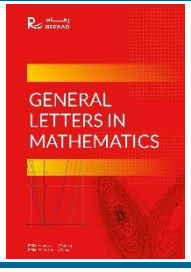
Available online at www.refaad.com
Gen. Lett. Math., 12(3) (2022), 126- 138
Research Article

General Letters in Mathematics (GLM)

Journal Homepage: https://www.refaad.com/Journal/Index/1

ISSN: 2519-9277 (Online) 2519-9269 (Print)



Comparison of Some Group Variable Selection Methods in High Dimensional Multiple Linear Regression via Simulation Study

Rejin Abdulqader Ibrahim[a], Hussein Abdulrahman Hashem[b,*]

[a] Department of Mathematics, Faculty of Sciences, University of Zakho, Kurdistan Region, Iraq.
Email: rejin.ibrahim@stud.uoz.edu.krd
[b] Department of Mathematics, College of Sciences, University of Duhok, Kurdistan Region, Iraq.
Email: hussein.hashem@uod.ac

Abstract

In many applications, covariates possess a grouping structure that can be incorporated into the analysis to select important groups as well as important members of those groups. In this paper, we reviewed some group variable selection methods in the penalized regression model. This paper investigates by comparing the performance of seven previously proposed group variable selection methods; the group Lasso estimates, the group Lasso net estimates, the sparse group Lasso estimates, the group scad estimates, the group scad net estimates, the group mcp estimates, and the group gel estimate via a simulation study. The simulation study is used in determining which methods are best in all of the linear regression scenarios.

Keywords: Variable Selection; Lasso; Group Lasso; Regularization; Simulation Study.

2010 MSC: 62J07, 62J20, 2J86.

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 [23] proposed Lasso (Least Absolute Shrinkage and Selection Operator), which minimizes the residual sum of squares subject to an L_1-norm constraint. The Lasso penalty results in some coefficients being estimated to completely zero, thus performing estimation and variable selection simultaneously. Following the seminal paper of Tibshirani [23], various extensions of the Lasso were developed, for example, adaptive Lasso [31] and Smoothly Clipped Absolute Deviation (SCAD) [7], etc.

* Corresponding author
Email addresses: hussein.hashem@uod.ac (Hussein Abdulrahman Hashem).
doi:10.31559/glm2022.12.3.3
Received: 27 July 2022 Revised: 3 Sept 2022 Accepted: 22 Sept 2022

Quantile regression, introduced by Koenker and Bassett [15], 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 [21] 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. [25], combining the idea of Least Absolute Deviance (LAD) and adaptive Lasso. Lambert-Lacroix and Zwald [16] 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 [1] implemented the dual algorithm and their implementation is available in the R package genLasso. Taddy [22] proposed the gamma Lasso (GL) algorithm which can be seen as a computationally more Attractive and multi-convex relaxation of best variable selection. Yi and Huang [27] 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. [20] proposed a method called Maximum Tangent Likelihood Estimation (MTE). Christidis et al. [5] 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. [30] 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 [24].

Group penalty can be applied when the grouping structure is unknown (thus must be estimated). In biological studies, genetic data usually have background scientific information. For example, genes with the same biological pathway are often located in a neighborhood forming a group.

Several penalties considering the grouping structure have been proposed. group Lasso which uses the L_2-norm of the coefficients within a group was proposed by Bakin [2] and extended by Yuan and Lin [28]. Huang *et al.* [11] showed group SCAD and group Minimax Concave Penalty (MCP) for covariates possessing a grouping structure to select important groups. In the context of the quantile regression models, Ciuperca [6] proposed an adaptive group, Lasso, that adaptive Lasso penalty and established the sparsity and asymptotic normality of their methods. Kato [13] considered high dimensional sparse quantile regression models with group Lasso penalty and attained a non-asymptotic error bound of the estimation error. The Group Lasso penalty was investigated for the classification problem by Hashem *et al.* [10]. McDonald [19] proposed a new R package for computing sparse group Lasso. Li et al. [18] proposed an adaptive sparse group Lasso penalty for Logistic regression and the proposed method is used for cancer data diagnosis.

In the next section, an overview will be given to some group variable selection methods in linear regression.

2. Method

We began with 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 [23] 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:

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

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 which 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 Group Lasso Methods

In some real data applications, predictors can be grouped in a natural way such that selecting groups of variables are of interest. Genetic data sometimes has this property. For example, data from genes can be grouped such that a group of genes corresponds to the same biological pathway. The group Lasso method [28] is ideal for this type of situation; it will shrink entire groups of predictors to zero or estimate the regression coefficients for the entire group. The regression coefficients of groups will either all be zero, or all be nonzero. For the group Lasso method, assume the predictor variables can be naturally grouped into $\boldsymbol{k}$ groups for $\boldsymbol{k = 1,\ldots,K}$, where each group consists of pk predictor variables such that $\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{p_k} = \boldsymbol{p}$. Within each group $\boldsymbol{k}$, there are $\boldsymbol{j}$ predictors for $\boldsymbol{j = 1,\ldots,p_k}$. The predictor variables should be standardized so that each $\boldsymbol{x_{ij}}$ has mean $\boldsymbol{0}$ and variance $\boldsymbol{1}$ for $\boldsymbol{j = 1,\ldots,p}$.The criterion to be minimized is:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k=1}^{K} x_{ik}\beta_k)^2 + n\lambda\sum_{k=1}^{K}\|\beta_k\|_2 \tag{5}$$

Where $\boldsymbol{\lambda \ge 0}$ is a tuning parameter, $\boldsymbol{y_i}$is the $\boldsymbol{ith}$ response, $\boldsymbol{x_{ik}}$ is a $\boldsymbol{1\ x\ p_k}$ vector of predictors in the $\boldsymbol{kth}$ group for the $\boldsymbol{ith}$ observation, and $\boldsymbol{\beta_k}$ is a $\boldsymbol{p_k\ x\ 1}$ vector of regression coefficients for group $\boldsymbol{k}$. As for the criterion above, for each group of predictors, minimize the sum of the squared distances while simultaneously shrinking unimportant groups with the Lasso penalty (the $\boldsymbol{L_2}$ the norm in this case). The tuning parameter $\boldsymbol{\lambda}$ controls the rate of shrinkage and can be chosen using cross-validation. In particular, Yuan and Lin [28] use a shrinkage parameter based on an approximate $\boldsymbol{C_p}$-type criterion. The Lasso method of simultaneous estimation and selection is ideal for predictors with little to no multicollinearity, but not for data with outliers. In particular, because it uses the Least Square Estimate (LSE), the group Lasso performs poorly in terms of robustness [14]. The computation of the group Lasso is based on the shooting algorithm [9]. Originally, this method was proposed for the Lasso method but was adapted for the group Lasso [28].

2.3 Group Descent Algorithms (grpreg)

Breheny and Huang [4] proposed the algorithms to fit models with grouped penalties which are either inefficient for models with large numbers of predictors, or limited to linear regression models, models in which the members of a group are orthogonal to each other, or both. Breheny and Huang [4] combine the ideas of coordinate descent optimization and local approximation of penalty functions to introduce a new, general algorithm for fitting models with grouped penalties. The resulting algorithm is stable and very fast even when the number of variables is much larger than the sample size. Breheny and Huang [4] apply the algorithm to models with grouped penalties, but note that the idea may be applied to other penalized regression problems in which the penalties are complicated but not necessarily grouped. All group-related methods can be found in the R package grpreg [4] except the ElasticNet, which is available *in* glmnet.

3. Simulation Study

In this section, we compare some group variable selection methods in low-dimensional with sparse and non-sparse coefficients($p = 50, n = 100$) 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. [29] and Li et al. [17], where $y = \beta_0 + x\beta + u$, with $\beta_0 = 0$ and we create a group structure by simulating 10 groups, each consisting of 10 covariates. The 100 variables are assumed to follow a multivariate normal distribution $N(0; \Sigma)$, with Σ having a diagonal block structure. Each block corresponds to one group and is defined by the matrix $r^{|i-k|}$, $i = 1,. \ . \ .,10$, $k = 1,. \ . \ .,10$. For the correlation r, we experiment both with $r = 0.95$ (well-defined group structure) and $r = 0.5$. For the β values we consider three cases:

(1) The values for the first three groups are given by $\beta_j =$ $(0.5, 1, 1.5, 2, 2.5, 2, 2, 2, 2, 2), (2, 2, 1, 1, 1, 1, 3, 3, 3, 3), (1, 1, 1, 2, 2, 2, 3, 3, 3, 3)$, and they are set to zero for all other groups, which corresponds to the sparse case with group structures in the predictors.

(2) $\beta_j = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$, and they are set to zero for all other groups, which corresponds to the very sparse case with group structures in the predictors.

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

For the error ϵ, we consider the following distributions:

- Normal: $N(0; 1)$
- Laplace distribution with location 0 and scale 1: $Laplace(0, 1)$
- A t distribution with 3 degrees of freedom: t_3
- Gamma distribution: $G(3,1)$
- A mixture of two normal distributions: $0.1N(0, 100) + 0.9N(0, 1)$
- A mixture of two Laplace distributions: $0.1Laplace\ (0, 1) + \ 0.9Laplace(0, 2)$
- Chi-square distributions: $\chi^2_{(3)}$

We compare the group variable selection methods described in the previous section, namely:

- "grp.lasso": group Lasso penalty [28].
- "grp.lasso.net": group Lasso penalty + L_2 penalty, extra parameters $alpha$ [12, 26].
- "sparse.grp.lasso": sparse group Lasso penalty (group Lasso + Lasso), extra parameters tau [12, 26].
- "grp.scad": group smoothly clipped absolute deviation, extra parameters $gamma$ [12, 26].
- "scad.net": smoothly clipped absolute deviation + L_2 penalty, extra parameters $gamma$ and $alpha$ [12, 26].
- "grp.mcp": group minimax concave penalty, extra parameters $gamma$ [12, 26].
- "gel": group exponential Lasso [3].

For the grp.lasso, grp.lasso.net, sparse.grp.lasso, and grp.scad.net methods we use the R package oem and for the grp.scad, grp.mcp and gel methods we use the R package grpreg.

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

In this section, we consider low-dimensional data with sparse coefficients set with $p = 50$ and $n = 100$. Table 1A, Table1B, Figure 1A, and Figure1B 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. Table 1A, Table1B, Figure1A, and Figure1B 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.

Our results show that: the grp.scad.net method does not perform well; the grp.lasso and the grp.lasso.net methods outperform all other methods when the predictors are highly correlated for most error distributions.

Table(1A): Average Median Model Error over 500 replications for the case: $p = 50, n = 100, r = 0.5$, and β values as in example 1 (case1), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	0.427	0.427	0.436	**0.304**	0.500	0.306	0.306
$Laplace$	0.851	0.851	0.871	**0.592**	0.985	**0.592**	**0.592**
t_3	1.113	1.113	1.137	**0.755**	1.277	0.757	0.761
$G(3,1)$	1.326	1.326	1.350	**0.889**	1.474	0.890	0.901
$Normal.M$	0.749	0.749	0.771	0.544	0.890	0.545	**0.541**
$Laplace.M$	2.914	2.914	2.964	**1.955**	3.214	1.972	1.975
$Chi(3)$	2.611	2.611	2.649	**1.736**	2.890	1.745	1.751

Table(1B): Average Median Model Error over 500 replications for the case: $p = 50, n = 100, r = 0.95$, and β values as in example 1(case1), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	**0.215**	**0.215**	0.248	0.313	0.323	0.314	0.310
$Laplace$	**0.430**	**0.430**	0.496	0.590	0.808	0.594	0.573
t_3	**0.564**	**0.564**	0.649	0.757	1.077	0.758	0.710
$G(3,1)$	**0.621**	**0.621**	0.727	0.886	1.220	0.890	0.843
$Normal.M$	**0.382**	**0.382**	0.443	0.555	0.658	0.556	0.540
$Laplace.M$	**1.501**	**1.501**	1.667	1.917	3.018	1.920	1.652
$Chi(3)$	**1.450**	**1.450**	1.621	1.791	2.853	1.791	1.538

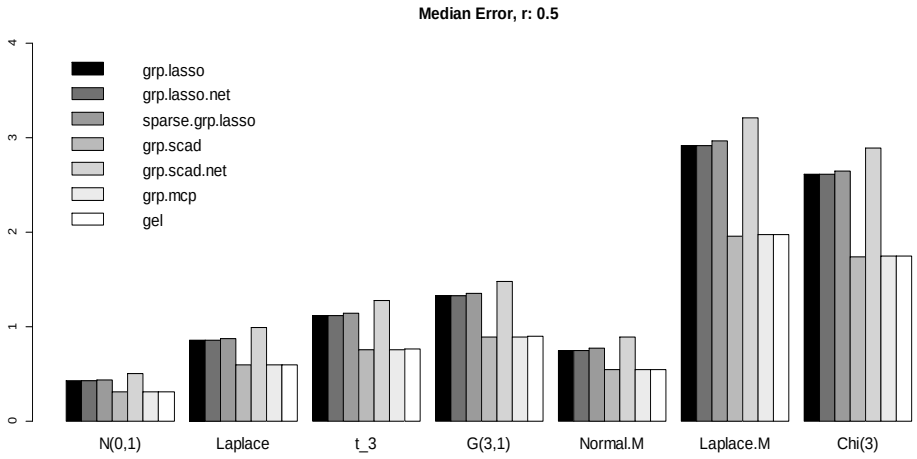


Figure (1A): Comparison of group variable selection methods under different error distributions, for low correlated predictors. The median model error over 500 replications for example 1 (case1), when $p = 50$ and $n = 100$.

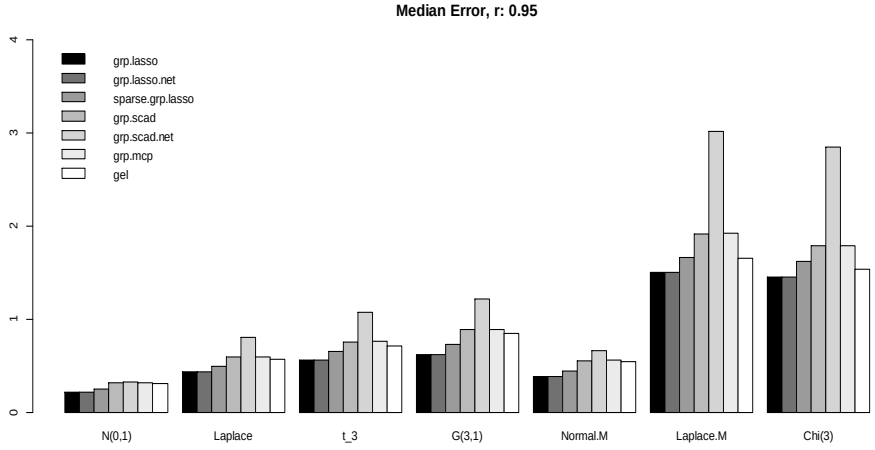


Figure (1B): Comparison of group variable selection methods under different error distributions, for highly correlated predictors. The median model error over 500 replications for example 1(case1), when $p = 50$ and $n = 100$.

3.2 Example 2: high-dimensional with 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 sparse coefficients with $p = 100$ and $n = 50$. Table 2A, Table2B, Figure2A, and Figure2B report the median model error over 500 replications, with the model error, computed in the same way as in Example 1.

Table (2A): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 100\,, \mathrm{n} = 50, \mathrm{r} = 0.5,$ and β values as in example 1 (case1), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	3.413	3.413	3.797	**0.598**	**0.598**	**0.598**	0.599
$Laplace$	2.983	2.983	3.150	**1.147**	**1.147**	1.150	1.156
t_3	3.573	3.573	3.793	**1.422**	**1.422**	1.424	1.423
$G(3,1)$	2.986	2.986	3.360	1.754	**1.752**	1.760	1.780
$Normal.M$	3.116	3.116	3.342	**1.066**	**1.066**	**1.066**	1.064
$Laplace.M$	4.616	4.616	5.149	3.847	**3.762**	3.844	3.786
$Chi(3)$	5.006	5.006	5.435	3.494	3.490	3.509	**3.447**

Table (2B): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 100, \mathrm{n} = 50, \mathrm{r} = 0.95,$ and β values as in example 1(case1), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	3.762	3.762	3.148	0.582	**0.496**	0.582	0.558
$Laplace$	4.606	4.606	3.988	1.170	1.136	1.169	**1.013**
t_3	3.438	3.438	2.937	1.358	**0.787**	1.486	1.166
$G(3,1)$	4.256	4.256	3.714	1.806	**0.990**	1.806	1.515
$Normal.M$	4.060	4.060	3.522	1.123	1.269	1.114	**0.966**
$Laplace.M$	4.233	4.233	3.791	3.655	**1.912**	3.657	2.825
$Chi(3)$	3.412	3.412	3.122	3.384	**2.523**	3.396	2.689

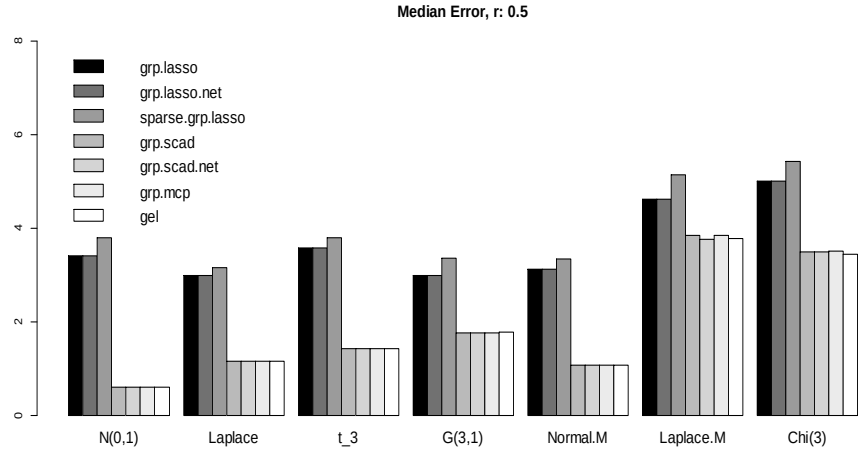


Figure (2A): Comparison of group variable selection methods under different error distributions, for low correlated predictors. The median model error over 500 replications for example 1 (case1), when $\mathrm{p} = 100$ and $\mathrm{n} = 50$.

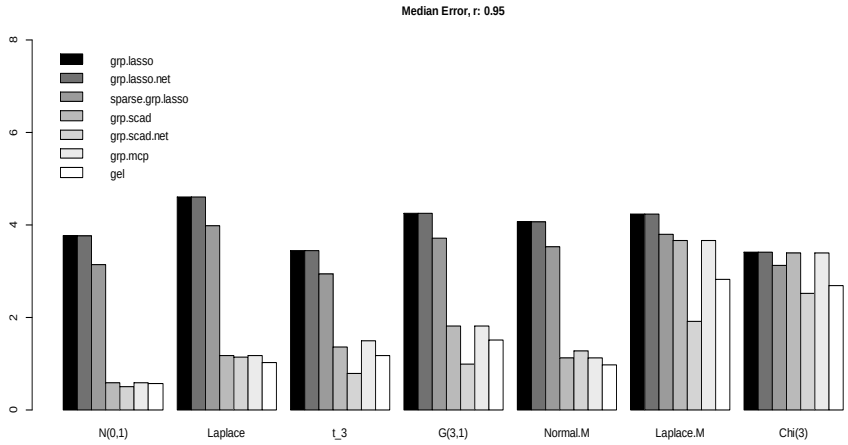


Figure (2B): Comparison of group variable selection methods under different error distributions, for highly correlated predictors. The median model error over 500 replications for example 1(case1), when $\mathrm{p} = 100$ and $\mathrm{n} = 50$.

The results support the performance of the methods: $\mathrm{grp.lasso}$ and $\mathrm{grp.lasso.net}$ do not perform well when the predictors are highly correlated, and the $\mathrm{grp.scad.net}$ method outperforms all other methods as departures from normality increase. This is particularly evident in the cases of the $G(3,1)$ and $Laplace.M$.

3.3 Example 3: low- dimensional with very sparse coefficients (Case 2)

To investigate the performance of group variable selection methods in example 1, we set up a new simulation where we have β_j as in case 2 that is a very sparse problem in which most of the coefficients are zero. Table 3A, Table3B, Figure3A, and Figure3B report the median model error over 500 replications, with the model error, computed in the same way as in Example 1.

Table (3A): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 50, \mathrm{n} = 100, \mathrm{r} = 0.5,$ and β values as in example 3(case2), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	0.418	0.418	0.421	0.102	0.495	0.101	**0.062**
$Laplace$	0.854	0.854	0.858	0.201	0.974	0.199	**0.131**
t_3	1.120	1.120	1.122	0.251	1.250	0.248	**0.158**
$G(3,1)$	1.291	1.291	1.295	0.285	1.438	0.286	**0.173**
$Normal.M$	0.772	0.772	0.778	0.184	0.896	0.180	**0.115**
$Laplace.M$	2.926	2.926	2.944	0.637	3.164	0.631	**0.399**
$Chi(3)$	2.674	2.674	2.683	0.609	2.885	0.585	**0.384**

Table (3B): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 50, \mathrm{n} = 100, \mathrm{r} = 0.95,$ and β values as in example 3 (case2), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	0.205	0.205	0.208	0.101	0.483	0.099	**0.053**
$Laplace$	0.522	0.522	0.540	0.201	0.985	0.198	**0.103**
t_3	0.681	0.681	0.701	0.260	1.271	0.261	**0.138**
$G(3,1)$	0.807	0.807	0.836	0.304	1.461	0.304	**0.152**
$Normal.M$	0.465	0.465	0.481	0.191	0.894	0.191	**0.097**
$Laplace.M$	2.097	2.097	2.153	0.664	3.177	0.671	**0.315**
$Chi(3)$	1.906	1.906	1.965	0.574	2.918	0.591	**0.288**

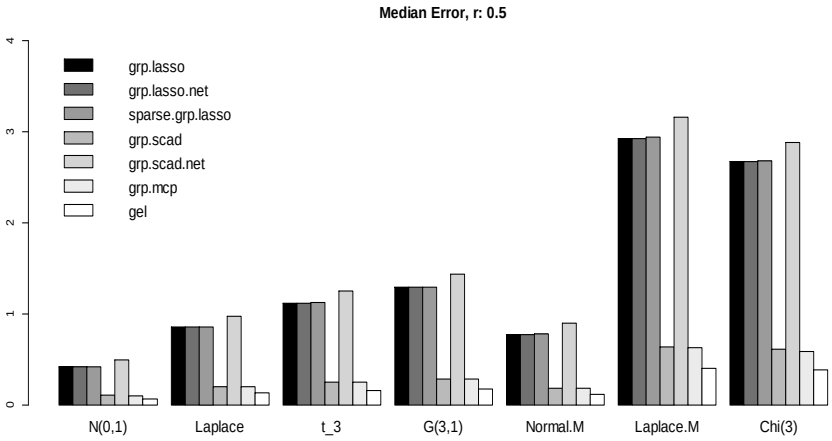


Figure (3A): Comparison of group variable selection methods under different error distributions, for low correlated predictors. The median model error over 500 replications for example 3 (case2), when $\mathrm{p} = 50$ and $\mathrm{n} = 100$.

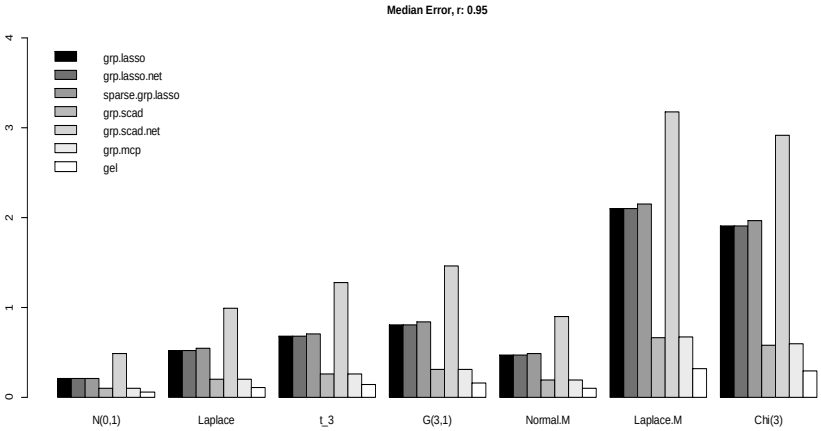


Figure (3B): Comparison of group variable selection methods under different error distributions, for highly correlated predictors. The median model error over 500 replications for example 3(case2), when $\mathrm{p} = 50$ and $\mathrm{n} = 100$.

The results in Table 3A, Table3B, Figure3A, and Figure3B in our simulation study confirm that the group exponential Lasso (gel) outperforms all other methods as departures from normality increase. This is particularly evident in the case when the predictors are highly correlated.

3.4 Example 4: high-dimensional with very sparse coefficients (Case 2)

We consider a similar setting to simulation 3.3 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 3A, Table4B, Figure4A, and Figure4B report the median model error over 500 replications, with the model error, computed in the same way as in Example 1.

Table (4A): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 100, \mathrm{n} = 50, \mathrm{r} = 0.5,$ and β values as in example 3 (case2), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	0.348	0.348	0.270	0.218	0.237	0.211	**0.145**
$Laplace$	0.626	0.626	0.543	0.433	0.495	0.421	**0.253**
t_3	0.676	0.676	0.615	0.510	0.609	0.505	**0.295**
$G(3,1)$	0.824	0.824	0.690	0.614	0.699	0.592	**0.373**
$Normal.M$	0.748	0.748	0.675	0.386	0.607	0.375	**0.259**
$Laplace.M$	2.366	2.366	2.401	1.385	3.127	1.389	**0.798**
$Chi(3)$	2.154	2.154	2.259	1.268	3.263	1.220	**0.770**

Table (4B): Average Median Model Error over 500 replications for the case: $p = 100, n = 50, r = 0.95$, and β values as in example 3(case2), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	0.635	0.635	0.364	0.204	0.187	0.209	**0.103**
$Laplace$	0.680	0.680	0.434	0.406	0.385	0.401	**0.185**
t_3	0.656	0.656	0.442	0.503	0.471	0.479	**0.226**
$G(3,1)$	0.677	0.677	0.484	0.658	0.619	0.625	**0.269**
$Normal.M$	0.560	0.560	0.373	0.391	0.368	0.382	**0.193**
$Laplace.M$	1.233	1.233	1.018	1.348	1.412	1.334	**0.558**
$Chi(3)$	1.027	1.027	0.827	1.256	1.325	1.183	**0.521**

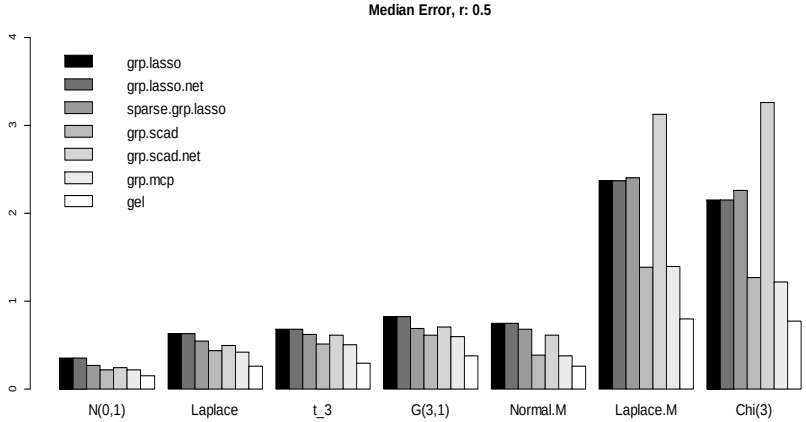


Figure (4A): Comparison of group variable selection methods under different error distributions, for low correlated predictors. The median model error over 500 replications for example 3 (case2), when $p = 100$ and $n = 50$.

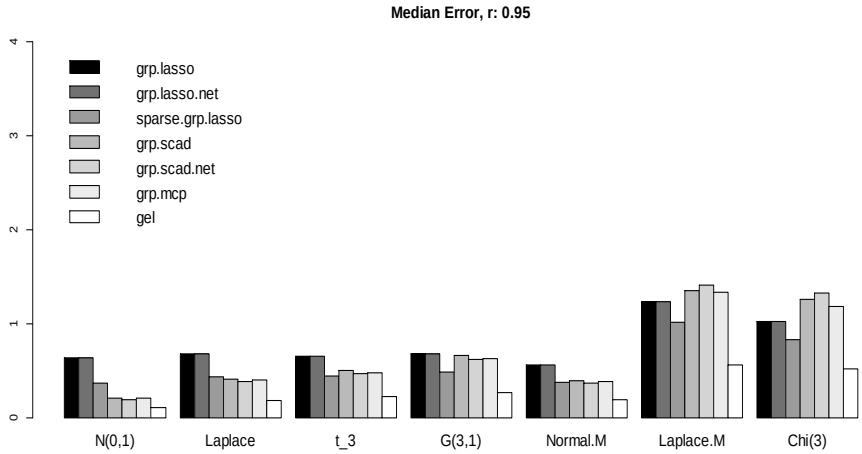


Figure (4B): Comparison of group variable selection methods under different error distributions, for highly correlated predictors. The median model error over 500 replications for example 3(case2), when $p = 100$ and $n = 50$.

From the results in Table 4A, Table4B, Figure4A, and Figure4B, it is shown that our simulation study confirms that the group exponential Lasso (**gel**) outperforms all other methods as departures from normality increase. This is particularly evident in the case when the predictors are highly correlated.

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

To investigate the performance of group variable selection methods in example 1, we set up a new simulation where we have β_j as in case 3, that is a non-sparse situation. Table 5A, Table5B, Figure5A and Figure5B reports the median model error over 500 replications for the cases $p = 50$ and $n = 100$.

Table (5A): Average Median Model Error over 500 replications for the case: $p = 50, n = 100, r = 0.5$, and β values as in example 5 (case3), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	0.485	0.485	0.487	**0.427**	0.498	0.446	0.441
$Laplace$	0.963	0.963	0.966	**0.617**	0.989	0.684	0.746
t_3	1.234	1.234	1.237	**0.700**	1.268	0.809	0.946
$G(3,1)$	1.440	1.440	1.444	**0.765**	1.477	0.856	0.837
$Normal.M$	0.868	0.868	0.870	**0.645**	0.894	0.745	0.724
$Laplace.M$	3.200	3.200	3.205	**1.044**	3.257	1.234	1.234
$Chi(3)$	2.898	2.898	2.902	**0.984**	2.950	1.034	1.018

Table (5B): Average Median Model Error over 500 replications for the case: $p = 50, n = 100, r = 0.95$, and β values as in example 5(case3), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	0.386	0.386	0.394	0.547	0.507	0.535	**0.182**
$Laplace$	0.809	0.809	0.820	1.008	0.954	1.008	**0.297**
t_3	1.046	1.046	1.063	1.288	1.231	1.274	**0.370**
$G(3,1)$	1.253	1.253	1.273	1.310	1.470	1.396	**0.401**
$Normal.M$	0.770	0.770	0.781	1.015	0.891	0.981	**0.273**
$Laplace.M$	2.931	2.931	2.958	2.141	3.324	2.489	**0.734**
$Chi(3)$	2.569	2.569	2.599	1.853	2.919	2.104	**0.698**

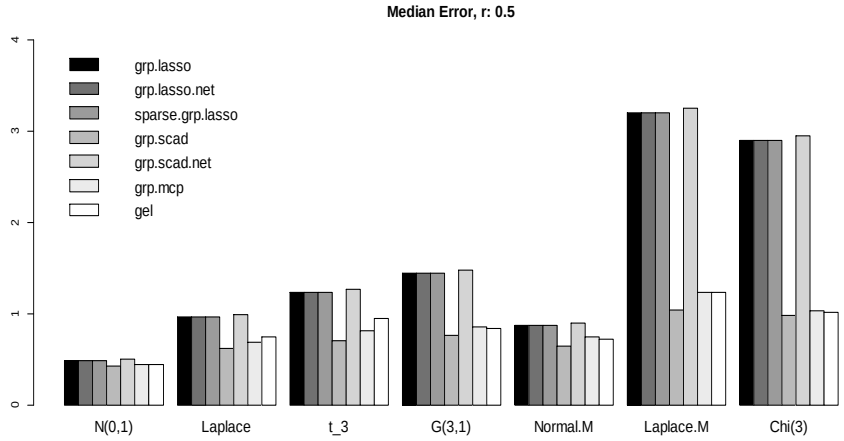


Figure (5A): Comparison of group variable selection methods under different error distributions, for low correlated predictors. The median model error over 500 replications for example 3 (case3), when $p = 50$ and $n = 100$.

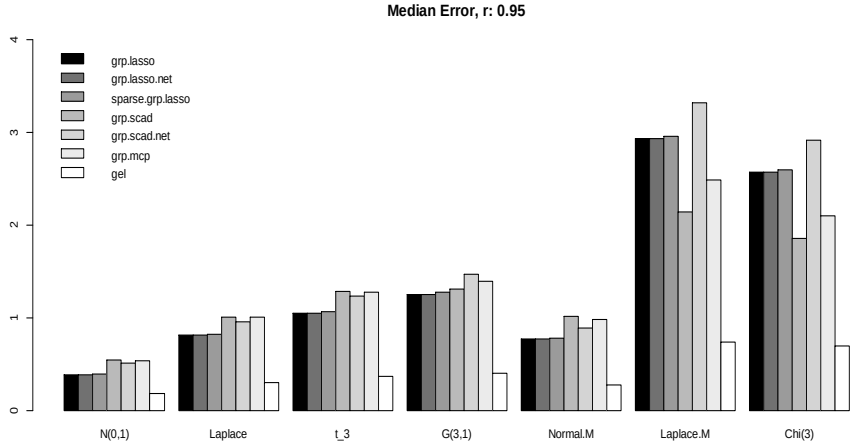


Figure (5B): Comparison of group variable selection methods under different error distributions, for highly correlated predictors. The median model error over 500 replications for example 3(case3), when $p = 50$ and $n = 100$.

From the results in Table 5A, Table5B, Figure5A, and Figure5B, our simulation study confirms that the group exponential Lasso (gel) outperforms all other methods as departures from normality increase. This is particularly evident in the case when the predictors are highly correlated.

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

To investigate the performance of group variable selection methods in example 2, we set up a new simulation where we have β_j as in case 3, that is a non-sparse situation. Table 6A, Table6B, Figure6A and Figure6B report the median model error over 500 replications for the cases $p = 100$ and $n = 50$.

Table (6A): Average Median Model Error over 500 replications for the case: $\mathrm{p} = 50, \mathrm{n} = 100, \mathrm{r} = 0.5,$ and β values as in example 5(case3), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	**0.672**	**0.672**	0.705	2.325	0.975	2.325	2.325
$Laplace$	**1.228**	**1.228**	1.290	2.559	1.865	2.986	2.727
t_3	**1.528**	**1.528**	1.598	1.733	2.333	2.208	1.725
$G(3,1)$	**1.929**	**1.929**	2.015	2.794	2.822	3.014	2.897
$Normal.M$	**1.211**	**1.211**	1.266	1.605	1.797	2.027	1.854
$Laplace.M$	4.140	4.140	4.336	**2.495**	6.117	2.945	2.654
$Chi(3)$	3.718	3.718	3.881	**2.365**	5.548	2.863	2.856

Table(6B): Average Median Model Error over 500 replications for the case: $p = 50, n = 100, r = 0.95,$ and β values as in example 3(case3), Best method indicated in bold.

	lasso	**lasso.net**	**sparse.lasso**	**scad**	**scad.net**	**mcp**	**gel**
$N(0,1)$	**0.328**	**0.328**	0.376	2.181	0.815	4.461	0.813
$Laplace$	**0.573**	**0.573**	0.634	3.044	1.177	3.867	1.546
t_3	**0.750**	**0.750**	0.829	3.036	2.013	6.052	1.535
$G(3,1)$	**0.895**	**0.895**	0.981	3.480	2.486	5.770	1.809
$Normal.M$	**0.567**	**0.567**	0.632	2.178	1.457	2.638	1.313
$Laplace.M$	**2.279**	**2.279**	2.466	5.050	5.689	7.090	3.035
$Chi(3)$	**1.654**	**1.654**	1.809	5.390	4.636	11.032	2.603

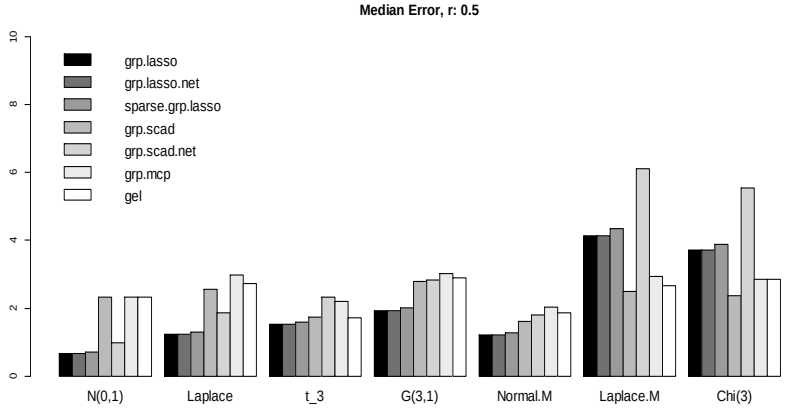


Figure (6A): Comparison of group variable selection methods under different error distributions, for low correlated predictors. The median model error over 500 replications for example 5 (case3), when $\mathrm{p} = 100$ and $\mathrm{n} = 50$.

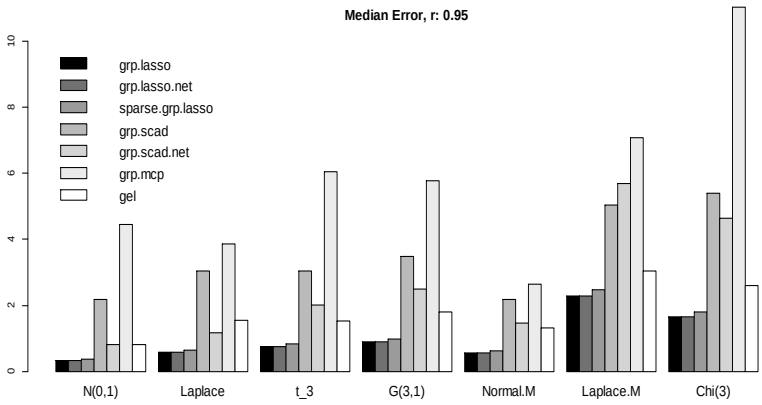


Figure (6B): Comparison of group variable selection methods under different error distributions, for highly correlated predictors. The median model error over 500 replications for example 5(case3), when $\mathrm{p} = 100$ and $\mathrm{n} = 50$.

From the results in Table 6A, Table6B, Figure6A, and Figure6B, our simulation study confirms the performances of the **grp.lasso** and the **grp.lasso.net** methods which outperform all other methods as departures from normality increase. Furthermore, the results show how **grp.mcp**the worst performing method in the case of departure from normality is especially when the predictors are highly correlated.

4. Conclusion

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

References

[1] T. B. Arnold & R. Tibshirani. Efficient Implementations of the Generalized Lasso Dual-path Algorithm. *Journal of Computational and Graphical Statistics*, 25(1), (2016), 1–27. https://doi.org/10.1080/10618600.2015.1008638

[2] S. Bakin. *Adaptive Regression and Model Selection in Data Mining Problems*, (Ph.D. thesis). The Australian National University, (1999).

[3] P. Breheny. The Group Exponential Lasso for Bi-level Variable Selection. *Biometrics*, 71 (3), (2015), 731-740. https://doi.org/10.1111/biom.12300

[4] P. Breheny & J. Huang. Group Descent Algorithms for Nonconvex Penalized Linear and Logistic Regression Models with Grouped Predictors. *Statistics and Computing*, 25(2), (2015), 173-187. https://doi.org/10.1007/s11222-013-9424-2

[5] A. A. Christidis, L. Lakshmanan, E. Smucler & R. Zamar, Split Regularized Regression. *Technometrics*, 62(3), (2020), 330–338. https://doi.org/10.1080/00401706.2019.1635533

[6] G. Ciuperca, Adaptive Group Lasso Selection in Quantile Models. *Statistical Papers*, 60 (1), (2019), 173–197. https://doi.org/10.1007/s00362-016-0832-1

[7] J. Fan & R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle. *Properties, Journal of the American Statistical Association*, 96(456), (2001), 1348–1360. https://doi.org/10.1198/016214501753382273

[8] H. Fujisawa & S. Eguchi, Robust Parameter Estimation with a Small Bias Against Heavy Contamination. *Journal of Multivariate Analysis*, 99(9), (2008), 2053-2081. https://doi.org/10.1016/j.jmva.2008.02.004

[9] W. J. Fu *Penalized Regressions: The Bridge Versus the Lasso*. Journal of Computational and Graphical Statistics, 7, 3, 397-416, (1998).

[10] H. Hashem, V. Vinciotti, R. Alhamzawi & K. Yu. Quantile regression with group Lasso for classification. *Advances in Data Analysis and Classification*, 10(3), (2016), 375–390. https://doi.org/10.1007/s11634-015-0206-x

[11] J.Huang, P. Breheny, & S. Ma, A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, 27(4), (2012), 481–499. https://doi.org/10.1214/12-sts392

[12] J.D. Huling, P. Chien, Fast Penalized Regression and Cross-Validation for Tall Data with the OEM Package. *J. Stat. Softw*, (2018).

[13] K. Kato, *Group Lasso for High Dimensional Sparse Quantile Regression Models*, arXiv:1103.1458 v2 [stat.ME],(2011).

[14] J. Y.Kim, A Lasso-type Robust Variable Selection for Time-Course Microarray Data. *Communications in Statistics-Theory and Methods*, 44 (7), (2015), 1411-1425. https://doi.org/10.1080/03610926.2013.770531

[15] R. Koenker, R. & G. W. Bassett, Regression Quantiles, *Econometrica*, 46, (1978), 33–50.

[16] S. Lambert-Lacroix, & L. Zwald, Robust Regression Through Huber's Criterion and Adaptive Lasso Penalty. *Electronic Journal of Statistics,* 5, (2011), 1015–1053. https://doi.org/10.1214/11-ejs635

[17] Q. Li, R. Xi, & N. Lin, Bayesian Regularized Quantile Regression. *Bayesian Analysis*, 5(3), (2010), 24. https://doi.org/10.1214/10-ba521

[18] J. Li, K. Liang & X. Song, Logistic Regression with Adaptive Sparse Group Lasso Penalty and its Application in Acute Leukemia Diagnosis. *Computers in Biology and Medicine*, 141, (2022), 105154. https://doi.org/10.1016/j.compbiomed.2021.105154

[19] D. J. McDonald, *sparsegl: An R Package for Estimating Sparse Group Lasso* (2022). https://arxiv.org/pdf/2208.02942.pdf.

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

[21] S. Rosset & J. Zhu, Piecewise Linear Regularized Solution Paths. *The Annals of Statistics*, 35 (3), (2007), 1012–1030. https://doi.org/10.1214/009053606000001370

[22] M.Taddy, One-Step Estimator Paths for Concave Regularization. *Journal of Computational and Graphical Statistics,* 26(3), (2017), 525- 536. https://doi.org/10.1080/10618600.2016.1211532

[23] R. Tibshirani, Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, Series B 58(1), (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[24] R. Tibshirani, & J. Taylor, The Solution Path of the Generalized Lasso. *Ann.Stat*., 39(3), (2011), 1335-1371. https://doi.org/10.1214/11-aos878

[25] H. Wang, G. Li & G. Jiang, Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3), (2007), 347 – 355. https://doi.org/10.1198/073500106000000251

[26] S. Xiong B. Dai, J. Huling & P.Qian, Orthogonalizing EM: A Design-Based Least Squares Algorithm. *Technometrics*, 58(3),(2016), 285-93. https://doi.org/10.1080/00401706.2015.1054436

[27] C. Yi & J. Huang, Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression. *Journal of Computational and Graphical Statistics,* 26(3), (2016), 547–557. https://doi.org/10.1080/10618600.2016.1256816

[28] M. Yuan & Y. Lin, Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society*. Series B (Statistical Methodology), 68(1), (2006), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[29] K. Yu, C. Cathy, C. Reed, & D. Dunson, Bayesian Variable Selection in Quantile Regression. *Statistics and Its Interface,* 6, (2013), 261–274.

[30] W. Zhu, C. L´evy-Leduc & N. Tern`es, A Variable Selection Approach for Highly Correlated Predictors in High-Dimensional Genomic Data. *Bioinformatics*, 37(16), (2021), 2238– 2244. https://doi.org/10.1093/bioinformatics/btab114

[31] H. Zou, The Adaptive Lasso, and its Oracle Properties. *Journal of the American Statistical Association*, 101 (476), (2006), 1418–1429. https://doi.org/10.1198/016214506000000735