Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

Adaptive Lad Lasso, Split Regularized Regression and DLasso: Simulation Study of Variable Selection

Hussein A. Hashem

Department of Mathematics, College of Science, University of Duhok, Kurdistan Region, Iraq.

Article information

Article history:
Received February 8, 2024
Revised: September 2,2024
Accepted September 20, 2024
Available online December 1, 2024

Keywords:
Variable Selection; Lasso; DLasso; Regularization.

Correspondence:
Hussein A. Hashem

hussein.hashem@uod.ac

Abstract

Variable selection is important for the analysis of high-dimensional data in many fields, including in the field of biology, signal processing and collaborative filtering. For example, microarray experiments allow one to measure thousands of variables (genes and proteins) simultaneously. The datasets resulting from these experiments are usually very large in terms of the number of predictors (P) and often small in terms of the number of biological samples (n). In regression analysis, this problem is often called the" big and small p-n problem " (p-n-n) and represents a major obstacle to traditional statistical methods. In this paper, we compare three different main methods for selecting variables for linear regression models: Adaptive Lad Lasso, Split Regularized Regression (SRR) and DLasso (AIC, GIC, BIC, CGV). In a simulation study, we show the performance of the methods considering the median model error. The case where the number of candidate variables exceeds the number of observations is considered as well. Also, the simulation study is used in determining which methods are best in all of the linear regression scenarios.

DOI 10.33899/IQJOSS.2024.185232

1. Introduction

Variable selection is important for high-dimensional data analysis in many research areas such as biology, signal processing, and collaborative filtering. For example, microarray experiments allow one to measure thousands of variables (genes, proteins) simultaneously. The data sets generated by these experiments are generally very large in terms of the number of predictors (p) and often small in terms of the number of biological samples (n). In regression analysis, this problem is often termed the "large p and small n problem" ($p \gg n$) and presents a major barrier to traditional statistical methods.

With the development of computer and data collection technologies, the database sizes continue to grow and various statistical methodologies have been developed over the past several decades to cope with the challenges presented by these data. In particular, there are major challenges in parameter estimation, model, and variable selection. Several regression methods have been proposed for fitting multiple regression models, especially for the case when $p \geq n$ where the least-squares method could not be used.

Tibshirani proposed Lasso (Least Absolute Shrinkage and Selection Operator) [1], which minimizes the residual sum of squares subject to an L_1-norm constraint. The Lasso penalty results in some coefficients being estimated to be completely zero, thus performing estimation and variable selection simultaneously. Following the seminal paper of Tibshirani [1], various extensions of the Lasso were developed, for example, the adaptive Lasso [2], Smoothly Clipped Absolute Deviation (SCAD) [3], etc.

Quantile regression, introduced by Koenker and Bassett [4], could be used when an estimate of the various quantiles (such as the median) of a conditional distribution is of interest. This allows one to see and compare how some quantiles of the response variable may be more affected by some predictor variables than other quantiles.

Some methods have combined regularized and robust regression methods to perform variable selection in high-dimensional data with outliers. For example, Rosset and Zhu [5] proposed the Huber Lasso method which combines Huber's criterion loss with a Lasso penalty. The LAD-adaptive Lasso method is proposed by Wang et al. [6], combining the idea of Least Absolute Deviance (LAD) and adaptive Lasso. Lambert-Lacroix and Zwald [7] developed Huber's Criterion with an adaptive Lasso which combines Huber's loss function and adaptive Lasso penalty.

Fujisawa and Eguchi [8] introduce the gamma divergence for regression. It measures the difference between two conditional probability density functions. Arnold and Tibshirani [9] implemented the dual algorithm and their implementation is available in the R package genLasso. Taddy [10] proposed the gamma Lasso (GL) algorithm which can be seen as a computationally more attractive, multi-convex relaxation of best variable selection. Yi and Huang [11] developed an algorithm, called Semismooth Newton Coordinate Descent (SNCD), to obtain better efficiency and scalability for computing the solution paths of penalized quantile regression. Qin et .al [12] proposed a method called Maximum Tangent Likelihood Estimation (MTE). Christidis et al. [13] introduced the Split Regularized Regression (SRR) method which can be seen as a computationally more attractive, multi-convex relaxation of best-split selection. Zhu et .al [14] proposed Whitening Lasso (WLasso) to remove the correlations by applying a whitening transformation to the data before using the generalized Lasso criterion designed by Tibshirani and Taylor [15].

In the next section, we will give an overview of some group variable selection methods in linear regression.

2. Material and methods

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

$$y = \boldsymbol{X}\beta + \epsilon \tag{1}$$

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

In general, OLS tends to give estimators with low biases but high variances, and better prediction accuracy can usually be obtained by lowering the variance with a little increased bias.

2.1 Lasso Regression

Tibshirani [1] proposed the Lasso penalty, a regularization technique for simultaneous estimation and variable selection for large data sets. The Lasso estimate $\widehat{\boldsymbol{\beta}}$ is defined by:

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

An equivalent form of the Lasso is,

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

or

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

lambda is the parameter deciding the weight on minimizing the RSS compared to the penalty term is the sum of the absolute value of coefficients.

The Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. In other words, Lasso is a regression shrinkage method typically used in models with a large number of variables but relatively few observations. The main purpose of Lasso is to perform variable selection while fitting the regression line to the data. This is done by shrinking certain coefficients but in addition setting some of the coefficients also to zero. Lasso performs a L_1 regularization by adding a penalty to the objective under optimization. This penalty is the sum of the absolute value of coefficients and determines which coefficients to shrink and how much.

2.2 Adaptive Lasso

Zou [2] proposed a new version of the Lasso, which is called the adaptive Lasso. The penalized least squares with the adaptive Lasso are defined as

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

Instead of simply using the absolute value of the parameters as the penalization, adaptive weights are added for penalizing different coefficients differently. Zou (2006) suggested the use of stated weights, $\widehat{w}_j = \frac{1}{|\hat{\beta}_j|^{\gamma}}$, where $\hat{\beta}_j$ comes from minimizing the OLS or Lasso and γ is a user-chosen constant. The choice of $\widehat{w}_j$ is very important and Zou [2] suggested using OLS while γ can is chosen by *K-fold* cross-validation. The adaptive Lasso selects the true set of nonzero coefficients with probability tending to one.

2.3 Lad Lasso

Wang et al. [6] developed combined methods from Least Absolute Deviation (LAD) regression that is useful for robust regression, and also Lasso which is a popular choice for shrinkage estimation and variable selection, becoming Lad Lasso. The LAD-Lasso can be written as (Wang et al., [6]).

$$\hat{\beta}_{Lad\ lasso} = \min_{\beta} \sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} \beta_j x_{ij}| + \lambda \sum_{j=1}^{p} |\beta_j| \quad (6)$$

As can be seen, the LAD- criterion combines the LAD criterion and the Lasso penalty, and hence the resulting estimator is expected to be robust against outliers and also to enjoy a sparse representation.

2.4 Adaptive Lad Lasso

The Lad estimator is more robust than the OLS estimator, especially when datasets are subject to heavy-tailed errors or outliers. Lasso is a popular choice for shrinkage estimation. Adaptive Lad-Lasso is combining the two classical ideas to put forward a robust detection method to estimate change points in the mean-shift model. The basic idea is to convert the change point estimation problem into a variable selection problem with a penalty. The Adaptive Lad-Lasso can be written as (Lambert-Lacroix and Zwald, [7])

$$\hat{\beta}_{ladl} = \min_{\beta} \sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} \beta_j x_{ij}| + \lambda \sum_{j=1}^{p} \widehat{w}_j^{ladl} |\beta_j| \quad (7)$$

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

2.5 Split Regularized Regression (SRR)

Christidis et al. [13] recently introduced the Split Regularized Regression (SRR) method which can be seen as a computationally more attractive, multi-convex relaxation of best-split selection. In high-dimensional regression, the proposed method builds an ensemble of models by splitting the set of covariates into different but possibly overlapping groups. A penalty term is introduced to encourage diversity between groups, and model stacking is used to generate accurate predictions.

The SRR-supervised variable clustering problem is to simultaneously estimate K models, one for each cluster, then predict the target based on the average coefficient vector across clusters. SRR estimates multiple sparse coefficient vectors and encourages these vectors to be as diverse as possible.

Although SRR does not explicitly search for variable clusters, they may be inferred from the coefficient vectors: a variable belongs to a given cluster if its coefficient in the corresponding coefficient vector is nonzero. This approach can identify overlapping clusters and does not force coefficients of variables in the same cluster to tend towards the same value. The objective function of SRR is

$$J(b_1, \ldots, b_K) = \sum_{k=1}^{K} \{ \frac{1}{2n} \|y - X b_k\|_2^2 + \delta \left[\alpha \sum_{j=1}^{p} |b_{jk}| + (1-\alpha) \sum_{j=1}^{p} b_{jk}^2 \right] + \lambda \sum_{g \neq k} \sum_{j=1}^{p} |b_{jk}| |b_{jg}| \} \quad (8)$$

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:

$\bar{b} = \frac{1}{K} \sum_{k=1}^{K} b_k$.

Like many of the coefficient-grouping methods, cluster assignment is more of a side effect than a principal objective of SRR. Moreover, the best solutions in terms of prediction error tend to be complex models with extensive cluster overlap. As a result, variables are assigned to clusters with less certainty, and the clustering performance is less optimal.

2.6 DLasso

Haselimashhadi and Vinciotti [16] proposed a new penalty term that is capable of producing similar results to other well-known penalty functions in the context of regularized. The new penalty is differentiable and this penalty opens up the possibility of using it in many contexts where differentiability plays a key role. For example, a differentiable objective function could lead to more effcient implementations of parameter estimation procedures for certain models or to improved model selection criteria by a more accurate estimation of the bias term. The method is implemented in the R package DLASSO freely available from CRAN, http://CRAN.R-project.org/package=DLASSO.

3. Simulation Study

In this section, we compare some regularized regression methods in low-dimensional with sparse and non-sparse coefficients($p = 15, n = 100$) and high-dimensional with sparse coefficients ($p = 100$, $n = 50$)settings. For the sparse settings, we use a classical simulation setting, e.g. Yu et al. [17] and Li et al.[18], where $y = \beta_0 + x\beta + u$. We draw the independent variables x from a multivariate normal distribution, $N(0, \Sigma_x)$. The pairwise covariance between x_i and x_j is set to be $(\Sigma_x)_{ij} = r^{|i-j|}$. For the error u, we choose a range of distributions in order to test the robustness of the methods to departures from normality. In particular, we consider the following cases: $u \sim N(0,1)$, We design a mixture normal distribution with large outliers, similar to Lambert-Lacroix and Zwald [7], by drawing 90% of the data from $a\ N(0,1)$ distribution and 10% from a $N(0,100)$ distribution, Laplace distribution, mixture of two Laplace distributions, t-distribution with 3 (t_3) degrees of freedom and Gamma$(3,1)$. Under all these cases, we compare the regularized regression methods described in the previous section, namely adaptive Lad Lasso (Xu and Ying, [19]; Lambert-Lacroix and Zwald, [7]), Split Regularized Regression (SRR) and DLasso (AIC, GIC, BIC, CGV). For the adaptive Lad Lasso we adapt some of the functions in the $parcor$ R package,for the SRR method, we use the R package $SplitReg$ and for the DLasso (AIC, GIC, BIC, CGV) methods, we use the R package $DLASSO$.For the correlation r, we experiment both with $r = 0.95$ and $r = 0.5$. For the β values we consider three cases:

(1) $\beta_j = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$, which corresponds to the very sparse case with structures in the predictors.

(2) $\beta_j = (1, 0, 0, 0, 5, 0, 1, 0, 0, 5, 0, 1, 0, \ldots, 0))$, which corresponds to the sparse case with structures in the predictors.

(3) $\beta_j = 0.1$ for all j, which corresponds to a dense case.

3.1 Simulation 1: low-dimensional with very sparse coefficients (Case 1)

In this section, we consider low-dimensional data with very sparse coefficients set with $p = 15$ and $n = 100$. Table 1A, Table 1B, and Figure1 report the results of the simulation. We consider both the case of low correlation ($r = 0.5$) and that of high correlation ($r = 0.95$) of the predictors. The top panels report the median model error over 500 iterations (similar results for the mean error), with the model error computed by$(\hat{\beta} - \beta)^T S_x (\hat{\beta} - \beta)$, where $\hat{\beta}$ are the estimated parameters and S_x the sample covariance. The bottom panels report the true positives which are the number of correctly found non-zero coefficients. Here three correspond to the case of all non-zero coefficients being correctly detected.

Our results show that: the DLasso (GIC, BIC) methods do not perform well when the predictors are highly correlated; the adaptive Lad Lasso and the Split Regularized Regression (SRR) methods outperform all other methods for most error distributions.

Table1A: Average Median Model Error over 500 replications for the case: $p = 15, n = 100, r = 0.5$, and β values as in simulation1, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.061	0.063	0.058	0.059	0.074	**0.046**
Normal. M	0.132	0.192	0.128	0.126	0.134	**0.070**
Laplace	0.139	0.222	0.140	0.133	0.146	**0.038**
Laplace. M	0.117	0.176	0.116	0.113	0.115	**0.041**
t_3	0.170	0.321	0.187	0.160	0.189	**0.050**
G(3,1)	0.976	1.898	1.846	1.003	0.218	**0.107**

Table1B: Average Median Model Error over 500 replications for the case: $p = 15, n = 100, r = 0.95$, and β values as in simulation1, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD

N(0,1)	0.084	0.089	0.081	0.081	0.059	**0.056**
Normal. M	0.157	0.217	0.159	0.149	**0.106**	0.122
Laplace	0.185	0.230	0.191	0.179	0.121	**0.055**
Laplace. M	0.154	0.191	0.141	0.137	0.097	**0.052**
t_3	0.252	0.314	0.259	0.233	0.172	**0.089**
G(3,1)	1.363	1.956	1.912	1.730	0.179	**0.163**

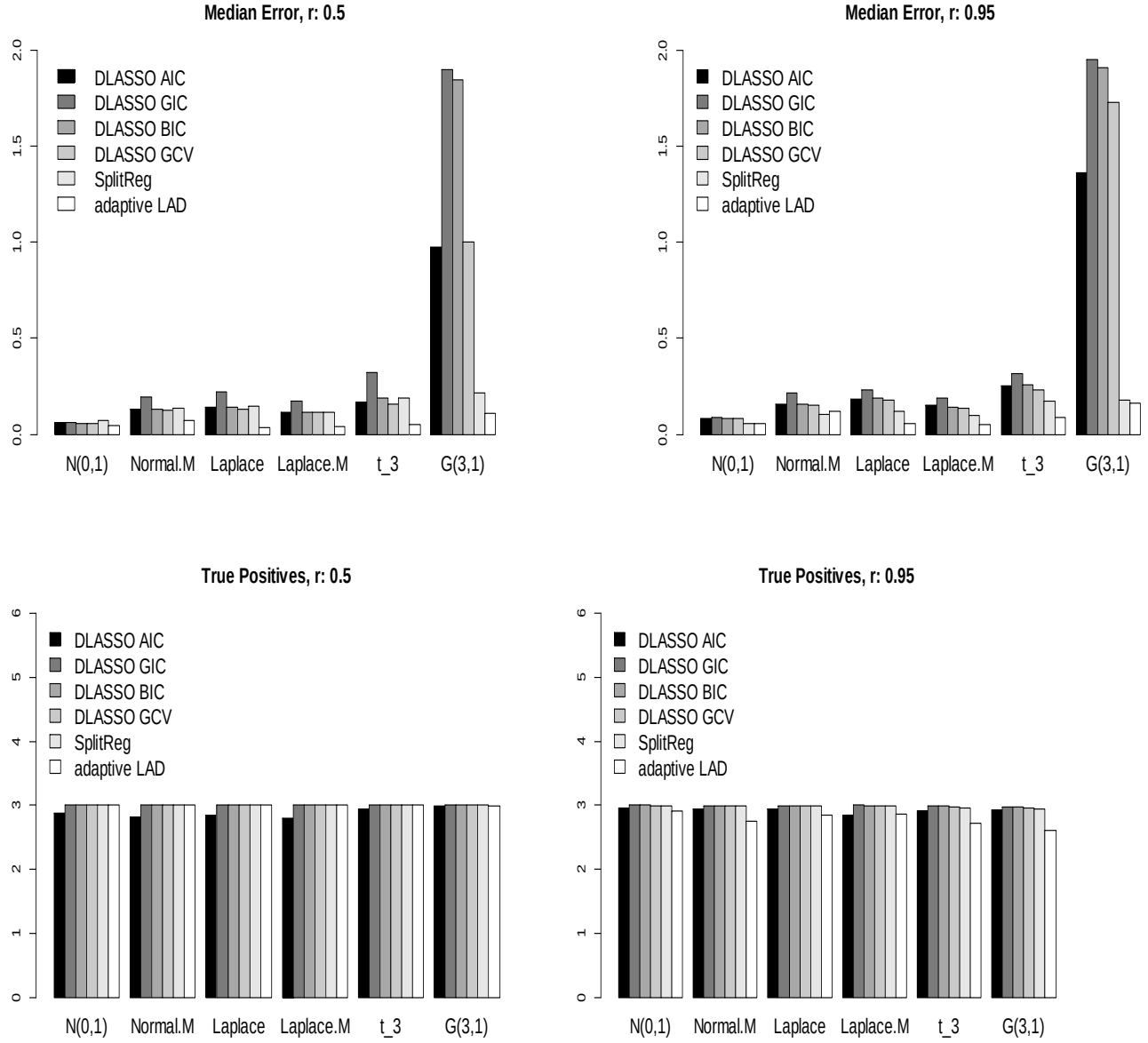


Figure1: Comparison of regularized regression methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for simulation1 and the bottom panels the average true positives when $p = 15$ and $n = 100$.

3.2 Simulation 2: high-dimensional with very sparse coefficients (Case 1)

We consider a similar setting to simulation 3.1 but with different sample size and several predictors. In particular, we consider a high-dimensional example with very sparse coefficients with $p = 100$ and $n = 50$.Table 2A, Table 2B and Figure 2 report the median model error over 500 replications, with the model error computed in the same way as in Figure1.

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

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.946	0.274	0.252	0.306	0.291	**0.081**
Normal. M	1.717	0.799	0.382	0.378	0.454	**0.134**
Laplace	1.928	1.063	0.573	0.528	0.490	**0.085**
Laplace. M	1.530	0.684	0.416	0.474	0.448	**0.069**
t_3	2.281	1.319	0.732	0.563	0.628	**0.123**
G(3,1)	2.799	2.634	2.413	2.031	**0.923**	2.039

Table2B: Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.95$, and β values as in simulation 2, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.628	0.234	0.184	0.203	**0.178**	0.335
Normal. M	1.373	0.849	0.370	0.329	0.330	**0.299**
Laplace	1.447	1.007	0.410	0.338	0.363	**0.141**
Laplace. M	1.087	0.668	0.287	0.269	0.297	**0.096**
t_3	2.004	1.411	0.673	0.450	0.448	**0.277**
G(3,1)	2.822	2.662	2.523	1.972	0.526	**0.421**

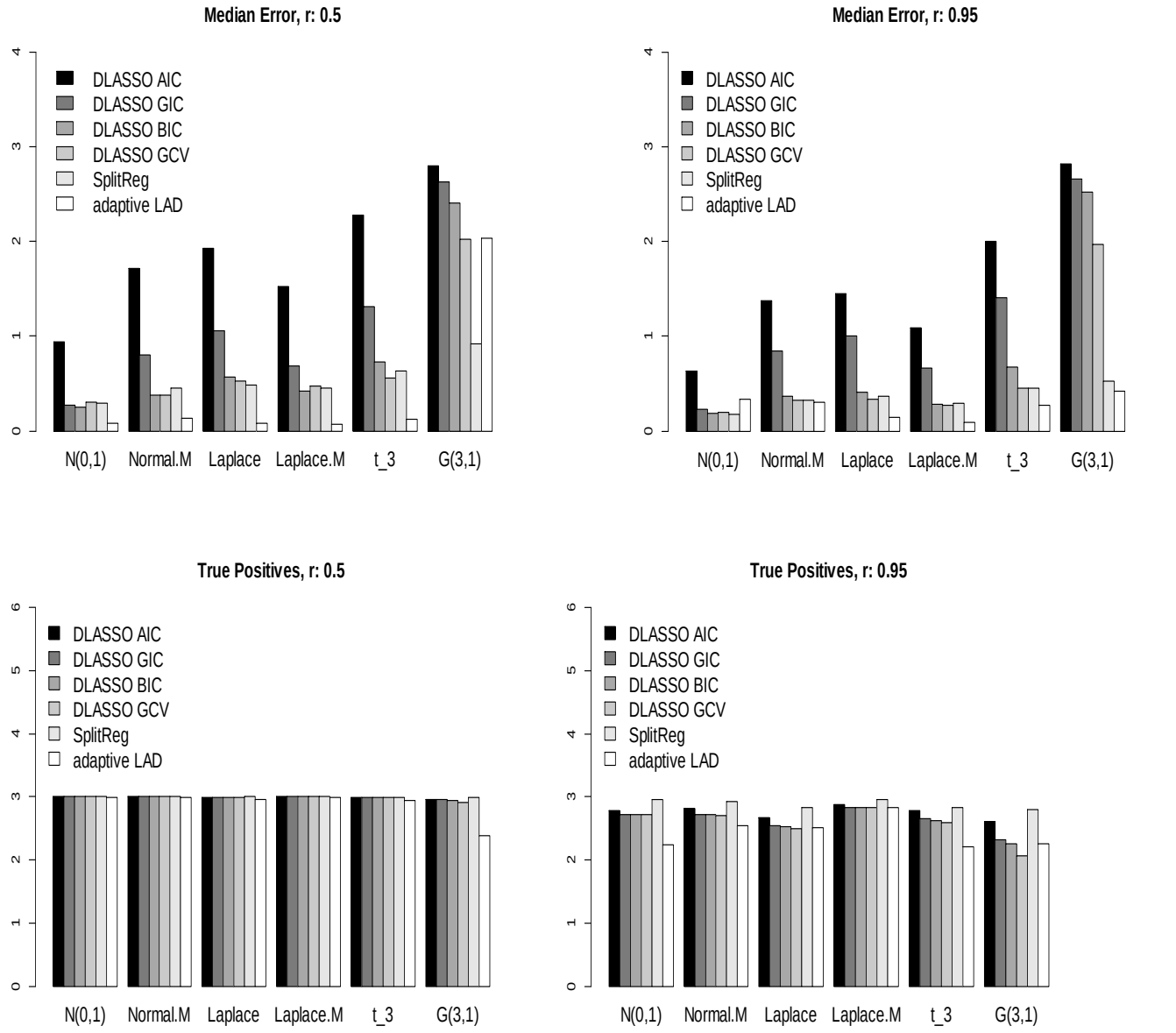


Figure 2: Comparison of regularized regression methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for simulation 2 and the bottom panels the average true positives when $p = 100$ and $n =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

The results support the performance of the methods: DLasso (AIC, GIC) do not perform well when the predictors are highly correlated, and the adaptive Lad Lasso method outperforms all other methods as departures from normality increase. This is particularly evident in the case $Laplace.M$ and t_3.

3.3 Simulation 3: low- dimensional with non-sparse coefficients (Case 2)

To investigate the performance of variable selection methods, we set up a new simulation where we have β_j as in case 2, that is a sparse situation. Table 3A, Table 3B and Figure 3 report the median model error over 500 replications for the cases $p = 50$ and $n = 100$

Table3.A: Average Median Model Error over 500 replications for the case: $p = 50, n = 100, r = 0.5$, and β values as in simulation 3, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.260	0.228	0.221	0.222	**0.159**	2.788
Normal. M	0.388	0.785	0.355	0.360	**0.291**	2.260
Laplace	0.400	0.851	0.366	0.365	**0.302**	2.776
Laplace. M	0.389	0.632	0.339	0.347	**0.250**	2.562
t_3	0.521	1.148	0.523	0.472	**0.409**	2.429
G(3,1)	1.537	3.618	3.566	1.657	**0.450**	2.400

Table3B: Average Median Model Error over 500 replications for the case: $p = 15, n = 100, r = 0.95$, and β values as in simulation3, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.757	0.428	0.460	0.513	**0.117**	0.366
Normal. M	0.968	0.819	0.744	0.826	**0.203**	0.712
Laplace	1.091	0.865	0.787	0.847	**0.209**	0.496
Laplace. M	1.036	0.741	0.684	0.738	**0.188**	0.488
t_3	1.133	1.107	0.977	0.933	**0.275**	0.426
G(3,1)	2.944	3.564	3.526	2.678	**0.352**	0.610

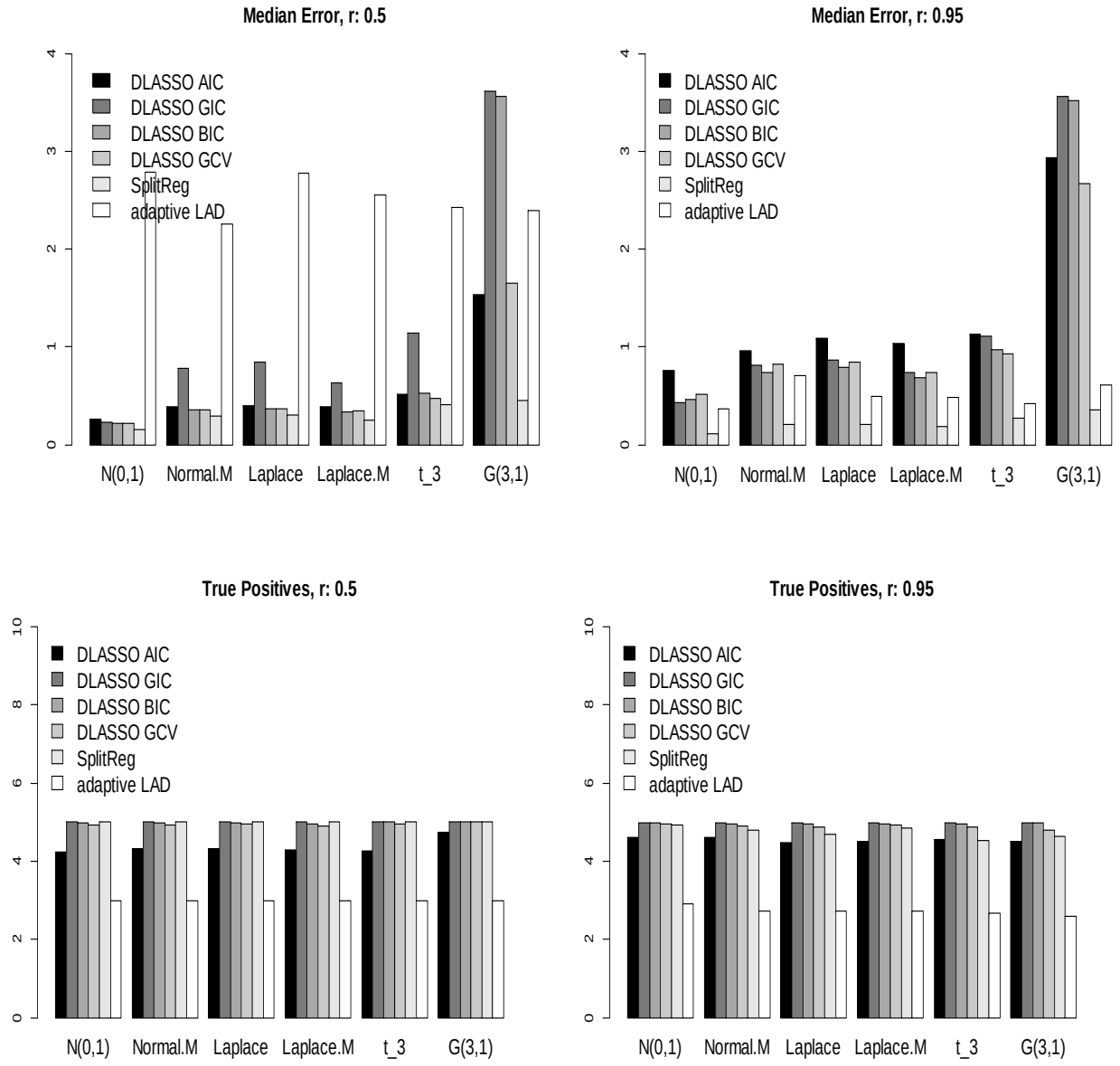


Figure 3: Comparison of regularized regression methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for simulation 3 and the bottom panels the average true positives when $p = 50$ and $n = 100$.

From the results in Table 3A, Table 3B and Figure 3, our simulation study confirms that the SRR outperforms all other methods as departures from normality increase. This is particularly evident in the case when the predictors are highly correlated.

3.4 Simulation 4: high-dimensional with non-sparse coefficients (Case 2)

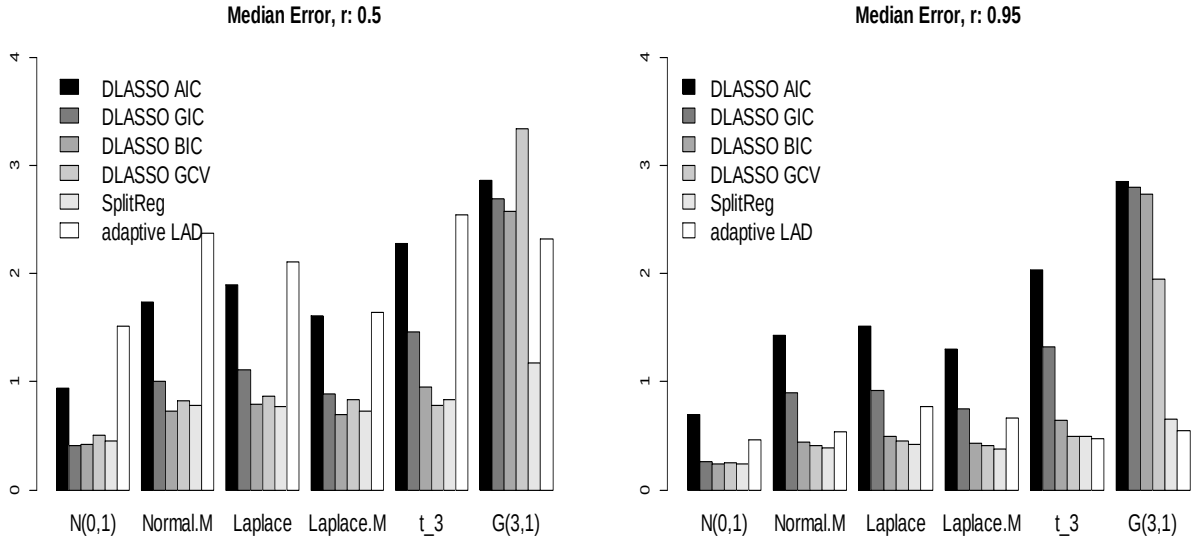
To investigate the performance of variable selection methods, we set up a new simulation where we have β_j as in case 2, that is a sparse situation. Table 4A, Table 4B and Figure 4 report the median model error over 500 replications for the cases $p = 100$ and $n = 50$.

Table 4A: Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.5$, and β values as in simulation 4, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.942	**0.410**	0.417	0.507	0.452	1.515
Normal. M	1.739	1.000	**0.731**	0.826	0.782	2.378
Laplace	1.893	1.107	0.793	0.863	**0.775**	2.108
Laplace. M	1.607	0.892	**0.693**	0.830	0.729	1.648
t_3	2.282	1.466	0.953	**0.786**	0.839	2.549
G(3,1)	2.864	2.694	2.578	3.347	**1.174**	2.328

Table 4B: Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.95$, and β values as in simulation 4, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.701	0.264	**0.236**	0.247	0.237	0.459
Normal. M	1.428	0.901	0.438	0.409	**0.388**	0.542
Laplace	1.514	0.919	0.491	0.449	**0.424**	0.772
Laplace. M	1.298	0.749	0.428	0.410	**0.379**	0.666
t_3	2.038	1.324	0.642	0.500	0.496	**0.469**
G(3,1)	2.852	2.799	2.742	1.949	0.658	**0.543**



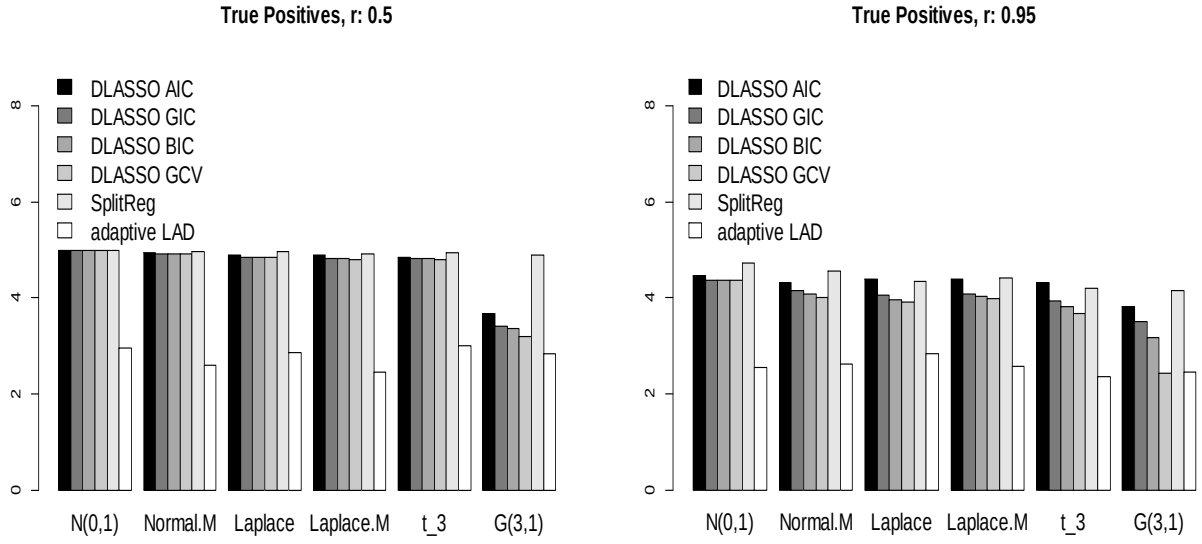


Figure 4: Comparison of regularized regression methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for simulation 4 and the bottom panels the average true positives when $p = 100$ and $n = 50$.

From the results in Table 4A and Table 4B and Figure 4 our simulation study confirms the performances of the adaptive Lad Lasso and the SRR methods outperform all other methods for most error distributions, Furthermore, the results show how DLasso (AIC) is the worst performing method in the case of departure from normality especially when the predictors are highly correlated.

3.5 Simulation 5: low- dimensional with non-sparse coefficients (Case 3)

To investigate the performance of variable selection methods, we set up a new simulation where we have β_j as in case 3, that is a non-sparse situation. Table5A, Table 5B and Figure 5 report the median model error over 500 replications for the cases $p = 15$ and $n = 100$.

Table5.A: Average Median Model Error over 500 replications for the case: $p = 15, n = 100, r = 0.5$, and β values as in simulation 5, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.387	0.139	0.304	0.386	**0.085**	0.219
Normal. M	0.456	0.230	0.292	0.456	**0.124**	0.256
Laplace	0.446	0.245	0.333	0.446	**0.132**	0.243
Laplace. M	0.430	0.217	0.287	0.429	**0.122**	0.249
t_3	0.377	0.310	0.344	0.377	**0.163**	0.221
G(3,1)	0.428	1.955	1.891	0.371	**0.198**	0.269

Table5B: Average Median Model Error over 500 replications for the case: $p = 15, n = 100, r = 0.95$, and β values as in simulation5, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.922	0.123	0.131	0.149	**0.040**	0.094
Normal. M	1.027	0.216	0.215	0.238	**0.066**	0.125
Laplace	0.740	0.243	0.246	0.316	**0.067**	0.091
Laplace. M	0.976	0.197	0.202	0.224	**0.059**	0.090
t_3	0.768	0.307	0.308	0.374	**0.083**	0.099
G(3,1)	0.564	1.064	0.917	0.570	**0.095**	0.141

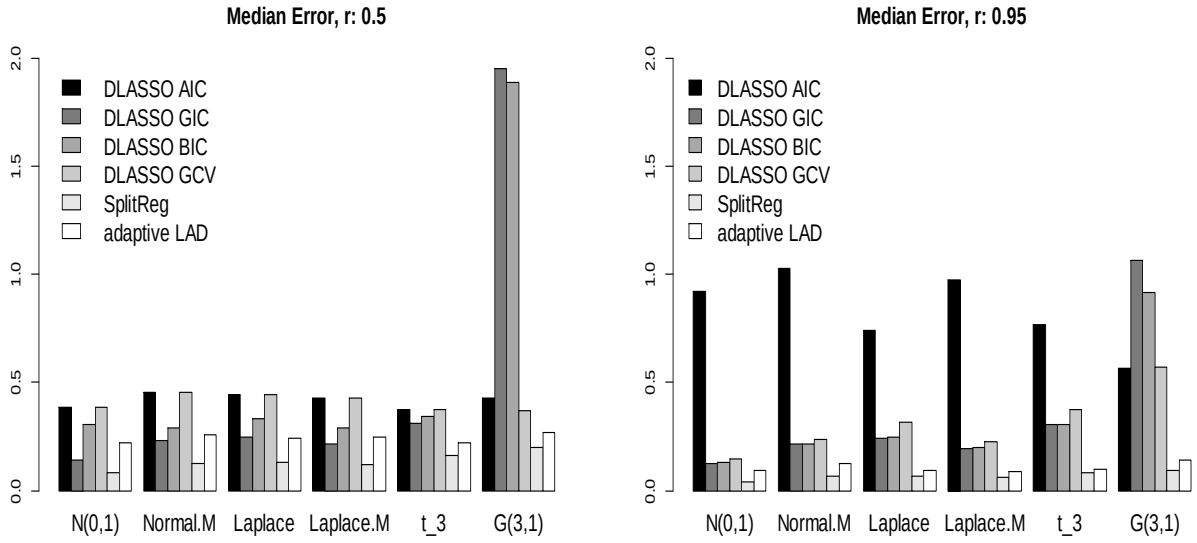


Figure 5: Comparison of regularized regression methods under different error distributions, for low (left) and high (right) correlated predictors. The median model error is plotted over 500 replications for simulation 5 when $p = 15$ and $n = 100$.

From the results in Figure 5, Table 5A and Table 5B our simulation study confirms that the SRR outperforms all other methods as departures from normality increase. This is particularly evident in the case when the predictors are highly correlated.

3.6 Simulation 6: high-dimensional with non-sparse coefficients (Case 3)

To investigate the performance of variable selection methods, we set up a new simulation where we have β_j as in case 3, that is a non-sparse situation. Table 6A, Table 6B and Figure 6 report the median model error over 500 replications for the cases $p = 100$ and $n = 50$.

Table 6.A: Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.5$, and β values as in simulation 6, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	0.949	**0.686**	0.828	1.780	0.779	1.675
Normal. M	1.720	1.160	**1.082**	2.043	1.170	2.220
Laplace	1.827	1.229	**1.041**	1.677	1.175	1.848
Laplace. M	1.610	1.077	**0.977**	1.794	1.265	1.723

t_3	2.304	1.534	**1.217**	1.536	1.234	1.584
G(3,1)	2.838	2.427	2.220	1.982	**1.732**	2.187

Table 6B: Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.95$, and β values as in simulation 6, Best method indicated in bold.

	DLasso AIC	DLasso GIC	DLasso BIC	DLasso CGV	SplitReg	adaptive LAD
N(0,1)	3.211	0.512	0.569	0.794	**0.428**	7.648
Normal. M	1.983	0.863	0.752	0.878	**0.674**	9.493
Laplace	2.289	0.956	0.816	0.942	**0.703**	10.072
Laplace. M	2.050	0.774	0.760	0.879	**0.633**	11.435
t_3	2.749	1.399	0.962	1.059	**0.844**	7.312
G(3,1)	2.745	2.676	2.591	2.264	**1.022**	7.942

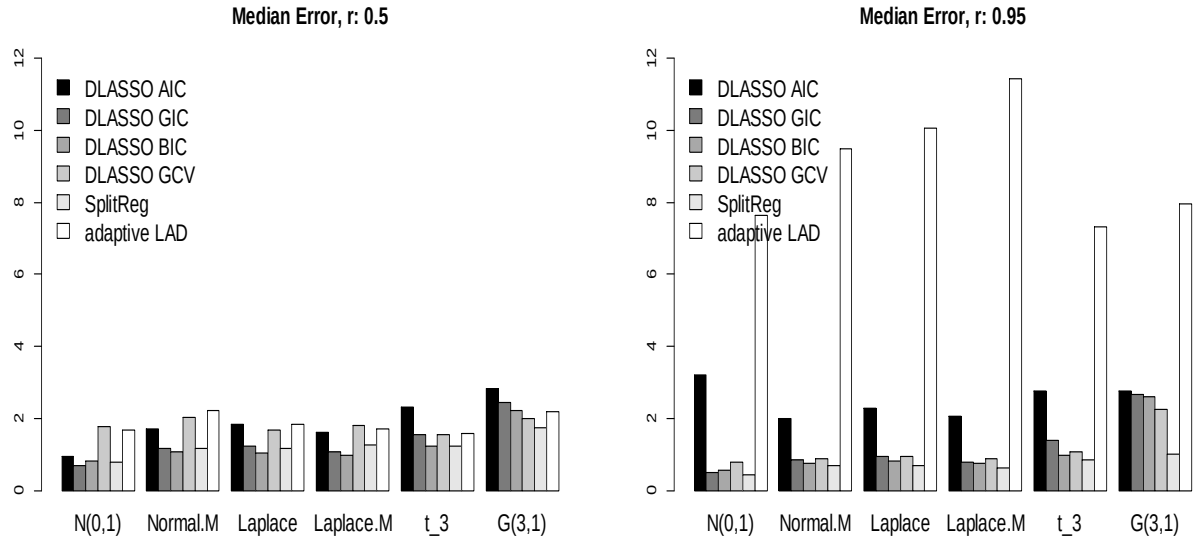


Figure 6: Comparison of regularized regression methods under different error distributions, for low (left) and high (right) correlated predictors. The median model error is plotted over 500 replications for simulation 6 when $p = 100$ and $n = 50$.

From the results in Table 6A and Table 6B and Figure 6 our simulation study confirms the performances of the SRR method outperform all other methods as departures from normality increase, Furthermore, the results show how adaptive Lad Lasso is the worst performing method in the case of departure from normality especially when the predictors are highly correlated.

4. Concluding remarks

Many approaches are developed in statistics that rely on the assumption of normality. These approaches are not suited to data that show clear departures from normality. This is often the case when data are contaminated, resulting in the presence of outliers. In this paper, we have considered recently developed variable selection methods, such as the Adaptive Lad Lasso, Split Regularized Regression (SRR) and DLasso. In a high dimensional setting, when $p \geq n$. In a simulation study, we show how the Adaptive Lad Lasso and the Split Regularized Regression (SRR) methods are superior to other methods, particularly for cases where there is a large departure from the normal distribution.

Acknowledgment

The author is very grateful to the University of Duhok, College of Science for their provided facilities, which helped improve this work's quality.

Conflict of interest

The author has no conflict of interest

References

1. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.
2. Zou, H. (2006). The adaptive Lasso and its oracle properties, Journal of the American Statistical Association 101, 1418–1429.
3. 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.
4. Koenker, R. and G. W. Bassett (1978). Regression quantiles, Econometrica 46, 33–50.
5. Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths, The Annals of Statistics 35 (3), 1012–1030.
6. 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.
7. Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through Huber's criterion and adaptive Lasso penalty, Electronic Journal of Statistics 5, 1015–1053.
8. Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination, Journal of Multivariate Analysis, 99(9), 2053-2081.
9. Arnold, T. B., and Tibshirani, R. J. (2016). Efficient implementations of the generalized Lasso dual-path algorithm, Journal of Computational and Graphical Statistics, 25(1):1–27.
10. Taddy, M. (2017). One-step estimator paths for concave regularization, *Journal of Computational and Graphical Statistics pp. 1–12*.
11 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.
12. Qin, Y., Li, S. and Yu, Y. (2017). Penalized Maximum Tangent Likelihood Estimation and Robust Variable Selection, https://arxiv.org/pdf/1708.05439.pdf.
13. Christidis, A.-A., Lakshmanan, L., Smucler, E., and Zamar, R. (2020)). Split regularized regression, Technometrics 62.3, pp. 330–338.
14. 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.
15. Tibshirani, R. J., and Taylor, J. (2011), The solution path of the generalized Lasso", Ann.Stat., 39(3), 1335-1371.
16. Haselimashhadi,H. and Vinciotti,V.(2016).A Differentiable Alternative to the Lasso Penalty, https://arxiv.org/abs/1609.04985#:~:text=Regularized%20regression%20has%20become%20very,inference%20where%20traditional%20methods%20fail.
17. Yu, K., C. Cathy, C. Reed, and D. Dunson (2013). Bayesian variable selection in quantile regression", Statistics and Its Interface 6, 261–274[17]
18. Li, Q., R. Xi, and N. Lin (2010), "Bayesian regularized quantile regression", Bayesian Analysis 5, 1–24.
19. Xu, J. and Ying, Z. (2010). Simultaneous estimation and variable selection in median regression using lasso-type penalty, Annals of the Institute of Statistical Mathematics 62, 487–514.

انحدار لاسو المكيف وانحدار المنظم المنفصل وانحدار دلاسو: دراسة محاكاة لاختيار المتغيرات

حسين عبد الرحمن هاشم
قسم الرياضيات، كلية العلوم، جامعة دهوك، دهوك، العراق

الخلاصة:في هذا البحث نقارن ثلاث طرق رئيسية لاختيار المتغيرات لنماذج الانحدار الخطي: انحدار لاسو المكيف ، وانحدار المنظم المنفصل **(SRR)**، و انحدار لاسو**(AIC, GIC, BIC, CGV)**. نعرض أداء هذه الأساليب في دراسة محاكاة من خلال النظر في خطأ النموذج المتوسط. نحن نعتبر أيضًا الحالة التي يتجاوز فيها عدد المتغيرات المستقلة عدد المشاهدات. تحدد دراسة المحاكاة الطرق الأفضل في جميع سيناريوهات الانحدار الخطي.

الكلمات المفتاحية: اختيار متغير ؛ لاسو ؛ لاسو ؛ تسوية.