International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 11 Issue 5 || Sep. – Oct., 2023 || PP 01-16

Hybrid Regression Estimation and Feature Selection Technique Using Robust Variable Screening Technique and Regularization

Adamu Buba[1], Umar Usman[2], Yakubu Musa[2], Murtala Muhammed Hamza[3]
[1]*Department of Mathematics/Statics, Federal University Birnin Kebbi, Kebbi State, Nigeria*
[2]*Department of Statics, Usmanu Danfodiyo University Sokoto, Sokoto State, Nigeria*
[3]*Department of Statics, Usmanu Danfodiyo University Sokoto, Sokoto State, Nigeria*

ABSTRACT*: The inability of conventional regularization techniques such as the Elastic-Net, SCAD and MCP to perform optimally in the presence of extremely large or ultra-high dimensional covariates has led to development and reliance on filtering technique like screening that have been consistently shown to outperform the usual form of regression analysis. These screening techniques (SIS, DC-SIS, and DC – RoSIS) also reduce the computational complexity in selecting important covariates from ultrahigh dimensional candidates. Several efforts have been made in this regard. In this paper, we combine some regularization techniques (ENET and SCAD) with a screening technique (DC – RoSIS) to form the hybrid methods with a view to achieving better dimension reduction and variable selection simultaneously.*

KEYWORDS*: Regularization Techniques, screening technique, ENET DC – RoSIS, ENET-M-DCRoSIS, SCAD DC – RoSIS*

Date of Submission: 06-09-2023 Date of acceptance: 18-09-2023

I. INTRODUCTION

Several approaches for Regression analysis is a form of predictive modeling technique mostly used in investigating relationship between a dependent variable and a set of predictors. It is a widely known technique for fitting models to data. For the end users, regression analysis is a reliable method of identifying which variables have impact on or greatly influence the problem of interest. To significantly explain the functional relationship between the predictor variables and the outcome variables one would need to select a parsimonious model in other to achieve a good prediction performance. When models are fitted by least squares regression each additional useful covariates adds to the actual variance of the final regression equation. In medical studies or clinical research, it is common to collect data with numerous variables, however the number of observations may be small due to cost or constraints. Datasets with more variables (features) are known as high dimensional. When the covariates dimension is high, it is natural to assume that some covariates are irrelevant. Specifically, when the number of covariates (predictors) p rivals or exceeds n (the number of observations), we often seek, for the sake of interpretation, a smaller set of variables. Hence, we want to our fitting procedure to make only a subset of the coefficients large and others small or even zero. These shortcomings are of high-dimensionality in regression setting. The traditional method (OLS) tends to over fit the model also the method becomes unusable as the coefficients estimate is no longer unique and its variance becomes infinite.

A practical approach to deal with such problem involves coefficient shrinkage (regularization) which requires fitting a model involving all p predictors. With regularization the estimated coefficients are shrunken towards zero relative to the least squares estimates. Depending on what type of shrinkage is performed, these procedures are capable of reducing the variance and can also perform variable selection. Some of these procedures e.g. the least absolute shrinkage selection Operator (LASSO) enable variable selection such that only the important predictor variables stay in the model (Szymczack, *et al*., 2009)[1]. Amongst other are, Screening and Elastic-Net, SCAD (smoothly clipped absolute deviation) (Fan and Li, 2001)[2] and the MCP (minimax concave penalty) (Zhang, 2010) [3].

With the emergence of modern technology for data collection, researchers are now able to collect data with extremely large or ultra-high dimensional covariates. Most conventional regularization techniques fail or may not perform well due to expediency and algorithmic stability (Fan, Samworth and Wu, 2009)[4]. These challenges call for a filtering technique like screening, which naturally focuses on the extremes and consistently outperform the usual form of regression analysis. These screening techniques further reduces the computational complexity in selecting important covariates from ultrahigh dimensional candidates. Such techniques are the SIS

(Sure Independence Screening) (Fan and Lv 2008)[5], DC-SIS (SIS based on Distance Correlation) (Li, Zhong and Zhu 2012)[6], DC – RoSIS (Robust SIS based on Distance Correlation) (Zhong et al, 2016)[7] .

When the covariate dimension is high in regression modelling, it is natural to assume that some covariates are irrelevant. The presence of irrelevant covariates may substantially deteriorate the precision of parameter estimation and the accuracy of response prediction (Altham, 1984)[8]. In the context of linear regression or generalized linear regression, many regularization methods and general penalty functions have been proposed to remove irrelevant covariates and simultaneously estimate the nonzero coefficients. However, when there are outliers in the response data, the above-mentioned techniques do not perform optimally. Freue et al (2019)[9] introduced penalized M-Estimation technique for high dimensional data with outliers in the response data. However, each of these methods have their shortcomings ranging from being impractical, poor performance, to algorithm instability. It is expected that incorporating screening with these methods will reduce the computational complexity in selecting important covariates from ultrahigh dimensional settings leading to improved performance and more stable computations. We perform extensive simulation and on real life data demonstration to evaluate the performance of the proposed techniques viz-a-viz existing alternatives.

II. METHODOLOGY

This section presents the methodology employed in this paper with a focus on the traditional linear regression techniques.

Linear Regression

Consider the multiple linear regression models where Y denote the response variable (also called the dependent variable) and $X_1, X_2 \ldots, X_p$, denote the explanatory variables (also called predictors, features or independent variables). The relationship between Y and $X_1, X_2 \ldots, X_p$ can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

The parameters $\beta_0, \beta_1 \ldots \beta_p$ are called regression coefficients and ε is the random error term

Given a data set $\{y_i, x_{i1}, x_{i2}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, each statistical unit can be expressed as

$$Y_i = \beta_0 + \beta_i X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \qquad 1,2, \ldots, n$$

Where y_i is the i^{th} response observation, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the unknown parameters and $\varepsilon_i \sim N(0, \delta_i^2)$. Often those n equations can be rewritten in vector form as

$$Y = X\beta + \varepsilon$$

- X *is called design matrix*
- Y *is called response* vector
- β *is the parameter vector*
- ε is the error vector

Assumptions of Multiple Linear Regression

1. **Linearity**: The relationship between the explanatory variables and the response variable is linear. This is the only restriction on the parameters (not explanatory variables), since the explanatory variables are regarded as fixed values.
2. **Independence**: There are two types of independence.
   - Each combination of explanatory variable and error is independent.
   - The error terms are independent. Therefore, $Cor(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
3. **Normality**: The error terms follow normal distribution.

$$\varepsilon_i \sim N(0, \delta_i^2),$$

where

$$\delta^2 = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix}$$

4. **Equal Variance**: Error terms are assume to have equal variances.

$$Var(\varepsilon_i) = \sigma^2 \; for \; all \; i$$
$$Var(Y_i) = \sigma^2 \; for \; all \; i$$

The ordinary Least Squares (OLS) is the traditional technique used to estimate the parameters of the multiple linear regression model. The OLS estimator, which minimizes the residual sum of squares,

$$RSS = \quad (Y - X\beta)^T (Y - X\beta)$$

is given as

$$\hat{\beta}_{0LS} = (X^T X)^{-1} X^T Y.$$

Penalization methods

We consider a linear regression model given with n observations on a dependent variable Y having p predictors. Penalized regression approaches have been used in cases where $p < n$, and in the case with $p \geq n$. In general, the Penalized Least Squares (PLS) is aimed at minimizing Residual Sum of Squares

$$(Y - X\beta)^T(Y - X\beta)$$

subject to $Pen(\beta) \leq t$, where $Pen(\beta)$ (specific penalty) is a function of $\beta = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T$ and t is a tuning parameter. This constrained optimization problem can be solved with the equivalent Lagrangian formulation which minimizes.

$$PLS = OLS + Penalty = (Y - X\beta)^T(Y - X\beta) + \lambda Pen(\beta),$$

where λ is a tuning parameter and controls the strength of shrinkage. For example, $\lambda = 0$, no penalty is applied and we have the ordinary least squares regression. When λ gets larger, more weight is given to the penalty term. Desirable properties of penalization include variable selection and grouping effect.

Elastic Net Penalty

The *Elastic Net* penalty which is based on a combined penalties of LASSO and Ridge regression penalties. For any fixed non-negative $\lambda_1 \; and \; \lambda_2$, we define the Elastic Net penalty as $Pen(\beta) = \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2$ and the Elastic Net estimator $\hat{\beta}_{EN}$ is the minimizer of

$$L(\lambda_1, \lambda_2, \beta) = (Y - X\beta)^T(Y - X\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2$$

where $\lambda_1 \; and \; \lambda_2$ are non-negative regularization parameters.

As in the case of LASSO regression procedure, the amount of shrinkage increases as λ_1 or λ_2 increases. This implies that when either $\lambda_1 \to \infty$ or $\lambda_2 \to \infty$, we have $\hat{\beta}_{EN} \to 0$. There is no explicit formula for the mean squared error for the Elastic Net estimator except when $\lambda_2 = 0$.

The Minimax Concave Penalty (MCP)

The MCP (Zhang et al, 2010) is defined as

$$Pen_{MCP}(\beta) = \sum_{i=1}^{p} \mathrm{p}_\lambda(\beta_i)$$

where,

$\mathrm{p}_\lambda(\beta_i) = \lambda\left(|\beta_i| - \frac{\beta_i^2}{2a\lambda}\right) I(0 \leq |\beta_i| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta_i| \geq a\lambda)$,

where a $>$ 1. Hence, the MCP estimator $\hat{\beta}_{MCP}$ is given as the minimizer of

$$L(\lambda_1, \lambda_2, \beta) = (Y - X\beta)^T(Y - X\beta) + Pen_{MCP}(\beta)$$

The Smoothly Clipped Absolute Deviation (SCAD)

The SCAD penalty (Fan and Li, 2001) is

$$Pen_{SCAD}(\beta) = \sum_{i=1}^{p} \mathrm{p}_\lambda(\beta_i)$$

where

$$\mathrm{p}_\lambda(\beta_i) = \lambda|\beta_i| I(O \leq \lambda) + \frac{a\lambda|\beta_i| - (\beta_i^2 + \lambda^2)/2}{a - 1} I(\lambda \leq |\beta_i| \leq a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta_i| > a\lambda), for \; some \; a > 2, \lambda > 0$$

where $I(\cdot)$ is the indicator function and a $=$ 3.7 is suggested by Fan and Li (2001).

The SCAD estimator $\hat{\beta}_{SCAD}$ is given as the minimizer of

$$L(\lambda_1, \lambda_2, \beta) = (Y - X\beta)^T(Y - X\beta) + Pen_{SCAD}(\beta).$$

Penalized M-Estimation

It is common to for the response variable in a regression problem to contain outliers. The OLS procedure and penalized methods discussed earlier do not perform adequately when there are outliers in the response data. One robust approach that handles the problem of outliers is M-Estimation. The letter M indicates that M estimation is an estimation of the maximum likelihood type. M estimation principle is to minimize the residual function

$$\hat{\beta}_M = \min_{\beta} \rho\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}{\sigma}\right),$$

where ρ is some function with the following properties:

i. $\rho(r) \geq 0$ for all r and has a minimum at 0

ii. $\rho(r) = \rho(-r)$ for all r
iii. $\rho(r)$ increases as r increases from 0, but doesn't get too large as r increases

If the ρ function can be differentiated, the M-estimator is said to be a ψ-type. Otherwise, the M-estimator is said to be a ρ-type. Note that the OLS estimator is a special case of the M-estimator.

Common ρ functions are the Tukey's bisquare, Andrew's and Huber's functions. Tukey's ρ function is given as

$$\rho(r_i) = \begin{cases} \frac{r_i^2}{2} - \frac{r_i^4}{2c^2} + \frac{r_i^6}{6c^4}, & if\ |r_i| \le c \\ \frac{c^2}{6}, & if\ |r_i| > c \end{cases},$$

where c is a constant.

Huber's ρ function is given as

$$\rho(r_i) = \begin{cases} \frac{1}{2} r_i^2, & if\ |r_i| < c \\ c|r_i| - \frac{1}{2}c^2, & if\ |r_i| \ge c \end{cases}.$$

Andrew's ρ function is given as

$$\rho(r_i) = \begin{cases} 1 - \cos(r_i), & if\ |r_i| \le \pi \\ 0, & if\ |r_i| > \pi \end{cases}.$$

The M-estimation algorithm using the Tukey's bisquare function is given as follows:

1. Estimate regression coefficients β^0 on the data using OLS.
2. Calculate residual value $e_i = y_i - \hat{y}_i$.
3. Calculate value $\hat{\sigma}_i = 1.4826\, MAD(e_1, \dots, e_n)$, where $MAD(e_1, \dots, e_n) = Median|e_i - Median(e_1, \dots, e_n)|$.
4. Calculate value $r_i = \frac{e_i}{\hat{\sigma}_i}$.
5. Calculate the weighted value
6. $w_i = \begin{cases} \left[1 - \left(\frac{r_i}{4.685}\right)^2\right]^2, & if\ |r_i| \le 4.685 \\ 0, & if\ |r_i| > 4.685 \end{cases}$
7. Calculate $\hat{\beta}_M$ using weighted least squares (WLS) method with weights w_i.
8. Repeat steps 2-6 to obtain a convergent value of $\hat{\beta}_M$. Note that at step 2, e_i is recalculated based on the fitted model in the current iteration.

While the M-estimation technique may be robust against outliers, it doesn't cater for other problems associated with regression such as high- dimensionality and multicollinearity (Freue et al, 2019). In order to solve the problem of high-dimensionality or multicollinearity a penalized M-Estimation procedure may be used.

A penalized M-Estimator is defined as the minimizer of

$$\rho\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}{\sigma}\right) + \lambda Pen(\beta),$$

Freue et al (2019) introduced efficient algorithms for penalized M-Estimators using the LASSO and Elastic-Net penalties. The pense R package contains implementation of M-Estimation using the LASSO and Elastic-Net penalties.

Robust Variable Screening based on Distance Correlation (DC-RoSIS)

In this study, a robust feature screening procedure for regression models using distance correlation proposed by Zhong et al (2016) will be adopted. The definition of distance correlation according to Szekely et al (2007) is given as follows: the distance covariance between random variables X and Y is

$$dcov^2(X, Y) = S_1 + S_2 - 2S_3,$$

where $S_1 = E\big(|X - \tilde{X}||Y - \tilde{Y}|\big), S_2 = E\big(|X - \tilde{X}||Y - \tilde{Y}|\big), S_3 = E\big(|X - \tilde{X}||Y - \tilde{Y}|\big)$, and $(\tilde{X}, \tilde{Y})$ is an independent copy of (X, Y). The distance correlation between X and Y is

$$dcorr(X, Y) = \frac{dcov(X, Y)}{\sqrt{dcov(X, Y)\, dcov(X, Y)}}$$

Szekely et al (2007) pointed out that $dcorr(X, Y) = 0$ if and only if X and Y are independent and $dcorr(X, Y)$ is strictly increasing in the absolute value of the Pearson correlation between X and Y. Motivated by these properties, Li et al (2012) proposed a sure independence screening to rank all predictors using their distance correlations with the response variable, termed DC-SIS, and proved its sure screening property for ultrahigh-dimensional data.

Following Zhong et al (2016), let X_k denote the k^{th} predictor with $k = 1,\ldots,p_n$, this work proposes to quantify the importance of X_k is through its distance correlation with the marginal distribution function of Y, denoted by $F(Y)$. That is,

$$\omega_k = dcorr\{X_K, F(Y)\},$$

where $F(y) = E\{\mathbf{1}(Y \leq y)\}$ *and* $\mathbf{1}(\cdot)$ denotes an indicator function. This is a modification of the marginal utility in Li et al (2012) in that here $F(Y)$ is used instead of Y.

The distance correlation has several advantages compared with existing measurements: $dcorr\{X_k, F(Y)\} = 0$ if and only if X_k and Y are independent, and following Li et al (2012), we can see that the screening procedure is model-free and hence is applicable for both dense and sparse situations ; since $F(Y)$ is a bounded function for all types of Y, it can be expected that the procedure has a reliable performance when the response is the heavy-tailed and when extreme values are present in the response values; If one suspects that the covariates also contain some extreme values, then one can use $\omega_k^b = dcorr\{F_k(X_k), F(Y)\}$ to rank the importance of the X_k, where $F_k(x) = E\{\mathbf{1}(X_k \leq x)\}$.

Zhong et al (2016) showed how to implement the marginal utility in the screening procedure as follows. Let $\{(\boldsymbol{X_i}, Y_i), i = 1,\cdots,n\}$ be a random sample from the population $(\boldsymbol{X}, Y)$. The distance covariance between X_k and $F(Y)$ is first estimated through the moment estimation method,

$$\widehat{dcov}^2\{X_k, F(Y)\} = \hat{S}_{k,1} + \hat{S}_{k,2} - 2\hat{S}_{k,3},$$

where

$$\hat{S}_{k,1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_{ik} - X_{jk}||F_n(Y_i) - F_n(Y_j)|,$$

$$\hat{S}_{k,2} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_{ik} - X_{jk}|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|F_n(Y_i) - F_n(Y_j)|,$$

and

$$\hat{S}_{k,3} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|X_{ik} - X_{lk}||F_n(Y_i) - F_n(Y_j)|$$

are the corresponding estimators of $S_{k,1}, S_{k,2}, S_{k,3}, and\ F_n(y) = n^{-1}\sum_{i=1}^{n}1(Y_i \leq y)$. We estimate ω_k with

$$\widehat{\omega}_k = \widehat{dcorr}\{X_k, F(Y)\} = \frac{\widehat{dcov}(X_k, F(Y))}{\sqrt{\widehat{dcov}(X_k, X_k)\,\widehat{dcov}\left(F(Y), F(Y)\right)}}$$

larger than a user-specified threshold. Let $\widehat{A} = \leq \{k : \widehat{\omega}_k \geq cn^{-\kappa}, for\ 1 \leq k \leq p_n\}$. The independence screening procedure retains the covariates with the ω_k values for some pre-specified thresholds $c > 0$ and $0 \kappa < 1/2$. The constants c and κ control the signal strength (see Zhong et al, 2016). Zhong et al (2016) referred to this approach as the distance correlation based robust independence screening procedure (DC-RoSIS).

Additionally, in this study, an estimate of $\widehat{\omega}_k^b$ which is based on the marginal distribution function of both X and Y is introduced and is defined as

$$\widehat{\omega}_k^b = \widehat{dcorr}\{F(X_K), F(Y)\} = \frac{\widehat{dcov}\left(F(X_K), F(Y)\right)}{\sqrt{\widehat{dcov}\left(F(X_K), F(X_K)\right)\,\widehat{dcov}\left(F(Y), F(Y)\right)}},$$

where,

$$\widehat{dcov}^2(F(X_K), F(Y)) = \hat{S}_{k,1}^b + \hat{S}_{k,2}^b - 2\hat{S}_{k,3}^b,$$

$$\hat{S}_{k,1}^b = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|F_n(X_{ik}) - F_n(X_{jk})||F_n(Y_i) - F_n(Y_j)|,$$

$$\hat{S}_{k,2}^b = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|F_n(X_{ik}) - F_n(X_{jk})|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|F_n(Y_i) - F_n(Y_j)|,$$

and

$$\hat{S}_{k,3}^b = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|F_n(X_{ik}) - F_n(X_{jk})||F_n(Y_i) - F_n(Y_j)|$$

The use of $\widehat{\omega}_k^b$ may produce better results if the covariates also contain some extreme values.

Sure screening property of DC-RoSIS

We first state the consistency of $\widehat{\omega}_k$ screening property of the DC-RoSIS procedure, which paves the road to proving the sure screening property of the DC-RoSIS procedure.

Theorem 1. Under the condition (C1) that there exist positive constants t_o and C such that $\max_{1\leq k\leq P_n} E\{\exp(t|X_k|)\} \leq C < \infty$, $for\ 0 < t \leq t_0$, for any $0 < \gamma < 1/2 - \kappa$, there exist positive constants c_1 and c_2 such that

$$P_r\left(\max_{1\leq k\leq p} |\widehat{\omega_k} - \omega_k| \geq cn^{-k}\right) \leq O\left(p\left[exp\{-c_1 n^{1-2(k+\gamma)}\} + n\exp(-c_2 n^{\gamma})\right]\right),$$

We remark here that to derive the consistency of the estimated marginal utility, we do not need any moment condition on the response. To prove the sure screening property, we make use of further assumption (C6) - the marginal utility satisfies $\min_{k\in A} \omega_k \geq 2cn^{-\kappa}$, for some constants $c > 0\ and\ 0 \leq \kappa < 1/2$.

Condition (C6) allows the minimal signal of the active covariates to converge to zero as the sample size diverges, while it requires the minimum signal of active covariates be not too small.

Theorem 2 (Sure Screening Property)**.** Under (C6) and the conditions in Theorem 1, it follows that $P_r(A \subseteq \hat{A}) \geq 1 - O(s_n[exp\{-c1^{n^{1-2(k+\gamma)}}\} + nexp(-c_2 n^{\gamma})])$, where s_n is the cardinality of A. Thus, $P_r(A \subseteq \hat{A}) \to 1$ as $n \to \infty$.

III. THE PROPOSED DC-ROSIS PENALIZED REGRESSION

The following gives detailed explanation of the proposed methods.

The ENET-DCRoSIS Penalized Regression

Considering that the earlier definitions of X, X_A and β_A remain unchanged. Then, the ENET-DCRoSIS estimator $\hat{\beta}_{\text{ENET-DCRoSIS}}$ is given as

$$\hat{\beta}_{\text{ENET-DCRoSIS}} = \text{argmin}_{\beta_A \epsilon R^p}(Y - X_A\beta_A)^T(Y - X_A\beta_A) + \lambda_1\sum_{i=1}^{p}|\beta_{A_i}| + \lambda_2\sum_{i=1}^{p}\beta_{A_i}^2, \quad (21)$$

which is a minimization problem that can be solved by an efficient optimization algorithm.

The SCAD-DCRoSIS Penalized Regression

Given that the earlier definitions of X, X_A and β_A remain unchanged. Then, the SCAD-DCRoSIS estimator $\hat{\beta}_{\text{SCAD-DCRoSIS}}$ is given as

$$\hat{\beta}_{\text{SCAD-DCRoSIS}} = \text{argmin}_{\beta_A \epsilon R^p}(Y - X_A\beta_A)^T(Y - X_A\beta_A) + \sum_{i=1}^{p}\text{p}_\lambda(\beta_{A_i}),$$

Where,

$$\text{p}_\lambda(\beta_{A_i}) = \lambda|\beta_{A_i}|I(0 \leq \lambda) + \frac{a\lambda|\beta_{A_i}| - \frac{\beta_{A_i}^2+\lambda^2}{2}}{a-1}I(\lambda \leq |\beta_{A_i}| \leq a\lambda) + \frac{(a+1)\lambda^2}{2}I(|\beta_{A_i}| > a\lambda),$$

for some $a > 2, \lambda > 0$ and $I(\cdot)$ is the indicator function. The minimization problem in (22) can be solved using coordinate descent algorithms.

The ENET-M-DCRoSIS Penalized Regression

Given that X, X_A and β_A are as earlier defined. Then, the ENET-M-DCRoSIS estimator $\hat{\beta}_{\text{ENET-M-DCRoSIS}}$ is given as

$$\hat{\beta}_{\text{LASSO-M-DCRoSIS}} = \text{argmin}_{\beta_A \epsilon R^p}\rho\left(\frac{Y - X_A\beta_A}{\sigma}\right) + \lambda_1\sum_{i=1}^{p}|\beta_{A_i}| + \lambda_2\sum_{i=1}^{p}\beta_{A_i}^2,$$

where $\rho(\cdot)$ is the Tukey's bisquare function.

The weighted ENET least squares technique proposed by Freue et al (2019) can be used to find the solution to the minimization problem.

IV. ANALYSIS AND RESULTS

This section presents details description of the proposed ENET-DCRoSIS, ENET-M-DCRoSIS and SCAD-DCRoSIS methods. The section also shows the results of the evaluation of the proposed hybrid methods against themselves and other classical methods under different sample size settings and outlier severity. It is worthy to note that all implementations of the methods, simulations and computations were carried out using R(R Core Team, 2019) while tables and plots are used to present the results.

Simulation Design

The performances of the ENET-DCRoSIS, ENET-M-DCRoSIS and SCAD-DCRoSIS for variable selection and estimation are evaluated via simulation at various sample sizes and level of contamination by outliers. Each simulated data consists of a training set for fitting the model, a validation set for selecting the tuning parameters, and a test set on which the test errors are computed for evaluation of performance. The notation $\cdot/\cdot/\cdot$ is used to represent the number of observations in the training, validation and test set, respectively.

Case 1

The true underlying regression model from which we simulate data is given by

$$Y = X^T\beta^* + \sigma^*\epsilon, \qquad \epsilon \sim N(0,1).$$

In this case, the simulated data sets consist of $n/10n/100$ observations and 200 predictors and we set $\beta = (\underbrace{5,\dots,5}_{20}, \underbrace{0,\dots,0}_{180})$, $n = 100$, $\sigma = 12$ and $\rho(i,j) = 0.5^{|i-j|}$ for all i, j.

Case 2

In this case, a linear model only is considered and is

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_7 X_{i7} + \varepsilon_i, i = 1,2,\dots,n.$$

$X = \left(X_1, X_2, \dots, X_p\right)^T$ was generated from $\mathcal{N}(0,\Sigma)$, where $\Sigma = \left(\sigma_{ij}\right)_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. Here, p was set to 1000 and $n = 50{,}100$ and 200. It should be noted that out of the 1000 generated covariates, only three (X_1, X_2 and X_7) are useful in the model. Hence, β was set such that $\beta = (3{,}1.5, 0{,}0{,}0{,}0{,}2{,}0 \dots, 0)^T$.

Case 3: In this case, the simulated data sets consist of $n/10n/200$ observations and 1000 predictors and we set $\beta = (\underbrace{0,\dots,0}_{485}, \underbrace{2,\dots,2}_{15}, \underbrace{0,\dots,0}_{485}, \underbrace{2,\dots,2}_{15})$, $n \in \{50, 100\}$, $\sigma = 2$ and $\rho(i,j) = 0.5^{|i-j|}$ for all i, j. In this case there are 1000 sparse grouped predictors with only 30 being relevant.

Case 4: In this case, the simulated data sets consisting of $n/10n/200$ observations and 1000 predictors and we set $\beta = (\underbrace{3,\dots,3}_{15}, \underbrace{0,\dots,0}_{985})$, $n \in \{50, 100\}$ and $\sigma = 15$. The predictors X are generated as follows:

$$X_i = Z_1 + w_i^x, \qquad Z_1 \sim N(0,1), \qquad i = 1,\dots,5,$$

$$X_i = Z_2 + w_i^x, \qquad Z_2 \sim N(0,1), \qquad i = 6,\dots,10,$$

$$X_i = Z_3 + w_i^x, \qquad Z_3 \sim N(0,1), \qquad i = 11,\dots,15.$$

X_i are independent identically distributed (iid) $N(0,1)$, for $i = 16,\dots,1000$ and w_i^x are iid $N(0,0.01)$. This setting implies there are three equally important groups with each containing 5 members. Under each case, the situation where the observations on the response variable Y contain outliers are also considered. In order to contaminate Y with outliers, the error ε_i, 90% of the errors were independently generated from $\mathcal{N}(0,1)$ and while the remaining 10% were generated from $N(20,2)$.

The proposed ENET-DCRoSIS, ENET-M-DCRoSIS and SCAD-DCRoSIS were applied to estimate β. To facilitate comparison, the classical LASSO, ENET and SCAD were applied too. The data simulation, variable screening and estimation were replicated 100 times and the performance of the technique is evaluated based on the following:

- S: the average number of non-zero estimated regression coefficients
- SE: the absolute difference between S and the actual size of the model defined here by $|S - TS|$, where TS is the true model size.
- C: the average number of truly non-zero coefficients correctly estimated to be non-zero
- IC: the average number of truly zero coefficients incorrectly estimated to be non-zero
- MSE_Y: prediction mean-squared errors defined as $\frac{1}{n_{test}}\left\|Y_{test} - X_{test}^T\hat{\beta}\right\|^2$
- MSE_β: mean-squared errors of estimates defined as $\left\|\hat{\beta} - \beta\right\|^2$
- AE: the total average absolute estimation error of $\hat{\beta}$, defined by $\sum_{j=1}^{p}\left|E(\hat{\beta}_j) - \beta_j\right|$.

Case 1

The simulation results are presented in this section. The results are based on 100 replications and the evaluation criteria are $S, SE, C, IC, MSE_Y, MSE_\beta$ and AE.

Table 4.1: Simulation results for case 1 at $n = 50, 100, 150, 200,$ with no outliers, based on 100 replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
$n = 50$	**ENET-DCRoSIS**	29	9	16	13	189.554	27.722	184.731
	SCAD-DCRoSIS	19	1	13	6	348.516	26.994	270.275
	ENET-M-DCRoSIS	26	6	16	10	192.257	26.131	179.420
	ENET	41	21	20	21	23.508	12.566	25.243
	SCAD	17	3	10	7	547.868	29.072	463.314
$n = 100$	**ENET-DCRoSIS**	34	14	20	14	3.050	4.859	6.867
	SCAD-DCRoSIS	20	0	20	0	2.050	3.271	5.810
	ENET-M-DCRoSIS	32	12	20	12	3.218	4.876	6.683
	ENET	41	21	20	21	3.176	3.319	6.943
	SCAD	20	0	20	0	1.638	1.790	5.288
$n = 150$	**ENET-DCRoSIS**	32	12	20	12	1.638	1.902	5.171
	SCAD-DCRoSIS	20	0	20	0	0.998	0.770	4.802
	ENET-M-DCRoSIS	34	14	20	14	1.578	2.369	5.176
	ENET	35	15	20	15	1.703	1.584	5.467
	SCAD	20	0	20	0	0.957	0.499	4.753
$n = 200$	**ENET-DCRoSIS**	30	10	20	10	1.079	1.171	4.801

	SCAD-DCRoSIS	20	0	20	0	0.751	0.514	4.445
	ENET-M-DCRoSIS	35	15	20	15	1.063	2.025	4.985
	ENET	32	12	20	12	1.073	1.097	4.766
	SCAD	20	0	20	0	0.741	0.439	4.447

Simulation results when there are no outliers in the response variable for case 1 are given in Table 4.1. The table contains medians of S, SE, C, IC, MSE_Y, AE and MSE_β over 100 replications at sample sizes 50, 100, 150 and 200. The true size of the model for this case is 20. At sample size of 100. In terms of variable selection SCAD and SCAD-DCRoSIS all select correctly the important variables and correctly leave out the unimportant ones. However, SCAD-DCRoSIS outperforms SCAD in terms of estimation and prediction at sample size 50. Also, ENET tend to select larger models compared to the proposed ENET-DCRoSIS and ENET-M-DCRoSIS. Similar behaviour can be observed at sample sizes 150 and 200.

Table 4.2: Simulation results for case 1 at $n = 50, 100, 150, 200,$ with 10% outliers in Y, based on 100 replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
$n = 50$	**ENET-DCRoSIS**	28	8	15	13	225.676	30.699	271.707
	SCAD-DCRoSIS	21	1	12	9	471.181	32.602	433.153
	ENET-M-DCRoSIS	24	4	16	8	154.718	14.976	149.378
	ENET	47	27	19	28	143.432	28.087	200.692
	SCAD	29	9	14	15	573.576	27.475	469.810
$n = 100$	**ENET-DCRoSIS**	34	14	20	14	31.044	8.751	73.527
	SCAD-DCRoSIS	27	7	20	9	70.970	8.346	84.054
	ENET-M-DCRoSIS	30	10	20	11	1.246	3.635	47.332
	ENET	40	20	20	20	27.107	7.626	70.118
	SCAD	41	21	20	21	93.324	10.848	101.754
$n = 150$	**ENET-DCRoSIS**	32	12	20	12	10.062	3.908	52.155
	SCAD-DCRoSIS	23	3	20	3	14.289	4.305	50.453
	ENET-M-DCRoSIS	32	12	20	12	0.526	1.526	44.541
	ENET	34	14	20	14	9.714	3..320	51.821
	SCAD	47	27	20	27	38.879	8.527	63.101
$n = 200$	**ENET-DCRoSIS**	31	11	20	11	5.602	2.731	47.991
	SCAD-DCRoSIS	20	0	20	0	7.503	2.627	46.542
	ENET-M-DCRoSIS	29	9	20	9	0.363	1.323	44.757

	ENET	30	10	20	10	5.297	2.478	48.238
	SCAD	31	11	20	11	10.934	5.009	49.401

Simulation results for case 1 with outliers introduced into the response are given in Table 4.2. At sample size of 100, ENET-M-DCRoSIS has the best performance in terms of prediction and estimation. SCAD-DCRoSIS outperforms SCAD in terms of estimation and prediction while SCAD seems to be strongly affected by presence of outliers. At sample sizes 150 and 200, ENET-M-DCRoSIS significantly outperforms others showing that it is superior when outliers are present.

Case 2

The simulation results are presented in this section. The results are based on 100 replications and the evaluation criteria are $S, SE, C, IC, MSE_Y, MSE_\beta$ and AE.

Simulation results when there are no outliers in the response variable for case 2 are given in Table 4.3. The true size of this model is 3. At sample size 50, ENET-M-DCRoSIS outperforms the rest in terms of prediction accuracy while the SCAD-DCRoSIS outperforms the others in terms of estimation accuracy variable selection. At sample sizes 100, 150 and 200, SCAD-DCRoSIS has the best performance in terms of variable selection, estimation and prediction. In this setting, all methods correctly selects the important variables into the model, however, larger models are selected by ENET and SCAD.

Table 4.4 present simulation results for case 2 with 10% outliers introduced into the response variable for case 2. Across all sample sizes ENET-M-DCRoSIS outperform the rest in terms of variable selection, prediction and estimation accuracy while SCAD produced the worst performance indicating that they don't do well in the presence of outliers. In this setting also, SCAD always select larger models while all the proposed methods always select more parsimonious models compared to existing methods.

Table 4.3: Simulation results for case 2 at $n = 50, 100, 150, 200,$ with no outliers, based on 100 replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
$n = 50$	**ENET-DCRoSIS**	10	7	3	8	2.104	3.480	6.483
	SCAD-DCRoSIS	9	6	3	6	1.799	2.304	5.485
	ENET-M-DCRoSIS	7	4	3	4	1.353	2.646	5.573
	ENET	17	14	3	14	1.826	3.503	5.925
	SCAD	17	14	3	14	2.481	2.603	5.737
$n = 100$	**ENET-DCRoSIS**	11	8	3	8	0.741	2.053	4.805
	SCAD-DCRoSIS	8	5	3	5	0.301	0.909	4.209
	ENET-M-DCRoSIS	9	6	3	6	0.525	1.500	4.555
	ENET	21	18	3	18	0.938	2.355	4.928
	SCAD	19	16	3	16	0.466	1.297	4.408
$n = 150$	**ENET-DCRoSIS**	13	10	3	10	0.519	1.564	4.437
	SCAD-DCRoSIS	6	3	3	3	0.109	0.420	4.108

	ENET-M-DCRoSIS	9	6	3	6	0.342	1.280	4.214
	ENET	17	14	3	14	0.500	1.633	4.400
	SCAD	12	9	3	9	0.181	0.846	4.305
$\boldsymbol{n = 200}$	**ENET-DCRoSIS**	12	9	3	9	0.336	1.340	4.230
	SCAD-DCRoSIS	20	0	20	0	7.503	2.627	46.542
	ENET-M-DCRoSIS	9	6	3	6	0.232	1.096	4.271
	ENET	17	14	3	14	0.346	1.405	4.436
	SCAD	9	6	3	6	0.110	0.480	4.086

Table 4.4: Simulation results for case 2 at $n = 50, 100, 150, 200,$ with 10% outliers in Y, based on 100 replications

		$\boldsymbol{S}$	$\boldsymbol{SE}$	$\boldsymbol{C}$	$\boldsymbol{IC}$	$\boldsymbol{MSE_\beta}$	**AE**	$\boldsymbol{MSE_Y}$
$\boldsymbol{n = 50}$	**ENET-DCRoSIS**	9	6	2	7	10.196	7.336	55.758
	SCAD-DCRoSIS	14	11	1	13	39.305	15.409	76.091
	ENET-M-DCRoSIS	6	3	3	3	0.338	1.654	45.207
	ENET	9	6	1	8	12.762	7.703	59.101
	SCAD	26	23	2	24	57.192	15.742	92.875
$\boldsymbol{n = 100}$	**ENET-DCRoSIS**	13	10	3	10	6.562	5.785	50.965
	SCAD-DCRoSIS	29	26	2	27	33.056	15.089	71.621
	ENET-M-DCRoSIS	8	5	3	5	0.101	0.795	44.190
	ENET	14	11	2	12	7.075	6.079	51.502
	SCAD	47	44	2	45	54.290	17.512	91.307
$\boldsymbol{n = 150}$	**ENET-DCRoSIS**	14	11	3	11	4.006	4.713	47.573
	SCAD-DCRoSIS	30	27	3	27	17.286	11.647	58.306
	ENET-M-DCRoSIS	9	6	3	6	0.065	0.642	43.487
	ENET	17	14	3	14	4.912	5.090	49.155
	SCAD	64	61	2	61	46.938	17.123	80.462
$\boldsymbol{n = 200}$	**ENET-DCRoSIS**	13	10	3	10	1.904	3.143	45.751
	SCAD-DCRoSIS	33	30	3	30	6.480	6.796	47.835

	ENET-M-DCRoSIS	9	6	3	6	0.049	0.478	44.033
	ENET	18	15	3	15	2.047	3.250	46.199
	SCAD	79	76	3	76	11.007	9.217	53.441

Case 3

Simulation results when there are no outliers in the response variable for case 3 are given in Table 4.5. The true size of this model is 30 but the values of the coefficients are relatively small and the importance of the corresponding predictors may be harder to detect. At sample size 50,100, and 150, the ENET outperforms the rest in terms of prediction, estimation accuracy and selection of important variables. However, the ENET always select larger models. At sample size 200, SCAD and SCAD-DCRoSIS have the best performance in terms of variable selection, estimation and prediction. In this setting, all the methods except ENET correctly selects the important variables into the model at small sample sizes. This is an indication that the ENET based methods are quite conservative in terms of variable selection.

Table 4.5: Simulation results for case 3 at $n = 50, 100, 150, 200,$ with no outliers, based on 100 replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
$n = 50$	**ENET-DCRoSIS**	28	2	10	17	117.944	50.0367	203.440
	SCAD-DCRoSIS	14	16	7	8	162.370	49.502	249.613
	ENET-M-DCRoSIS	25	5	10	14	119.573	44.590	182.222
	ENET	71	41	19	52	94.920	48.960	168.271
	SCAD	18	12	7	9	125.117	53.103	249.119
$n = 100$	**ENET-DCRoSIS**	48	18	23	25	48.661	24.048	57.091
	SCAD-DCRoSIS	29	1	17	12	101.739	23.268	91.621
	ENET-M-DCRoSIS	41	11	23	18	51.888	20.196	53.237
	ENET	82	52	30	53	18.195	16.543	22.953
	SCAD	34	4	15	19	145.221	36.416	125.094
$n = 150$	**ENET-DCRoSIS**	55	25	28	27	18.253	11.594	19.518
	SCAD-DCRoSIS	37	7	27	10	18.981	6.790	14.961
	ENET-M-DCRoSIS	47	17	28	19	16.013	8.433	15.629
	ENET	77	47	30	47	4.576	7.208	8.567
	SCAD	50	20	22	28	71.154	21.348	39.160
$n = 200$	**ENET-DCRoSIS**	55	25	30	25	4.711	5.525	7.049
	SCAD-DCRoSIS	33	3	30	3	1.875	2.657	5.695
	ENET-M-DCRoSIS	49	19	29	20	7.274	5.117	8.364
	ENET	73	43	30	43	2.684	4.927	6.330

	SCAD	32	2	30	2	1.170	0.832	4.804

Table 4.6 present simulation results for case 3 with 10% outliers introduced into the response variable for case 2. Across all sample sizes ENET-M-DCRoSIS outperform the rest in terms of prediction and estimation accuracy while SCAD the worst performance.

Table 4.6: Simulation results for case 3 at $n = 50, 100, 150, 200,$ with 10% outliers in Y, based on 100 replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
$n = 50$	**ENET-DCRoSIS**	23	7	8	14	126.691	53.096	280.850
	SCAD-DCRoSIS	18	12	6	13	269.223	65.477	390.465
	ENET-M-DCRoSIS	19	11	9	9	108.432	36.328	196.746
	ENET	109	79	20	89	102.616	80.202	222.437
	SCAD	36	6	0	36	829.179	93.528	1045.335
$n = 100$	**ENET-DCRoSIS**	48	18	22	26	67.297	30.256	125.231
	SCAD-DCRoSIS	38	8	16	22	176.564	37.161	176.564
	ENET-M-DCRoSIS	36	6	22	14	45.357	12.452	80.393
	ENET	86	56	27	60	57.251	32.977	124.411
	SCAD	62	32	2	60	706.323	100.389	875.419
$n = 150$	**ENET-DCRoSIS**	56	26	27	29	33.883	17.597	78.167
	SCAD-DCRoSIS	47	17	22	25	79.892	21.809	102.168
	ENET-M-DCRoSIS	44	14	27	17	17.282	6.795	55.770
	ENET	85	55	30	55	30.211	19.243	77.232
	SCAD	80	50	7	73	498.179	80.622	576.196
$n = 200$	**ENET-DCRoSIS**	57	27	29	28	13.239	9.682	55.718
	SCAD-DCRoSIS	50	20	29	22	25.882	9.305	59.041
	ENET-M-DCRoSIS	49	19	29	19	5.486	3.556	46.806
	ENET	74	44	30	44	12.825	9.727	55.680
	SCAD	85	55	21	63	96.376	26.282	103.093

Case 4

Table 4.7: Simulation results for case 4 at $n = 50, 100, 150, 200,$ with no outliers, based on 100 replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
$n = 50$	**ENET-DCRoSIS**	18	3	15	3	0.570	4.375	4.908
	SCAD-DCRoSIS	3	12	3	0	531.851	6.855	4.635
	ENET-M-DCRoSIS	17	2	14	4	37.796	7.733	5.027
	ENET	21	6	15	6	0.344	1.653	4.786
	SCAD	3	12	3	0	538.689	71.909	4.378
$n = 100$	**ENET-DCRoSIS**	16	1	15	1	0.073	0.286	4.281
	SCAD-DCRoSIS	3	12	3	0	538.511	7.292	4.318
	ENET-M-DCRoSIS	19	4	15	4	3.712	7.721	4.416
	ENET	17	2	15	2	0.066	0.277	4.151
	SCAD	3	12	3	0	539.963	71.878	4.371
$n = 150$	**ENET-DCRoSIS**	16	1	15	1	0.055	0.266	4.220
	SCAD-DCRoSIS	3	12	3	0	537.645	8.817	4.023
	ENET-M-DCRoSIS	19	4	15	4	2.710	6.232	4.376
	ENET	16	1	15	1	0.043	0.256	4.024
	SCAD	3	12	3	0	538.411	71.925	4.025
$n = 200$	**ENET-DCRoSIS**	15	0	15	0	0.031	0.163	3.970
	SCAD-DCRoSIS	3	12	3	0	537.110	5.815	4.046
	ENET-M-DCRoSIS	21	6	15	8	2.127	3.816	4.220
	ENET	16	1	15	1	0.031	0.300	4.083
	SCAD	15	0	7	8	302.801	5.810	4.188

Simulation results when there are no outliers in the response variable for case 4 are given in Table 4.6. The true size of this model here is 15 and the important predictors are divided into three groups such that predictors within each group are strongly correlated. Across all sample sizes ENET, ENET-DCRoSIS and ENET-M-DCRoSIS outperforms the rest in terms of variable selection and estimation. However, all the methods perform similarly with respect to prediction. Also, SCAD and SCAD-DCRoSIS tend to select one of the important variables in each group. Only, ENET, ENET- DCRoSIS and ENET-M-DCRoSIS have the ability to do group selection.

Table 4.8: Simulation results for case 4 at $n = 50, 100, 150, 200,$ with 10% outliers in Y, based on 100 replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
$n = 50$	**ENET-DCRoSIS**	18	3	15	3	3.767	6.484	51.826
	SCAD-DCRoSIS	3	12	3	0	510.214	12.619	45.612
	ENET-M-DCRoSIS	11	4	6	5	292.831	6.807	44.241
	ENET	29	14	15	14	7.010	9.927	61.795
	SCAD	3	12	3	0	537.854	73.312	43.118
$n = 100$	**ENET-DCRoSIS**	16	1	15	1	0.391	0.654	45.451
	SCAD-DCRoSIS	3	12	3	0	537.374	12.021	42.660
	ENET-M-DCRoSIS	18	3	15	3	2.996	4.215	43.917
	ENET	18	3	15	3	0.599	1.128	45.146
	SCAD	3	12	3	0	543.058	72.047	41.198
$n = 150$	**ENET-DCRoSIS**	16	1	15	1	0.203	0.386	44.564
	SCAD-DCRoSIS	3	12	3	0	530.967	10.459	41.530
	ENET-M-DCRoSIS	20	5	15	5	2.397	4.500	44.101
	ENET	16	1	15	1	0.203	0.561	44.648
	SCAD	3	12	3	0	537.219	71.810	40.594
$n = 200$	**ENET-DCRoSIS**	16	1	15	1	0.124	0.198	44.222
	SCAD-DCRoSIS	3	12	3	0	537.457	12.747	41.320
	ENET-M-DCRoSIS	20	5	15	5	1.501	3.806	43.967
	ENET	16	1	15	1	0.131	0.235	44.447
	SCAD	3	12	3	0	543.027	71.980	40.471

Table 4.8 present simulation results for case 4 with 10% outliers introduced into the response variable for case 4. Across all sample sizes ENET- DCRoSIS outperforms the rest in terms of variable selection and estimation while SCAD has the worst performance in all criteria. Just like when there were no outliers, SCAD and SCAD-DCRoSIS select one of the important variables in each group while ENET, ENET- DCRoSIS and ENET-M-DCRoSIS are able to do group variable selection.

V. CONCLUSION

The presence of outliers can prevent the ENET and existing screening techniques from performing optimally. This creates the need to generate new hybrid approaches that improve the performance of legacy screening techniques. To achieve this, the ENET and SCAD have been combined with a robust screening technique that can do well in the presence of outliers. Thus, achieving better dimension reduction and variable selection simultaneously. Our numerous simulations show that our proposed ENET-M-DCRoSIS significantly

outperforms the rest in terms of prediction and estimation accuracy showing that it is superior when outliers are present while SCAD has the worst performance.

REFERENCES

[1]. Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H. and Sun, Y. V. (2009). Machine learning in genome-wide association studies. Genet. Epidemiol., 33, S51–S57.

[2]. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and it oracle properties. J. Amer. Statist. Assoc. **96**, 1348-1360.

[3]. Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. **101**, 1418-1429.

[4]. Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. J. Machine Learn. Res. **10**, 1829-1853.

[5]. Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). J. Roy. Statist. Soc. Ser. B **70**, 849-911.

[6]. Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. J. Amer. Statist. Assoc. **107**, 1129-1139.

[7]. Zhong, W., Zhu, L., Li, R. and Cui, H. (2016). Regularized Quantile Regression and Robust Feature Screening for Single Index Models. Statistica Sinica, 26, 69-95.

[8]. Altham, P. M. (1984). Improving the precision of estimation by fitting a model. Journal of the Royal Statistical Society: Series B (Methodological), 46(1), 118-119.

[9]. Freue, G. V. C., Kepplinger, D., Salibián-Barrera, M. and Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. The Annals of Applied Statistics, 13(4), 2065-2090.