonstrate that the adaptive Least Absolute Deviation (LAD) and Split Regularized Regression (SRR) approaches outperform the other resilient methods, especially when there is a considerable departure from normality.

REFERENCES

Arnold, T. B., and Tibshirani, R. J. (2014).Efficient implementations of the generalized Lasso dual-path algorithm. *Journal of Computational and Graphical Statistics, 25(1):1–27, 2*016.

Bradic, J. and J. Fan (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society, B 73* (3), 325–349.

Christidis, A.-A., Lakshmanan, L., Smucler, E., and Zamar, R. (2020). Split regularized regression. *Technometrics 62.3, pp. 330–338.*

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

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination, *Journal of Multivariate Analysis, 99(9), 2053-2081*.

Koenker, R. and G. W. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through Huber's criterion and adaptive Lasso penalty. *Electronic Journal of Statistics 5*,

Qin, Y., Li, S. and Yu, Y. (2017). Penalized Maximum Tangent Likelihood Estimation and Robust Variable Selection.https://arxiv.org/pdf/1708.05439.pdf .

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics 35 (3)*, 1012–1030.

Taddy, M. (2017). One-step estimator paths for concave regularization, *Journal of Computational and Graphical Statistics pp. 1–12*.

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

Tibshirani, R. J., and Taylor, J. (2011). The solution path of the generalized Lasso. *Ann.Stat., 39(3), 1335-1371*.

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

Yi, C. Huang, J. (2016). Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression. *Journal of Computational and Graphical Statistics 3.* 547–557

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zhu, W., L´evy-Leduc, C., and Tern`es, N. (2021). A variable selection approach for highly correlated predictors in high-dimensional genomic data. *Bioinformatics, 37(16), 2238–2244.*