Pertanika J. Sci. & Technol. 29 (2): 1053 - 1070 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Penalized LAD-SCAD Estimator Based on Robust Wrapped Correlation Screening Method for High Dimensional Models

Ishaq Abdullahi Baba[1,3], Habshah Midi[1,2]*, Leong Wah June [1,2] and Gafurjan Ibragimove[1,2]

[1] Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia
[2] Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor
[3] Department of Mathematical Sciences, Faculty of Science, Taraba State University Jalingo, Taraba State Nigeria

ABSTRACT

The widely used least absolute deviation (LAD) estimator with the smoothly clipped absolute deviation (SCAD) penalty function (abbreviated as LAD-SCAD) is known to produce corrupt estimates in the presence of outlying observations. The problem becomes more complicated when the number of predictors diverges. To overcome these problems, the LAD-SCAD based on sure independence screening (SIS) technique is put forward. The SIS method uses the rank correlation screening (RCS) algorithm in the pre-screening step and the traditional Pathwise coordinate descent algorithm for computing the sequence of the regularization parameters in the post screening step for onward model selection. It is now evident that the rank correlation is less robust against outliers. Motivated by these inadequacies, we propose to improvise the LAD-SCAD estimator using robust wrapped correlation screening (WCS) method by replacing the rank correlation in the SIS method with robust wrapped correlation. The proposed estimator is denoted as WCS+LAD-SCAD and will be employed for variable selection. The simulation study and real-life data examples show that the proposed procedure produces more efficient results compared to the existing methods.

Keywords: LAD-SCAD estimators, robust screening, ultrahigh dimensional data, variable selection

ARTICLE INFO

Article history:
Received: 19 August 2020
Accepted: 6 January 2021
Published: 30 April 2021

DOI: https://doi.org/10.47836/pjst.29.2.19

E-mail addresses:
ishaqbaba@yahoo.com; ishaqabdullahibaba@gmail.com (Ishaq Abdullahi Baba)
habshah@upm.edu.my (Habshah Midi)
leongwj@upm.edu.my (Leong Wah June)
ibragimov@gmail.com (Gafurjan Ibragimove)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

The variable selection has become the key ingredient in building reliable and reproducible prediction models and hence, is fundamental to scientific studies in areas such as text processing, gene expression analysis, epidemiology, and combinatorial chemistry, among others. The main objective of variable selection is to find a subset of the predictor variables with the highest predictive power and better interpretability. In practice, most of the real-life dataset contains a huge number of the predictor variables, though some may not be significant to the target response. As a result, the accuracy of the model selection poses a major challenge due to the fact that redundant variables may affect the selection frequency of the authentic predictor variables (George, 2000; Heinze et al., 2018; Uraibi et al., 2017). Stepwise selection procedures are often used to find the most relevant variables which influence the value of the dependent variable. However, this method is known to produce inconsistent and biased estimates in addition to poor prediction and therefore, it is considered impractical for variables selection (Whittingham et al., 2006; Desboulets, 2018). In line with this, penalized least square estimators such as the least absolute shrinkage and selection operator are developed to remedy these shortcomings (Tibshirani, 1996). The attractive features of Lasso are that they can simultaneously perform both estimation and variable selection, and they can also be applied to the high dimensional dataset. Furthermore, Leng et al. (2006) and Meinshausen and Bühlmann (2006) exemplified the inconsistency of the Lasso condition. Fan and Li (2001) and Fan and Peng (2004) noted that the Lasso did not enjoy the oracle property, which was the ability to correctly estimate the insignificant coefficient with probability converging to one. This led to the development of the Adaptive Lasso estimator which had been proven to enjoy the oracle properties under regularity conditions (Zou, 2006). It is noted that the two objective functions of Lasso and Adaptive Lasso are convex; therefore, they do not achieve the closed form numerical solutions. Conversely, smoothly clipped absolute deviation (SCAD) penalized least squared objective function is concave (Xie & Huang, 2009) and note that SCAD penalized estimators can achieve sparse estimates and unbiased solution for the large coefficients. There are several interesting penalized estimators in the literature which include the bridge regression (Frank & Friedman, 1993), Dantzig selector (Candes & Tao, 2007), and the Elastic Net (Zou & Hastie, 2005). The difficulty of using penalized methods becomes obvious when the dimensionality is ultrahigh. To address this problem, Fan and Lv (2008) introduced the concept of sure independence screening (SIS) to reduce dimensionality from ultrahigh scale to moderate which was below the sample size and then selected the most significant variables into the linear model. As mentioned in Fan and Song (2010), it is explicit that SIS is computed based on the ordinary least squares (OLS) and is heavily dependent on the joint normality assumption between predictors and the response. Consequently, compared with Lasso and smoothly clipped absolute deviation

(SCAD) method, SIS proves to produce accurate prediction and interpretable model. Li et al. (2012) and Fan et al. (2009) presented feature screening-based distance correlations. The extended ultrahigh dimensional sure independence screening for the generalized linear model was discussed in Saldana and Feng (2018) and Ahmed and Bajwa (2019). However, since the preceding studies created a synergy between the Lasso type estimators and the sure screening methods for ultrahigh dimensional data, the Lasso type based sure screening methods had been studied by Ghaoui et al. (2010), Tibshirani et al. (2012), Xiang and Ramadge (2012) as cited in Ahmed and Bajwa (2019).

All the aforesaid estimators are known to be inefficient when the errors come from a heavy tailed distribution and/or in the presence of outliers in the response, as all of them technically utilize either the OLS loss function and/or the classical correlation learning algorithm. To address this issue, Wang et al. (2007) and Wu and Liu (2009) had put forward the least absolute deviation lasso (LAD lasso) and the LAD-SCAD estimators, respectively. Fan and Li (2001) observed that the penalized SCAD satisfied oracle only when the numbers of parameters were finite. Fan et al. (2009) noted that using a combination of the penalized estimators with the SIS procedure substantially improved the performance of the penalized estimators. The SIS is computed based on the Pearson correlation between the dependent and independent variables; however, it is known to be sensitive to outliers or heavy tailed errors (Li et al., 2011; Li et al., 2012). However, the authors suggested the use of rank correlation screening (RCS) method based on Kendall instead of the Pearson correlation in the pre-screening step of the SIS method. Li et al. (2011) combined the RCS with the penalized LAD-SCAD to achieve robust variable selection and parameter estimation when the number of predictor variables diverged.

Li et al. (2011) and Li et al. (2012) innovative approach and outstanding results are challenged at both the pre-screening and post screening steps as their screening algorithms and computation of the regularization parameter are based on Kendall rank correlation and the traditional path descent algorithm, which are both not robust against outlying observations (Wang et al., 2015). As an alternative to the existing correlation learning algorithm, Raymaekers & Rousseeuw (2019) advocated a robust and fast wrapped correlation algorithm which was based on the concept of g-product moment transformation. A comparison between this approach and other correlation algorithms such as the Kendall tau, Gnanadesikan Kettenring (GK) and scale estimator, among others, had clearly shown that the wrapped correlation algorithm was more robust against outliers than other existing methods (Raymaekers & Rousseeuw, 2019). Inspired by these, we proposed incorporating the wrapped correlation learning algorithm in the LAD-SCAD method to serve as a screening algorithm to reduce dimensionality from high to below sample size. The proposed method is expected to be more efficient than the existing methods in this study.

MATERIALS AND METHODS

Consider the linear regression Equation 1

$$y_i = X_i^T \beta_0 + e_i, \tag{1}$$

where β_0 is the vector of the regression parameters, X_i represents p_n the dimensional vectors of the predictors, y_i is the vector of the response variable, and e_i is independent identically distributed vector of the random errors with mean 0 and the constant variance σ^2. Under these conditions, the classical linear regression estimators are applied. In the presence of anomalous observations in a dataset, and/or when the distribution of the errors is not normal, the robust regression loss function is applied to shrink the effect of outlying observations on the estimates of the regression (Bai & Wu, 1997; Maronna et al., 2019). In the application of this concept, the relationship between the vector of the response variable y_i and the set of predictors $X_i = (x_1, \ldots, x_p)$ can be modelled by minimizing the following objective function (Equation 2):

$$\sum_{i=1}^{n} \rho(y_i - X_i^T \beta_0), \tag{2}$$

where ρ is a continuous symmetric function called the objective function with unique minimum at 0 (Rousseeuw & Leroy,1987). Various choices of ρ functions have been suggested in the robust works (Bai & Wu, 1997; Stuart, 2011) for linear regression; for example, the quantile regression with $\rho_\alpha(x) = \alpha x^+ + (1-\alpha)(-x)^+, 0 < \alpha < 1,$ where $x^+ = \max(x,0)$ and the L_q regression estimates with $\rho(x) = |x|^q, 1 \le q \le 2$. If $q = 1$, the minimizer in Equation 2 is generally named the least absolute deviation (LAD) estimator. Conversely, for sparse regression model, namely $\beta_0 = (\beta_a^T, \beta_b^T)^T$ where β_a is a $k_n \times 1$ vector of the significant parameters $(\beta_a \neq 0)$ and β_b is an $m_n \times 1$ vector of the insignificant parameters $(\beta_b = 0)$ such that $k_n + m_n = p_n$, we define correctly fitted model as a model which has k_n significant and m_n insignificant coefficients. Throughout this paper we used λ_n instead of λ to accentuate the dependency of λ on the sample size n (Fan & Li, 2001). Following Wang et al. (2015) the Huber penalized loss function can be expressed as Equation 3:

$$\sum_{i=1}^{n} \rho(y_i - X_i^T \beta_n) + n \sum_{j=1}^{p_n} p_\lambda(|\beta_{nj}|), \tag{3}$$

where p_n is the dimension of β_n, $p_\lambda(\cdot)$ is the SCAD penalty function which depend on the regularization parameter λ as defined in Fan and Li (2001) and Wang et al. (2015). There are various versions of regularization functions in the literature, but the Lasso and SCAD penalties have been cited to be more efficient. Comparison of the two functions by different scholars (Fan & Li, 2001; Xie & Huang, 2009) has shown that SCAD penalty enjoys all the three desirable properties of good penalty function: continuity, sparsity, and biasness while the Lasso penalty generates estimation bias but enjoys sparsity and continuity properties (Fan & Li, 2001; Li et al., 2011). Motivated by these properties, Huang & Xie (2007) and Li et al. (2011) proposed the penalized least squares SCAD and SCAD penalized M estimators for estimation and variable selection. More recently, Wang et al. (2015) suggested the penalized LAD-SCAD estimator with divergence number of predictors. By following Wang et al. (2015), Equation 3 can be written as Equation 4:

$$\sum_{i=1}^{n}\left|y_i - X_i^T \beta_n\right| + n\sum_{j=1}^{p_n} p'_{\lambda_n}\left(\left|\beta_{nj}^0\right|\right)\left|\beta_{nj}\right|, \tag{4}$$

Where $p'_{\lambda_n}(\cdot)$ presents a vector of $p_n \times 1$ dimension whose j^{th} elements are the first derivative of the SCAD penalty function defined by Equation 5:

$$p'_{\lambda_n}\left(\beta_{nj}^0, a\right) = \begin{cases} \operatorname{sgn}\left(\beta_{nj}^0\right)\lambda, & \left|\beta_{nj}^0 \leq \lambda,\right| \\ \operatorname{sgn}\left(\beta_{nj}^0\right)\left(a\lambda - \left|\beta_{nj}^0\right|\right)/(a-1), & \lambda < \left|\beta_{nj}^0\right| \leq a\lambda, \\ 0, & \left|\beta_{nj}^0\right| > a\lambda. \end{cases} \tag{5}$$

and $\beta_{nj}^0 = \left(\beta_{n1}^0, \ldots, \beta_{np_n}^0\right)^T$ represents a vector of an initial estimates, which is usually obtained by minimizing the unpenalized objective function in Equation 2 when $q = 1,$ and as a result, Equation 4 followed. Hence, the objective function in Equation 4 is continuous and differentiable which can achieve global minimum. To simplify the selection of regularization parameter, Fan and Li (2001) recommended using $a = 3.7$ and $\lambda > 0$ as the tuning parameter in Equation 5 for the SCAD regularization function. The minimization function in Equation 4 is called penalized LAD-SCAD estimator.

RESULTS AND DISCUSSION

Asymptotic Properties of the LAD-SCAD Estimator

In this section, we discuss some important asymptotic properties of the LAD-SCAD estimator when the number of predictor variables diverges with increasing number of samples. Pursuing Wang et al. (2015), the following definitions and assumptions are adapted:

Assumption 1. The errors are continuous and has a positive density at origin with median 0.
Assumption 2. There exists a positive fixed value $M < \infty : \max_{1 \leq i \leq n, 1 \leq j \leq p_n}\left|X_{ij}\right| \leq M.$
Assumption 3. $p_n^3/n \to 0$ as $n \to \infty$
Assumption4. There exists fixed values $0 < \rho_1 < \rho_2 < \infty$ and $0 < \tau_1 < \tau_2 < \infty : \rho_1 \leq \rho_{n1} \leq \rho_{n2} \leq \rho_2$ and $\tau_1 \leq \tau_{n1} \leq \tau_{n2} \leq \tau_2.$
Assumption5. $\sqrt{n/p_n}\lambda_n \to \infty$ and $\lambda_n \to 0$

The preceding notations and assumptions have been used in the literature to study the asymptotic normality, consistency, and sparsity properties of the penalized LAD-SCAD estimators (Li et al., 2011; Wang et al., 2015). To establish asymptotic properties of the LAD-SCAD, the following important notations and definitions are required.

Partitioning the $p_n \times 1$ vector of parameters β_n as $\beta_n = \left(\beta_{1n}^T, \beta_{2n}^T\right)^T$ in the same way as β_0, where $\beta_{1n} = \left(\beta_1, \dots, \beta_{p_0}\right)^T$ represents the vector corresponding to all significant coefficients and $\beta_{2n} = \left(\beta_{p_0+1}, \dots, \beta_p\right)$ represents a vector corresponding to all the insignificant coefficients. Let $\beta_S = \left(\hat{\beta}_{S1n}^T, \hat{\beta}_{S2n}^T\right.$ be the corresponding LAD-SCAD estimator. We also divided the predictors x_i for $i = 1, \dots, n$ as $x_i = \left(x_{i1n}^T, x_{i2n}^T\right)$ with $x_{i1n} = \left(x_{i1}, \dots, x_{ip_0}\right)^T$ and $x_{i2n} = \left(x_{p_0+1}, \dots, x_p\right)^T$. Furthermore, define $a_{1n} = \max\left\{p'_{\lambda_n}\left(\left|\beta_{nj}^0\right|\right): 1 \leq j \leq p_0\right\}$ and $b_{2n} = \min\left\{p'_{\lambda_n}\left(\left|\beta_{nj}^0\right|\right): p_0 < j \leq p\right\}$, where $p'_{\lambda_n}(\cdot)$ is a function of sample size *n* as mentioned in (Fan & Peng, 2004; Li et al., 2011). Based on the preceding assumptions and notations, the following theorems can be established for the modified penalized LAD-SCAD estimator with divergence number of predictors.

Theorem 1. (Consistency). Assuming that $\left(y_i, x_i\right)$, $i = 1, \dots, n$ are i.i.d. and the conditions given in Assumption1-5 are satisfied, then there exists a penalized LAD-SCAD estimator $\hat{\beta}_n : \left\|\hat{\beta}_n - \beta_0\right\| = \mathrm{O}p\left(\sqrt{p_n/n}\right)$

Theorem 2. (Oracle property). Assume that $k_n^3 p_n^3 / n \to 0, E_n\left(r\left(\hat{\theta}_n\right)\right) = op\left(1/\sqrt{p_n}\right)$, and the assumptions 1-5 are satisfied, then the penalized LAD-SCAD estimator $\hat{\beta}_n = \left(\hat{\beta}_{1n}^T, \hat{\beta}_{2n}^T\right)^T$ satisfies the following:

i. (Sparsity) $\hat{\beta}_{2n} = 0$ with $\Pr\left(\hat{\beta}_{2n} = 0\right) \to 1$ as $n \to \infty$.

ii. (Asymptotic normality)

$$n^{1/2} \alpha^T \Sigma_{n,11}^{1/2} \left(\hat{\beta}_{1n} - \beta_{10}\right) = \frac{-1}{2 f(0) \sqrt{n}} \alpha^T \Sigma_{n,11}^{-1/2} \sum_{i=1}^{n} \left(I\left(e_i < 0\right) - I\left(e_i > 0\right)\right) w_i \overset{D}{\to} N\left(0, 0.25\left(f(0)\right)^{-2}\right),$$

Here α is a random $k_n \times 1$ vector with $\|\alpha\| = 1$, and the notation $\underset{D}{\to}$ means the convergence in distribution. The proofs of the foregoing theorems follow from the adaptation of the proofs in Wang et al. (2015).

Wrapping Sure Independent Screening for Ultrahigh Dimensional Data

In the preceding section, we discussed the desirable properties of the penalized LAD-SCAD estimator when the number of predictors is less than the number observations. In the present section, we will give special attention on the ultrahigh dimensional scenario which is $p_n > n$. Fan and Lv (2008) noted that the traditional penalized selection procedure like Lasso or SCAD or Dantzing selector tended to misbehave when the number of the predictor variables p_n was too large compare to the number of observations. This had motivated Fan et al. (2009) and Wang et al. (2015) to combine the penalized and the SIS procedures to have a two steps procedure called penalized sure screening for high dimensional linear regression models. In the first step, SIS is implemented to reduce the dimension from the ultrahigh to below sample size and then used a suitable penalized estimator to accomplish the final variable selection and parameters estimation concurrently. The SIS is computed according to the Pearson correlation between the dependent variable and the predictors.

Li et al. (2012) noted that Pearson correlation was affected by outliers and or heavy tailed errors. This led to the procedure of LAD-SCAD given in Wang et al. (2015) where it employed rank correlation screening method based on rank correlation, denoted as RCS+ LAD-SCAD. The appealing features of the combined procedure includes: (i) It is easy to apply because the computational burden for huge or large scale problems can easily be handled by any well-known correlation algorithm for example Pearson correlation and (ii) it enjoys the oracle properties (Fan & Li, 2001). Despite these striking properties, the existing combined procedure that is RCS+LAD-SCAD still have two shortfalls: (i) it uses rank based correlation which may or may not be preferable in a given application as they measure monotonicity instead of the linear relationship in addition to being too sensitive to multiple outliers (Croux & Dehhon, 2010; Raymaekers & Rousseeuw, 2019) (ii) the RCS combined with LAD-SCAD estimator, does not take into account the effect of outlier in the computation of the sequence of regularization parameters which is considered as the key determinant factor in selecting predictable and interpretable model (Zhang et al., 2010; Shevlykov & Simironov, 2011). Recently, Raymaekers and Rousseeuw (2019) had shown that the rank correlation based on Kendall was liable in the presence of contamination and or heavy tailed errors. Comparisons between the robust and the non-robust correlation methods can be found in Shevlykov and Simironov (2011). The latter works motivate us to improvise the RCS+LAD-SCAD by employing the wrapped correlation algorithm to achieve robust sure independent screening and compute the set of the regularization parameters for model selection for onward model selection and prediction. We call this method WCS+LAD-SCAD estimator. Following Fan and Lv (2008), let $S_* = \{1 \le p : \beta_j \ne 0\}$ be the true model parameter values with non-sparsity size $s << n$. Under the assumption of sparsity, the other $p - s$ variables can also be associated with the response variable by linkage to the predictor variables that are enclosed in the model. Let $w = (w_1, \ldots, w_p)$ be a $p_n \times 1$ vector computed based on the component wise regression that is Equation 6.

$$w = X^T y \tag{6}$$

For any given $\gamma \in (0,1)$, the component wise magnitude of the vector w is sorted in decreasing order and a submodel can be defined as:

$S_\gamma = \{1 \le j \le p : | w_j |$ is among the first $[\gamma n]$ largest of all$\}$, where $[\gamma n]$represents the integer part of γn. This is the commonly used procedure to reduce the full model from high down to a submodel S_γ with a size $d_n = [\gamma n] < n$. In similar style, Wang et al. (2015), replaced Equation 6 with Kendall correlation learning algorithm while maintaining the same selection criteria to better estimates since the latter is proved to be non-robust estimator. The Kendall correlation values can be obtained using the Equation 7:

$$w_k = \frac{1}{n(n-1)} \sum I(x_{ij} < x_{ij}) I(y_i < y_i) - \frac{1}{4}, \quad j = 1, \ldots p_n \tag{7}$$

Using the same approach as in Fan and Lv (1996) and Wang et al. (2015), but with fast and more robust correlation estimator built upon the wrapped correlation algorithm of Raymaekers and Rousseeuw (2019), we will replace the correlation formula in Equation 6 and 7 with the wrapped correlation algorithm introduced by Raymaekers and Rousseeuw (2019) while maintaining the selection procedure in Fan and Lv (1996) and Wang et al. (2015). The wrapping correlation procedure can be defined as follows: let the correlation values of the wrapped variables be defined by the entries as Equation 8.

$$w_j = cor(X^*, y^*) = cor(\psi_{b,c}\left(\frac{x_{ij} - \hat{\mu}_{xj}}{\hat{\sigma}_{xj}}\right), \left(\frac{y_i - \hat{\mu}_{yj}}{\hat{\sigma}_y}\right)) \tag{8}$$

where X^* and y^* are the transformed variables, $\hat{\mu}_{xj}$ and $\hat{\mu}_y$ are the estimates of location computed based on the one step M estimator of location with the wrapping function $\psi_{b,c}$, with $b = 1.5$ and $c = 4$, $\hat{\sigma}_{xj}$ and $\hat{\sigma}_y$ are computed using the MAD estimator (Raymaekers & Rousseeuw, 2019). We will present the broader steps involved for the implementation of the proposed WCS+LAD-LASSO in the next section.

Computational Procedures

The LAD-SCAD estimator can be computed easily by augmenting a dataset and using any suitable existing software; for example, Matlab or Phyton or R for quantile regression. Define an augmented dataset $\{(\tilde{y}_i, \tilde{X}_i), i = 1, \ldots, n, n+1, \ldots + n + p_n\}$, where $\tilde{y}_i = y_i / n, \tilde{X}_i = X_i / n$ for $i = 1, \ldots, n$ and $\tilde{y}_i = 0, \tilde{X}_i = p'_{\lambda_n}(|\beta^0_{nj}|)\tilde{e}_i$ for $i = n+1, \ldots, n + p_n$ and $\tilde{e}_i$ is a $p_n \times 1$ dimensional vector with i^{th} term component equal to 1 such that Equation 4 can be expressed as Equation 9.

$$\sum_{i=1}^{n} \left| y_i - X_i^T \beta_n \right| + n \sum_{j=1}^{p_n} p'_{\lambda_n}(|\beta^0_{jn}|) |\beta_{jn}| = \sum_{i=1}^{n+p_n} \left| \tilde{y}_i - \tilde{X}_i^T \beta_n \right| \tag{9}$$

Equation 9 is equivalent to the traditional LAD objective function for the dataset $(\tilde{y}_i, \tilde{X}_i)$ for $i = 1, 2, \ldots, n, n+1, \ldots, n + p_n$. In this paper, an R software package quantreg is used for solving Equation 9. Furthermore, in the robust Lasso type regression, the selection of appropriate regularization parameter controls the complexity and improves the prediction accuracy of the selected model (Wang & Zhu, 2011; Friedman et al., 2010; Gao & Huang, 2010). Although there are several methods for selecting the best regularization parameter in the literature, the issue of robustness and computational efficiency desire special attention as well. Friedman et al. (2010) suggested a coordinate descent algorithm for computing the regularization parameters which is adapted in Wang et al. (2015). The computation of the sequence of *K* values of λ is a function of $\lambda_{\min}$ and $\lambda_{\max}$ where $\lambda_{\min} = \in \lambda_{\max}$, $\lambda_{\max} = \max |\langle X_i, y \rangle| / n, \in = 0.001$ and $K = 100$ (Friedman et al., 2010). Following this concept, we replace $|\langle X_i, y \rangle|$ with the wrapped correlation to robustly estimate the maximal correlation between the columns of *X* matrix and the vector *y* based on the idea of the product moment transformation of the dataset as backed in Raymaekers and Rousseeuw (2019). Our reason for this is to reduce the effect of outliers before applying the selection criteria for onward model selection. This is followed by selecting a suitable model selection criterion such as the type considered in Wang et al. (2007) and Chang et al. (2018). In this paper, we consider the BIC type criteria used in Gao and Huang (2010) as cited in Wang et al. (2015). This is defined by Equation 10:

$$BIC = \log\left(\frac{1}{n}\sum_{i=1}^{n}\left|y_i - X_i^T\hat{\beta}_n\right|\right) + d_\lambda \frac{\log(n)}{n}, \tag{10}$$

where d_λ is the number of significant regression coefficients. The computation procedure for the proposed WCS+LAD-SCAD estimator can be summarized as follows:
Step 1. For a given dataset (X, y) where $X \in \Re^{n\times p}$ is the design matrix and $y \in \Re^n$ is the response variable and $d_n \in \mathrm{Z}_+$

1. Compute the robust initial scale $\hat{\sigma}_{xj}$ for $1,2,\ldots,p$ based on the median absolute deviation (MAD) for the matrix $X \in \Re^{n\times p}$ and $\hat{\sigma}_y$ for the vector of response $y_i \in \Re^{n\times p}$.
2. Compute a one-step M location estimator $\hat{\mu}_{xj}$ for the matrix $X \in \Re^{n\times p}$ and $\hat{\mu}_y$ for the vector of response $y_i \in \Re^{n\times p}$ with wrapping function $\psi_{b,c}$ where $b = 1.5, c = 4$
3. Let the variables X^* and y^* denote the transformed of the original variables X and y , then we can compute w_j for $j = 1,\ldots p$ based on Equation 8 and select the subset of predictor variables $d_n < n$ such that $\left|w_j\right|$ is among the first largest ones. Then apply the LAD-SCAD penalized estimator to Equation 1 by employing the following steps.

Step 2. Compute the sequence of regularization parameter λ defined on the interval $[\lambda_{\max}, \lambda_{\min}]$, where $\lambda_{\max} = \max\left|\langle g_X(X)\ g_y(y)\rangle\right|$ and $\lambda_{\min} = \in \lambda_{\max}$ with $\in = 0.001$

Step 3. Compute the initial estimates $\hat{\beta}$ by minimizing the unpenalized lad objective function $\sum_{i=1}^{n}|y - x\beta|$. Using the initial estimates in step 3 and the set of regularization parameter λ in step2, compute the SCAD penalty function as in Equation 5.

Step 4. Form the augmented dataset $(\tilde{X}, \tilde{y})$ for $i = 1,2,\ldots,n,n+1,\ldots n+p,$ where $(\tilde{y}, \tilde{X}) = \left(\frac{y_i}{n}, \frac{X_i}{n}\right)$, $i = 1,\ldots,n$ and $(\tilde{y}, \tilde{X}) = \left(0, p'_\lambda\left(\left|\beta^0_{nj}\right|\right)\tilde{e}_i\right)$ for $i = 1,2,\ldots,n,n+1,\ldots n+p,$ and $\tilde{e}_i$ is a vector of p by 1 with all the i^{th} component equal to 1.

Step 5. Use any of the LAD regression procedure, for example quantreg package in R to compute the lad regression estimators for the augmented dataset $(\tilde{X}, \tilde{y})$ for $i = 1,2,\ldots,n,n+1,\ldots n+p,$

Step 6. Select the best model by computing the formula in Equation 9 for each value of λ. The model that corresponds to the minimum value of λ is considered as the best.

Numerical Evaluation

To assess the performance of the proposed WCS+LAD-SCAD estimator as explained in the proceeding sections, we carried a simulation study and analysis three real datasets namely NIR, octane and cookie dataset.

Simulation Study

A simulation study is carried out to compare the variable selection properties and prediction accuracy of the proposed method and the existing RCS+LAD-SCAD estimator with divergent number of predictors. As per Wang et al. (2015) the following three cases are considered:

Case 1: $y = X_i^T \beta + e_i$, with $e_i \sim N(0,1), i = 1, \ldots, n$
Case 2: $y = X_i^T \beta + e_i$, with $e_i \sim t(3), i = 1, \ldots, n$
Case 3: $y = X_i^T \beta + e_i$, with $e_i \sim 0.9N(0,1) + 0.1Cauchy(3), i = 1, \ldots, n$

Here, for each case, we set the vector of coefficients β such that $\beta_1 = 3, \beta_2 = 1.5, \beta_5 = 2$ and $\beta_j = 0$ for $j \notin \{1,2,5\}$. The vector of predictors $X_i = (x_1, \ldots, x_p)^T$ are generated from the multivariate normal distribution with mean 0 and covariance matrix $\Sigma = \sigma_{ij}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The number of predictors p_n and the number of sample size *n* are set to $p_n = (500,1000)$ and $n = (100,200)$ which are repeated 200 times in each case. We used the same seed number, 1234 throughout this paper both in the simulation and real date examples. The threshold $d_n = 4n/\log(n)$ is used with wrapped correlation screening WCS method to reduce the dimension from p_n to d_n. Table 1 to 3 exhibit the results based on the average number of zero coefficients correctly estimated as zero (NC), the average number of non-zero coefficients incorrectly estimated to zero (NIC). Following Aslan (2012) and Wang et al. (2015), the average median estimation error (MEE) defined as $\left\|\hat{\beta} - \beta\right\|$ is also used as performance measures of the various estimators. The values in parenthesis are for $n = 200$ and $p_n = 1000$

Table 1

Results for normal distributed errors with $n = 100$, $p = 500$; $n = 200$, $p = 1000$ *in parenthesis*

Method	NC	NIC	MEE	%Efficiency
RCS+LAD-Lasso	495.050(995.435)	0.000(0.000)	0.585(0.467)	37.43(31.69)
RSC+LAD-SCAD	496.570(996.975)	0.003(0.000)	0.281(0.160)	77.94(92.50)
WSC+LAD-Lasso	495.275(995.510)	0.000(0.000)	0.553 (0.398)	39.60(37.19)
WSC+LAD-SCAD	496.615(996.980)	0.000(0.000)	0.261(0.150)	83.91(98.66)
Oracle	497.000(997.000)	0.000(0.000)	0.219(0.148)	100(100)

Table 2

Results for t_3 *distributed errors with* $n = 100$ *,* $p = 500$ *;* $n = 200$ *,* $p = 1000$ *in parenthesis*

Method	NC	NIC	MEE	%Efficiency
RCS+LAD-Lasso	495.175(995.985)	0.000(0.000)	0.720(0.580)	31.81(31.55)
RSC+LAD-SCAD	495.775(996.815)	0.004(0.000)	0.520(0.214)	44.04(85.51)
WSC+LAD-Lasso	495.525(996.705)	0.000(0.000)	0.654(0.465)	35.02(39.35)
WSC+LAD-SCAD	496.061(997.000)	0.003(0.000)	0.482(0.183)	47.51(100.00)
Oracle	497.000(997.000)	0.000(0.000)	0.229(0.183)	100(100)

Table 3

Results for normal errors with 10% contaminated observations, $n = 100$ *,* $p = 500$ *;* $n = 200$ *,* $p = 1000$ *in parenthesis*

Method	NC	NIC	MEE	%Efficiency
RCS+LAD-Lasso	495.110(996.175)	0.006(0.001)	0.623(0.521)	36.44(29.37)
RSC+LAD-SCAD	496.310(996.970)	0.051(0.002)	0.381(0.163)	59.58(93.87)
WSC+LAD-Lasso	495.465(995.895)	0.005(0.000)	0.597(0.424)	38.02(36.08)
WSC+LAD-SCAD	496.430(997.000)	0.030(0.000)	0.360(0.156)	63.06(98.08)
Oracle	497.000(997.00)	0.000(0.000)	0.227(0.153)	100(100)

It can be observed From Table 1 to 3 that the RCS+LAD-Lasso and WSC+LAD-Lasso provide the worst results in terms of having NC and MEE values far away from the oracle values. Although their results are close, the values of MEE for WSC+LAD-Lasso are consistently smaller than those of the RCS+LAD-Lasso which indicate that WSC screening algorithm tends to increase the estimation accuracy. On the other hand, the values of the NC, NIC and MEE of WCS+LAD-SCAD estimator tend to be closer to the Oracle estimator. Nevertheless, the WSC+LAD-SCAD estimator provides NC, NIC and MEE values which are in best agreement with the oracle values especially when the $n = 200$ and $p = 1000$. The performance of RSC-LAD-SCAD is quite good both in terms of NC and MEE values; however, its accomplishment cannot outperform the WSC+LAD-SCAD. The performances of WSC-LAD-SCAD compared to other estimators is further assesed based on efficiency criterion (Dhhan et al., 2017) defined as follows:

Efficiency = (MEE of the (oracle estimator)/MEE of the (oracle competitors)) x 100%.

It can be clearly seen from Table 1 to 3 that our proposed WSC-LAD-SCAD has the highest value of efficiency, followed by RSC+LAD-SCAD, WSC+LAD-Lasso and RCS+LAD-Lasso. The results seem to be consistent for each sample $n = (100{,}200)$, and the number of predictors, $p_n = (500{,}1000)$, respectively. Hence, the WSC+LAD-SCAD may give a better alternative estimator for handling ultrahigh dimensional data in the presence of outliers. Furthermore, the results incline to suggest that by using the WSC screening algorithm, the efficiency of the LAD-SCAD estimator can substantially be improved.

Real Life Application

In this section, several real-life examples are presented to show the applicability and merit of our proposed method. These datasets include near infrared spectroscopy (NIR), cookies, and octane. The dataset can be obtained from the chemometrics, ppls, and rrcov packages in R, which have been formerly analysed in Brown et al. (2001), Liebmann et al. (2009), and Hubert et al. (2005), respectively. The octane data set consists of 39 samples and 226 predictors. After some preliminary investigation of the response observations based on Hubert and Van der Veeken (2008) as used in https://github.com/marcellodo/univOutl, we detected observation 25, 26, 36-39 as outliers. This procedure was repeated on the responses of the cookies and NIR data set, but all detected zero outlying points. The dimension for NIR data is 166 by 235 and 72 by 700 for the cookies data which represent number of sample and predictors, respectively. Motivated by the skewness property of the octane data plus our motive to assess the vigour of the proposed procedure, we contaminated the responses of NIR and cookies data by multiplying 0.05 percent of randomly selected observation by 100 after splitting the data into training and test set (the percentage are given in parenthesis) as given in Table 4. Applying the computational procedure described in the previous sections produced the results as shown in Table 4 for test data set, where the second column before the last column, represents the number of variables selected (NVS), the column before the last represents the average median absolute error, and the last represents the average robust R^2 statistics over 100 repeated simulations, as used in (Liu et al, 2016). Here, we employed the robust R^2 which is adapted from Wang et al. (2007) as defined Equation 11:

$$R^2 = 1 - \left(\frac{med\left(\left|y_i - \hat{y}_i\right|\right)}{mad\left(y_i\right)} \right)^2 \qquad [11]$$

Generally speaking, the range of R^2 value start from 0 to 1 for a reasonable fit, R^2=1 signifies perfect fit and $R^2 < 0$ corresponds to bad model fit. From Table 4, it is interesting to observe that the proposed technique outperformed its competitors on all the data set, and is considered as our method to select a reasonable number of variables with minimum median absolute error (MAE) and better R^2 statistics. The threshold value of d=n/log (n) is used in all the data sets. A boxplot of the median absolute error and R^2 statistics for the original data (raw values) as shown in Figure 1 to 2 explicitly demonstrate the effectiveness of our method against its competitors.

Table 4

Real data application

No.	Dataset	Method	#train (%)	#test (%)	NVS	MAE	R^2
1	NIR (166×235)	RCS+LAD-Lasso	120(72)	46(28)	1	0.659	-0.8597
		RSC+LAD-SCAD			3	0.253	0.2222
		WSC+LAD-Lasso			1	0.368	-0.1848
		WSC+LAD-SCAD			8	0.212	0.3472
2	Octane (39×226)	RCS+LAD-Lasso	25(64)	14(36)	3	0.008	0.9994
		RSC+LAD-SCAD			4	0.007	0.9999
		WSC+LAD-Lasso			6	0.005	0.9995
		WSC+LAD-SCAD			6	0.004	1.0000
3	Cookie (72×700)	RCS+LAD-Lasso	40(56)	32(44)	1	0.433	0.7589
		RSC+LAD-SCAD			8	0.237	0.9281
		WSC+LAD-Lasso			1	0.432	0.7559
		WSC+LAD-SCAD			10	0.143	0.9697

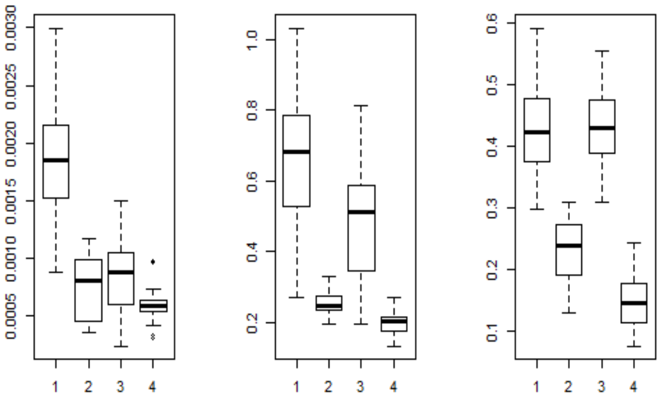

Figure 1. Boxplot of the median absolute error for the octane, NIR and cookies data set with RCS+LAD-Lasso=1, RCS+LAD-SCAD=2, WCS+LAD-Lasso =3 and WCS+LAD-SCAD =4 respectively.

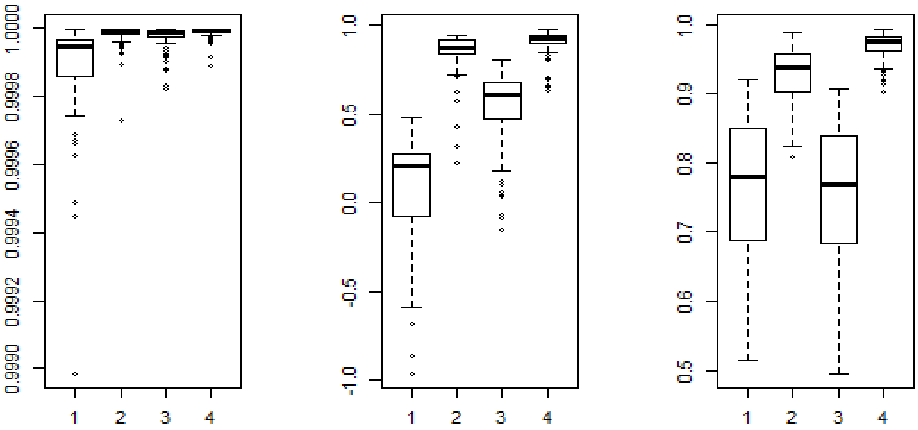

Figure 2. Boxplot of the R^2 statistics for the octane, NIR and cookies data set with RCS+LAD-Lasso=1, RCS+LAD-SCAD=2, WCS+LAD-Lasso =3 and WCS+LAD-SCAD =4, respectively.

CONCLUSION

Inspired by the robust rank correlation sure screening-based LAD-SCAD, we proposed a wrapped based sure screening LAD-SCAD to achieve better robust estimates. The main advantage of our method is that it deals with outliers in both the pre-screening and post-screening step by using the robust wrapped transformation in the computation of best regularization parameter. The proposed procedure shows more success as it appears to be more robust and efficient than the existing RCS+LAD-SCAD method for solving linear regression in the presence of outlying observations. Therefore, the proposed procedure can be used by practitioners for parameter estimations and variable selection when the response observation contains some outliers. Future work will consider the impact of both vertical and horizontal outliers.

ACKNOWLEDGEMENT

This research was partially supported by the Fundamental Research Grant Scheme (FRGS) under Ministry of Education with project number FRGS/1/2019/STG06/UPM/01/1 and Petroleum Technology Development Fund of Nigeria.

REFERENCE

Ahmed, T., & Bajwa, W. U. (2019). ExSIS: Extended sure independence screening for ultrahigh-dimensional linear models. *Signal Processing*, *159*, 33-48. https://doi.org/10.1016/j.sigpro.2019.01.018

Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis, 56*(6), 1952-1965. https://doi.org/10.1016/j.csda.2011.11.022

Bai, Z. D., & Wu, Y. (1997). General M-estimation. *Journal of Multivariate Analysis*, *63*(1), 119-135. https://doi.org/10.1006/jmva.1997.1694

Brown, P. J., Fearn, T., & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, *96*(454), 398-408. https://doi.org/10.1198/016214501753168118

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, *35*(6), 2313-2351. https://doi.org/10.1214/009053606000001523

Chang, L., Roberts, S., & Welsh, A. (2018). Robust Lasso Regression Using Tukey's Biweight Criterion. *Technometrics*, *60*(1), 36-47. https://doi.org/10.1080/00401706.2017.1305299

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, *19*(4), 497-515. https://doi.org/10.1007/s10260-010-0142-z

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*, *6*(4), Article 45. https://doi.org/10.3390/econometrics6040045

Dhhan, W., Rana, S., & Midi, H. (2017). A high breakdown, high efficiency and bounded influence modified GM estimator based on support vector regression. *Journal of Applied Statistics*, *44*(4), 700-714. https://doi.org/10.1080/02664763.2016.1182133

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348-1360. https://doi.org/10.1198/016214501753382273

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928-961. https://doi.org/10.1214/009053604000000256

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, *38*(6), 3567-3604.

Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, *10*, 2013-2038.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109-135.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Gao, X., & Huang, J. (2010). Asymptotic analysis of high-dimensional LAD regression with LASSO. *Statistica Sinica*, 1485-1506.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, *95*(452), 1304-1308.

Ghaoui, L. E., Viallon, V., & Rabbani, T. (2010). Safe feature elimination for the lasso and sparse supervised learning problems. *Machine Learning, 2000*, 1-31.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431-449. https://doi.org/10.1002/bimj.201700067

Huang, J., & Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. In *Asymptotics: Particles, Processes and Inverse Problems* (pp. 149-166). Institute of Mathematical Statistics. https://doi.org/10.1214/074921707000000337

Hubert, M., & Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *22*(3-4), 235-246. https://doi.org/10.1002/cem.1123

Hubert, M., Rousseeuw, P. J., & Branden, K. V. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, *47*(1), 64-79. https://doi.org/10.1198/004017004000000563

Leng, C., Lin, Y., & Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, 1273-1284.

Li, G., Peng, H., & Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica*, 391-419.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129-1139. https://doi.org/10.1080/01621459.2012.695654

Liebmann, B., Friedl, A., & Varmuza, K. (2009). Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Analytica Chimica Acta*, *642*(1-2), 171-178. https://doi.org/10.1016/j.aca.2008.10.069

Liu, J., Wang, Y., Fu, C., Guo, J., & Yu, Q. (2016). A robust regression based on weighted LSSVM and penalized trimmed squares. *Chaos, Solitons & Fractals*, *89*, 328-334. https://doi.org/10.1016/j.chaos.2015.12.012

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods (with R)*. John Wiley & Sons.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436-1462. https://doi.org/10.1214/009053606000000281

Raymaekers, J., & Rousseeuw, P. J. (2019). Fast robust correlation for high-dimensional data. *Technometrics*, 1-15. https://doi.org/10.1080/00401706.2019.1677270

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley.

Saldana, D. F., & Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software*, *83*(2), 1-25. https://doi.org/10.18637/jss.v083.i02

Shevlyakov, G., & Smirnov, P. (2011). Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics*, *40*(1&2), 147-156. https://doi.org/10.17713/ajs.v40i1&2.206

Stuart, C. (2011). *Robust regression*. Durham University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., & Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(2), 245-266. https://doi.org/10.1111/j.1467-9868.2011.01004.x

Uraibi, H. S., Midi, H., & Rana, S. (2017). Selective overview of forward selection in terms of robust correlations. *Communications in Statistics: Simulation and Computation, 46*(7), 5479-5503. https://doi.org/10.1080/03610918.2016.1164862

Wang, H., Li, G., & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, *25*(3), 347-355. https://doi.org/10.1198/073500106000000251

Wang, M., Song, L., & Tian, G. L. (2015). SCAD-penalized least absolute deviation regression in high-dimensional models. *Communications in Statistics-Theory and Methods*, *44*(12), 2452-2472. https://doi.org/10.1080/03610926.2013.781643

Wang, T., & Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, *102*(7), 1141-1151. https://doi.org/10.1016/j.jmva.2011.03.007

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182-1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x

Wu, Y., & Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica, 19*(2), 801-817.

Xiang, Z. J., & Ramadge, P. J. (2012). Fast lasso screening tests based on correlations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2137-2140). IEEE Conference Publication. https://doi.org/10.1109/ICASSP.2012.6288334

Xie, H., & Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, *37*(2), 673-696. https://doi.org/10.1214/07-AOS580

Zhang, Y., Li, R., & Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*(489), 312-323. https://doi.org/10.1198/jasa.2009.tm08013

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429. https://doi.org/10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, *67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x