A Robust Variable Screening Approach With Application To Gene Expression Data

Authors: MOHAMMAD KAZEMI [iD] ✉
            – Department of Statistics, Faculty of Mathematical Sciences,
            University of Guilan,
            Rasht, Iran
            m.kazemi@guilan.ac.ir, m.kazemie64@yahoo.com

Abstract:

• The presence of outlier observations may lead to misleading results in variable screening problems. To address this issue, this paper presents a new robust variable screening method using the L_1 loss and the Huber loss. As an extension, we also develop an effective iterative procedure to improve the finite sample performance of the presented method. The effectiveness of the proposed methods is illustrated through simulation studies and real data analysis to show their capabilities. Numerical studies show that the proposed methods work well with ultrahigh-dimensional data sets, which may contain outliers, and perform better than some competing methods.

Keywords:

• *Gene selection; independence screening; NP dimensionality; outliers; sparsity.*

AMS Subject Classification:

• 62F07, 62F35.

✉ Corresponding author

1.   INTRODUCTION

Recent research in statistical science has focused on developing effective and useful techniques for analyzing gene expression data. In such ultrahigh-dimensional data, the number of genes is usually in the order of thousands or millions and exponentially larger than the available cases or subjects. For instance, thousands of gene expression profiles can be used in disease classification; millions of single-nucleotide polymorphisms are available for genome-wide association studies between genotypes and phenotypes. If we are to relate these ultrahigh dimensional genes to a response variable in a regression set-up, we need to perform variable selection.

A relevant family of methods for prediction of the response based on the high dimensional gene expression data are penalized linear regression models. Consider the linear regression model with response variable Y and p explanatory variables (e.g. gene expressions) as predictors. Given the responses $y_1, \ldots, y_n$ from n independent samples and the corresponding predictor values, say $x_{ij}, i = 1, \ldots, n$ for the j-th covariate for $j = 1, \ldots, p$, this model can be written in matrix form as

$$(1.1) \qquad y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ and ε_is are independent following $N(0, \sigma^2)$ for $i = 1, \ldots, n$. The model parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and σ^2 need to be estimated from the data. In the ultrahigh dimensional case with $p \gg n$, we need to assume sparsity of the regression coefficient $\boldsymbol{\beta}$ to achieve identifiability of the estimators, i.e., we assume that only a few of the components of $\boldsymbol{\beta}$ are non-zero. Under the sparsity assumption, estimation of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ is performed through penalized estimation procedures with appropriate penalties which can successfully recover all and only the truly important variables asymptotically with probability tending to one. There are plenty of such penalized regression procedures available in recent literature, starting from the LASSO ([16]) and its refinements([18], [21]) to more advanced procedures based on penalties like SCAD ([1]) or MCP ([19]) and many more, which work well in moderately high dimensions. However, a common problem with these methods in ultrahigh dimensional set-ups is their computational cost and numerical issues, which has led to development of simpler variable screening methods at the initial stage to reduce the number of genetic predictors from the order of potentially millions to an order of a few hundred (often lesser than the sample size as well) and then apply an appropriate penalization method to obtain final model estimates from the reduced set of covariates. The most popular method for such screening purposes is the Sure Independence Screening (SIS) proposed by Fan and Lv [2] which has a simple interpretation and theoretical guarantees along with fast computation. Even with its simple structure, the method yet enjoys the model selection oracle property under ultrahigh-dimensional set-ups where $\log(p) = O(n^\zeta)$ for some $0 < \zeta < 1$. An iterative extension, ISIS, is also proposed in [2] to tackle

the issue of collinearity among covariates. The SIS and ISIS are routinely being applied in ultrahigh dimensional applications and have also been extended to more complex models ([3], [4], [5], [6], [11], [12], [13], [10], [15], [22], for instance). Each of those proposals focuses on a specific model, and its performance is based upon the belief that the imposed working model is close to the true model. Zhu et al. [20] proposed a sure independent ranking and screening procedure which avoids the specification of a particular model structure. Motivated by this work, He et al. [7] proposed a framework called quantile-adaptive model-free screening. However, one major drawback of the SIS or ISIS is their non-robust nature against data contamination as indicated already in the discussion of the original paper itself. This issue can be crucial when applying the method for screening of important genes from large-scale genomic data, which are often prone to at least a few outliers.

In this paper, we develop a new robust screening procedure, in the context of ultrahigh dimensional linear regression, where the number of covariates p may grow exponentialy with the sample size n, using robust loss functions such as the L_1 loss and the Huber loss ([8]). A robust version of ISIS along the same lines will also be discussed to tackle the correlations among covariates. The suggested methods will be applied to the riboflavin data example.

The plan of the paper is as follows. In Section 2, a brief description of the new screening method is presented. We also introduce an iterative approach to enhance the finite sample performance of the proposed screening procedure. In Section 3, simulation studies are carried out to assess the performance of the suggested approaches, and the riboflavin data set is analyzed.

2.   METHODOLOGY

In this section, we develop a marginal utility for robust variable screening based on the L_1 loss and the Huber loss [8] to reduce the dimensionality.

Suppose that we are interested in exploring the relationship between $\mathbf{X} = (X_1, ..., X_p)^T$ and Y. A general robust framework is to minimize an objective function

$$(2.1) \qquad Q(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ is a parameter vector, and $L(.,.)$ is a robust loss function such as the L_1 loss, $L(Y_i, \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}) = |Y_i - \beta_0 - \mathbf{X}_i^T \boldsymbol{\beta}|$ or the Huber loss, defined as

$$(2.2) \quad L_\delta(Y_i, \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}) = \begin{cases} \frac{1}{2}(Y_i - \beta_0 - \mathbf{X}_i^T \boldsymbol{\beta})^2 & \text{if } |Y_i - \beta_0 - \mathbf{X}_i^T \boldsymbol{\beta}| \leq \delta \\ \delta \left(|Y_i - \beta_0 - \mathbf{X}_i^T \boldsymbol{\beta}| - \frac{1}{2}\delta \right) & \text{if } |Y_i - \beta_0 - \mathbf{X}_i^T \boldsymbol{\beta})| > \delta \end{cases},$$

where δ is a parameter that controls the robustness level, and a smaller value of δ usually leads to more robust estimation. In implementation, we consider $\delta = 1.345$, the value commonly used in robust regression that produces 95% efficiency for normal errors (see Huber [9]). Here, $L(.,.)$ can be regarded as the loss of using $\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}$ to predict Y_i.

We consider the problem of robust variable screening in ultrahigh dimensional data. The goal is to rapidly reduce the number of the predictors from p to a moderate scale via a computationally convenient procedure. Consider the marginal utility of the jth predictor as

$$(2.3) \qquad L_j = \min_{\beta_0, \beta_j} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + X_{ij}\beta_j), \qquad j = 1, \ldots, p,$$

which minimizes the loss function. The idea of SIS in this framework is to compute the vector of marginal utilities $\mathbf{L} = (L_1, \ldots, L_p)^T$ and rank the predictors according to the marginal utilities: the smaller the more important. Note that in order to compute L_j, we need only fit a model with two parameters, β_0 and β_j, so computing the vector $\mathbf{L}$ can be done very quickly and stably, even for an ultrahigh dimensional problem.

The predictor X_j is selected by SIS if L_j is one of the d smallest components of $\mathbf{L}$. Fan and Lv [2] suggested setting $d = [n/\log(n)]$, where $[a]$ refers to the integer part of a. Further, Zhu et al. [20] proposed a combination of hard and soft thresholding strategies to obtain the cutoff point that separates the active and inactive predictors. We refer to the screening procedures described above as LAD-SIS and Huber-SIS, corresponding to the L_1 and the Huber loss functions, which means that the LAD loss and the Huber loss are applied to screen the truly important predictors.

Suppose that d predictors are selected in the screening step. Now we further knock out unimportant predictors among them using a more refined penalized method, as we now describe. By reordering the predictors if necessary, we may assume without loss of generality that $X_1, \ldots, X_d$ are the variables recruited by screening. We let $\mathbf{X}_{i,d} = (X_{i1}, \ldots, X_{id})^T$ and redefine $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$. In the penalized robust approach, we seek to minimize

$$(2.4) \qquad \min_{\beta_0, \beta_j} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{X}_{i,d}^T \boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} |\beta_j|,$$

where $\lambda > 0$ is a regularization parameter, which may be chosen by five-fold cross-validation, for example. This two-stage method is summarized in Algorithm 1.

Algorithm 1 Robust SIS

Input: Data matrix $\mathbf{X}$, vector of responses $\mathbf{Y}$, model size d

Steps:

- **Step 1**(*screening*): For each $j = 1, \ldots, p$, compute the marginal utility of the jth predictor as

$$L_j = \min_{\beta_0, \beta_j} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + X_{ij}\beta_j),$$

where $L(.,.)$ is either the L_1 loss or the Huber loss. The screened sub-model $\hat{\mathcal{M}}_d$ is the indices of the d smallest entries of the marginal utilities $\mathbf{L} = (L_1, \ldots, L_p)$.

- **Step 2**(*post-screening variable selection*): Apply a robust penalized regression model, either LAD-LASSO or Huber-LASSO, to the screened variables $\mathbf{X}_{\hat{\mathcal{M}}_d} = \{X_j : j \in \hat{\mathcal{M}}_d\}$ to obtain an estimated coefficient vector, say $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_{d0}, \hat{\beta}_{dr_1}, \ldots, \hat{\beta}_{dr_d})^T$.

Output: The final estimated model $\hat{\mathcal{M}} = \{1 \leq k \leq d, \hat{\beta}_{dr_k} \neq 0\}$ along with the parameter estimates $\hat{\boldsymbol{\beta}}_d$.

The key idea of Algorthim 1 is to use different marginal utilities to screen predictor variables. As such, it can suffer from the same potential issues as the usual SIS. First, some unimportant predictors that are highly correlated with the important ones can have higher marginal utilities and thus priority to be selected than other important ones that are relatively weakly related to the response. Second, some important predictors that are jointly correlated but marginally uncorrelated with the response can be missed after screening. Such cases occur mostly due to strong correlation between the important and unimportant predictors. To address these issues, we next briefly discuss an iterative extension of Algorithm 1 that enables us to exploit more fully the joint information among the predictors.

Our the iterative extension, is motivated by the idea of two-scale learning with the iterative SIS (ISIS) in [2] and [5]. It works as follows by applying large-scale screening and moderate-scale selection in an iterative fashion. First, apply LAD-SIS (or, Huber-SIS) to the original sample $(\mathbf{X}_i, Y_i)_{i=1}^{n}$ to select k_1 variables with index set $\mathcal{A}^{(0)} = \{i_1, \ldots, i_{k_1}\}$, and then employ the LAD-LASSO (or, Huber-LASSO) to obtain a subset $\hat{\mathcal{M}}^{(0)}$ of these indices. In the second step, we compute the residuals from the fitted regression model of the response $\mathbf{Y}$ on the selected predictors in $\hat{\mathcal{M}}^{(0)}$. The LAD-SIS (or, Huber-SIS) screening is again applied taking these residuals as our new response to select another k_2 predictors from the pool of predictors with index set $\hat{\mathcal{M}}_c^{(0)} = \{1, \ldots, p\} \setminus \hat{\mathcal{M}}^{(0)}$; let us denote the index set of these k_2 selected predictors as $\mathcal{A}^{(1)}$. Then apply the LAD-LASSO (or, Huber LASSO) to predictors with indices in $\mathcal{I} = \hat{\mathcal{M}}^{(0)} \cup \mathcal{A}^{(1)}$ to obtain a set

---

Algorithm 2 Robust ISIS

Input: Data matrix $\mathbf{X}$, vectore of responses $\mathbf{Y}$, model size d.
Steps:

1. Apply LAD-SIS (or, Huber-SIS) to pick a set $\mathcal{A}^{(0)}$ of indices of size $k_1 = [2d/3]$, and then employ the LAD-LASSO regression (or, Huber-LASSO) to select a subset $\hat{\mathcal{M}}^{(0)}$ of these indices.

2. Set $t = 1$.

3. For each $j \in \hat{\mathcal{M}}_c^{(t-1)} = \{1, \ldots, p\} \setminus \hat{\mathcal{M}}^{(t-1)}$, compute

$$L_j^{(2)} = \min_{\beta_0, \beta_j} \frac{1}{n} \sum_{i=1}^{n} L(r_i^{(t-1)}, \beta_0 + \beta_j X_{ij}),$$

where $r_i^{(t-1)} = Y_i - \mathbf{x}_{i,\hat{\mathcal{M}}^{(t-1)}}^T \hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}^{(t-1)}}$ is the residual from the previous step of fitting and $\mathbf{x}_{i,\hat{\mathcal{M}}^{(t-1)}}$ is the sub-vector of $\mathbf{x}_i$ consisting of those elements in $\hat{\mathcal{M}}^{(t-1)}$. After ordering $\{L_j^{(2)} : j \in \hat{\mathcal{M}}_c^{(t-1)}\}$, we form the set $\mathcal{A}^{(t)}$ consisting of the indices corresponding to the smallest $k_2^{(t)} = d - |\hat{\mathcal{M}}^{(t-1)}|$ elements.

4. Apply the LAD-LASSO (or, Huber-LASSO) regression to variables with indices in $\mathcal{I} = \hat{\mathcal{M}}^{(t-1)} \cup \mathcal{A}^{(t)}$ to obtain the indices of the coefficients $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_{d0}, \hat{\beta}_{dr_1}, \ldots, \hat{\beta}_{dr_d})^T$ that are non-zero yield a new estimated set, $\hat{\mathcal{M}}^{(t)}$, of active indices.

5. If obtained a set of indices $\hat{\mathcal{M}}^{(t)}$, which either has reached the size d, or satisfies $\hat{\mathcal{M}}^{(t)} = \hat{\mathcal{M}}^{(t-1)}$, break and go to output stage. Otherwise go to Step 6.

6. Change t to $t + 1$ and go to Step 3.

Output: The final estimated model $\hat{\mathcal{M}} = \{1 \leq k \leq d, \hat{\beta}_{dr_k} \neq 0\}$ along with the parameter estimates $\hat{\boldsymbol{\beta}}_d$.

$\hat{\mathcal{M}}^{(1)}$ of active indices. We further proceed repeating these steps to generate the index sets $\hat{\mathcal{M}}^{(1)}, \ldots, \hat{\mathcal{M}}^{(l)}$ of selected predictors in the subsequent stages till we reach our target model size, say d, i.e., till the smallest l for which $|\hat{\mathcal{M}}^{(l)}| = d$. Considering its similarity with the ISIS, we refer to this robust iterative variable screening procedure as LAD-ISIS or Huber-ISIS, corresponding to the L_1 and the Huber loss functions, which is presented schematically in Algorithm 2.

Note that in Step 3 of Algorithm 2, $L_j^{(2)}$ can be interpreted as the additional contribution of predictor X_j given the existence of predictors in $\hat{\mathcal{M}}^{(t-1)}$. In our implementation, we chose $k_1 = [2d/3]$, and thereafter at the t-th iteration, we took $k_2^{(t)} = d - |\hat{\mathcal{M}}^{(t-1)}|$. This ensures that the iterated versions of the Robust SIS method takes at least two iterations to terminate.

3.   NUMERICAL STUDIES

In this section, we consider some numerical experiments to illustrate the usefulness of the suggested methods in the linear regression model. We first analyze three simulated data sets for more illustrative purposes. Then, we analyze the performance of the proposed screening procedures in a real-world example related to the riboflavin production.

3.1.   Monte-Carlo simulation

In this subsection, three simulation examples including different models with various scenarios have been conducted to assess the finite sample performance of the proposed methods. The first example is allocated to our proposed non iterative independence screening procedures, while in the second example, the aim is to examine the influence of the percentage of outliers as well as the impact of sample size on the performance of the proposed methods. In the third example, the iterative ISIS procedure is applied to improve the proposed SIS methods in the situation where SIS fails. The performance of the robust alternatives are compared with the existing competitors, such as DC-SIS ([15]), SIRS ([20]), (I)SIS ([2]) and NIS ([4]).

To evaluate the performance of the proposed methods, three criteria are considered. The first criterion is the minimum model size (denoted by M), that is the smallest number of predictors needed to ensure that all the important predictors are selected. To get better inference, the 5%, 25%, 50%, 75% and 95% quantiles of M out of 500 replications were also presented. The second criterion (denoted by P_j) is the empirical probability that the important predictor X_j is selected, when the threshold $d = 2\lceil n/\log(n) \rceil$ is adopted. The last criterion is the proportion (denoted by S) that all important predictors are selected for a given model size in 500 replications, when the threshold $d = 2\lceil n/\log(n) \rceil$ is adopted.

Note that the first criterion does not need to specify a threshold. The more reliable screening procedure, the closer M value to the number of important predictors and also the closer S and P_j value to 1. In an ideal situation, both S and P_j are equal to one.

Example 3.1.   Consider the following linear model:

$$Y = c\beta^T X + \sigma\varepsilon,$$

where $\boldsymbol{\beta} = (0, 1, 0.8, 0.6, 0.4, 0.2, 0, \ldots, 0)^T$ takes grid values and $\sigma^2 = 6.83$. This model is adapted from [20]. We varied the constant c to control the signal-to-noise ratio. In this example, X_1, X_2, X_3, X_4, X_5 are important predictors and remaining ones $(X_6, \ldots, X_p)$ are not relevant.

We choose $c = 1$ and 2, with the corresponding $R^2 = 50\%$ and 80%. The vector of covariates $X = (X_1, \ldots, X_p)$ was generated from the multivariate normal distribution with mean 0 and the covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.8^{|i-j|}$ for $i \neq j$. We set the sample size $n = 200$ and the total number of predictors $p = 2000$ considering two error ε distributions, $N(0,1)$ and $t(1)$ (t-student), and then repeat each scenario 500 times. The results are given in Table 1.

ε	c	method	M 5%	25%	50%	75%	95%	P_j 1	2	3	4	5	S
N	1	DC-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		NIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		SIRS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		LAD-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		Huber-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
	2	DC-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		SIRS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		LAD-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		Huber-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
$t(1)$	1	DC-SIS	5	5	5	5	54.50	0.975	0.995	0.990	0.970	0.940	0.940
		SIS	5.00	105.50	722.50	1584.75	1944.05	0.450	0.435	0.405	0.310	0.260	0.195
		NIS	11.95	218.50	846.50	1602.00	1933.85	0.245	0.275	0.265	0.205	0.125	0.095
		SIRS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		LAD-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		Huber-SIS	5	5	5	5	6	1.000	1.000	1.000	1.000	1.000	1.000
	2	DC-SIS	5	5	5	5	5	0.995	0.995	0.995	0.995	0.990	0.990
		SIS	5	5	41.50	726.25	1872.90	0.650	0.685	0.640	0.620	0.555	0.485
		NIS	5	14.75	157	1017.50	1901.20	0.510	0.550	0.510	0.480	0.385	0.335
		SIRS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		LAD-SIS	5	5	5	5	5	1.000	1.000	1.000	1.000	1.000	1.000
		Huber-SIS	5	5	5	5	6	1.000	1.000	1.000	1.000	1.000	1.000

Table 1:   Five quantiles of minimum model size M, the empirical probability P_j and the proportion of S in Example 3.1.

From Table 1, it can be seen that when the random error has a normal distribution, all six screening methods perform quit well, with the proportion S equal to 1. However SIS and NIS break down for the heavy-tailed error distribution $t(1)$, while other methods continue to perform well. As expected, when the error distribution is $t(1)$, the proportions S for the LAD-SIS and Huber-SIS are equal to 1, which support the assertion that the LAD-SIS and Huber-SIS process the sure screening property. In addition, the SIRS also performs very well. This is because the SIRS is robust to the outliers since it only uses the ranks of the observed response values.

Example 3.2. In this example, we consider the impact of both the percentage of outliers and the sample size on the performance of the proposed methods. To this end, we introduce various outlier percentages to assess how the proposed robust methods perform under these different conditions. We also vary the sample size from 50 to 200 for the fixed dimension p, and generate the predictors in the same way as Example 1. The observations for the response variable are determined by

$$Y = \beta^T X + \varepsilon,$$

where $\boldsymbol{\beta} = (0, 1, 1, 1, 1, 1, 0, \dots, 0)^T$ and the noise ε is independent of the predictors, and is generated from standard normal distribution. We have investigated various types of contamination schemes, all of which have produced similar results. Hence, for the sake of brevity, we present the results for one particular contamination scheme where the responses are contaminated by replacing its value y by $y - 30$; This choice is arbitrary but simulates a situation of response contamination that arises quite frequently in practice. The contamination proportion is taken as such as 5%, 10%, and 20%, resulting in mild, moderate and heavy contaminations, respectively. For each simulation set-up, we have applied the proposed screening procedures to select the important predictors. The process is replicated 500 times to report some performance measures including the proportion S, when the model size $d = 2\lceil n/\log(n) \rceil$ is chosen; median of the number of true positives selected (TP); median of the minimum model size required to select all four important covariates (M). The simulation results are summarized in Table 2.

From Table 2, we can draw the following conclusions:

a) For each values of (n, p), as the fraction of contamination increases, the performance of all six methods in accurately detecting the true model diminishes.

b) When the sample size is very small ($n = 50$) and the data are contaminated with 20% outliers, values of zero for both S and T indicate that SIS, NIS and DC-SIS can only select important variables by chance. In contrast, SIRS, LAD-SIS and Huber-SIS exhibit better performance in such scenarios.

c) For larger samples, as the percentage of outlier data increases up to 20%, it becomes evident that SIS, NIS, and DC-SIS are less effective. Conversely, SIRS, LAD-SIS and Huber-SIS demonstrate highly satisfactory performance in identifying important predictors.

d) In all scenarios, SIRS, LAD-SIS and Huber-SIS consistently outperform the other methods. Specifically, SIRS and LAD-SIS, followed by Huber-SIS, exhibit the best performance.

e) For each fixed value of p and fixed percentage of contamination, as expected, an increase in sample size leads to an improvement in the performance of all six methods.

			5%			10%			20%		
p	n	method	M	TP	S	M	TP	S	M	TP	S
1000	50	DC-SIS	10	5	0.730	70	4	0.210	197	1	0.005
		SIS	231.5	2	0.050	552	1	0.005	734	0	0.000
		NIS	337	1	0.000	688	0	0.000	758.5	0	0.000
		SIRS	7	5	0.830	12	5	0.655	40.5	4	0.355
		LAD-SIS	6	5	0.885	10.5	5	0.700	30.5	4	0.400
		Huber-SIS	15	5	0.670	26.5	4	0.470	66	3.5	0.120
	100	DC-SIS	5	5	1.000	9	5	0.915	52.5	4	0.390
		SIS	61	4	0.410	193	3	0.170	396.5	2	0.040
		NIS	145	4	0.190	348.5	2	0.050	544.5	1	0.010
		SIRS	5	5	1.000	5	5	1.000	6	5	0.960
		LAD-SIS	5	5	1.000	5	5	0.995	6	5	0.940
		Huber-SIS	5	5	0.985	6	5	0.975	11.5	5	0.915
	200	DC-SIS	5	5	1.000	5	5	1.000	9	5	0.995
		SIS	7	5	0.945	40	5	0.645	145	4	0.310
		NIS	20	5	0.840	108	4	0.390	302	3	0.120
		SIRS	5	5	1.000	5	5	1.000	5	5	1.000
		LAD-SIS	5	5	1.000	5	5	1.000	5	5	1.000
		Huber-SIS	5	5	1.000	5	5	1.000	5	5	1.000
2000	50	DC-SIS	16	5	0.615	118	3	0.110	361.5	0	0.000
		SIS	343.5	2	0.040	1040.5	0.5	0.000	1361.5	0	0.000
		NIS	577.5	0	0.000	1225.5	0	0.000	1520.5	0	0.000
		SIRS	8	5	0.760	21	5	0.525	71	4	0.315
		LAD-SIS	7	5	0.795	16	5	0.620	58.25	4	0.325
		Huber-SIS	24	5	0.520	41.5	4	0.265	133.5	3	0.055
	100	DC-SIS	5	5	0.980	12	5	0.830	88.5	4	0.160
		SIS	90.5	4	0.320	465	3	0.060	885	1	0.020
		NIS	277.5	3	0.090	700	1	0.005	1195.5	1	0.000
		SIRS	5	5	1.000	5	5	0.995	6	5	0.910
		LAD-SIS	5	5	1.000	5	5	0.995	7	5	0.920
		Huber-SIS	5	5	1.000	5	5	0.975	15	5	0.805
	200	DC-SIS	5	5	1.000	5	5	1.000	11.5	5	0.975
		SIS	8	5	0.920	48	5	0.600	258	4	0.235
		NIS	30	5	0.715	160	4	0.285	499.5	3	0.090
		SIRS	5	5	1.000	5	5	1.000	5	5	1.000
		LAD-SIS	5	5	1.000	5	5	1.000	5	5	1.000
		Huber-SIS	5	5	1.000	5	5	1.000	5	5	0.995

Table 2:   Median of the minimum model size (M), median of the true positive (TP) and the proportion of S in Example 3.2, considering the percentage of contamination 5%, 10% and 20%.

Example 3.3.   In this example, we compare the empirical performance of the LAD-ISIS and Huber-ISIS with SIS, ISIS, DC-SIS and SIRS in a linear model with weak signal-to-noise ratio, which has the form of

$$(3.1) \qquad Y = 2.5X_1 + 2.5X_2 + 2.5X_3 - 7.5\sqrt{\rho}X_4 + \varepsilon.$$

This model was first considered by Zhong et al. [22]. We adopted exactly the same settings as in [22]. In this model, $\mathbf{X} = (X_1, \ldots, X_{1000})^T$, each X_k is generated from a normal distribution with zero mean and unit variance. All X_ks except X_4 are equally correlated with the Pearson correlation coefficient ρ, while X_4 has the Pearson correlation $\sqrt{\rho}$ with all other $p - 1$ predictors. We draw ε independently from $N(0,1)$ and $t(1)$ (t-student). We set the sample size $n = 200$ and $d = 2[n/\log n] = 74$ for the LAD-SIS and Huber-SIS procedures with LASSO penalty function. Then, we repeat the simulations 500 times and summarize the results in Table 3.

		t(1)					N(0,1)				
		P_j					P_j				
ρ	Error	1	2	3	4	S	1	2	3	4	S
0.2	SIS	1.000	1.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000
	ISIS	1.000	1.000	1.000	0.890	0.890	1.000	1.000	1.000	1.000	1.000
	SIRS	1.000	1.000	1.000	0.025	0.025	1.000	1.000	1.000	0.030	0.030
	DC-SIS	1.000	1.000	1.000	0.040	0.040	1.000	1.000	1.000	0.020	0.020
	LAD-SIS	1.000	1.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000
	LAD-ISIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Huber-SIS	1.000	0.990	0.995	0.000	0.000	0.985	0.990	0.995	0.000	0.000
	Huber-ISIS	1.000	1.000	1.000	0.995	0.995	1.000	1.000	1.000	0.995	0.995
0.5	SIS	1.000	0.995	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000
	ISIS	1.000	1.000	1.000	0.550	0.550	1.000	1.000	1.000	0.995	0.995
	SIRS	1.000	0.995	1.000	0.000	0.000	1.000	0.990	1.000	0.010	0.010
	DC-SIS	1.000	1.000	0.995	0.000	0.000	1.000	0.980	0.990	0.010	0.010
	LAD-SIS	0.985	0.965	0.980	0.000	0.000	0.995	0.990	0.995	0.000	0.000
	LAD-ISIS	1.000	1.000	1.000	0.990	0.990	1.000	1.000	1.000	1.000	1.000
	Huber-SIS	0.925	0.945	0.950	0.000	0.000	0.945	0.955	0.960	0.000	0.000
	Huber-ISIS	0.995	1.000	1.000	0.970	0.965	0.995	1.000	0.995	0.985	0.975
0.8	SIS	0.855	0.845	0.845	0.000	0.000	0.940	0.915	0.935	0.000	0.000
	ISIS	1.000	1.000	1.000	0.625	0.625	0.995	0.995	0.995	0.245	0.245
	SIRS	0.910	0.910	0.920	0.000	0.000	1.000	0.990	1.000	0.010	0.010
	DC-SIS	0.890	0.900	0.925	0.000	0.000	1.000	0.980	0.990	0.010	0.010
	LAD-SIS	0.790	0.770	0.805	0.000	0.000	0.845	0.845	0.860	0.000	0.000
	LAD-ISIS	0.995	0.990	0.995	0.930	0.910	0.995	1.000	1.000	0.935	0.930
	Huber-SIS	0.730	0.715	0.730	0.000	0.000	0.790	0.785	0.780	0.000	0.000
	Huber-ISIS	1.000	0.990	1.000	0.930	0.920	0.995	1.000	0.990	0.910	0.895

Table 3: The proportions of P_j and S given the model size $d = 2[n/\log n]$ in Example 3.3.

In this example, X_4 is jointly important but marginally independent to the response Y, so the marginal screening methods (SIS, DC-SIS, SIRS, LAD-SIS and Huber-SIS) can work badly and hardly detect important predictor X_4. According to Table 3, when the error distribution is normal and correlations among predictors are not strong, i.e. $\rho = 0.2, 0.5$, ISIS selects X_4 with high empirical probability; in other scenarios, ISIS dose not perform well to detect the marginal signal of X_4. In contrast, the proposed LAD-ISIS and Huber-ISIS are able to select X_4 effectively for all different cases in both error distributions. For example, when $\rho = 0.5$ and error distribution is $t(1)$, LAD-ISIS and Huber-ISIS can select all truly important predictors in the model with the empirical probability 99% and 96.5%, respectively, while the ISIS only has 55%. Similarly, when $\rho = 0.8$, the LAD-ISIS and Huber-ISIS can detect all truly important predictors with the empirical probability 91% and 92%, respectively, while the ISIS only has 62.5%. Thus, due to model misspecification, ISIS can not perform as well as in the Example 3.1. However, our LAD-ISIS and Huber-ISIS are still able to identify all important predictors with an overwhelming probability. These once again confirm the capabilities of LAD-ISIS and Huber-ISIS.

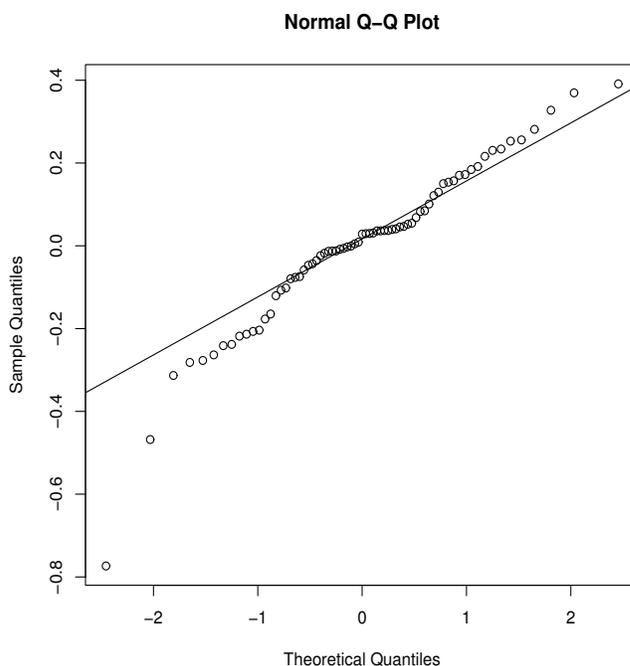### 3.2.   Application to riboflavin production data set

---

To support our assertions, we consider the data set about riboflavin (vitamin B2) production in Bacillus subtilis ([14], [17]), which can be found in R package "hdi". There is a single real valued response variable which is the logarithm of the riboflavin production rate and $p = 4088$ explanatory variables measuring the logarithm of the expression level of 4088 genes. There is one rather homogeneous data set from $n = 71$ samples that were hybridized repeatedly during a fed batch fermentation process where different engineered strains and strains grown under different fermentation conditions were analyzed.

Figure 1 shows the normal Q-Q plot based on the LASSO regression for the riboflavin production data set. Also, the bivariate boxplot for some of the selected genes of this data is depicted in Figure 2. The bivariate boxplot is a two-dimensional analogue of the boxplot for univariate data. This diagram is based on calculating robust measures of location, scale, and correlation; it consists essentially a pair of concentric ellipses, one of which (the hinge) includes 50% of the data and the other (called the fence) delineates potentially troublesome outliers. In addition, robust regression lines of both response on predictor and vice versa are shown, with their intersection showing the bivariate location estimator. The acute (large) angle between the regression lines will be small (large) for a large (small) absolute value of correlations. Figures 1 and 2 clearly reveal that the data contains some outliers. Now, if we are screening the genes via correlation with response in SIS or ISIS, these outliers will have an erroneous effects.

To analyze this data set, we first apply LAD-(I)SIS and Huber-(I)SIS robust procedures to shrink the dimension from $p = 4088$ down to $d = 2[n/\log(n)] = 32$ genes. After the variable screening, we fit the data by the penalized robust methods such as LAD-LASSO and Huber-LASSO models with the selected genes. On the other hand, LASSO is applied directly to $p = 4088$ genes without screening procudure. For the purpose of comparison, we also implement the (I)SIS with LASSO penalty to select most relevant genes. We compared their performance in terms of the adjusted R^2 which is defined as $R^2_{adj} = 1 - [(n-1)/(n-k-1)](1-R^2)$, where k is the number of predictors in the model, excluding the intercept, and R^2 is coefficient of determination defined as $R^2 = 1 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. The results are displayed in Table 4, in which the column labelled "Genes" stands for the number of the genes selected and the column of "R^2_{adj}" for the adjusted R^2.

Next, to measure the prediction accuracy of proposed estimators, the leave-one-out cross-validation (CV) criterion was used, which is defined by

$$(3.2) \qquad\qquad CV = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

Figure 1: Q-Q plot based on the LASSO regression for the riboflavin production
data set.

where $\hat{Y}_i$ is the predicted value of response variable where i-th observation left out of the estimation of the parameters $\boldsymbol{\beta}$. We also compute the correlation between the true and predicted response values $\hat{Y}_i$ obtained from leave-one-out cross-validation.

From the values of R^2_{adj}, we can infer that although LASSO model can explain 90.45% of the total variation of the logarithm of the riboflavin production rate, it does not perform as well as the screening competitors, beacuse the two-step screening procedures use much fewer genes while giving smaller CV error and higher correlation between the true and predicted values of response variable.

In addition, it can be seen that the iterative screening procedures (ISIS, LAD-ISIS and Huber-ISIS) outperform the corresponding noniterative screening methods (SIS, LAD-SIS and Huber-SIS) with the larger R^2_{adj} and $Corr(Y, \hat{Y})$, and smaller CV error, indicating that the iterative procedures identify some genes missed by the noniterative screening methods. Those important genes missed by the SIS (or,LAD-SIS and Huber-SIS) either may be closely marginally independent of the response or have relatively weaker marginal signals than some unimportant genes which are highly correlated with some strong active genes.

Moreover, the number of genes selected by SIS and ISIS are fewer than those of the robust alternatives, while because of the existence of outliers in the
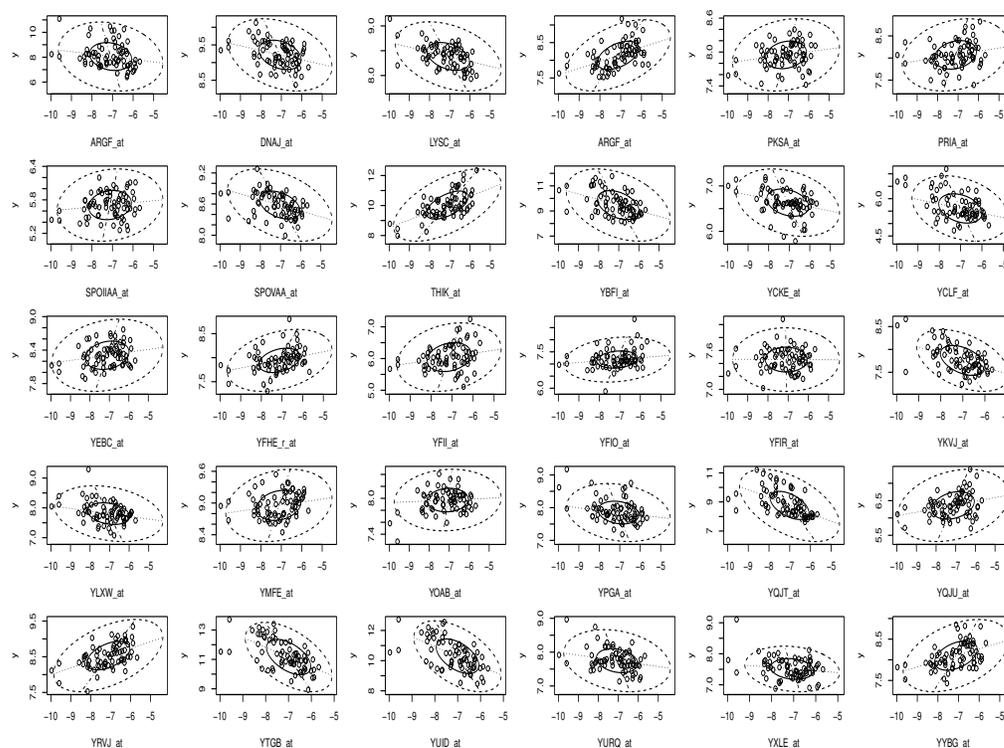
Figure 2: Bivariate boxplot of the riboflavin production data set for some of the effective genes.

data set, it can be seen that CV error and R^2_{adj} of the robust type methods are more acceptable than those of the non-robust type screening procedures.

In sum, we can clearly conclude from Table 4 that the LAD-ISIS performs the best with the smallest CV error 0.2309, the largest $Corr(Y, \hat{Y})$ value 0.852, the largest adjusted R^2 value 80.06% (compared to the other robust approaches), indicating that the LAD-ISIS is the best method for riboflavin production data set analysis.

Model	Genes	R^2_{adj}	CV	Corr
LASSO	42	90.45%	0.3973	0.730
SIS	11	76.90%	0.4461	0.687
ISIS	13	78.83%	0.3566	0.760
LAD-SIS	22	73.12%	0.3004	0.800
LAD-ISIS	21	80.06%	0.2309	0.852
Huber-SIS	20	73.99%	0.3362	0.779
Huber-ISIS	22	78.63%	0.3175	0.799

Table 4: Riboflavin production data analysis results.

CONCLUSIONS

In this paper, we have studied a robust variable screening methodology for ultrahigh dimensional data using robust loss functions. This technique uses the L_1 loss and the Huber loss which we refer to the LAD-SIS and Huber-SIS. We examined the finite sample performance of the proposed procedures via Monte Carlo studies, and illustrated the proposed methodology through the riboflavin production data set. In our numerical studies, both proposed robust methods (LAD-SIS and Huber-SIS) and SIRS perform equally well and behave better than SIS and DC-SIS in presence of outliers.

Similar to the SIS, the proposed technique may fail to identify some important predictors that are marginally independent of the response. Motivated by this, we introduced an iterative robust sure independence screening procedure. We examined its finite-sample performance via intensive simulations. The simulation results indicate that the iterative robust approach can significantly improve the LAD-SIS and Huber-SIS in the presence of truly important predictors that are marginally independent of the response and unimportant predictors that have relatively stronger marginal signals than some important predictors. Our empirical results indicate that the LAD-ISIS is the best approach among the selected competitors. We used only the LASSO penalty but other penalties such as adaptive LASSO ([21]), SCAD ([1]) and MCP ([19]) could also be applied.

ACKNOWLEDGMENTS

REFERENCES

[1] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 465, 1348–1360.

[2] FAN, J. and 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.

[3] FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality, *The Annals of Statistics*, **38**, 6, 3567–3604.

[4] FAN, J.; FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultrahigh-dimensional additive models, *Journal of the American Statistical Association*, **106**, 494, 544–557.

[5] FAN, J.; SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model, *The Journal of Machine Learning Research*, **10**, 2013–2038.

[6] FAN, J.; MA, Y.B. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models, *Journal of the American Statistical Association*, **109**, 507, 1270–1284.

[7] HE, X.; WANG, L. and HONG, H. (2013). Quantile-adaptive model-free variable screening for highdimensional heterogeneous data, *The Annals of Statistics*, **41**, 1, 342–369.

[8] *Huber, P.* (1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics*, **35**, 1, 73–101.

[9] HUBER, P.J. and RONCHETTI, E.M. (2009). *Robust Statistics*, Wiley: Hoboken, NJ, USA.

[10] KAZEMI, M. (2024). Support vector machine in ultrahigh dimensional feature space, *Journal of Statistical Computation and Simulation*, **94**, 3, 517–535.

[11] KAZEMI, M. (2020). Partial correlation screening for varying coefficient models, *Journal of Mathematical Modeling*, **8**, 4, 363–376.

[12] KAZEMI, M.; SHAHSAVANI, D. and ARASHI, M. (2018). Variable selection and structure identification for ultrahigh-dimensional partially linear additive models with application to cardiomyopathy microarray data, *Statistics, Optimization and Information Computing*, **6**, 3, 373–382.

[13] KAZEMI, M.; SHAHSAVANI, D. and ARASHI, M. (2019). A sure independence screening procedure for ultra-high dimensional partially linear additive models, *Journal of Applied Statistics*, **46**, 8, 1385–1403.

[14] LEE, JM.; ZHANG, S.; SAHA, S.; ANNA, SS.; JIANG, C. and PERKINS, J. (2001). RNA expression analysis using an antisense Bacillus subtilis genome array, *Journal of Bacteriology*, **183**, 24, 7371–7380.

[15] LI, R.Z.; ZHONG, W. and ZHU, L.P. (2012). Feature screening via distance correlation learning, *Journal of the American Statistical Association*, **107**, 499, 1129–1139.

[16] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **58**, 1, 267–288.

[17] ZAMBONI, N.; FISCHER, E.; MUFFLER, A.; WYSS, M.; HOHMANN, HP. and SAUER, U. (2005). Transient expression and flux changes during a shift from high to low riboflavin production in continuous cultures of Bacillus subtilis, *Biotechnology and Bioengineering*, **89**, 2, 219–232.

[18] ZHANG, CH. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression, *Annals of Statistics*, **36**, 4, 1567–1594.

[19] ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty, *Annals of statistics*, **38**, 2, 894–942.

[20] ZHU, L. P.; LI, L.; LI, R. and ZHU, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data, *Journal of the American Statistical Association*, **106**, 496, 1464–1475.

[21]   Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 476, 1418–1429.

[22]   Zhong, W. and Zhu, L. (2015). An iterative approach to distance correlation-based sure independence screening, *Journal of Statistical Computation and Simulation*, **85**, 11, 2331–2345.