A robust variable screening procedure for ultra-high dimensional data

Abhik Ghosh[1] and Magne Thoresen[2]

[1]Indian Statistical Institute, Kolkata, India. *abhik.ghosh@isical.ac.in*

[1]University of Oslo, Oslo, Norway. *magne.thoresen@medisin.uio.no*

Abstract

Variable selection in ultra-high dimensional regression problems has become an important issue. In such situations, penalized regression models may face computational problems and some pre-screening of the variables may be necessary. A number of procedures for such pre-screening has been developed; among them the sure independence screening (SIS) enjoys some popularity. However, SIS is vulnerable to outliers in the data, and in particular in small samples this may lead to faulty inference. In this paper, we develop a new robust screening procedure. We build on the density power divergence (DPD) estimation approach and introduce DPD-SIS and its extension iterative DPD-SIS. We illustrate the behavior of the methods through extensive simulation studies and show that they are superior to both the original SIS and other robust methods when there are outliers in the data. Finally, we illustrate its use in a study on regulation of lipid metabolism.

Keywords: Variable selection; NP dimensionality; Independence screening; Minimum density power divergence estimator; Influence Function; Gene selection.

1 Introduction

The introduction of the Omics technologies has led to a revolution in medical research, leading to an increased knowledge of the biological background of many diseases and paving the way for personalized therapies. A characteristic feature of data arising from the Omics technologies is its high dimensionality, which is a challenge for the statistical analysis. If we are to relate these high-dimensional features to some outcome variable in a regression set-up, we need to perform some sort of variable selection [1–4].

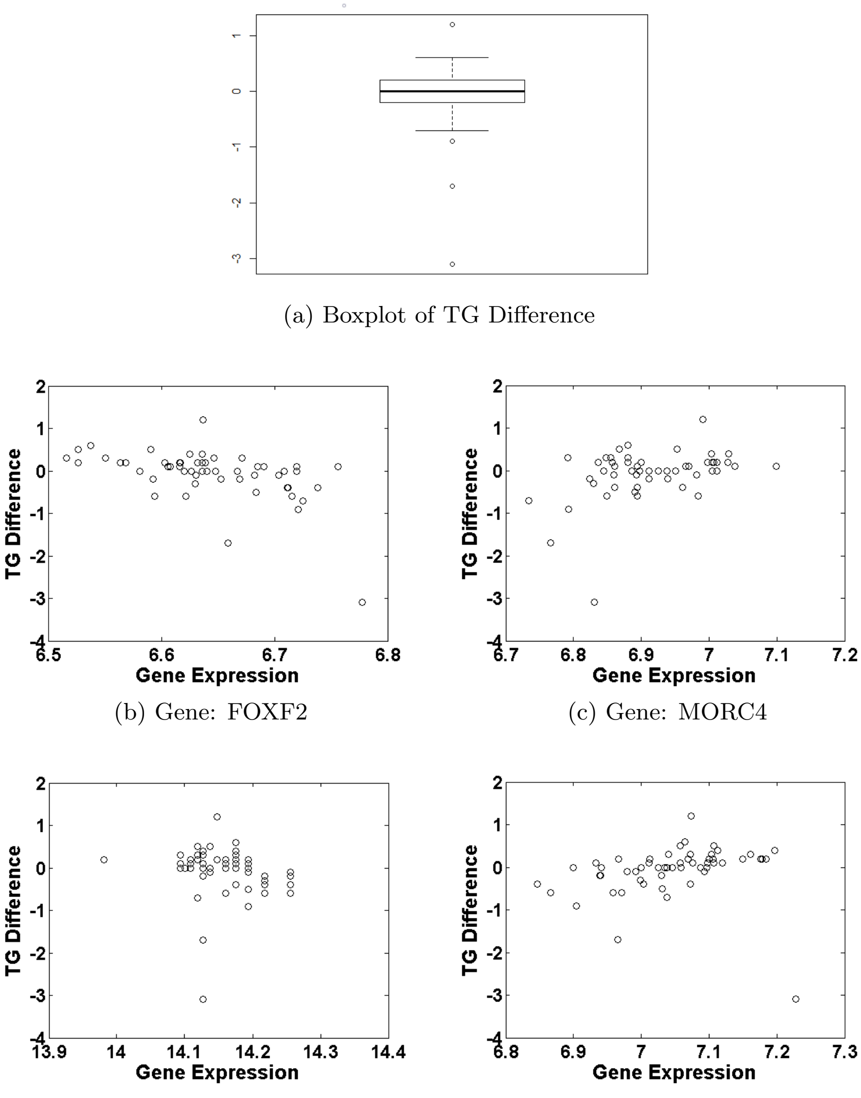
The most commonly used method for identifying important predictor variables in a high-dimensional regression model is to fit a penalized model. Consider the linear regression model with response variable Y and p explanatory variables (e.g. gene expressions) as covariates. Given the responses $y_1, \dots, y_n$ from n independent samples and the corresponding covariate values, say x_{ij}, $i = 1, \dots, n$, for the j-th covariate for $j = 1, 2, \dots, p$, this model can be written in matrix form as

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^T$ and ϵ_is are independent following $N(0, \sigma^2)$, for $i = 1, \dots, n$. The model parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ and σ^2 need to be estimated from the data. In the ultra-high dimensional case with $p \gg n$, e.g., Omics data, 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. Without loss of generality, we may assume that the true model parameter values are $(\boldsymbol{\beta}_0, \sigma_0^2)$ where $\boldsymbol{\beta}_0^T = (\beta_0, \boldsymbol{\beta}_{01}, \mathbf{0}_{p-s})$ with $\boldsymbol{\beta}_{01}$ being the non-sparse part of size $s \ll n$. Under sparsity assumptions, estimation of the 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 (corresponding to non-zero β_j) asymptotically with probability tending to one. There are plenty of such penalized regression procedures available in recent literature, starting from the LASSO method of Tibshirani [5] and its refinements [e.g., 6, 7] to more advanced procedures based on penalties like SCAD [8] or MCP [9], and many more, which work well in moderately high dimensions. However, a common problem with these methods in ultra-high 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 model size (e.g., number of genetic features) 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 Independent Screening (SIS) proposed by Fan and Lv [10] which has a simple interpretation and theoretical guarantees along with fast computation. Even with its simple structure (the SIS ranks the covariates based on their correlation with the response), the method yet enjoys the model selection oracle property under ultra-high dimensional set-ups where $\log(p) = O(n^l)$ for some $0 < l < 1$. An iterative extension, ISIS, is also proposed in [10] to tackle the issue of collinearity among covariates. The SIS and ISIS are routinely being applied in ultra-high dimensional applications and have also been extended to more complex models [see, e.g., 11–20, and the references therein]. 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 Omics data, which are often prone to at least a few outliers.

Motivating Example: In our motivating example (to be further described in Sec. 4) we are analyzing data from a small randomized study ($n = 54$) where the subjects received either fish oil, oxidized fish oil or sunflower oil for a period of seven weeks, and serum triglyceride (TG) levels were measured at baseline and after seven weeks. Our goal is to relate TG response (the difference between the two measurements of serum TG levels) to gene expressions measured at baseline. Gene expressions were measured using microarray technology, and we have available data from in total $p = 21236$ probes. Thus, an initial variable screening to reduce the model size is needed.

From the box-plot of the TG response (Fig 1a), it is clearly justified to fit a normal error distribution in the linear regression model (1) except for three outlying values. Now, if we are screening the genes via correlation with response in SIS or ISIS, these outliers will have an erroneous effects. Note that, it is not justified to remove these outliers at the start of the screening procedure, since they are outliers in the univariate distribution of response only but may or may not be outliers in the bivariate distribution of the response with any covariate. A few such examples are given in Fig 1b–1e; the outliers seem more legitimate in the bivariate space for the first case (Figure 1b), having no effect on the significance of the associated regression slope, but this is not true for the other cases. In Figure 1c, the outliers make the relation between the response Y and the corresponding gene look significant and hence it will come up towards the top of the selected gene list through usual SIS, although there is clearly no association between these variables after removing the outliers. The situation is more serious in the last two cases (Figures 1d-1e); there are actually strong associations between the response variable and both these genes which get masked by the presence of outliers and hence, these genes will not be among the top selected genes in SIS or even through ISIS (we have checked up to three steps of



(a) Boxplot of TG Difference

(b) Gene: FOXF2

(c) Gene: MORC4

(d) Gene: EEF1A1

(e) Gene: ZSCAN12

Figure 1: Box-plot of Response and Scatter plots of the response against different Genes

ISIS). Further, Figure 1d also presents a new outlier in the covariate space (in the gene expressions) and the same outlier may be present in several other gene expressions as well; such a scrutiny for each gene is clearly not feasible with larger sets of Omics data with potentially millions of features. Even if performed (with a huge time effort), this might leave us with very few cases left for performing any reasonable (joint) inference. A robust screening method which would ignore the effect of such outliers would be of great help in such ultra-high dimensional problems.

In this paper, we will develop a new robust screening procedure, an extension of the usual SIS, using the popular density power divergence (DPD) based estimation approach (briefly described in Section 1 of the Online Supplement). The DPD measure was originally proposed by Basu et al. [21] in the context of robust estimation in IID data. It has recently become very popular for robust inference in general and is widely applied on different types of data; see, e.g., [22]. The same approach has also been used for high-dimensional penalized linear regression and variable selection more recently [23, 24] and has been shown to be extremely useful under data contamination where it still gives consistent estimates and the oracle selection property still holds. However, the computation is still a concern in ultra-high dimensional set-ups and a robust version of SIS along the same line would be a useful approach to analyze such data more robustly, with robust screening at an initial stage followed by the robust DPD based penalized regression method to the reduced low or moderately high dimensional set of covariates. In the current work we fill the gap in the literature for the first (screening) stage by proposing a robust screening method based on the DPD for ultra-high dimensional linear regression models and illustrate its claimed robustness property theoretically as well as numerically. A robust version of ISIS along the same line using DPD will also be discussed to tackle the correlations among covariates. The suggested method will be applied to our motivating data example.

We also compare our method with the existing state-of-the-art robust screening procedures for ultra-high dimensional linear models [25–29, etc.] through extensive simulation studies. The major advantages of our proposed DPD based SIS and ISIS methods can be summarized as follows.

- Most (if not all) of the existing robust screening procedures are non-parametric in nature. It is well-known that, when a parametric model can be assumed, parametric inference is statistically more efficient than the non-parametric approach. In practice, it may often be the case that the assumed (parametric) linear regression model is at least approximately correct, and we may loose efficiency by using the existing non-parametric screening procedures. In this spirit, we consider for the first time the more efficient parametric approach to develop a robust sure screening procedure (in Section 2) via an appropriate robust parameter estimation technique. Consequently, our proposed screening procedure enjoys a significantly improved performance over the existing non-parametric robust versions of SIS (see Section 3).

- As a consequence of the parametric approach considered in our proposal, our proposed screening method also estimates the error variance (σ^2) from the data in each step (see Section 2.1), rather than just assuming it to be known or ignoring it as in most existing (non-parametric) approaches. In practice, the error variance is mostly unknown and has a significant impact in subsequent inference via the signal-to-noise ratio. Thus, one unique feature of our proposed robust variable selection procedure is data-based estimation of the error variance which, in turn, provides better control of the signal-to-noise ratio in each step of the screening, eventually leading to improved variable selection behavior.

2 Proposed Robust Variable Screening Procedures

2.1 The DPD-SIS

We now consider the linear regression model (1) with ultra-high dimensional covariates and the true sparse regression coefficient $\boldsymbol{\beta}_0$ as described in Section 1; let us denote the true sparse model as $\mathcal{M}_0 = \{1 \leq j \leq p : \beta_{0j} \neq 0\} = \{1, 2, \ldots, s\}$. Recall that the SIS method [10] can also be considered as ordering the absolute value of the slope in marginal regression models of the response with individual (standardized) covariates. Given values of the j-th covariate X_j for each $j = 1, \ldots, p$, we consider the j-th marginal model

$$y_i = \gamma_j + \beta_j x_{ij} + \epsilon_{ij}, \qquad i = 1, \ldots, n, \tag{2}$$

where the ϵ_{ij}s are IID for $i = 1, \ldots, n$, each having distribution $N(0, \sigma_j^2)$. We estimate the parameters $\boldsymbol{\theta}_j = (\gamma_j, \beta_j, \sigma_j)^T$ by usual MLE or OLS based methods, say, $(\widehat{\gamma}_j, \widehat{\beta}_j, \widehat{\sigma})$. Note that, when all covariates are standardized, ranking them in order of (absolute) correlation with the response is equivalent to ordering the estimated slopes $|\widehat{\beta}_j|$. However, this method is clearly non-robust since the estimates $\widehat{\beta}_j$s are so for MLE/OLS.

Here, we will propose to use the same approach as in the usual SIS, but with robust estimates for β_j in the marginal model using the DPD approach. Let us fix a $j \in \{1, 2, \ldots, p\}$ and an $\alpha > 0$. Since, given covariate values, $y_i \sim N(\gamma_j + \beta_j x_{ij}, \sigma_j^2)$, it belongs to the non-homogeneous set-up [30] discussed in Section 1 of the Online Supplement and hence, we can define the MDPDE of the parameters $\boldsymbol{\theta}_j$ via the objective function there. For the marginal model (2), one can easily simplify the MDPDE objective function as to have the form $H_{n,\alpha}(\boldsymbol{\theta}_j) = \frac{1}{n}\sum\limits_{i=1}^{n} l_\alpha\left(y_i, \gamma_j + \beta_j x_{ij}, \sigma\right)$, where

$$l_\alpha\left(y, \eta, \sigma\right) = \frac{1}{\sigma^\alpha (2\pi)^{\alpha/2}}\left(\frac{1}{\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha} e^{-\frac{\alpha(y-\eta)^2}{\sigma^2}}\right) + \frac{1}{\alpha}, \tag{3}$$

Then, we define the MDPDE of $\boldsymbol{\theta}_j$ for the marginal model (2) as

$$\widehat{\boldsymbol{\theta}}_j^M = (\widehat{\gamma}_j^{M\alpha}, \widehat{\beta}_j^{M\alpha}, \widehat{\sigma}_j^{M\alpha}) = \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} l_\alpha\left(y_i, \gamma_j + \beta_j x_{ij}, \sigma\right). \tag{4}$$

This is a much simpler optimization problem with only three parameters (compared to any penalized estimation problem), but we need to run it p times (once for each $j = 1, \ldots, p$). However, the overall computation time is still much lower than the penalized regression procedure with ultra-high dimensional p. Based on these MDPDEs for a given $\alpha > 0$, we can now choose the important variables in order of the values of $|\widehat{\beta}_j^{M\alpha}|$, which we refer to as the DPD-SIS procedure; for given index d we select the estimated model $\widehat{\mathcal{M}}_\alpha(d)$. Once $\widehat{\mathcal{M}}_\alpha(d)$ is obtained, one can then apply any suitable penalized regression method on the reduced set of covariates from $\widehat{\mathcal{M}}_\alpha(d)$ to obtain the final sparse estimate of the parameters in the original model (1). Therefore, our proposed robust screening procedure, the DPD-SIS, can be summarized in the following algorithm.

Algorithm 1: DPD-SIS(α)

1. **Input:** n-vector of responses $\boldsymbol{y}$; $n \times p$ matrix of (standardized) covariates $\boldsymbol{X}$; model size d.

2. For each $j = 1, \ldots, p$, compute the marginal MDPDE $\widehat{\beta}_j^{M\alpha}$ via (4).

   (This is an optimization in three parameters only and can be performed either by a standard optimization function in some software or by standard numerical techniques).

3. Sort $|\widehat{\beta}_j^{M\alpha}|$ in decreasing order for $j = 1, \ldots, p$. Set $r_k = j$, if $|\widehat{\beta}_j^{M\alpha}|$ has rank k, for $k = 1, \ldots, p$.

4. Construct the estimated model set $\widehat{\mathcal{M}}_\alpha(d) = \{r_1, \ldots, r_d\}$, with indices corresponding to the top d values of (absolute) marginal MDPDEs.

5. Run a robust penalized regression model (low or moderate dimensional) with the covariates

selected in $\widehat{\mathcal{M}}_\alpha(d)$ to obtain an estimated coefficient vector, say $\widehat{\boldsymbol{\beta}}_d = (\widehat{\beta}_{d0}, \widehat{\beta}_{dr_1}, \ldots, \widehat{\beta}_{dr_d})^T$. (We suggest to use the DPD based method of Ghosh and Majumdar [24] with the same α, which also gives an estimate $\widehat{\sigma}^2$ of the overall model error variance σ^2.)

6. **Output:** The final estimated model $\widehat{\mathcal{M}} = \left\{1 \leq k \leq d : \widehat{\beta}_{dr_k} \neq 0\right\}$ along with the parameter estimates $\widehat{\boldsymbol{\beta}}_d$ (and the estimate $\widehat{\sigma}^2$ of σ^2, if available).

Note that, at $\alpha = 0$ (in a limiting sense), the marginal MDPDE of regression coefficients coincides with the MLE and, hence, with the OLS as well. Thus the proposed DPD-SIS algorithm at $\alpha = 0$ becomes exactly the same as the usual SIS of Fan and Lv [10]. The extent of robustness of the DPD-SIS increases with increasing $\alpha > 0$.

For brevity in presentation, the theoretical properties of the DPD-SIS are presented in an appendix which includes the theoretical justifications of the claimed robustness of DPD-SIS through the influence function analyses in Section A.1 and a brief (non-technical) discussion of the sure-screening property and the oracle consistency of the final model and estimator obtained from Algorithm 1 (DPD-SIS) in Section A.2.

2.2 Iterative DPD-SIS

It has been noted that the usual SIS fails to pick up a variable having weak marginal correlation but significant joint relation with the response; on the other hand, it might pick up a variable having stronger marginal correlation but no joint relation with the response. Such cases occur mostly due to strong correlation between the important and unimportant predictor variables. To solve these issues, Fan and Lv [10] also proposed an iterative extension of SIS, namely the ISIS, which selects the truly important variables even under the above situations. Later, several extensions of the original ISIS have also been proposed [31]. As a robust extension of SIS, the DPD-SIS also suffers from the above issues, and fails to provide optimal results when covariates are strongly correlated (see Section 3) and an iterative extension in the line of ISIS is required.

The DPD-SIS can be easily extended through iterations to avoid the strong effects of correlation among predictors by considering, in subsequent iterations, the residuals from the fitted regression with predictors picked up in earlier stages. More explicitly, we start with DPD-SIS (Algorithm 1) in the first step to select k_1 variables with index set $\mathcal{A}_1 = \{i_1, \ldots, i_{k_1}\}$. Then, in the second step, we compute the residuals from the fitted regression model of the response $\boldsymbol{y}$ on the selected covariates in $\mathcal{A}_1$. The DPD-SIS screening is again applied taking these residuals as our new response to select another k_2 variables from the pool of variables with index set $\{1, 2, \ldots, p\} \setminus \mathcal{A}_1$; let us denote the index set of these k_2 selected variables as $\mathcal{A}_2$. We further proceed repeating these steps to generate the index sets $\mathcal{A}_3, \ldots, \mathcal{A}_l$ of selected variables in the subsequent stages till we reach our target model size, say d, i.e., till the smallest l for which $| \cup_{i=1}^{l} \mathcal{A}_i| = d$. Considering its similarity with the ISIS, we refer to this robust iterative variable screening procedure as Iterative DPD-SIS or, in short, DPD-ISIS, which is presented schematically in the following algorithm.

Algorithm 2: DPD-ISIS(α)

1. **Input:** n-vector of responses $\boldsymbol{y}$; $n \times p$ matrix of (standardized) covariates $\boldsymbol{X}$; model size d.

2. Set $i = 1$, $\boldsymbol{y}^{(1)} = \boldsymbol{y}$ and index set of available covariates as $\mathcal{W}_1 = \{1, \ldots, p\}$

3. **DPD-SIS with model size d':**

   (a) For each $j \in \mathcal{W}_1$, compute the marginal MDPDE $\widehat{\beta}_j^{M\alpha}$ via (4) with response $\boldsymbol{y}^{(i)}$ and covariate X_j.

   (b) Sort $|\widehat{\beta}_j^{M\alpha}|$ in decreasing order for $j \in \mathcal{W}_i$ and set $r_k = j$, if $|\widehat{\beta}_j^{M\alpha}|$ has rank k.

   (c) Construct the estimated model set $\widehat{\mathcal{M}}_\alpha^{(i)} = \{r_1, \ldots, r_{d'}\}$, with indices corresponding to the top d' values of (absolute) marginal MDPDEs.

4. Run any suitable (fast) robust penalized regression model (e.g., RLARS [32]) with the main response $\boldsymbol{y}$ and the covariates selected in $\cup_{k=1}^{i} \widehat{\mathcal{M}}_\alpha^{(k)}$ to get estimated coefficient vector $\widehat{\boldsymbol{\beta}}^{(i)}$. Let us assume that, at this i-th stage, the number of covariates selected in $\cup_{k=1}^{i} \widehat{\mathcal{M}}_\alpha^{(k)}$ is k_i and denote them as $\{j_1, \ldots, j_{k_i}\}$ so that the estimated coefficient vector has the form $\widehat{\boldsymbol{\beta}}^{(i)} = (\widehat{\beta}_0^{(i)}, \widehat{\beta}_{j_1}^{(i)}, \ldots, \widehat{\beta}_{j_{k_i}}^{(i)})^T$. Denote $\mathcal{A}_i = \left\{ j_a : \widehat{\beta}_{j_a}^{(i)} \neq 0, a = 1, \ldots, k_i \right\} \subset \mathcal{W}_1$.

5. If a specified stopping criterion (see discussion below) is satisfied, go to step 8. Otherwise go to Step 6.

6. Compute the residuals $\boldsymbol{r}^{(i)} = \boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_i}\widehat{\boldsymbol{\beta}}^{(i)}$.

7. Set $\boldsymbol{y}^{(i+1)} = \boldsymbol{r}^{(i)}$ and the index set of available covariates as $\mathcal{W}_i = \mathcal{W}_1 \setminus \mathcal{A}_i$.

   Change i to $i+1$ and go to Step 3.

8. Run a robust penalized regression model (low or moderate dimensional) with the covariates selected in $\mathcal{A}_i$ to get estimated coefficient vector, say $\widehat{\boldsymbol{\beta}}_d = (\widehat{\beta}_{d0}, \widehat{\beta}_{dr_1}, \ldots, \widehat{\beta}_{dr_d})^T$.

9. **Output:** The final estimated model $\widehat{\mathcal{M}} = \left\{1 \leq k \leq d : \widehat{\beta}_{dr_k} \neq 0\right\}$ along with the parameter estimates $\widehat{\boldsymbol{\beta}}_d$ (and the estimate $\widehat{\sigma}^2$ of σ^2, if available).

A few remarks related to the above algorithm is in order before further discussions. Firstly, the most straightforward stopping criterion (required in Step 5) could be $|\mathcal{A}_i| < d$. Step 8 assumes that the size of $\mathcal{A}_i$, at the end of the last iteration, is exactly d, which may not always be the case. When $|\mathcal{A}_i| > d$, we may work with all those selected variables or remove the extra variables having lower values of the marginal MDPDEs at the last stage of iteration. Alternatively the DPD-ISIS can also be terminated after a pre-fixed number of iterations (say $i = i_{\max}$) or when the size of the active set does not change from its value in the previous iteration (*i.e.*, $|\mathcal{A}_i| = |\mathcal{A}_{i-1}|$).

Secondly, in step 4 of DPD-ISIS, any fast robust penalized regression method, like RLARS, may be used without hampering the basic structure of DPD-ISIS. However, we strongly suggest to use the DPD based penalized regression method of Ghosh and Majumdar [24] with the same α in Step 8 to obtain the final model; as in DPD-SIS, it makes the whole procedure structurally consistent and also provides an estimate $\widehat{\sigma}^2$ of the overall error (unexplained) variance σ^2 in our final model.

Finally, it is worthwhile to note that our algorithm of DPD-ISIS is more similar to an extension of ISIS, namely Van-ISIS described in [31], rather than its original version proposed in [10]. The difference is mainly in Step 4 of the algorithm, where we consider all the covariates selected till the i-th iteration in the penalized joint regression model in the i-th stage, as in Van-ISIS; hence a variable which has been selected in an earlier stage could have been removed at the i-th stage due to insertion of new variables in the model. The original version of Fan and Lv [10] considered the penalized regression to be run with only the variables selected in that i-th iteration (and not the previously selected covariates) and hence a false positive selected at one iteration cannot be removed at any subsequent iteration. In this method the model size continue to increase whereas, in our approach, it may grow or shrink depending on the joint relationship of all the variables selected, reducing the number of false-positive covariates.

Although the DPD-SIS at $\alpha = 0$ is the same as the usual (non-robust) SIS, the DPD-ISIS($\alpha = 0$) is slightly different from its usual non-robust counterpart van-ISIS. Due to the use of RLARS within iterations, the DPD-ISIS($\alpha = 0$) is slightly more robust than van-ISIS and additionally DPD-ISIS at any α (including 0) estimates the error variance in a marginal regression setting whereas the usual van-ISIS uses marginal correlation based screening. However, the DPD-ISIS at $\alpha = 0$ is not yet acceptable as a robust method since the estimates of the marginal regression coefficients are still non-robust (MLE). As α increases, the DPD-ISIS becomes more robust.

3 Simulation Studies

3.1 Experimental Plans

We have performed extensive simulation studies to study and illustrate the performance of our proposed DPD based screening procedures. For each set-up we have simulated a random sample of size n from a linear regression model (1) of dimension $p \gg n$ where the $(p-1)$ covariates, except the intercept, are generated from a multivariate normal distribution having mean vector $\mathbf{0}$ and some specified covariance matrix, say Σ_x. After generating covariate values and error components for some fixed σ^2, the responses are computed based on specified true values of the regression coefficient $\boldsymbol{\beta} \in \mathbb{R}^p$; these true values are taken to be sparse with only the first $s = 5$ components being non-zero and the rest being zero. So, other than the intercept, only four covariates are significantly related to the response variable and the rest are noise covariates in all our simulation set-ups. Additionally, to study the robustness, a part (say, $100\epsilon\%$ for some ϵ specified later) of the samples are contaminated. All the parameters in the simulations are considered as follows.

- Two possible sample sizes are considered; $n = 50$ and $n = 100$. For each case, the model dimension is taken as $p = 5000$ to mimic the common ultra-high dimensional set-ups appearing in real life. Recall $s = 5$. These set-ups are clearly more extreme with regard to dimensionality compared to the set-ups studied in the SIS literature, but we believe they are closer to the true scenarios in practical Omics data analysis.

- The first $s = 5$ non-zero coefficient values are all taken as 1. Three different values of the error variance are considered; given by $\sigma = 0.2, 1, 2$, which yield three different signal-to-noise (SN) values. We refer to them, respectively, as *strong* (SN=5), *moderate* (SN=1) and *weak* (SN=0.5) signals.

- Different correlation structures are considered among the covariates via different Σ_x. In particular, we consider *independent* covariates with Σ_x being identity, and two types of correlated cases with the (i, j)-th elements of Σ_x being $\rho^{|i-j|}$, and $\rho I(i \neq j)$. We will refer to these two cases, respectively, as the case of *autoregressively (AR) correlated* and *strongly correlated* covariates. Several values of ρ have been studied but only the SIS performance results corresponding to $\rho = 0.5$ (in both cases) are reported in the paper for brevity.

- We have also studied different types of contamination schemes which all yield similar (in spirit) results. Hence, for 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 yields a (testing) situation of distant contamination in response which arise quite commonly in practice. The contamination proportion is taken as $\epsilon = 0.05, 0.1, 0.2$, resulting in mild, moderate and heavy contaminations, respectively.

For each simulation set-up, we have applied the proposed DPD-SIS procedure to select the important variables and different performance measures are computed in order to study the results. The whole process is replicated 300 times to report some stable summary of the performance measures. In particular, the performance measures considered are

IC	:	Indicator if all (4) important covariates are selected in a model of size $(n - 1)$.
TP	:	The number of true positives selected when a model of size $(n - 1)$ is chosen.
MMS	:	Minimum model size required to select all (4) important covariates.

Note that the average IC over all 300 replications yields the percentage of times the full model is selected in a model of size $(n-1)$. This is reported in the tables. However, for a deeper understanding, resulting values of TP and MMS from 300 replications are presented in terms of box-plots and histograms, respectively. Additionally, a run-time comparison is provided towards the end.

Along with studying our proposed DPD-SIS, the above performance measures are also used to compare our proposal with existing parametric and nonparametric competitive screening procedures as described below; the first is the usual SIS approach (non-robust) and the remaining four are robust non-parametric extensions of SIS available in the literature.

- **SIS**: The usual SIS of Fan and Lv [10] which use the Pearson correlation between the response and covariates for screening.
- **Rank-SIS**: A robust extension of SIS obtained by using non-parametric rank correlation in place of Pearson correlation in SIS [26].
- **GK-SIS**: A robust extension of SIS obtained by using a robust correlation measure proposed by Gnanadesikan and Kettenring [33] in place of Pearson correlation [34].
- **dCor-SIS**: A robust extension of SIS obtained by using a distance based correlation measure from Szekely et al. [35] in place of Pearson correlation [27, 29].
- **MCP-SIS**: A robust extension of SIS based on a robust measure of association, namely the median of component-wise products (MCPs) introduced by Mu and Xiong [28], which is used to rank the covariate importance.

Another non-parametric robust screening procedure is available in the literature, based on the bivariate winsorized (BW) correlation estimator of Khan et al. [32] in place of the usual correlation in SIS; we have not considered this BW based SIS, since Mu and Xiong [28] have already shown it to have similar performance as the MCP-SIS considered here.

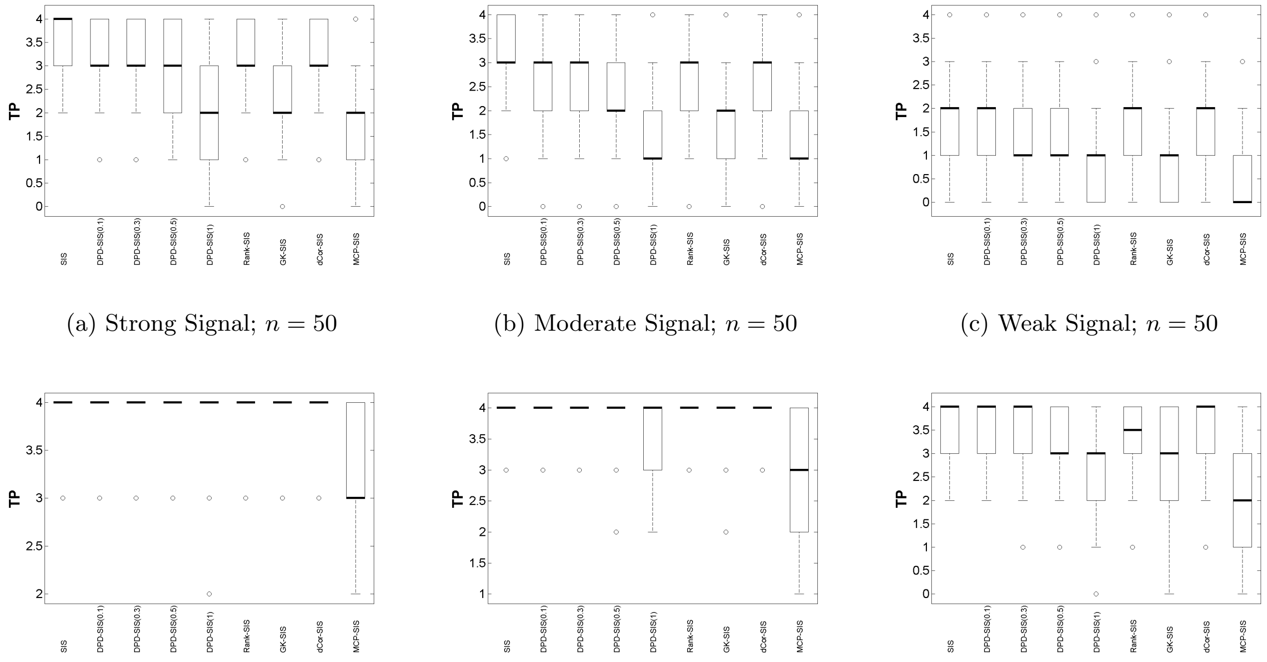
3.2 Performance of the DPD-SIS without contamination

Let us first illustrate the performance of the DPD-SIS under pure data containing no outliers. For all the simulation set-ups without contamination as described in the previous subsection, the percentage of times the full (correct) model is selected (average IC) by different SIS approaches with target model size $d = n - 1$ is reported in Table 1. One can immediately see that, as expected, all the SIS methods fails in case of strongly correlated covariates (except for large sample sizes and strong signal strength) and we need to use appropriate ISIS is such cases; for brevity, we will illustrate the performance of our proposed DPD-ISIS in the Online Supplement. For the other two types of covariates, the performance of our proposed DPD-SIS under pure data deteriorate slightly with increasing values of α (due to the loss in efficiency of MDPDE under pure data), but the DPD-SIS at $\alpha = 0.1, 0.3$ are pretty much comparable with the usual SIS is most cases and also significantly better compared to the existing non-parametric robust SIS approaches. For all the AR correlated cases as well as independent cases with moderate to strong signals and $n = 100$, the DPD-SIS provides the correct full model in over 90% of the replications which decreases as the signal strength becomes weaker or sample size becomes smaller. Among the two types of covariates, the performance is far better when some amount of correlation is present compared to the fully independent covariates when we have weaker signal strength and/or smaller sample sizes. This is somewhat surprising. The good behavior of the methods in the case with AR correlated data is caused by the way we simulated our data, with a cluster of important variables at the start of the X-sequence. When these important variables are randomly distributed in the sequence, the results are less good (as expected). This holds for all methods, but their relative behavior is again observed to be the same as in the present case. So, to keep our focus on comparison between the models, we have not presented the results for randomly distributed important covariates for brevity.

The methods can be compared further via the other performance measures, TP and MMS. With regard to TP, our simulations show that the median of the true positives selected by the usual SIS and our DPD-SIS at $\alpha \leq 0.5$ with a target model size of $d = n - 1$ are all equal four (the true active set size) for the cases of AR correlated covariates under pure data and hence, they are comparable in these cases (data not shown). For the independent covariate cases, the box-plots of the obtained true-positives are presented in Figure 2 where we can see that the results are again very similar for $n = 100$. For smaller sample size $n = 50$, however, the results are not that good; the usual SIS has median true-positive values of 4, 3 and 2, respectively, for strong, moderate and weak signals, whereas

Table 1: Percentage of times the full (correct) model is selected (average IC) by different SIS approaches with target model size $d = n - 1$ for pure data

Signal	Sample	Non-robust	Proposed DPD-SIS(α)				Existing Robust (Non-parametric) SIS			
Strength	size (n)	SIS	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 1$	Rank-SIS	GK-SIS	dCor-SIS	MCP-SIS
Independent Covariates										
Strong	50	55.0	48.3	41.7	27.7	8.0	41.7	12.7	43.7	1.7
	100	99.0	99.7	98.7	98.3	90.7	97.7	93.7	97.7	49.3
Moderate	50	25.3	18.7	15.7	10.0	1.0	14.7	5.3	15.3	0.3
	100	94.3	94.7	93.3	91.3	72.0	91.0	77.3	91.3	26.3
Weak	50	2.0	1.0	1.0	0.7	0.3	1.3	0.7	0.7	0.0
	100	58.7	57.3	53.7	44.7	23.3	50.0	26.7	50.7	5.3
AR Correlated Covariates with $\rho = 0.5$										
Strong	50	99.3	99.7	99.7	99.0	86.3	98.3	78.0	98.7	49.3
	100	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.7
Moderate	50	98.7	99.0	98.7	97.3	73.7	95.7	68.0	97.7	37.7
	100	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	95.4
Weak	50	86.7	85.3	82.3	75.7	38.3	79.0	42.0	80.7	20.3
	100	99.7	100.0	100.0	100.0	99.7	99.7	97.3	100.0	85.0
Strongly Correlated Covariates with $\rho = 0.5$										
Strong	50	14.7	1.7	1.0	0.7	0.0	5.7	0.0	9.0	0.0
	100	82.3	48.0	39.0	25.7	7.3	59.7	2.7	66.3	1.0
Moderate	50	5.0	1.0	0.7	0.7	0.0	2.3	0.0	2.7	0.0
	100	31.7	17.4	4.4	6.5	6.5	43.0	1.0	50.0	0.4
Weak	50	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0
	100	25.0	8.7	8.3	5.3	1.3	14.0	0.7	16.0	0.3



(a) Strong Signal; $n = 50$

(b) Moderate Signal; $n = 50$

(c) Weak Signal; $n = 50$

(d) Strong Signal; $n = 100$

(e) Moderate Signal; $n = 100$

(f) Weak Signal; $n = 100$

Figure 2: Box-Plots of the true-positives (TP) obtained by different SIS approaches with target model size $d = n - 1$ for independent covariates with pure data

the median true positives obtained by DPD-SIS at $\alpha = 0.1$ are 3, 3, and 2, respectively. The values of true-positives generally seem to decrease with increasing α in DPD-SIS under pure data scenarios but $\alpha = 0.3$ also gives very competitive results in most cases. As for the other (non-parametric) robust methods, the Rank-SIS and the dCor-SIS also perform reasonably well with regard to this measure.

We have further investigated MMS, the minimum target model size (d) required to select all the four true positives by different SIS approaches. Whenever SIS performs well, e.g., AR correlated covariates and/or strong signals, the MMS values are pretty low, often less than 10 with a median of about 4-6. For brevity, we only present the results (histogram) on MMS for two extreme cases with independent covariates in the Online Supplement, namely for strong signal with $n = 100$ (one of the best performing cases) and weak signal with $n = 50$ (one of the worst performing cases). The range (and median) of MMS differ widely in both cases but the general trend is the same (which is also the same in all other cases not reported here). The median MMS for DPD-SIS increases with increasing values of α and are generally higher than the usual SIS in pure data, but those obtained by DPD-SIS at $\alpha = 0.1, 0.3$ are very close to the values obtained by the usual SIS and often significantly better than the other existing non-parametric SIS approaches.

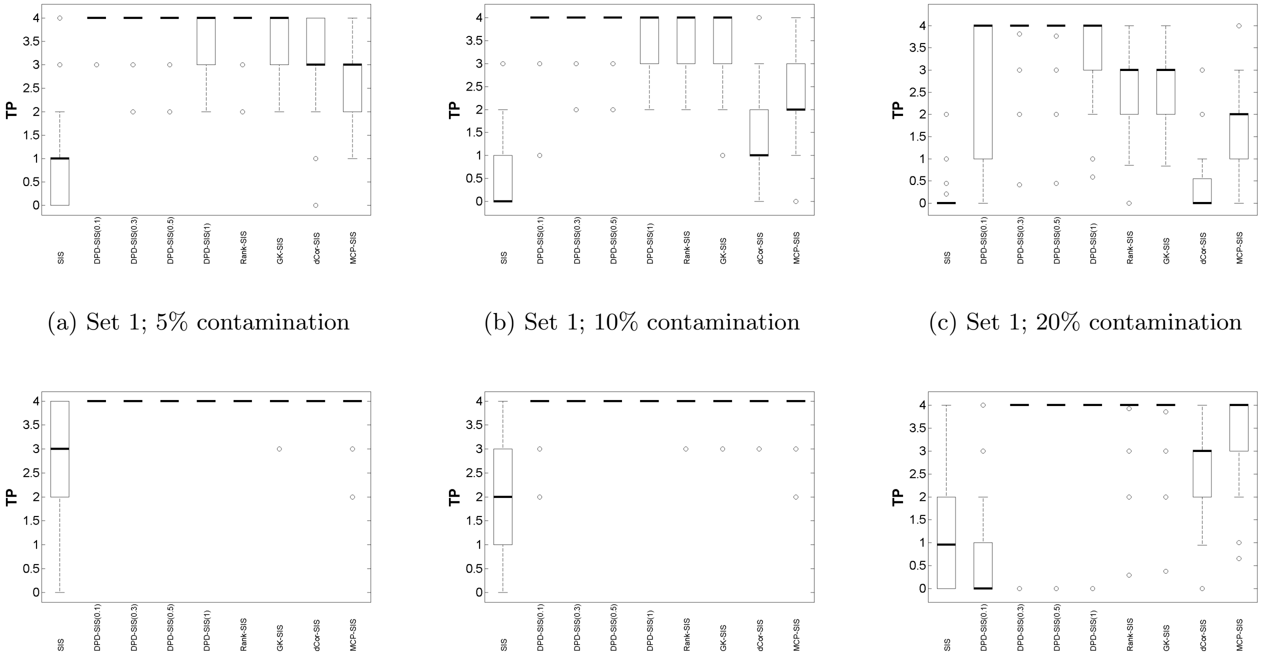
In summary, under pure data, usual SIS performs the best as expected, but there is only a slight loss in performance by the proposed DPD-SIS with smaller values of $\alpha > 0$. We will see next that, with this small price in case of pure data, we gain significant improvement over the usual SIS by using DPD-SIS under data contamination. Having a parametric nature, the proposed DPD-SIS naturally performs better than the existing non-parametric SIS approaches.

3.3 Performance of the DPD-SIS under data contamination

Let us now illustrate the performance of our DPD-SIS under data contamination and investigate the claimed improvements over the existing SIS and non-parametric robust SIS approaches. Due to the similarity in the patterns of results across all the cases considered (the only difference being in the magnitude of the performance measures, as in the pure data cases), we here only present the results for a representative case of $n = 100$ and moderate signal strength for both the independent and AR correlated covariates. For these cases, the percentage of times the full model is selected and the box-plots of the actual numbers of true-positives selected by different SIS approaches with target model size $d = n - 1$ are presented in Table 2 and Figure 3, respectively.

Table 2: Percentage of times the full (correct) model is selected (average IC) by different SIS approaches with target model size $d = n - 1$ for contaminated data with sample size $n = 100$, moderate signal strength and different contamination proportion (ϵ). Corresponding values for pure data are also given for comparison.

	Non-robust	Proposed DPD-SIS(α)				Existing Robust (Non-parametric) SIS			
$100\epsilon\%$	SIS	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 1$	Rank-SIS	GK-SIS	dCor-SIS	MCP-SIS
Independent Covariates									
0%	94.3	94.7	93.3	91.3	72.0	91.0	77.3	91.3	26.3
5%	0.3	94.3	91.0	89.0	73.7	77.3	63.0	32.7	18.0
10%	0.0	91.3	89.0	85.7	70.3	55.3	51.3	2.7	12.3
20%	0.0	59.3	82.5	77.5	64.6	25.1	23.1	0.0	3.7
AR Correlated Covariates with $\rho = 0.5$									
0%	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	95.4
5%	31.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0	94.3
10%	6.3	93.3	100.0	100.0	100.0	99.7	97.3	96.7	85.7
20%	2.0	1.7	99.7	99.7	99.7	92.8	85.8	21.7	60.3



(a) Set 1; 5% contamination

(b) Set 1; 10% contamination

(c) Set 1; 20% contamination

(d) Set 2; 5% contamination

(e) Set 2; 10% contamination

(f) Set 2; 20% contamination

Figure 3: Box-Plots of the true-positives (TP) obtained by different SIS approaches with target model size $d = n - 1$ for independent covariates (Set 1) and AR correlated covariates (Set 2) with $n = 100$ and moderate signal strength under different amount of contamination in data.

It can be noted that the usual SIS performs extremely poorly under any amount of contamination. Even at only 5% contamination, they select all the true positives in only about 30% of the cases with AR correlated data and almost never for the independent covariates, although these numbers were 99.7% and about 94-96%, respectively, under no contamination. As the contamination proportion increases, their performance becomes even worse and the same poor performance can also be seen in terms of the median number of true positives selected by these methods in Figure 3. Our proposed DPD-SIS with $\alpha > 0$ shows a much more stable performance under data contamination. In terms of percentages of full model selection, DPD-SIS with $\alpha \approx 0.3$ yields the best performance under heavy contamination (20%) and are quite competitive to the choice of $\alpha = 0.1$ also at milder contamination of 5%. A similar improved performance of our DPD-SIS is observed over the usual SIS in terms of selected true positives as well. More interestingly, our DPD-SIS with $\alpha \in [0.3, 0.5]$ often outperforms the existing non-parametric robust SIS approaches and the improvement becomes more significant at higher contamination level and for the cases of independent samples (or weaker signals). For the AR correlated covariates, the non-parametric Rank-SIS and GK-SIS performs quite good with a median true positive equal to four (the actual active set size) but have an overall worse performance (more outlying cases with lower number of true-positives selected) compared to DPD-SIS with moderate α values.

We have also studied the minimum target model size (MMS) required to select all four true positives under contamination which further illustrates the huge advantage of the proposed DPD-SIS over existing SIS approaches. The results for 20% contamination under the representative cases are shown in the Online Supplement. Note that the median values of the MMS required by the usual SIS are of the order 3950 and 2150, respectively, for the independent and AR correlated covariates. These become heavily improved by the existing non-parametric robust SIS approaches with Rank-SIS and GK-SIS yielding better performance compared to the other two. But, still for these two cases of independent or AR correlated covariates, they reach the minimum median MMS of 286 (by GK-SIS) and six (by Rank-SIS), respectively. Our proposed DPD-SIS with $\alpha \geq 0.3$ clearly outperforms all these existing methods yielding even lower values of MMS with a median of four (the minimum possible value) for the AR correlated case. For the independent covariates the improvement is even more significant with the best performance of DPD-SIS at $\alpha = 0.3$ which provides a median MMS of 20 only (in comparison with the minimum value of 286 obtained by existing approaches). The results for all other simulation experiments, not presented here for brevity, have indicated the similar advantages of our proposed DPD-SIS under different types and amounts of data contamination with the improvements being larger for the more vulnerable cases of heavy contamination or weaker signal strength or smaller sample sizes.

A comparison of our proposed DPD-SIS with the existing procedures in terms of their median runtime is provided in the Online Supplement.

3.4 On the Choice of robustness parameter α in DPD-SIS or DPD-ISIS

Our proposed DPD-SIS (and also the DPD-ISIS) depends on a tuning parameter α which is seen to control the trade-off between the asymptotic efficiency of the underlying MDPDE under pure data and its robustness under contamination (see Appendix A.1). In terms of variable screening as well, similar trade-offs are observed through our extensive empirical experiments. When there is no contamination in the data, the usual SIS (which is DPD-SIS at $\alpha = 0$) has the best performance, which deteriorates for DPD-SIS(α) as α increases although the loss is seen to be acceptable for smaller values of $\alpha \leq 0.3$. On the other hand, under contamination, the performance of the DPD-SIS becomes more and more stable with increasing values of α while the performance of the usual SIS breaks down completely even in presence of small amounts of contamination. Considering these trade-offs, it has been observed from our simulation studies that DPD-SIS with $\alpha = 0.3$ performs the best under data contamination in all the scenarios considered and it also clearly outperforms all the existing non-parametric methods. Based on these experiments, we recommend $\alpha \approx 0.3$ to be a good empirical suggestion to use in most practical applications of DPD-SIS (or DPD-ISIS).

It is worthwhile to note that, in usual practice with statistical procedure depending on a tuning parameter, an adaptive data-driven choice of the underlying tuning parameter is expected and seems to provide the best results in each cases. For the underlying MDPDE used in our DPD-SIS, such data-driven selection procedures for the robustness tuning parameter are available. In the context of linear regression, one such algorithm for selecting the optimal α is explored by Ghosh and Basu [36].

However, in the present case of DPD-SIS, we are using MDPDE for each marginal regression model and a data-based algorithm will often produce different values of α for each such marginal model, since the amount of contamination is often different across covariates. Working with different α values in one application of DPD-SIS is not useful and would break the coherence of the analysis – one should use the same α across all the steps of DPD-SIS in one application to get consistent inference. Additionally, data-driven selection of α would also increase the computation time, which is not an attractive feature in variable screening situations. We believe our empirical suggestion should work well in most applications.

4 Analysis of Triglyceride Data

In this section, we will apply our suggested variable screening method to our motivating example described in the Introduction, and show how this helps us in the variable selection process. Intake of marine omega-3 fatty acids may reduce the risk of cardiovascular disease (CVD), especially in high-risk individuals. Elevated serum triglyceride (TG) levels are strongly associated with increased risk of CVD, and the CVD risk reducing effect of marine omega-3 fatty acids is thought to be mainly mediated through reduction of serum TG levels. However, it is well known that there is large individual variation with regard to TG response in relation to intake of fatty acids, and an improved understanding of such individual variation would be beneficial. As described in the Introduction, we have data from 54 individuals who underwent an intervention with intake of capsules of either fish oil, oxidized fish oil or sunflower oil for a period of seven weeks. The study is presented in Ottestad et al. [37]. Fasting TG levels were measured at baseline and after seven weeks of intervention. In addition, we have gene expression measured in Peripheral blood mononuclear cells (PBMC). These are immune system cells and because they are circulating cells, they are exposed to nutrients, metabolites and peripheral tissues and may therefore reflect whole-body health. We are interested in relating TG change (seven weeks minus baseline) to gene expressions at baseline and our main goal is to identify genes that may be associated with TG response. Thus, we are primarily interested in variable selection.

As we have relatively few subjects, outliers might have a profound effect on the result, and hence, we are interested in performing a robust variable screening. Our analysis strategy is as follows: We will perform three iterations of the proposed robust DPD-ISIS (Algorithm 2) with RLARS in each iteration (Step 4), followed by a robust L_1-penalized regression, the DPD-LASSO method of Ghosh and Majumdar [24] to be consistent (in Step 8), to produce our list of selected genes. In each iteration of DPD-ISIS we select the $d = n/\log(n) \approx 13$ top variables, while we use penalization parameter $\lambda = \sqrt{\log(p)/n}$ in the final DPD-LASSO. A penalization parameter of this order has been shown to have certain optimality properties, see e.g. page 296, Hastie et al. [38]. We will do this for $\alpha = 0$ (which is not the usual ISIS as discussed in Section 2.2), 0.1, 0.3 and 0.5 and compare the lists of selected genes. We have also performed the usual correlation based Van-ISIS [31] described in Section 2.2 as our benchmark of comparison for the proposed procedures. When applying Van-ISIS, we observe that the estimated active set size (number of selected genes) does not change after three iterations, and we have used this as our stopping criterion. For the sake of comparison, we have also performed exactly three iterations of our proposed DPD-ISIS for each α. In the final penalized regression model, we also include treatment group and body mass index. However, this does not change the results significantly for any of the procedures. We present the results on the number of genes selected in the final model obtained by each procedure in Table 3; the detailed gene list and estimated regression coefficients in the final model are only presented, for brevity, in case of the usual non-robust ISIS (benchmark) and our recommended choice $\alpha = 0.3$ in Table 4.

Table 3: Numbers of Genes selected by different ISIS for the Triglyceride data

	Usual	DPD-ISIS with α			
	van-ISIS	0	0.1	0.3	0.5
Genes selected by ISIS	21	18	26	23	30
Genes selected in the final joint model	7	9	18	**21**	20

Two observations are worth discussing. First, three times as many genes are selected with the robust procedure (21 vs. 7) as with the non-robust ISIS. Second, there is very little overlap between the two gene sets (only three of the genes selected with $\alpha = 0.3$ are selected by van-ISIS). If we have a look at the number of genes selected as a function of α, we observe that the numbers are increasing with increasing α, more or less. This is somewhat counterintuitive, as the efficiency of the procedure is reduced with increasing α. However, as pointed out earlier, the stability (in terms of robustness) is increasing. We see this as a strong indication of problems with outliers in this rather small dataset,

Table 4: Detailed list of Genes selected by the usual ISIS and the proposed DPD-ISIS and associated estimated regression coefficients ($\widehat{\beta}_j$) in the final model for the Triglyceride data

Usual van-ISIS			Proposed DPD-SIS($\alpha = 0.3$)		
Genes	Prob id	$\widehat{\beta}_j$	Genes	Prob id	$\widehat{\beta}_j$
HNRNPK	ILMN3260017	-0.004	FOXF2	ILMN1674934	0.081
NA	ILMN1896699	0	HNRNPK	ILMN3260017	0.051
NA	ILMN1910805	-0.042	HYAL1	ILMN1739813	-0.399
NA	ILMN1712784	-0.063	UTY	ILMN3233091	0.309
NA	ILMN1679106	0	NA	ILMN1772136	-0.227
NA	ILMN1687707	-0.025	RPS27	ILMN1660498	-0.088
MORC4	ILMN1795463	0	SCARA3	ILMN1723358	-0.058
RPS16	ILMN1651850	0	SLITRK5	ILMN1789040	0.060
ZP3	ILMN1672378	0.006	ZP3	ILMN1672378	-0.427
FOXF2	ILMN1674934	-0.006	ZSCAN12	ILMN1786281	0.738
PKLR	ILMN1725172	0	SEZ6L2	ILMN2413780	-0.224
NA	ILMN1881212	0	FAM161A	ILMN3238106	-0.332
NA	ILMN3242572	0	EEF1A1	ILMN3201843	-0.091
TTC8	ILMN2401927	0	ABCD1	ILMN3237161	-0.264
XCR1	ILMN1764034	0	ALG1	ILMN1787954	-0.197
TBX1	ILMN2248112	0	AASS	ILMN1678323	-0.248
PPT2	ILMN1750664	0	EML1	ILMN1729455	-0.152
KLHL26	ILMN1805330	0	NA	ILMN1839740	-0.313
SCARA3	ILMN1723358	0	CDCA2	ILMN1660654	0.163
NA	ILMN1880704	-0.037	NA	ILMN1698246	-0.459
FGB	ILMN2114972	0	EPB41L4A	ILMN1791867	-0.096
			SLC7A11	ILMN1655229	0
			TMEM47	ILMN2129234	0

as illustrated in the Introduction. The fact that there is very little overlap between the two gene sets in Table 4 can also be seen as an illustration of this problem; with small sample size, outliers are dominating the analysis to a rather large extent. It is worth pointing out that the single gene with the strongest effect by DPD-ISIS (ZSCAN12, illustrated in Fig. 1e) is not even on the list of selected genes for van-ISIS.

If we think in direction of biological interpretation of the findings, we observe that a clear majority of the genes have an estimated coefficient with a negative sign, indicating a down-regulation of TG. It should also be commented that three of the identified probes do not map to a known gene, and hence, their function is unclear.

5 Discussions

In this paper, we have proposed a new robust variable screening procedure for ultra-high dimensional data using the marginal linear regression approach and the minimum density power divergence estimator for the regression parameter. This is extremely important in modern statistical analyses of large scale data from medical, biological and other applied sciences. We have also proposed an iterative version of our DPD based sure independence screening procedure in line with ISIS that is helpful for robust variable screening in the presence of correlated covariates. In this paper we have concentrated on linear relationships between the response and all the available covariates and hence, our proposed procedure is robust against data contamination (e.g., outliers, or leverage points) whenever the assumed linear regression model is approximately correct. The robustness of the proposed DPD-SIS is justified theoretically through use of influence functions and sensitivity analyses and also empirically through an extensive simulation study. It has been empirically shown, based on a first (limited) simulation study, that the proposed DPD-SIS at suitably chosen robustness tuning parameter α provides the best performance under data contamination and is superior compared to the usual SIS as well as several existing robust non-parametric screening procedures under most critical scenarios. We have applied our proposal for the robust analyses of data on triglyceride response to identify the important genes that may cause the variation in triglyceride response between different subjects.

We have primarily focused on the methodological and practical aspects of the proposed procedures and illustrated it through extensive simulation studies. It is worthwhile to emphasize here that the size of our empirical experiments is considered more realistic compared to scenarios arising in medical sciences, including the data example we aim to analyze using the proposed method. In particular, we have considered $p = 5000$ covariates with a sample size as small as $n = 50$ and different possible signal-to-noise ratios along with correlated covariates. Such an experimental set-up is more extreme compared to most (if not all) other existing work on SIS or its extensions for the linear regression model. Hence, our numerical illustrations of the finite-sample performance of all the screening methods provide more confidence about their performance in real life applications.

Through our extensive simulation studies, the proposed DPD based screening procedure is shown to perform well with suitably chosen tuning parameter value (α). Under pure data, as expected with any robust procedure, the performance of the proposed screening procedures become slightly inferior compared to the usual SIS; however, the loss is small for smaller values of the underlying tuning parameter α and additionally, they significantly outperform all the existing (non-parametric) robust procedures. Under contamination, our proposal performs the best, selecting the most stable set of variables even when the contamination is as heavy as 20%, whereas the usual SIS breaks down even in the presence of 5% contamination. A few existing non-parametric procedures like those based on ranks or G-K correlation [33] provide robust results under contaminations which are competitive to our proposal, but our proposed screening procedure even outperform them under more vulnerable cases like heavier contaminations, weak signal-to-noise ratios or smaller sample sizes. This makes our proposal even more advantageous over all the existing variable screening procedures.

We have implemented the proposed DPD-based procedure in `R` for all the simulations and real data analyses. The relevant codes are made available in a `GitHub` repository (`R` package) titled `dpdSIS`[1], which can be used by any practitioner for robust analyses of their experimental datasets from real-life studies.

With the promising and encouraging performance of the proposed DPD based robust variable screening procedure, this paper opens up several important directions of future research. Besides developing the technical details of the theory of DPD-SIS and DPD-ISIS, it would be practically important to extend them to more general parametric regression settings, like generalized linear models [39]. For the practical applications, in order to ensure stability of the final solution, it would make sense to apply some relevant additional procedure, e.g. stability selection [40, 41] on top of the final penalized regression. We hope to pursue these important research extensions in our future works.

A Theoretical properties of the proposed DPD-SIS

A.1 On Robustness of the DPD-SIS

The proposed DPD-SIS (and also the DPD-ISIS) depends crucially on the MDPDEs $\widehat{\boldsymbol{\beta}}_j^M$ from each marginal regression model and hence, the same is true for their robustness. If these marginal estimates are robust with respect to any outliers or noise contamination in either the response or the

[1] https://github.com/abhianik/dpdSIS

corresponding covariate, their ordering (in absolute value) is also expected to be robust under data contamination, leading to correct and stable variable screening performance of the DPD-SIS.

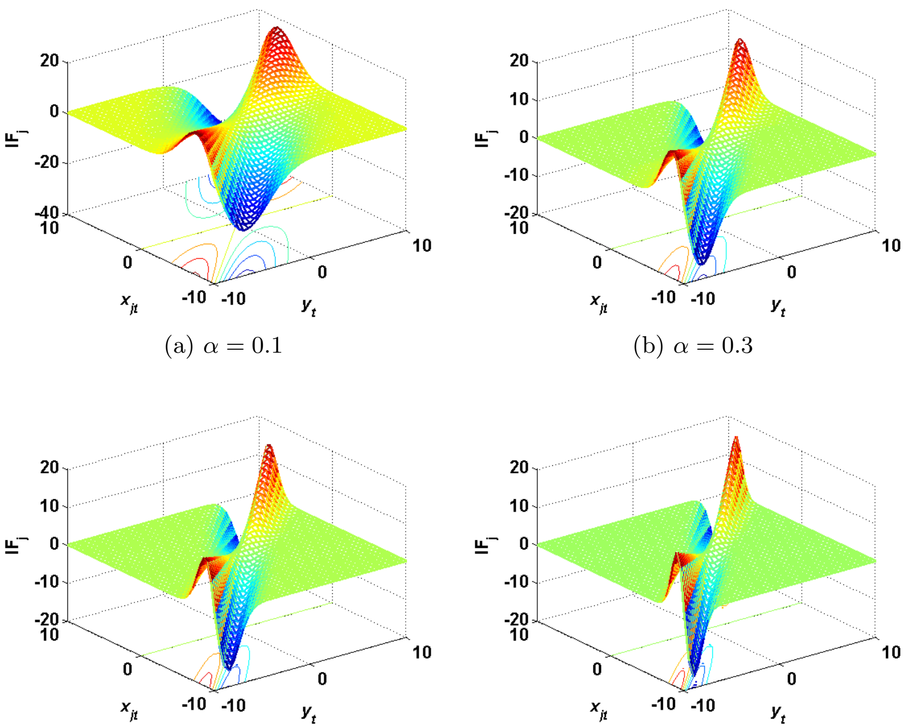
The robustness of the MDPDE under any parametric set-up, including the linear regression model, is well-studied in the literature [22, 30]; it is known to crucially depend on the choice of α. This can be examined theoretically through the concept of influence function (IF) and gross-error sensitivity. The IF measures the asymptotic (standardized) bias of the estimator caused by an infinitesimal contamination through the degenerate distribution at a distant outlier point. For our j-th marginal regression model (2), the IF of the MDPDE estimator $\widehat{\beta}_j^{M\alpha}$ of the regression coefficients β_j with respect to the contamination point, say y_t, in the response for a given covariate value, say x_{jt}, can be obtained from the general results of Ghosh and Basu [30]. When the assumed linear model is true with parameter values $(\gamma_j^{(0)}, \beta_j^{(0)}, \sigma_j^{(0)})$, it has a simplified form given by

$$\mathcal{IF}_j^{(\alpha)}(y_t|x_{jt}) \quad = \quad (1+\alpha)^{3/2} \frac{(x_{jt} - E(X_j))}{Var(X_j)} \left(y_t - \gamma_j^{(0)} - \beta_j^{(0)} x_{jt}\right) e^{-\frac{\alpha\left(y_t - \gamma_j^{(0)} - \beta_j^{(0)} x_{jt}\right)^2}{2(\sigma_j^{(0)})^2}}.$$

To study its nature, in Figure 4 we plot $\mathcal{IF}_j^{(\alpha)}(y_t|x_{jt})$ over the contamination point y_t for different $\alpha > 0$, by taking $(\gamma_j^{(0)}, \beta_j^{(0)}, \sigma_j^{(0)}) = (0, 1, 1)$. We assume $E(X_j) = 0$ and $Var(X_j) = 1$. For the case $\alpha = 0$, the IF simplifies to a linear function of y_t, and hence, it is unbounded with respect to the contamination point y_t, for all possible covariate values x_{jt} which justifies the well-known non-robust nature of the MLE (MDPDE at $\alpha = 0$). However, at any $\alpha > 0$, the IF of the corresponding marginal estimator $\widehat{\beta}_j^{M\alpha}$ is bounded in y_t for all values of x_t, indicating the claimed robust nature.

Further, we also study their self-standardized gross-error sensitivity, which is the maximum of the L_2-norm of the IFs, standardized by the variance of the MDPDE, over all possible contamination points. It is seen from Figure 5a that these sensitivity measures decrease with increasing $\alpha > 0$ for any given value of $\delta = (x_{jt} - E(X_j))^2/Var(X_j)$. Thus the extent of robustness of the MDPDE increases with increasing $\alpha > 0$ and hence, the same is also expected for the DPD-SIS with increasing $\alpha > 0$.

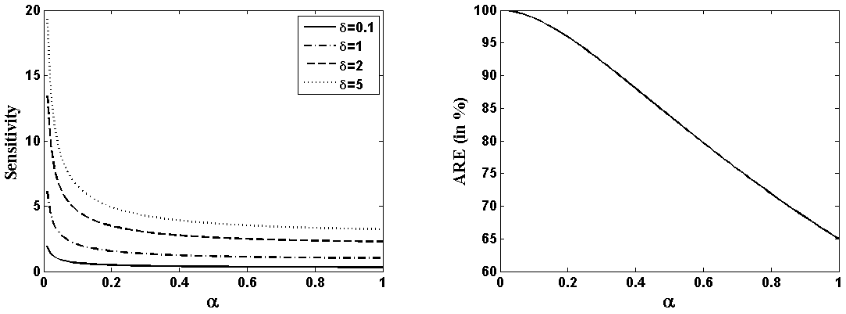
Remark A.1 *Although we have seen that the robustness increases as $\alpha > 0$ increases, we cannot use the larger values of α in every cases. This is because, when there is no outlier (pure data) with respect to the assumed (marginal) regression model, the asymptotic variance of the MDPDE $\widehat{\beta}_j^{M\alpha}$*



(a) $\alpha = 0.1$ (b) $\alpha = 0.3$

(c) $\alpha = 0.5$ (d) $\alpha = 1$

Figure 4: Influence functions ($\mathcal{IF}_j$) of the marginal MDPDEs for different values of $\alpha > 0$



(a) Sensitivity for different values of δ (b) Asymptotic relative efficiency(ARE)

Figure 5: The trade-off between robustness and efficiency of the marginal MDPDEs over α

is known to increase with increasing α values and hence their asymptotic relative efficiency (ARE) compared to the (most efficient but non-robust) OLS/MLE decreases as α increases (see Figure 5b). Therefore, in summary, the tuning parameter α provides a trade-off between efficiency of the MDPDE under pure data and its robustness under data contamination (which is the case with most common *robust inference approaches) so that we need to choose α carefully for any practical application, ideally depending on the amount of expected contamination in the data as discussed in Section 3.4.*

A.2 On Asymptotics of the DPD-SIS

It can be shown that the proposed DPD-SIS method asymptotically satisfies the sure screening property, and we have indeed done so in [39]. In particular, under appropriate assumptions, we have proved the following asymptotic properties of the DPD-SIS.

(R1) At the population level, for any $j = 1, \ldots, p$, the j-th marginal MDPDE functional corresponding to the j-th covariate X_j is zero if and only if X_j is uncorrelated with the response variable.

(R2) For some optimally chosen convergent sequence $R_n \to 0$, we have

$$P\left(\mathcal{M}_0 \subset \widehat{\mathcal{M}}\right) \geq 1 - sR_n, \qquad \text{and} \quad P\left(|\widehat{\mathcal{M}}| \leq O(n^{\kappa}\lambda)\right) \geq 1 - pR_n,$$

for some constants $\kappa > 0$ and $\lambda > 0$. The first result provides the sure screening property of the DPD-SIS, whereas the second one proves its control of the false-discovery rate.

(R3) Combining above results in (R2), for any $\alpha \geq 0$, we have $P\left(\widehat{\mathcal{M}} = \mathcal{M}_0\right) = 1 - o(1)$, i.e., DPD-SIS with any given $\alpha \geq 0$ satisfies the model selection consistency.

Finally, based on the above sure screening property of the DPD-SIS, and the consistency of the DPD-based penalized regression estimators from Ghosh and Majumdar [24], we can easily argue the consistency of the final estimator obtained by the proposed DPD-SIS Algorithm 1. A more general result in this regard is presented in the following theorem, justifying the use of the DPD-SIS Algorithm 1 for ultra-high dimensional linear regression problems.

Theorem A.1 *Assume the conditions required for Results (R2) and the conditions of Theorem 4 of [24] hold true for a given $\alpha \geq 0$. Let $\widehat{\mathcal{M}}$ be the final selected model by the DPD-SIS Algorithm 1 based on the final parameter estimate $\widehat{\boldsymbol{\beta}}_d = (\widehat{\beta}_{d0}, \widehat{\beta}_{dr_1}, \ldots, \widehat{\beta}_{dr_d})^T$ and $\widehat{\sigma}^2$ of σ^2. Then, we have the following results with probability tending to one:*

$$\widehat{\mathcal{M}} = \mathcal{M}, \quad ||\widehat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_{01}|| = O(\sqrt{s/n}), \qquad \textit{and} \quad |\widehat{\sigma} - \sigma_0| = O(n^{-1/2}).$$

The above result is exactly similar to the asymptotic properties of the usual SIS (Theorem 5, [10]) and hence illustrates that asymptotically the proposed DPD-SIS has the same optimal variable selection and parameter consistency properties under appropriate assumptions.

DATA ACCESSIBILITY

The data used in this analysis are from a third party and are not available for distribution. The relevant codes are made available in a `GitHub` repository '`DPD-SIS`' (https://github.com/abhianik/DPD-SIS).

ACKNOWLEDGMENTS

The authors wish to thank Prof. Stine Ulven, Department of Nutrition, University of Oslo, for providing the Triglyceride dataset and guiding us in biological interpretation of the results. A major part of this research work has been done while the first author (AG) was visiting University of Oslo, Norway. Research of AG is also partially supported by the INSPIRE Faculty Research Grant from the Department of Science and Technology, Government of India.

References

[1] Segal MR, Dahlquist KD, Conklin BR. Regression approaches for microarray data analysis. J Comput Bio. 2003; 10(6):961-980.

[2] Lusa L, Korn EL, McShane LM. A class comparison method with filtering-enhanced variable selection for high-dimensional data sets. Stat Med. 2008; 27(28):5834–5849.

[3] Fu L, Wang YG. Variable selection in rank regression for analyzing longitudinal data. Stat Methods Med Res. 2018; 27(8):2447-2458.

[4] Jung Y, Zhang H, Hu J. Transformed low-rank ANOVA models for high-dimensional variable selection. Stat Methods Med Res. 2019; 28(4):1230-1246.

[5] Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc B. 199; 58:267–288.

[6] Zhang CH, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Ann Stat. 2008; 36(4):1567–1594.

[7] Zou H. The Adaptive Lasso and Its Oracle Properties. J Amer Statist Assoc. 2006; 101:1418-1429.

[8] Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. J Amer Stat Assoc. 2001; 96:1348-1360.

[9] Zhang CH. Nearly Unbiased Variable Selection under Minimax Concave Penalty. Ann Statist. 2010; 38:894–942.

[10] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J Royal Stat Soc B . 2008; 70(5):849–911.

[11] Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Stat. 2010; 38(6):3567-3604.

[12] Zhao SD, Li, Y. . Principled sure independence screening for Cox models with ultra-high-dimensional covariates. J Mult Anal. 2012; 105(1):397-411.

[13] Barut E, Fan J, Verhasselt A. Conditional sure independence screening. J Amer Stat Assoc. 2016; 111(515):1266-1277.

[14] Zhou Y, Liu J, Hao Z, Zhu L. Model-free conditional feature screening with exposure variables. Stat Its Interface. 2019;12(2):239-51.

[15] He Y, Zhang L, Ji J, Zhang X. Robust feature screening for elliptical copula regression model. J Mult Anal. 2019; 173:568-82.

[16] Wang Y, Van Aelst S. Robust variable screening for regression using factor profiling. Stat Anal Data Mining. 2019; 12(2):70-87.

[17] Edelmann D, Hummel M, Hielscher T, Saadati M, Benner A. Marginal variable screening for survival endpoints. Biometrical J. 2020; 62(3):610-26.

[18] Lai P, Chen Y, Zhang J, Dai B, Zhang Q. Robust model-free feature screening for ultrahigh dimensional surrogate data. J Stat Comput Simul. 2020; 90(3):550-69.

[19] Song F, Chen Y, Lai P. Conditional distance correlation screening for sparse ultrahigh-dimensional models. Appl Math Model. 2020; 1;81:232-52.

[20] Wang G, Guan G. Weighted Mean Squared Deviation Feature Screening for Binary Features. Entropy. 2020; 22(3):335.

[21] Basu A, Harris IR, Hjort NL, Jones MC. Robust and efficient estimation by minimising a density power divergence. Biometrika. 1998; 85: 549–559.

[22] Basu A, Shioya H, Park C. Statistical Inference: The Minimum Distance Approach. 2011; Chapman & Hall, Boca de Raton.

[23] Zang Y, Zhao Q, Zhang Q, et al. Inferring gene regulatory relationships with a high-dimensional robust approach. Genet Epidemiol. 2017; 41(5):437–454.

[24] Ghosh A, Majumdar S. Ultrahigh-dimensional Robust and Efficient Sparse Regression using Non-Concave Penalized Density Power Divergence. IEEE Trans Info Theory, 2020; 66(12): 7812–7827.

[25] Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. J Comput Graphical Stat. 2009; 18(3):533-550.

[26] Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. Ann Stat. 2012; 40(3):1846-1877.

[27] Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Amer Statist Assoc. 2012; 107(499):1129-1139.

[28] Mu W, Xiong S. Some notes on robust sure independence screening. J App Stat. 2014; 41(10):2092–2102.

[29] Wang T, Zheng L, Li Z, Liu H. A robust variable screening method for high-dimensional data. J App Stat. 2017; 44(10):1839-1855.

[30] Ghosh A, Basu A. Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. Electron J Stat. 2013; 7:2420–2456.

[31] Saldana DF, Feng Y. SIS: An R package for sure independence screening in ultrahigh-dimensional statistical models. J Stat Software. 2018; 83(2):1–25.

[32] Khan JA, Van Aelst S, Zamar RH. Robust linear model selection based on least angle regression. J Amer Statist Assoc. 2007; 102:1289–1299.

[33] Gnanadesikan R, Kettenring JR. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics. 1972; 81-124.

[34] Gather U, Guddat C. Comment on "Sure Independence Screening for Ultrahigh Dimensional Feature Space" by Fan, JQ and Lv, J. J Royal Stat Soc B. 2008; 70:893-895.

[35] Szekely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007; 35:2769–2794.

[36] Ghosh A, Basu A. Robust Estimation for Non-Homogeneous Data and the Selection of the Optimal Tuning Parameter: The DPD Approach *J Appl Stat.* 2015; 42(9):2056–2072.

[37] Ottestad I, Vogt G, Retterstl K, Myhrstad MC, et al. Oxidised fish oil does not influence established markers of oxidative stress in healthy human subjects: a randomised controlled trial. British J Nutr. 2012; 108(2):315–26.

[38] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. 2015; CRC press.

[39] Ghosh A, Ponzi E, Sandanger T, Thoresen M. Robust Sure Independence Screening for Non-polynomial dimensional Generalized Linear Models. ArXiv preprint. 2021; arXiv:2005.12068v2.

[40] Meinshausen N, Buhlmann P. Stability selection. J Royal Stat Soc B. 2010; 72(4): 417–473.

[41] Shah RD, Samworth RJ. Variable selection with error control: another look at stability selection. J Royal Stat Soc B. 2013; 75(1):55–80.