ISSN No:-2456-2165

On Some New Hybridized Regression Estimation and Feature Selection Techniques

Adamu Buba[1]
[1]Department of Mathematics/Statistics, Federal University Birnin Kebbi, Kebbi State, Nigeria

Umar Usman[2]; Yakubu Musa[3]
[2,3]Department of Statistics, Usmanu Danfodiyo University Sokoto, Sokoto State, Nigeria

Murtala Muhammed Hamza[4]
[4]Department of Mathematics, Usmanu Danfodiyo University Sokoto, Sokoto State, Nigeria

Abstract:- Conventional regularization techniques like LASSO, SCAD and MCP have been shown to perform poorly in the presence of extremely large or ultra-high dimensional covariates. This has created the need for and led to the development and reliance on filtering technique like screening. Screening techniques (such as SIS, DC-SIS, and DC – RoSIS) have been shown to reduce the computational complexity in selecting important covariates from ultrahigh dimensional candidates. To this end, there have been various attempts to hybridize the conventional regularization techniques. In this paper, we combine some regularization techniques (LASSO and SCAD) with a screening technique (DC – RoSIS) to form new hybrid methods with a view to achieving better dimension reduction and variable selection simultaneously. Extensive simulation results and real life data performance show that the proposed methods perform better than the conventional methods.

Keywords:- *Regularization Techniques, Screening Technique, LASSO DC–RoSIS, SCAD DC – RoSIS.*

I. INTRODUCTION

Regression analysis, a form of predictive modeling technique mostly used in investigating relationship between a dependent variable and a set of predictors, is a widely known technique for fitting models to data. It 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.

To deal with such problems, coefficient shrinkage (regularization) is employed to shrink the estimated coefficients 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 like the least absolute shrinkage selection Operator (LASSO), SCAD (smoothly clipped absolute deviation) (Fan and Li, 2001)[2] and the MCP (minimax concave penalty) (Zhang, 2010) [3] enable variable selection such that only the important predictor variables stay in the model (Szymczack, *et al.*, 2009)[1].

The high volume of data currently processed due to the great evolution in social media and other data intensive tasks has led to the collection of extremely large or ultra-high dimensional covariates. This makes conventional regularization techniques fail or underperform well due to expediency and algorithmic stability (Fan, Samworth and Wu, 2009)[4]. This has led to the use filtering techniques 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 …, X_p, denote the explanatory variables (also called predictors, features or independent variables). The relationship between Y and X_1, X_2 …, X_p can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

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

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

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

$$Y = X\beta + \varepsilon \quad (3)$$

- 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*

- *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.

- *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 $i \neq j$.

- *Normality:*

The error terms follow normal distribution.

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

Where

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

- *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,

$$RSS = \quad (Y - X\beta)^T (Y - X\beta) \tag{4}$$

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.

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

$$+\lambda Pen(\beta) \tag{5}$$

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.

➢ *LASSO Penalty*

The Least Absolute Shrinkage and Selection Operator (LASSO) regression method was introduced by Tibshirani (1996) as an estimation and variable selection method. It is also called L_1 penalized regression. The LASSO is a penalized least squares procedure that minimizes RSS subject to the non-differentiable constraint expressed in terms of the L_1 norm of the coefficients. The penalty function is given by

$$Pen(\beta) = \lambda \sum_{i=1}^{p} |\beta_i| \tag{6}$$

The objectives is to minimize

$$\hat{\beta}_{LASSO} = argmin_{\beta \epsilon R^p} (Y - X\beta)^T (Y - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \tag{7}$$

Where λ is a non-negative regularization parameter.

Since the LASSO penalty term is no longer quadratic, there is no explicit formula for the mean squared error of the LASSO estimator. Generally, the $Bias(\hat{\beta}_{LASSO})$ also increases as the tuning parameter λ increases, while the variance, $Var(\hat{\beta}_{LASSO})$ decreases. For instance

Where $\lambda = 0$
$MSE(\hat{\beta}_{LASSO}) = MSE(\hat{\beta}_{OLS})$.
And when $\lambda \to \infty$
$MSE(\hat{\beta}_{LASSO}) = trace\left(Var(\hat{\beta}_{LASSO})\right) + Bias^T(\hat{\beta}_{LASSO})Bias(\hat{\beta}_{LASSO}) \to 0.$

Since $Bias^T(\hat{\beta}_{LASSO})Bias(\hat{\beta}_{LASSO}) and \ \ trace\left(Var(\hat{\beta}_{LASSO})\right)$ move to opposite directions as the tuning parameter λ increases, thus, we can choose an optimal value of the parameter λ that minimizes $MSE(\hat{\beta}_{LASSO})$.

- *The Smoothly Clipped Absolute Deviation (SCAD)*

The SCAD penalty (Fan and Li, 2001) is

$$Pen_{SCAD}(\beta) = \sum_{i=1}^{p} \mathrm{p}_{\lambda}(\beta_i) \tag{8}$$

Where

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

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

$$L(\lambda_1, \lambda_2, \beta) = (Y - X\beta)^T(Y - X\beta) + Pen_{SCAD}(\beta) \tag{9}$$

- *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.

$$\hat{\beta}_M = \min_{\beta} \rho\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}{\sigma}\right), \tag{10}$$

Where ρ is some function with the following properties:

- $\rho(r) \ge 0$ for all r and has a minimum at 0
- $\rho(r) = \rho(-r)$ for all r
- $\rho(r)$ 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

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

Andrew's ρ function is given as

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

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

- Estimate regression coefficients β^0 on the data using OLS.
- Calculate residual value $e_i = y_i - \hat{y}_i$.
- Calculate value $\hat{\sigma}_i = 1.4826\ MAD(e_1, \dots, e_n)$, where $MAD(e_1, \dots, e_n) = Median|e_i - Median(e_1, \dots, e_n)|$.
- Calculate value $r_i = \frac{e_i}{\hat{\sigma}_i}$.
- Calculate the weighted value
- $w_i = \begin{cases} \left[1 - \left(\frac{r_i}{4.685}\right)^2\right]^2, & if\ |r_i| \leq 4.685 \\ 0, & if\ |r_i| > 4.685 \end{cases}$
- Calculate $\hat{\beta}_M$ using weighted least squares (WLS) method with weights w_i.
- 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

$$\rho\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}}{\sigma}\right) + \lambda Pen(\beta), \quad (11)$$

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

$$dcov^2(X, Y) = S_1 + S_2 - 2S_3, \quad (12)$$

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

$$dcorr(X, Y) = \frac{dcov(X, Y)}{\sqrt{dcov(X, Y)\ dcov(X, Y)}} \quad (13)$$

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 $k = 1, \dots, p_n$, 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 $F(Y) = E\{\mathbf{1}(Y \leq y)\}$. 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,

$$\widehat{dcov}^2\{X_k, F(Y)\} = \hat{S}_{k,1} + \hat{S}_{k,2} - 2\hat{S}_{k,3}, \tag{14}$$

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

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

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 $\omega_k$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

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

Where,

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

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

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

And

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

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 $\max_{1\le k\le P_n} E\{\exp(t|X_k|)\} \le C < \infty, \; for \; 0 < t \le t_0$, for any $0 < \gamma < 1/2 - \kappa$, there exist positive constants c_1 and c_2 such that

$$P_r\left(\max_{1\le k\le p}|\widehat{\omega_k} - \omega_k| \ge cn^{-k}\right) \le O\left(p\left[exp\left\{-c_1 n^{1-2(k+\gamma)}\right\} + n\exp(-c_2 n^{\gamma})\right]\right), \quad (15)$$

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 $\min_{k\in A}\omega_k \ge 2cn^{-\kappa}$, for some constants $c > 0 \; and \; 0 \le \kappa < 1/2$.

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 $P_r(A \subseteq \hat{A}) \ge 1 - O(s_n[exp\{-c1^{n^{1-2(k+\gamma)}}\} + nexp(-c_2 n^{\gamma})])$, 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

In this paper, we propose some new estimators by combining the DC-RoSIS with some penalized regression estimators, namely, the LASSO, SCAD and MCP. We achieve this by first utilizing the DC-RoSIS to select $d = 2\left[\frac{n}{log(n)}\right]$ (see Zhong et al, 2016) top ranked covariates and then applying penalized linear regression to estimate the direction of β. The combination gives rise to LASSO-DCRoSIS, SCAD-DCRoSIS and LASSO-M-DCRoSIS. Hence, the proposed method is a two-stage method. First, DCRoSIS is used to to reduce the covariate dimension to a moderate scale and then, based on the reduced model, penalized linear regression further estimates and refines selection of important covariates.

The need for this hybridization stems from the fact that from a practical perspective, when the covariate dimension is extremely large, it is hoped that the DCRoSIS offers a useful complement to penalized regression since it helps to reduce the computational complexity in selecting important covariates from ultrahigh dimensional candidates. More so in our previous work (Buba, Usman, Musa, and Hamza, 2023), the hybridization of Elastic Net, SCAD and MCP gave rise to some visible improvements.

From a practical perspective, when the covariate dimension is extremely large, it is hoped that the DCRoSIS offers a useful complement to penalized regression since it helps to reduce the computational complexity in selecting important covariates from ultrahigh dimensional candidates.

Since F is bounded and monotone, we can reasonably expect that the procedure still works in the presence of outliers or extreme values in the covariate or response variable. It is computationally efficient and hence offers a useful complement, rather an alternative, to the penalized regression approach since the proposed independence screening can precede the penalized regression when the latter fails to produce a reliable solution within a tolerable time. Zhong et al (2016) showed that this new independence screening procedure has the sure screening property even when p is ultrahigh.

➢ *The LASSO-DCRoSIS Penalized Regression*

Considering the model given by (3), X is the matrix with p columns representing all the predictors. The DCRoSIS technique is used to compute $\widehat{\omega}_j$ (or $\widehat{\omega}_k^b$), $j = 1, \dots, p$. Thereafter, the $\widehat{\omega}_j$'s are ranked. Let X_A denote the matrix with columns containing the top d predictors corresponding to the top d ranked $\widehat{\omega}_j$'s. Also, let $\beta_A =$

$\left(\beta_{A_0}, \beta_{A_1}, \beta_{A_2}, \dots, \beta_{A_p}\right)$ denote the regression coefficients associated with X_A. Then,

The minimization problem given by (20) can be solved by a number of algorithms including as coordinate descent (Fu, 1998), proximal methods (Beck and Teboulle, 2009) and quadratic solver (Grandvalet et al, 2017).

➢ *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

$$\hat{\beta}_{\text{SCAD-DCRoSIS}} = \text{argmin}_{\beta_A \epsilon R^p}(Y - X_A\beta_A)^T(Y - X_A\beta_A) + \sum_{i=1}^{p} \text{p}_\lambda\left(\beta_{A_i}\right), \quad (16)$$

Where,

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

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 LASSO-M-DCRoSIS Penalized Regression*

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

$$\hat{\beta}_{\text{LASSO-M-DCRoSIS}} = \text{argmin}_{\beta_A \epsilon R^p} \rho\left(\frac{Y - X_A\beta_A}{\sigma}\right) + \lambda \sum_{i=1}^{p} \left|\beta_{A_i}\right|, \quad (17)$$

Where $\rho(\cdot)$ is the Tukey's bisquare function defined in section 3.3.

The minimization problem in (24) can be solved by a weighted LASSO least squares technique proposed by Freue et al (2019).

IV. ANALYSIS AND RESULTS

This section presents details description of the proposed LASSO-DCRoSIS, LASSO-M-DCRoSIS and SCAD-DCRoSIS. 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 LASSO-DCRoSIS, LASSO-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:

$$X_i = Z_1 + w_i^x, \quad Z_1 \sim N(0,1), \quad i = 1,\dots,5,$$
$$X_i = Z_2 + w_i^x, \quad Z_2 \sim N(0,1), \quad i = 6,\dots,10,$$
$$X_i = Z_3 + w_i^x, \quad Z_3 \sim N(0,1), \quad i = 11,\dots,15.$$

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 LASSO-DCRoSIS, LASSO-M-DCRoSIS and LASSO-DCRoSIS were applied to estimate β. To facilitate comparison, the classical LASSO 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.

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 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
$n = 50$	**LASSO-DCRoSIS**	26	6	16	10	207.348	30.567	213.200
	SCAD-DCRoSIS	19	1	13	6	348.516	26.994	270.275
	LASSO-M-DCRoSIS	23	3	16	8	166.630	18.358	129.842
	LASSO	28	8	20	8	57.500	21.075	65.991
	SCAD	17	3	10	7	547.868	29.072	463.314
$n = 100$	**LASSO-DCRoSIS**	41	21	20	21	3.108	3.680	6.856
	SCAD-DCRoSIS	20	0	20	0	2.050	3.271	5.810
	LASSO-M-DCRoSIS	31	11	20	11	1.116	3.566	5.018
	LASSO	41	21	20	21	3.183	3.571	7.029
	SCAD	20	0	20	0	1.638	1.790	5.288
$n = 150$	**LASSO-DCRoSIS**	42	22	20	22	1.658	2.637	5.609
	SCAD-DCRoSIS	20	0	20	0	0.998	0.770	4.802
	LASSO-M-DCRoSIS	31	11	20	11	0.623	1.890	4.496
	LASSO	43	23	20	23	1.802	2.726	5.697
	SCAD	20	0	20	0	0.957	0.499	4.753
$n = 200$	**LASSO-DCRoSIS**	36	16	20	16	1.145	1.033	4.830
	SCAD-DCRoSIS	20	0	20	0	0.751	0.514	4.445
	LASSO-M-DCRoSIS	31	11	20	11	0.466	1.428	4.396
	LASSO	45	25	20	25	1.186	2.197	5.155
	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 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. In terms of variable selection SCAD and SCAD-DCRoSIS correctly select the important variables and correctly leave out the unimportant ones. However, SCAD-DCRoSIS outperforms the SCAD in terms of estimation and prediction at sample size 50. Also, LASSO tend to select larger models compared to the proposed LASSO-DCRoSIS and LASSO-M-DCRoSIS. Similar behaviour can be observed at sample sizes 150 and 200.

Table 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
$n = 50$	**LASSO-DCRoSIS**	29	9	17	13	228.839	35.226	270.052
	ENET-DCRoSIS	28	8	15	13	225.676	30.699	271.707
	SCAD-DCRoSIS	21	1	12	9	471.181	32.602	433.153
	MCP-DCRoSIS	19	1	12	8	476.287	32.941	438.463
	LASSO-M-DCRoSIS	24	4	16	8	156.260	14.845	149.378
	ENET-M-DCRoSIS	24	4	16	8	154.718	14.976	149.378
	LASSO	44	24	18	27	224.143	30.550	270.027
	ENET	47	27	19	28	143.432	28.087	200.692
	SCAD	29	9	14	15	573.576	27.475	469.810
	MCP	25	5	14	12	598.667	25.885	471.252
$n = 100$	**LASSO-DCRoSIS**	37	17	20	14	52.831	10.445	75.917
	SCAD-DCRoSIS	27	7	20	9	70.970	8.346	84.054
	LASSO-M-DCRoSIS	26	6	19	7	31.108	4.283	52.914
	LASSO	42	22	20	22	33.792	11.560	77.480
	SCAD	41	21	20	21	93.324	10.848	101.754
$n = 150$	**LASSO-DCRoSIS**	33	13	20	14	10.855	6.426	53.967
	SCAD-DCRoSIS	23	3	20	3	14.289	4.305	50.453
	LASSO-M-DCRoSIS	27	7	20	7	0.502	1.526	45.011
	LASSO	43	23	20	23	11.621	7.129	54.537
	SCAD	47	27	20	27	38.879	8.527	63.101
$n = 200$	**LASSO-DCRoSIS**	35	15	20	15	6.025	4.435	49.315
	SCAD-DCRoSIS	20	0	20	0	7.503	2.627	46.542
	LASSO-M-DCRoSIS	29	9	20	9	0.363	1.323	44.757
	LASSO	42	22	20	22	6.611	5.178	50.391
	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 2. SCAD-DCRoSIS outperforms SCAD in terms of estimation and prediction. SCAD seems to be strongly affected by the presence of outliers. At sample sizes 150 and 200, LASSO-M-DCRoSIS significantly outperform others showing that they are 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 3. The true size of this model is 3. At sample size 50, LASSO-M-DCRoSIS outperforms the rest in terms of prediction and estimation accuracy but SCAD-DCRoSIS has the best performance in terms of 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 LASSO and SCAD.

Table 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
$n = 50$	**LASSO-DCRoSIS**	13	10	3	10	1.797	3.296	6.049
	SCAD-DCRoSIS	9	6	3	6	1.799	2.304	5.485
	LASSO-M-DCRoSIS	7	4	3	4	0.462	1.629	4.524
	LASSO	21	18	3	18	2.333	3.691	6.452
	SCAD	17	14	3	14	2.481	2.603	5.737
$n = 100$	**LASSO-DCRoSIS**	14	11	3	11	0.867	2.045	4.816
	SCAD-DCRoSIS	8	5	3	5	0.301	0.909	4.209
	LASSO-M-DCRoSIS	9	6	3	6	0.251	1.066	4.212
	LASSO	19	16	3	16	0.871	2.167	4.862
	SCAD	19	16	3	16	0.466	1.297	4.408
$n = 150$	**LASSO-DCRoSIS**	12	9	3	9	0.439	1.493	4.375

ISSN No:-2456-2165

	SCAD-DCRoSIS	6	3	3	3	0.109	0.420	4.108
	LASSO-M-DCRoSIS	9	6	3	6	0.156	0.844	4.251
	LASSO	19	16	3	16	0.503	1.730	4.605
	SCAD	12	9	3	9	0.181	0.846	4.305
	LASSO-DCRoSIS	12	9	3	9	0.322	1.269	4.353
	SCAD-DCRoSIS	6	3	3	3	0.092	0.314	4.075
$n = 200$	**LASSO-M-DCRoSIS**	9	6	3	6	0.129	0.764	4.080
	LASSO	21	18	3	18	0.364	1.442	4.374
	SCAD	9	6	3	6	0.110	0.480	4.086

Table 4 Simulation Results for Case 2 at $n = 50, 100, 150, 200,$ with 10% Outliers in Y, based on 100 Replications

		S	SE	C	IC	MSE_β	AE	MSE_Y
	LASSO-DCRoSIS	6	3	1	5	12.092	7.055	58.209
	SCAD-DCRoSIS	14	11	1	13	39.305	15.409	76.091
$n = 50$	**LASSO-M-DCRoSIS**	6	3	3	3	0.312	1.605	45.207
	LASSO	10	7	1	9	11.663	7.303	57.176
	SCAD	26	23	2	24	57.192	15.742	92.875
	LASSO-DCRoSIS	10	7	2	8	7.132	5.809	51.648
	SCAD-DCRoSIS	29	26	2	27	33.056	15.089	71.621
$n = 100$	**LASSO-M-DCRoSIS**	8	5	3	5	0.101	0.787	44.190
	LASSO	14	11	2	12	7.512	6.147	51.868
	SCAD	47	44	2	45	54.290	17.512	91.307
	LASSO-DCRoSIS	12	9	3	9	5.067	4.861	49.362
	SCAD-DCRoSIS	30	27	3	27	17.286	11.647	58.306
$n = 150$	**LASSO-M-DCRoSIS**	9	6	3	6	0.064	0.632	43.472
	LASSO	16	13	3	13	4.951	5.122	49.192
	SCAD	64	61	2	61	46.938	17.123	80.462
	LASSO-DCRoSIS	13	10	3	10	2.048	3.208	46.450
	SCAD-DCRoSIS	33	30	3	30	6.480	6.796	47.835
$n = 200$	**LASSO-M-DCRoSIS**	9	6	3	6	0.049	0.477	44.033
	LASSO	20	17	3	17	2.325	3.537	46.323
	SCAD	79	76	3	76	11.007	9.217	53.441

Table 4 present simulation results for case 2 with 10% outliers introduced into the response variable for case 2. Across all sample sizes LASSO-M-DCRoSIS outperformed 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 selects larger models while all the proposed methods always select more parsimonious models compared to existing methods.

- *Case 3*

Table 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
	LASSO-DCRoSIS	22	8	9	12	118.178	51.043	219.490
	SCAD-DCRoSIS	14	16	7	8	162.370	49.502	249.613
$n = 50$	**LASSO-M-DCRoSIS**	19	11	11	8	110.334	34.909	151.395
	LASSO	35	5	24	11	56.869	20.174	61.112
	SCAD	18	12	7	9	125.117	53.103	249.119
	LASSO-DCRoSIS	43	13	22	21	57.359	25.915	66.021
	SCAD-DCRoSIS	29	1	17	12	101.739	23.268	91.621
$n = 100$	**LASSO-M-DCRoSIS**	36	6	23	13	44.165	13.223	38.862
	LASSO	76	46	30	46	13.545	17.941	16.593
	SCAD	34	4	15	19	145.221	36.416	125.094
	LASSO-DCRoSIS	52	22	28	24	16.553	11.280	16.756
$n = 150$	**SCAD-DCRoSIS**	37	7	27	10	18.981	6.790	14.961
	LASSO-M-DCRoSIS	44	14	28	16	13.315	6.187	12.927

	LASSO	85	55	30	55	4.580	7.534	8.864
	SCAD	50	20	22	28	71.154	21.348	39.160
$n = 200$	**LASSO-DCRoSIS**	58	28	29	29	6.668	6.567	7.901
	SCAD-DCRoSIS	33	3	30	3	1.875	2.657	5.695
	LASSO-M-DCRoSIS	50	20	29	20	5.814	3.837	7.347
	LASSO	83	53	30	53	2.520	5.449	6.707
	SCAD	32	2	30	2	1.170	0.832	4.804

Simulation results when there are no outliers in the response variable for case 3 are given in Table 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 LASSO outperforms the rest in terms of prediction, estimation accuracy and selection of important variables. However, at sample size 200, SCAD followed by SCAD-DCRoSIS have the best performance in terms of variable selection, estimation and prediction. In this setting, all the methods except LASSO correctly selects the important variables into the model at small sample sizes. This is an indication that the LASSO is quite conservative in terms of variable selection.

Table 6 present simulation results for case 3 with 10% outliers introduced into the response variable for case 2. Across all sample sizes LASSO-M-DCRoSIS outperform the rest in terms of prediction and estimation accuracy while SCAD produced the worst performance.

Table 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
$n = 50$	**LASSO-DCRoSIS**	17	13	7	10	125.476	56.116	287.257
	SCAD-DCRoSIS	18	12	6	13	269.223	65.477	390.465
	LASSO-M-DCRoSIS	19	11	9	9	109.214	35.836	196.114
	LASSO	30	0	13	17	113.895	53.630	214.896
	SCAD	36	6	0	36	829.179	93.528	1045.335
$n = 100$	**LASSO-DCRoSIS**	40	10	19	21	79.008	34.611	147.307
	SCAD-DCRoSIS	38	8	16	22	176.564	37.161	176.564
	LASSO-M-DCRoSIS	36	6	22	14	45.706	12.422	80.393
	LASSO	83	53	23	60	77.679	33.938	132.499
	SCAD	62	32	2	60	706.323	100.389	875.419
$n = 150$	**LASSO-DCRoSIS**	52	22	27	25	41.602	22.631	84.648
	SCAD-DCRoSIS	47	17	22	25	79.892	21.809	102.168
	LASSO-M-DCRoSIS	44	14	27	17	17.283	6.794	55.770
	LASSO	79	49	28	51	41.121	22.821	89.129
	SCAD	80	50	7	73	498.179	80.622	576.196
$n = 200$	**LASSO-DCRoSIS**	57	27	29	28	17.282	13.351	60.556
	SCAD-DCRoSIS	50	20	29	22	25.882	9.305	59.041
	LASSO-M-DCRoSIS	49	19	29	19	5.486	3.556	46.806
	LASSO	84	54	30	54	15.058	13.895	61.126
	SCAD	85	55	21	63	96.376	26.282	103.093

- *Case 4*

Table 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
$n = 50$	**LASSO-DCRoSIS**	10	5	5	6	404.660	6.619	5.062
	SCAD-DCRoSIS	3	12	3	0	531.851	6.855	4.635
	LASSO-M-DCRoSIS	9	6	5	3	373.798	7.040	4.395
	LASSO	28	13	4	23	393.286	23.812	7.249
	SCAD	3	12	3	0	538.689	71.909	4.378
$n = 100$	**LASSO-DCRoSIS**	13	2	5	8	369.792	7.686	4.587
	SCAD-DCRoSIS	3	12	3	0	538.511	7.292	4.318
	LASSO-M-DCRoSIS	11	4	6	5	334.217	6.225	4.292
	LASSO	24	9	5	18	379.969	6.6767	5.134
	SCAD	3	12	3	0	539.963	71.878	4.371
$n = 150$	**LASSO-DCRoSIS**	14	1	6	8	343.050	7.792	4.505

ISSN No:-2456-2165

	SCAD-DCRoSIS	3	12	3	0	537.645	8.817	4.023
	LASSO-M-DCRoSIS	14	1	7	7	282.709	7.059	4.082
	LASSO	27	12	6	21	327.214	7.855	4.770
	SCAD	3	12	3	0	538.411	71.925	4.025
$n = 200$	**LASSO-DCRoSIS**	15	0	7	8	302.801	5.810	4.188
	SCAD-DCRoSIS	3	12	3	0	537.110	5.815	4.046
	LASSO-M-DCRoSIS	16	1	7	9	251.832	5.095	4.176
	LASSO	28	13	6	22	303.060	4.790	4.470
	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 7. 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. All the methods perform similarly with respect to prediction. However, LASSO, LASSO-DCRoSIS, LASSO-M-DCRoSIS, SCAD, and SCAD-DCRoSIS tend to select one of the important variables in each group with none having the ability to do group selection.

Table 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
$n = 50$	**LASSO-DCRoSIS**	9	6	3	6	434.349	15.012	58.394
	SCAD-DCRoSIS	3	12	3	0	510.214	12.619	45.612
	LASSO-M-DCRoSIS	8	7	5	3	392.046	7.673	44.489
	LASSO	24	9	3	21	76.333	15.370	368.988
	SCAD	3	12	3	0	537.854	73.312	43.118
$n = 100$	**LASSO-DCRoSIS**	11	4	4	7	416.201	12.853	51.732
	SCAD-DCRoSIS	3	12	3	0	537.374	12.021	42.660
	LASSO-M-DCRoSIS	12	3	6	6	323.962	5.098	43.871
	LASSO	22	7	4	18	412.615	10.085	55.300
	SCAD	3	12	3	0	543.058	72.047	41.198
$n = 150$	**LASSO-DCRoSIS**	12	3	4	8	407.867	8.031	49.090
	SCAD-DCRoSIS	3	12	3	0	530.967	10.459	41.530
	LASSO-M-DCRoSIS	13	2	7	6	277.646	4.997	44.103
	LASSO	22	7	4	19	410.733	11.271	50.458
	SCAD	3	12	3	0	537.219	71.810	40.594
$n = 200$	**LASSO-DCRoSIS**	12	3	4	8	432.346	9.929	46.403
	SCAD-DCRoSIS	3	12	3	0	537.457	12.747	41.320
	LASSO-M-DCRoSIS	14	1	7	7	252.355	4.848	43.932
	LASSO	23	8	4	19	413.570	14.299	46.930
	SCAD	3	12	3	0	543.027	71.980	40.471

Table 8 present simulation results for case 4 with 10% outliers introduced into the response variable for case 4. Across all sample sizes, SCAD has the worst performance in all criteria and just like when there were no outliers, LASSO, LASSO-DCRoSIS, LASSO-M-DCRoSIS, SCAD, and SCAD-DCRoSIS, select one of the important variables in each group.

➢ *Application to Real Life Datasets*

In this section, application of the proposed methods (LASSO-DCRoSIS, LASSO-M-DCRoSIS, and SCAD-DCRoSIS on a real life dataset is considered. The dataset is the gene expression data from the microarray experiments on 120 mammalian eye tissue samples by Scheetz et al. (2006). The dataset consist of 200 predictors which represents 200 gene probes of 120 rats. The response is the expression level of TRIM32 gene.

Firstly, the data were randomly split into a training set with 100 observations, and a test set with 20 observations. The training dataset were used for model fitting and selection of tuning parameters by10-fold cross validation. The performance of the methods are then compared based on their prediction mean squared error (MSEy) on the test dataset and number of non-zero coefficients. The process of data splitting, model fitting and computation of MSEy were repeated 100 times. The results for both datasets are summarized in Table 6.

The boxplot and the histogram of Y (TRIM32 gene) are displayed in Figures 1 and 2. Both indicate that the response distribution may be heavy-tailed and the data contain outliers.

ISSN No:-2456-2165

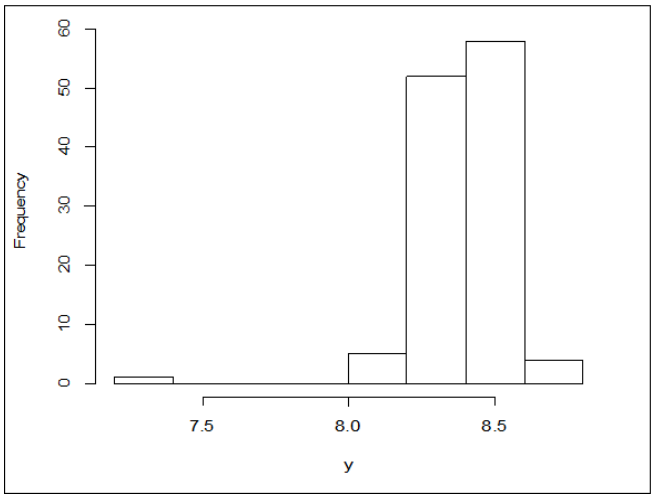

Fig 1 Histogram of the Response Variable (the Expression Level of TRIM32 Gene) for the Gene Expression Data

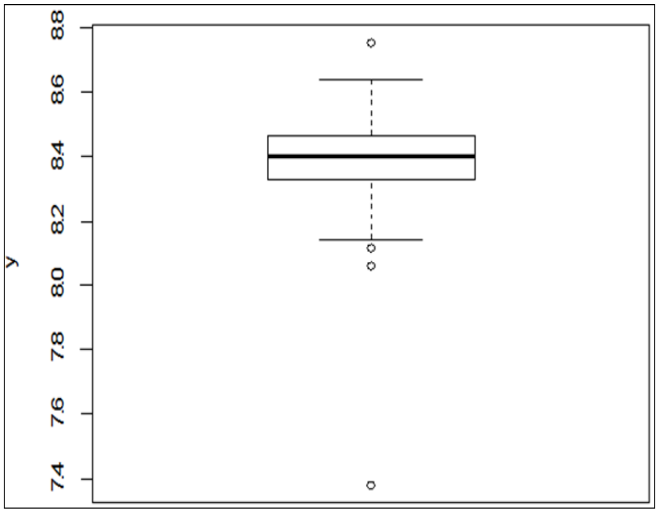

Fig 2 Boxplot of the Response Variable (the Expression Level of TRIM32 Gene) for the Gene Expression Data

After the DC-RoSIS screening, only 50 predictors were left in the model. The penalized regression was further used to select important predictors and estimate the coefficients using the considered penalty functions. Table 9 gives the size (number of predictors selected) of the model produced by the considered penalty functions, the selected predictors and corresponding estimates.

Table 9 Median mean squared errors of prediction ($\mathbf{MSE_Y}$) and median estimated model sizes (**S**), based on 100 replications

Method	**Eye Tissue Data**	
	$\boldsymbol{MSE_Y}$	$\boldsymbol{S}$
LASSO-DCRoSIS	0.0091	13
SCAD-DCRoSIS	0.0101	7
LASSO-M-DCRoSIS	0.0077	10
LASSO	0.0077	25
SCAD	0.0094	11

Table 9 obviously shows that the proposed methods select more sparse models compared to the corresponding existing version with no substantial loss in prediction accuracy. The results indicate that the LASSO and LASSO-M-DCRoSIS yielded the same prediction error which is the lowest but the LASSO-M-DCRoSIS selected fewer predictors underscoring the superiority of the proposed method for this data. The LASSO-DCRoSIS selected 12 less predictors than the LASSO without significant loss in prediction accuracy.

The results from this section further show that the proposed methods perform considerably well for prediction and variable selection.

V. CONCLUSION

Evidence abound in literature to show that traditional screening techniques perform poorly in the presence of outliers, necessitating the need to generate new approaches that improve the performance of legacy screening techniques. In this paper, we attempt to enhance the performance of traditional approaches (LASSO and SCAD) we combined them with the robust screening technique (DCRoSIS) that can do well in the presence of outliers with a view to achieving better dimension reduction and variable selection simultaneously. The simulation and performance on real life data show that our proposed LASSO-DCRoSIS performs better than the rest in both circumstances.

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.

[10]. Buba, A., Usman, U., Musa, Y., & Hamza, M. M.(2023). Hybrid Regression Estimation and Feature Selection Technique Using Robust Variable Screening Technique and Regularization. *International Journal of Mathematics and Statistics Invention (IJMSI)*, 11(5), 10-16.