Variable Screening
- via Distance Correlation and Complete Least Squares

Abstract

Variable selection is the process of filtering out irrelevant variables and selecting the relevant ones. However, when the size of predictors gets much larger than the sample size, which can be defined to be an ultrahigh dimensional setting, variable selection techniques will give results that are noisy and unreliable. This thesis considers a new screening technique under this circumstance. We developed a screening procedure based on distance correlation and complete least squares (DC-CLS). DC-CLS is an extension of the sure independence screening procedure based on Pearson correlation proposed by Fan and Lv (2008) [1] and based on distance correlation (DC-SIS) proposed by Li, Zhong and Zhu (2012) [2]. It differs in that DC-CLS takes into consideration the correlations between the response and predictors as well as that within the predictors. We conducted simulation studies to assess the performance of this proposed method. While our procedure showed promise in a proof of concept example, our simulation studies revealed it is not competitive for general use.

1 Introduction

Variable selection, also known as feature selection, is a procedure that selects relevant variables (features). With large data sets becoming more common, variable selection becomes an important process to effectively select a subset of relevant variables to produce more accurate estimations. These techniques include, but are not limited to, LASSO [3] (Tibshirani, 1996), Elastic Net [4] (Zou and Hastie, 2005), the adaptive LASSO (Zou, 2006) [5], and the LARS algorithm (Tibshirani, Johnston, Hastie, Efron, 2004) [6]. These methods perform well with high dimensional data, under the condition that the number of predictors is on the order of or larger than the sample size [2].

With the advancement of technology, the data grows to possibly have an ultrahigh dimension when the number of predictors in the data is much higher than the sample size. Under an ultrahigh dimensional setting, the variable selection techniques mentioned above may not perform very well and may give noisy and inaccurate results due to the challenges including singularity that is caused by more columns than rows in the design matrix, ill-functioning variance-covariance matrix, possible decay of estimators to noise level, and inaccurate distributions (Fan and Lv, 2008) [1]. Using similar simulation conditions as shown by Fan and Lv (2008) [1]:

Let X be the $n \times p$ design matrix with the rows consisting of independent realizations from a multivariate Normal distribution $N_p(0, 1)$. We consider (1) $n = 60, p = 1000$ and (2) $n = 60, p = 5000$. Let the response $Y \sim N(0, 2)$ and independent of X.

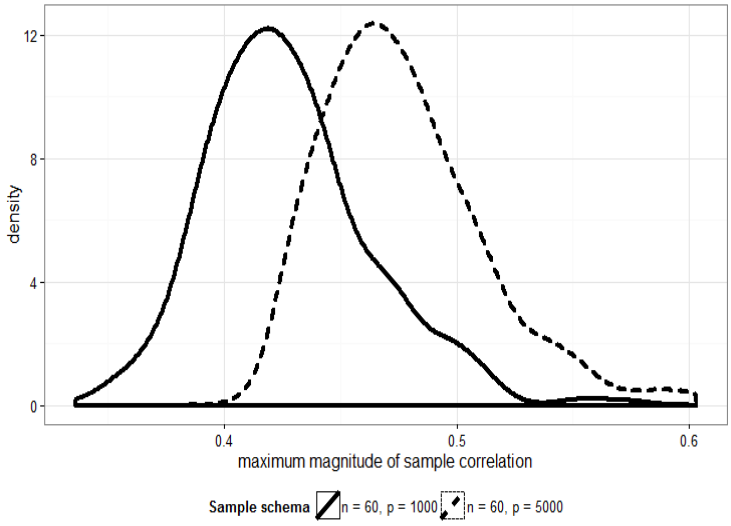

Figure 1: *Density of maximum magnitude of sample correlation under the circumstance that the response is independent from the predictor, and the design matrix has a size of (1) n=60, p=1000 (2) n=60, p=5000 [1]*

Replicating the simulation for each set-up above for 500 times and recording the maximum magnitude of sample correlations, we were able to plot the density of the values. From Figure 1, it can be shown that although during the simulation, the response was set to be independent from the predictors, the marginal correlations are in the range between 0.3 and 0.6, which is away from the true value of 0. This condition illustrates the noise introduced under an ultrahigh

dimensional setting and the challenges faced by the variable selection techniques mentioned above.

With the challenges introduced and possibility that the techniques mentioned above would give noisy and inaccurate results, new statistical methods are proposed for ultrahigh dimensional data. Variable screening is one of the methods that has been developed. Variable screening is a step that is usually applied to the data prior to variable selection with ultrahigh dimensional data, and this step conducts a first filter on the predictors in the pool such that the number of potential predictors can be reduced to be on the order of the sample size. With this reduction in predictor size, the variable selection techniques can be applied to offer less noisy and more accurate results.

Fan and Lv (2008) [1] proposed Sure Independence Screening (SIS) that ranks each predictor based on the magnitude of its marginal Pearson's correlation with the response, and selects a number of top ranked predictors as the relevant predictors for further processing. It is shown to have a sure screening property that the probability of the important variables selected from the screening process belonging to the true model tends toward 1 [1]. However, as mentioned by Fan and Lv (2008) [1] in the paper, using Pearson's correlation has the drawback that although it can identify linear relationships very well, it is not able to capture the nonlinear relationship between response and predictors. For example:
Let the predictor, X, be a sequence of number, such that $X_i = 0, 0.5, ..., 15$
Let the response, Y, be a nonlinear function of the predictor such that

$$Y_i = sin(X_i) + \epsilon_i$$

where $\epsilon_i \sim N(0, 0.2)$. Figure 2 shows one realization from this model.

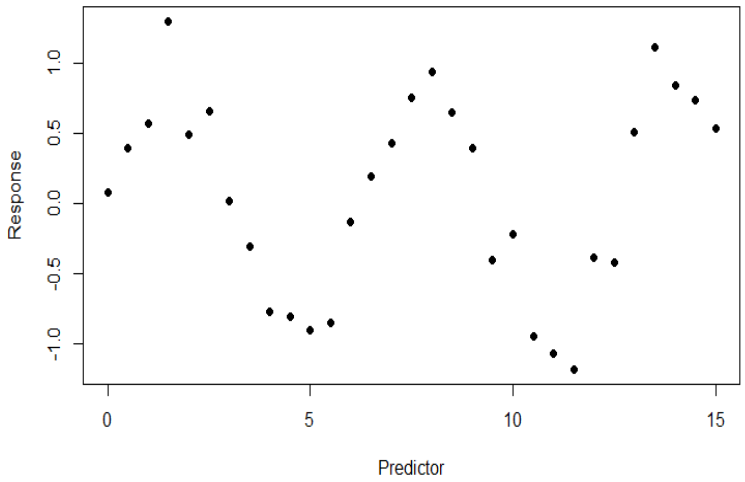

Figure 2: *Response vs. Predictor*

Both the function above, and the scatter plot generated between response and predictor indicate that there is a strong relationship between the response and predictor. However, Pearson's correlation between X and Y is -0.02255. This low correlation computed by Pearson's correlation indicates that if there exists a nonlinear relationship within the sample, then it is possible that marginal Pearson's correlation cannot capture it, which poses a possible shortcoming for SIS.

With the concerns posed by Pearson's Correlation, other screening methods have been developed based on SIS and other correlation measures. Li, Zhong and Zhu (2012) proposed SIS using distance correlation (DC-SIS) rather than Pearson's correlation [2] since distance correlation is able to capture nonlinear relationships. To be specific, using a procedure similar to SIS, DC-SIS ranks each predictor based on its marginal distance correlation with the response. As shown by numerical simulations (Li, Zhong and Zhu, 2012) [2], DC-SIS has a much better performance than SIS in different settings.

However, one potential shortcoming for both SIS and DC-SIS is that both approaches rank only the marginal correlations, and the correlations between the predictors are not considered, which may cause an issue in certain models. To be specific, let X_1 and X_2 be two predictors, and Y be the response. It is possible that X_1 is strongly correlated with Y and X_2 is independent from Y while X_1 and X_2 are strongly correlated. Under this condition, when considering only the marginal correlations, it is possible for SIS and DC-SIS to select both variables instead of only X_1. Consider the following example. Suppose that there are two predictors and a response as follows:

$$X_1 \sim N(1, 0.5)$$
$$X_2 \sim \begin{cases} N(2, 0.5) & X_1 \geq 1 \\ X_1 & Otherwise \end{cases}$$
$$y \sim \begin{cases} N(2, 0.5) & X_1 \geq 1 \\ N(1, 0.5) & Otherwise \end{cases}$$

Based on the data simulated above, an obvious relationship can be observed through a scatter plot (as shown below on the left) for response, y, vs. X_2, from which either SIS or DC-SIS may pick X_2 as a relevant variable for the response. However, if we group by X_1, then, as shown below on the right, no obvious trend can be observed for each group of the variables indicating that only X_1 is relevant to the response. This manifests the potential issue that can be caused by only considering marginal correlations.

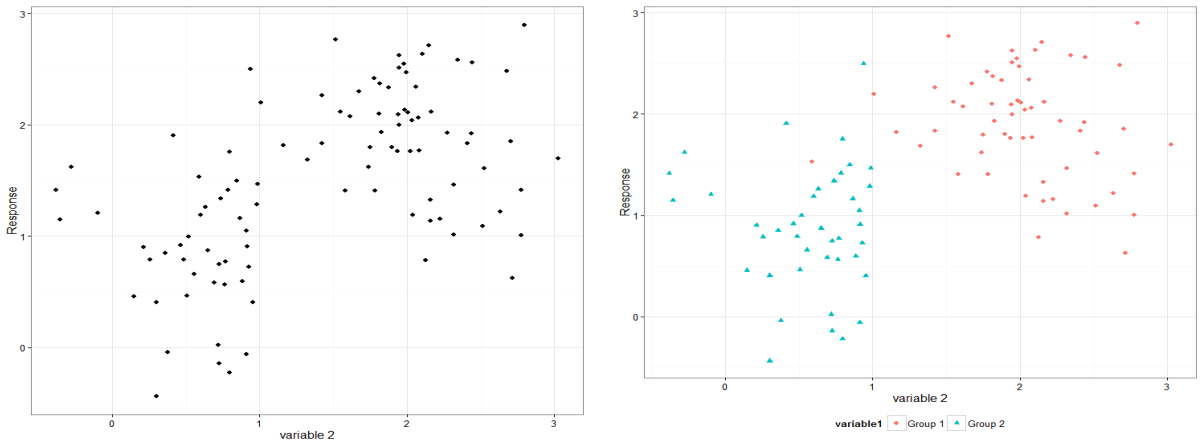

Figure 3: ***Left**: Scatter plot of response vs. predictor without considering groups **Right**: Scatter plot of response vs. predictor considering groups*

In this paper, we propose a new feature selection method using distance correlation and complete least squares (DC-CLS) for ultrahigh dimensional data settings. Distance correlation has been proposed as a new measurement of dependence. It has been shown that if both of the two random vectors for computation have finite first moments, distance correlation has a range from 0 to 1, and it is 0 if and only if the two vectors are independent [7]. Reyes (2012) proposed Complete Least Squares (CLS) as a new method of estimation and showed the benefits of using CLS in variable selection and variable screening; namely, it has a higher stability compared to other techniques. Furthermore, it has been shown by simulation studies that CLS can be generally competitive with other commonly used approaches [8]. These properties of distance correlation and CLS motivate us to combine them in our approach.

To assess the performance of DC-CLS, we did a proof of concept, which showed promising results that it can be generally competitive with SIS and DC-SIS. Further, we conducted simulation studies using different models. Our simulation studies suggested that although DC-CLS showed promise in a proof of concept example, it is not competitive for general use. However, further research is needed to fully develop this approach.

In the next section, we introduce feature screening via distance correlation and complete least squares (DC-CLS). Also, we explain in detail complete least squares (CLS) and distance correlation (DC), and how feature screening should work by the combination of the two methods.

2 DC-CLS

2.1 Complete Least Squares (CLS)

Let **X** be a $n \times p$ design matrix and **y** be a $n \times 1$ response that follows a linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ where for $i = 1...n$, ϵ_i are independently and identically distributed such that they are centered at 0 with a unknown variance σ^2.

Assuming that **X** and **y** are scaled and centered, Ordinary Least Squares (OLS) estimator for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{OLS} = argmin_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$$

is well-known to be the best linear unbiased estimator. However, when there exists a high correlation within the predictors, OLS does not preform very well. Moreover, in high dimensional settings, the OLS estimator is not uniquely defined. Under these conditions, other biased estimators, such as LASSO [3] (Tibshirani, 1996), are preferred. Complete Least Squares (CLS) has also proved to be a biased estimator that shares similarity with other biased estimators (Reyes, 2012) [8].

As his motivating example for the objective function of CLS, Reyes (2012) used a data set from a study on 442 diabetic patients. The data set contains three different predictors of each patient, including the age ($\mathbf{x}_{(1)}$), body mass index (BMI) ($\mathbf{x}_{(2)}$), and average blood pressure ($\mathbf{x}_{(3)}$). Also, the data contains "a quantitative measure of disease progression one year after baseline" as the response of interest, **y**. And the question of interest is to estimate the parameters, $\{\beta_1, \beta_2, \beta_3\}$, for each corresponding predictor. Since the relationship between the response and predictors is unknown before examining the data set, there are multiple possible linear models with different predictors included. Then, all possible objective functions can be listed as follows [8]:

$$\begin{aligned}
\text{1-Variable}: &\begin{cases} ||\mathbf{y} - \mathbf{X}_{(1)}\beta_1||^2 \\ ||\mathbf{y} - \mathbf{X}_{(2)}\beta_2||^2 \\ ||\mathbf{y} - \mathbf{X}_{(3)}\beta_3||^2 \end{cases} \\
\text{2-Variable}: &\begin{cases} ||\mathbf{y} - \mathbf{X}_{(1)}\beta_1 - \mathbf{X}_{(2)}\beta_2||^2 \\ ||\mathbf{y} - \mathbf{X}_{(2)}\beta_2 - \mathbf{X}_{(3)}\beta_3||^2 \\ ||\mathbf{y} - \mathbf{X}_{(3)}\beta_3 - \mathbf{X}_{(1)}\beta_1||^2 \end{cases} \\
\text{3-Variable}: &\ ||\mathbf{y} - \mathbf{X}_{(1)}\beta_1 - \mathbf{X}_{(2)}\beta_2 - \mathbf{X}_{(3)}\beta_3||^2
\end{aligned}$$

With no prior information about which model is correct to pick, Reyes (2012) suggested "determining a value of $\boldsymbol{\beta}$ that is 'good' across all seven models"; further, he proposed constructing an estimator that minimizes the sum of all possible objective functions for different linear models, which, in particular, is:

$$\begin{aligned}
Q(\boldsymbol{\beta}) = &||\mathbf{y} - \mathbf{X}_{(1)}\beta_1||^2 + ||\mathbf{y} - \mathbf{X}_{(2)}\beta_2||^2 + ||\mathbf{y} - \mathbf{X}_{(2)}\beta_2||^2 \\
&+ ||\mathbf{y} - \mathbf{X}_{(1)}\beta_1 - \mathbf{X}_{(2)}\beta_2||^2 + ||\mathbf{y} - \mathbf{X}_{(2)}\beta_2 - \mathbf{X}_{(3)}\beta_3||^2 + ||\mathbf{y} - \mathbf{X}_{(3)}\beta_3 - \mathbf{X}_{(1)}\beta_1||^2 \\
&+ ||\mathbf{y} - \mathbf{X}_{(1)}\beta_1 - \mathbf{X}_{(2)}\beta_2 - \mathbf{X}_{(3)}\beta_3||^2
\end{aligned}$$

$Q(\boldsymbol{\beta})$ is defined to be the CLS Objective function and the estimator that minimizes this objective function is the CLS estimator under this given situation [8].

Reyes (2012) showed the objective function for CLS can be written in a compact form. First, define

$$Q_{p,k} = \sum_{\mathbf{s}\in\mathbf{S}_{p,k}} ||\mathbf{y} - \mathbf{X}\mathbf{D}_s\beta||^2$$

where $\mathbf{S}_{p,k}$ indicates "the set of all p-dimensional vectors with exactly k elements taking value 1, and exactly $p-k$ elements taking value 0" with $1 \leq k \leq p$ and $\mathbf{D}_s$ is a matrix with the diagonal being the values of s vector. In the case when $k = p$, the CLS objective function will turn out to be the same as the objective function for a full-model OLS as all variables are included inside of the model with size to be included being exactly p [8]. Moreover, a weight vector can be introduced into the objective function such that different linear models in the summation have different emphasis under different cases. Having the weight introduced, he defined the CLS objective function as

$$Q_p(\boldsymbol{\beta},\boldsymbol{\omega}) = \sum_{k=1}^{p} \omega_k Q_{p,k}(\boldsymbol{\beta})$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_p)^T$ is the weight vector, in which each value in the vector is the weight for the model size at the corresponding position and the weight is always between 0 and 1. Then, the objective function can be simplified to

$$Q_p(\boldsymbol{\beta},\boldsymbol{\omega}) = \lambda_0\mathbf{y}^T\mathbf{y} - 2\lambda_1\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T(\lambda_2\mathbf{X}^T\mathbf{X} + (\lambda_1 - \lambda_2)\mathbf{D}_{\mathbf{X}^T\mathbf{X}})\boldsymbol{\beta}$$

where for $i = 0, 1, 2$, $\lambda_i = \sum_{k=1}^{p} \omega_k \binom{p-i}{k-i}$, $\mathbf{D}_{\mathbf{X}^T\mathbf{X}}$ is a diagonal matrix such that the values on the diagonal of the matrix is the same as the diagonal of matrix $\mathbf{X}^T\mathbf{X}$.

From the objective function defined above, the CLS estimator (Reyes, 2012), the value of $\boldsymbol{\beta}$, which minimizes the CLS objective function, is

$$\hat{\boldsymbol{\beta}}_{CLS} = (\tau\mathbf{X}^T\mathbf{X} + (1-\tau)\mathbf{D}_{\mathbf{X}^T\mathbf{X}})^{-1}\mathbf{X}^T\mathbf{y}$$

where $\tau = \lambda_2/\lambda_1$ [8].

A few remarkable properties of CLS motivate us to use it in variable screening. The first one is the relationship between CLS estimator and univariate marginal estimators and that between CLS estimator and OLS estimator. With different choices on τ, the CLS estimator tends to move toward either the full model OLS estimators ($\tau = 1$) or univariate marginal estimators ($\tau = 0$). [8]

The second one is that if the design matrix and response are centered and scaled properly such that $\mathbf{y}^T\mathbf{1} = 0$, $\mathbf{y}^T\mathbf{y} = 1$, $\mathbf{X}^T\mathbf{1} = 0$, and $\mathbf{X}^T\mathbf{X} = \mathbf{R}$, where $\mathbf{R}$ is a valid correlation matrix within the predictors, then the CLS estimator can be rewritten as:

$$\hat{\boldsymbol{\beta}}_{CLS} = (\tau\mathbf{R} + (1-\tau)\mathbf{I})^{-1}\mathbf{R}_{\mathbf{Xy}}$$

where $\mathbf{R}_{\mathbf{Xy}}$ is a vector of correlations between response and each predictor.

Being able to substitute a part of the estimator to be a valid correlation matrix allows us to handle both the correlation among the predictors and the correlations between response and predictors [8].

2.2 Distance Correlation (DC)

Distance correlation is proposed as a new approach to assess independence between two random vectors by Szkely, Rizzo and Bakirov (2007). To be specific, the distance correlation between two random vectors is a weighted Euclidean distance between the two characteristic functions of the two random vectors. Let $g_{\mathbf{x}}(t)$ and $g_{\mathbf{y}}(s)$ be the characteristic function of two random vectors, **x** and **y**, respectively. Let $g_{\mathbf{x},\mathbf{y}}(t,s)$ be the joint characteristic function of **x** and **y**. Let $dim(\mathbf{x})$ and $dim(\mathbf{y})$ denote the dimensions of **x** and **y**. Then, the distance covariance (dCov) between **x** and **y** with a weight function $w(t,s)$ is

$$\begin{aligned} dCov(\mathbf{x},\mathbf{y}) &= ||g_{\mathbf{x},\mathbf{y}}(t,s) - g_{\mathbf{x}}(t)g_{\mathbf{y}}(s)||_w^2 \\ &= \int_{\mathbb{R}^{dim(\mathbf{x})+dim(\mathbf{y})}} |g_{\mathbf{x},\mathbf{y}}(t,s) - g_{\mathbf{x}}(t)g_{\mathbf{y}}(s)|^2 w(t,s)dtds \end{aligned}$$

where $w(t,s) = [c_{dim(\mathbf{x})}c_{dim(\mathbf{y})}||t||_{dim(\mathbf{x})}^{1+dim(\mathbf{x})}||s||_{dim(\mathbf{y})}^{1+dim(\mathbf{y})}]^{-1}$, $c_d = \pi^{\frac{1+d}{2}}[\Gamma(\frac{1+d}{2})]^{-1}$ and $||\mathbf{c}||$ represents the Euclidean norm of **c** in the expression. Then the distance correlation between the two random vectors, **x** and **y**, can be defined as [7]

$$dCorr(\mathbf{x},\mathbf{y}) = \begin{cases} \frac{dCov(\mathbf{x},\mathbf{y})}{\sqrt{dCov(\mathbf{x},\mathbf{x})dCov(\mathbf{y},\mathbf{y})}} & dCov(\mathbf{x},\mathbf{x})dCov(\mathbf{y},\mathbf{y}) > 0 \\ 0 & dCov(\mathbf{x},\mathbf{x})dCov(\mathbf{y},\mathbf{y}) = 0 \end{cases}$$

Moreover, Szkely, Rizzo and Bakirov (2007) gave a definition of empirical distance covariance

$$dCov(\mathbf{x},\mathbf{y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}$$

where $A_{kl} = a_{kl} - \overline{a}_{k*} - \overline{a}_{*l} + \overline{a}_{**}$ with $a_{kl} = |\mathbf{x}_k - \mathbf{x}_l|_p$, $\overline{a}_{k*} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}$, $\overline{a}_{*l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}$, $\overline{a}_{**} = \frac{1}{n^2}\sum_{l,k=1}^{n} a_{kl}$, and B_{kl} shares a similar description with A_{kl} but for **y** random vector. This allows the computation of distance correlation without knowing the characteristic functions. Also, it allows us to find a possible way to extend the existing computation package for our purposes.

A remarkable property of distance correlation motivates us to use it in variable screening. The property is that if the two random vectors have a finite first moment, then the distance correlation between the two vectors is between 0 and 1. Moreover, the distance correlation is equal to 0 if and only if the two vectors are independent. Distance correlation will perform better than Pearson's correlation if there exist nonlinear relationships between the two random vectors [7]. We demonstrate this property by using two random variables that are non-linearly related. Let **X** be a random vector that follows a normal distribution with mean of 0 and variance of 1. Let **Y** be another random vector that follows: $\mathbf{Y} = \mathbf{X}^2 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is an error term. Figure 4 illustrates the relationship.

We computed both Pearson's correlation and distance correlation between the two vectors and obtained a Pearson's correlation of -0.06415 and a distance correlation of 0.54625. The Pearson's correlation has a magnitude that is close to zero indicating a very weak or no relationship between the two vectors. However, the distance correlation gives a stronger magnitude suggesting the existence of relationship between the two random vectors.

In the next section, we will demonstrate a DC-and-CLS-based variable screening technique. Via two small simulations as proof of concept, we will illustrate how a DC-CLS can have a potential to perform better than DC-based and general sure screening techniques.

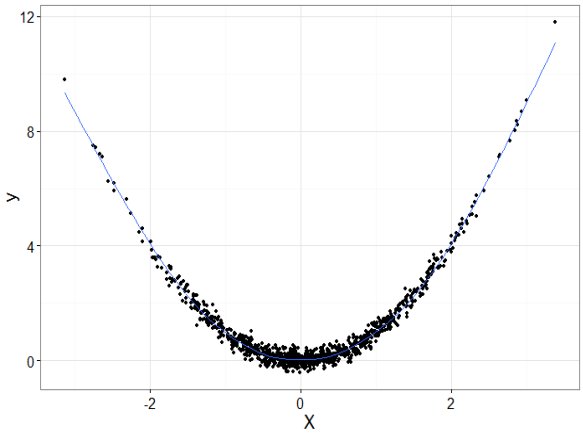

Figure 4: *Relationship for* $\boldsymbol{Y} \sim \boldsymbol{X}$*, when* $\boldsymbol{Y} = \boldsymbol{X}^2 + \boldsymbol{\epsilon}$

2.3 DC-CLS

2.3.1 Screening Procedure

In this section, we propose a variable screening procedure that is based on DC and CLS. Let $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ be the response vector. Let $\mathbf{X} = (\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, ..., \mathbf{X}_{(p)})^T$ be the design matrix, where $\mathbf{X}_{(i)}$, for $i = 1, 2, ..., p$, denotes each predictor column with a size of $n \times 1$ of the design matrix. With an ultrahigh-dimensional setting, the total number of predictor columns, p, is much greater than the sample size, n. Assuming that there is a small number of predictors that are actually relevant and important to the response, then we can define two categories of variables. Let S_I denote the set of relevant predictors and S_U denote the set of irrelevant predictors. Hence,

$$\mathbf{X}_{(i)} \in S_I \text{ if } \mathbf{y} \text{ depends on } \mathbf{X}_{(i)}$$
$$\mathbf{X}_{(i)} \in S_U \text{ if } \mathbf{y} \text{ is independent from } \mathbf{X}_{(i)}$$

To find the set of relevant predictors in the screening, we use the CLS estimator mentioned earlier so that both the relationship among the predictors and that between the response and predictors are captured.

$$\hat{\boldsymbol{\beta}}_{CLS} = (\tau \mathbf{R_{XX}} + (1 - \tau) I)^{-1} \mathbf{R_{Xy}}$$

where $\mathbf{R_{XX}}$ and $\mathbf{R_{Xy}}$ are substituted with the corresponding distance correlation matrices.

Since CLS estimator allows both positive and negative values, we assume that the estimator is symmetric about zero so that larger magnitude indicates a higher importance while smaller magnitude (i.e. close to 0) suggests weak or no relevance to the model. Hence, we consider ranking the relevance of each predictor by using the magnitude of the CLS estimators. Then, we can choose a set of relevant predictors whose magnitude of CLS estimators are ranked at the top $n - 1$. In the next section, we will illustrate our proposed approach using two small simulations as proof of concepts.

2.3.2 Proof of Concepts

To illustrate our proposed approach and assess whether this approach can possibly perform better than the other screening procedures, we performed two simulations as proof of concept.

Case 1:

The first case is a similar case as mentioned in the introduction. Let X_1, X_2 and X_3 be three predictors with size 1000×1, and Y be a 1000×1 response such that

$$\begin{gathered} X_1 \sim N(1, 0.5) \\ X_2 \sim N(\mathbb{1}(X_1 \geq 0) + [1 - \mathbb{1}(X_1 \geq 0)](-1), 1) \\ X_3 \sim N(0, 1) \\ Y \sim X_1 + N(0, 0.5) \end{gathered}$$

Based on the simulation data above, we computed the estimators using all three screening techniques, including SIS, DC-SIS, and DC-CLS (using $\tau = 0.7$).

Table 1: Relevance From Three Screening Methods (Proof of Concept 1)

	SIS	DC_SIS	DC_CLS
x1	0.691	0.702	0.607
x2	0.478	0.493	0.187
x3	0.003	0.054	0.030

As shown in the table above, it can be seen that all three approaches pick up X_1 with a large relevance level from 0.691 to 0.702. Also, all three approaches have a small relevance level (i.e. close to zero) for X_3, which indicates that the response shares a weak or no relationship with X_3. This is a reasonable selection since X_3 is neither directly related to other predictors nor directly related to the response. However, for X_2, SIS and DC-SIS show a relevance level at around 0.478 to 0.493, which suggests that X_2 is a fairly important and promising variable in the model. Compared with the relevance levels obtained from SIS and DC-SIS, the relevance level obtained from DC-CLS is relatively smaller and closer to zero at around 0.187.

From the original model indicated above, it can be seen that response Y is directly a function of X_1, which suggests that Y functionally depends on X_1. However, X_2 is a function of X_1 but not directly related to the response, as illustrated in Figure 5.

Although it can be observed that there exists a marginal relationship between response Y and X_2 from the plot on the left, if we group X_2 by X_1, no obvious trends can be observed for each group of X_2 indicating that the trend seen may be driven by X_1 and only X_1 is relevant to the response.

Intuitively, DC-CLS is similar to a regression analysis, which determines importance conditional on the remaining terms in the model. Using a linear model fit of the response vs. the three predictors, we found that only X_1 is relevant to the response with a significance level of 0.05.

According to the analysis above, the results from SIS and DC-SIS can be misleading since it indicates that X_2 is a fairly promising predictor in the model by giving a level around 0.5. However, DC-CLS gave a relevance result that is much smaller than given by the other two to be

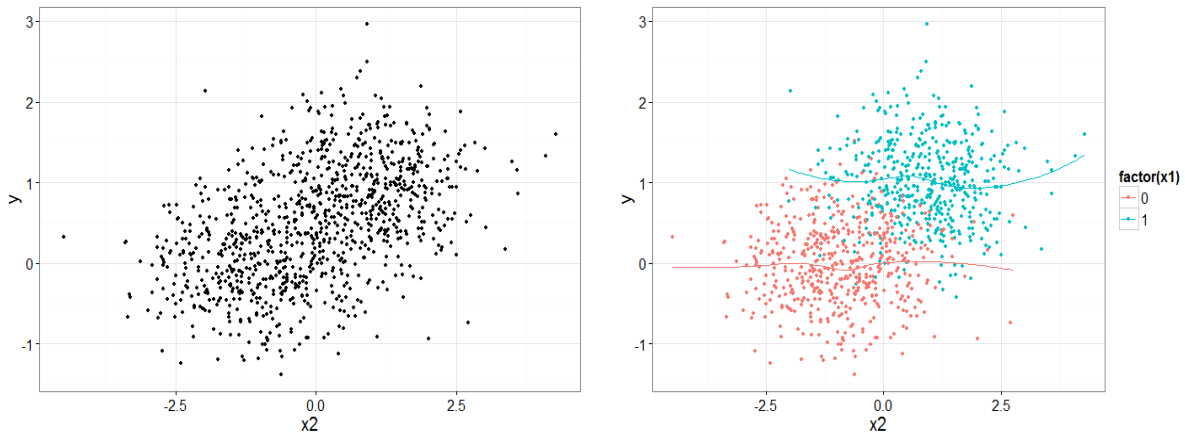

Figure 5: *Proof of Concept 1:* ***left:*** *response* Y *vs.* X_2 *without grouping by* X_1*,* ***right:*** *response* Y *vs.* X_2 *with grouping by* X_1

Table 2: Summary of Linear Model Fit

	Estimate	Std. Error	t value	P-value
x1	0.9813	0.0449	21.85	0.0000
x2	-0.0116	0.0159	-0.73	0.4646
x3	0.0338	0.0323	1.05	0.2954

around 0.2, which indicates the relationship is fairly weak between the response and X_2. Although the rank of each predictor from each approach is the same, DC-CLS gave a more reliable estimation of the importance level of X_2.

Case 2:

The second case uses a model that contains both nonlinear relationships and correlations within predictors. Let X_1, X_2 and X_3 be three predictors with size 1000×1, and Y be a 1000×1 response such that

$$\begin{gathered} X_1 \sim N(0,1) \\ X_2 \sim -sin(X_1) + N(0, 0.5) \\ X_3 \sim N(0,1) \\ Y \sim sin(X_1) + X_2 + N(0, 0.5) \end{gathered}$$

Again, based on the simulation data above, we computed the estimators using all three screening techniques, including SIS, DC-SIS, and DC-CLS (using $\tau = 0.7$).

Table 3: Relevance From Three Screening Methods (Proof of Concept 2)

	SIS	DC_SIS	DC_CLS
v1	0.037	0.051	0.213
v2	0.437	0.375	0.488
v3	0.004	0.050	0.044

As shown in the table above, it can be seen that all three approaches pick up X_2 with a large relevance level from 0.375 to 0.488. Also, all three approaches have a small relevance level (i.e.

close to zero) for X_3, which indicates that the response shares a weak or no relationship with X_3. This is a reasonable selection since X_3 is neither directly related to other predictors nor directly related to the response. However, for X_1, SIS and DC-SIS show a close-to-zero relevance level indicating that X_1 should not be in the model. Compared with the relevance levels obtained from SIS and DC-SIS, the relevance level obtained from DC-CLS is relatively larger at around 0.213.

From the original model indicated above, it can be seen that response Y is directly a function of X_2, which suggests that Y functionally depends on X_2. Moreover, since the response depends on X_1 with a sine function, the response also functionally depends on X_1. However, $E(Y|X_1) = 0$ suggesting that only considering the marginal relationship would be misleading. This is shown in Figure 6.

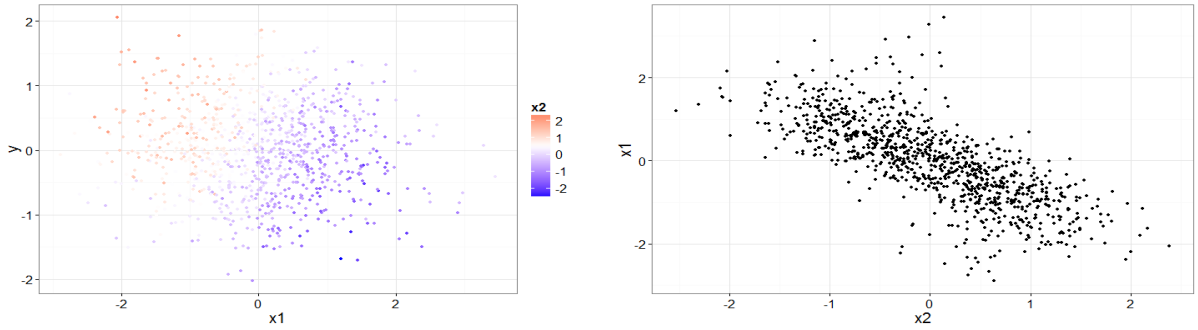

Figure 6: *Proof of Concept 1:* ***left:*** *response* Y *vs.* X_1 *with grouping by* X_1*,* ***right:*** X_1 *vs.* X_2

It can be observed that there is no obvious trend between the response and X_1. However, there is a strong correlation between X_2 and X_1. Considering the conditional relationship allows us to detect both variables.

Moreover, using a linear model fit of the response vs. the three predictors (Table 4), we found that both X_1 and X_2 are relevant to the response at a significance level of 0.05.

Table 4: Summary of Linear Model Fit

	Estimate	Std. Error	t value	P-value
x1	0.4401	0.0255	17.26	0.0000
x2	0.7723	0.0315	24.52	0.0000
x3	-0.0040	0.0166	-0.24	0.8090

Based on the analysis above, the relevance results from SIS and DC-SIS can be misleading since it indicates that X_1 should not be included in the model. However, DC-CLS gave a relevance result that is relatively larger at around 0.2, which indicates a possibility that X_1 should be included. DC-CLS gave a more reliable estimation of the importance level of X_1.

Through the simulations as proof of concept above, we conclude that DC-CLS can perform better than the other screening procedures under certain given models. Hence, we further conducted simulation studies on different models. In the next section, we will demonstrate our results from the simulation studies.

3 Simulation Studies

In this section, we assess the performance of DC-CLS by conducting simulation studies.

In the simulation studies, we generated a design matrix (**X**) having a size of $n \times p$, where n is the number of rows and p is the number of columns. Each row of **X** is generated such that $\mathbf{X}_i^T$ follows a multivariate normal distribution with zero mean and variance-covariance matrix $\mathbf{\Sigma} = (\sigma_{i,j})_{p\times p}$, where, for all $1 \leq i \leq p$ and $1 \leq j \leq p$,

$$(\mathbf{\Sigma})_{ij} = \rho^{\mathbb{1}(i\neq j)}$$

where ρ is a given constant correlation, and $\mathbb{1}(u)$ is an indicator for taking value 1 if event u occurs and is 0 otherwise.

With the design matrix, **X**, generated, we obtained the response, **y**, based on the three models as follows with an error term $\boldsymbol{\epsilon}$ following a $N(0, \sigma_y^2)$, where σ_y^2 is computed based on a given R^2 value. Consider the case that $\mathbf{y}_i = \mathbf{X}_i^T\boldsymbol{\beta} + \epsilon_i$. Then,

$$\begin{aligned} Var(y_i) &= Var(\mathbf{X}_i^T\boldsymbol{\beta}) + Var(\epsilon_i) \\ &= \boldsymbol{\beta}^T\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} + \sigma_y^2 \end{aligned}$$

Since R^2 is defined as

$$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST},$$

the theoretical R^2 value can be represented as

$$R^2 = 1 - \frac{\sigma_y^2}{\boldsymbol{\beta}^T\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}+\sigma_y^2}.$$

Then, given constant R^2, the variance, σ_y^2, for the error term in the response can be computed.

The three models are chosen to assess the performance of DC-CLS under different cases that contain nonlinear terms. The idea of these models are similar to the models tested in DC-SIS [2]. To be specific, Model(1) has the nonlinear term on $\mathbf{X}_{12}$. Model(2) contains both an interaction term X_1X_2 and a nonlinear-interaction term $\mathbb{1}(X_{12} > 0)X_{22}$. Model(3) contains nonlinear term on $\mathbf{X}_{12}$ and $\mathbf{X}_{22}$ as well as an interaction term X_1X_2.

$$\begin{aligned} \text{Model(1): } \mathbf{E}(\mathbf{y}|\mathbf{X}) &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3\mathbb{1}(X_{12} > 0) + \beta_4X_{22} \\ \text{Model(2): } \mathbf{E}(\mathbf{y}|\mathbf{X}) &= \beta_0 + \beta_1X_1X_2 + \beta_2\mathbb{1}(X_{12} > 0)X_{22} \end{aligned}$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (1, 4, 2, 2, 1)$, and R^2 = 0.6. For each model, we considered two different covariance matrices defined by $\rho = 0, 0.5$ with $n = 60, 200$ and $p = 1000, 3000$.

For each replication, we computed the SIS, DC-SIS, and DC-CLS ($\tau = 0.5$, $\tau = 0.9$) estimators. Then, the predictors were ranked by their magnitudes. To assess how well each method evaluates each predictor in the design matrix, we computed the minimum model size that is required to contain all important predictors, and found the $5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}$ percentile of the set of minimum model sizes out of 500 replications for each model.

For Model(1), the result quantiles are given below.

Table 5: Model 1: $n = 60, p = 1000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	88.00	304.75	521.00	735.00	953.15	113.75	290.50	481.50	686.25	890.00
DC_SIS	112.00	345.75	579.00	779.25	965.10	119.95	299.00	476.00	688.00	906.05
DC_CLS_0.5	496.60	691.75	838.00	931.25	988.00	211.95	496.75	706.00	864.00	978.05
DC_CLS_0.9	458.85	707.75	839.50	929.25	984.05	491.10	697.75	832.00	926.25	985.00

Table 6: Model 1: $n = 60, p = 3000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	278.75	854.50	1465.00	2221.25	2842.00	342.00	998.50	1477.50	2075.00	2655.55
DC_SIS	303.55	955.25	1611.00	2221.00	2839.30	339.00	907.75	1431.00	2078.50	2667.25
DC_CLS_0.5	1388.80	2146.50	2538.50	2777.25	2970.00	1359.55	2069.75	2463.00	2756.25	2965.00
DC_CLS_0.9	1387.90	2070.00	2501.00	2796.75	2960.25	1371.35	2098.75	2558.50	2789.00	2953.05

Table 7: Model 1: $n = 200, p = 1000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	9.95	46.00	138.50	312.00	797.05	33.95	126.75	273.50	482.25	749.10
DC_SIS	9.95	50.00	153.50	381.25	816.65	34.95	124.25	252.50	459.25	733.00
DC_CLS_0.5	9.95	61.00	256.00	595.25	936.00	40.95	181.00	393.00	723.50	950.10
DC_CLS_0.9	435.35	692.00	813.00	915.25	977.00	122.90	375.50	609.00	837.25	975.00

Table 8: Model 1: $n = 200, p = 3000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	26.95	143.00	413.50	963.25	2097.15	105.95	369.75	792.50	1402.75	2381.15
DC_SIS	24.00	163.50	433.00	1087.50	2493.95	115.95	371.50	790.00	1400.00	2251.20
DC_CLS_0.5	1479.30	2125.25	2532.50	2799.50	2964.00	160.95	671.50	1394.50	2255.00	2868.10
DC_CLS_0.9	1258.80	2006.25	2475.00	2774.50	2952.15	1425.35	2121.75	2510.00	2795.00	2956.15

From the tables above, it can be seen that the minimum model size for DC-CLS that is required to contain all important predictors can be much larger than that required for SIS and DC-SIS, which indicates that DC-CLS is not as efficient as SIS and DC-SIS under this model setting. However, note that when the sample size increases, all three procedures tend to perform better for different number of predictors. This is also verified with a density plot of minimum model size.

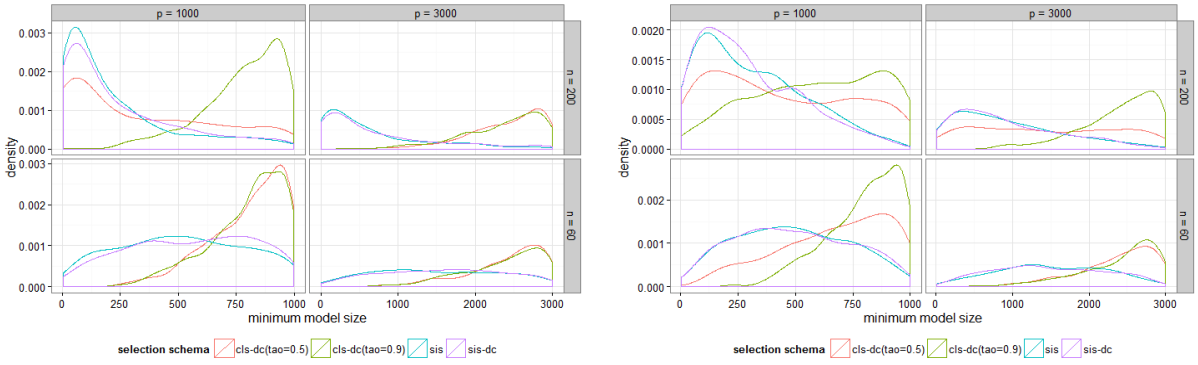

Figure 7: *Model 1: density of minimum model size* ***left:*** *with* $\rho = 0$, ***right:*** *with* $\rho = 0.5$

From the plot above, we can see that under Model(1), the blue and purple curves, the curves for SIS and DC-SIS respectively, are always farthest to the left and farther left than the curve for DC-CLS, which indicates that DC-CLS is not as efficient as the other two procedures. This corresponds well with the results from the results percentile tables above.

For Model(2), the density plot for the minimum model size and result quantiles tables are given below.

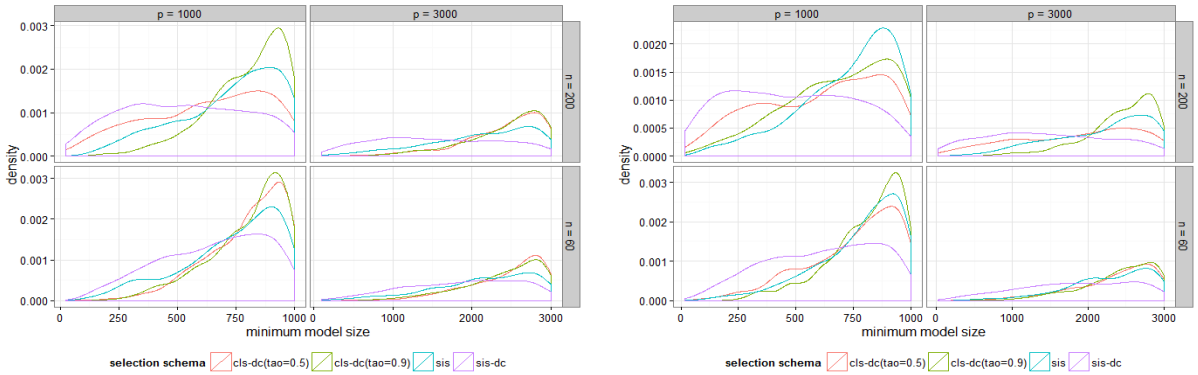

Figure 8: *Model 2: density of minimum model size* ***left:*** *with* $\rho = 0$*,* ***right:*** *with* $\rho = 0.5$

Table 9: Model 2: $n = 60, p = 1000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	306.95	614.00	778.50	903.00	980.05	405.85	677.00	824.50	923.00	987.00
DC_SIS	252.95	474.75	683.50	838.75	962.00	208.00	432.75	640.00	827.25	954.00
DC_CLS_0.5	482.80	687.50	830.50	928.25	988.05	374.90	621.00	798.00	912.25	984.05
DC_CLS_0.9	464.90	706.00	846.50	922.00	985.05	470.95	721.75	844.00	930.00	986.00

Table 10: Model 2: $n = 60, p = 3000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	859.95	1759.00	2258.50	2674.00	2915.30	1233.25	1935.50	2394.50	2744.25	2953.15
DC_SIS	647.65	1325.00	1961.50	2482.00	2854.10	539.05	1316.00	1890.00	2481.50	2862.40
DC_CLS_0.5	1448.40	2127.00	2562.50	2812.25	2961.00	1223.35	2040.75	2474.00	2762.75	2960.15
DC_CLS_0.9	1411.65	2122.25	2501.50	2799.50	2965.05	1434.85	2148.75	2515.50	2805.00	2960.05

Table 11: Model 2: $n = 200, p = 1000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	312.85	580.00	773.00	898.25	986.00	301.40	592.50	776.50	886.50	975.00
DC_SIS	110.00	324.75	532.00	748.50	956.10	80.95	251.75	483.00	708.50	919.20
DC_CLS_0.5	165.65	419.25	661.50	834.00	971.00	165.80	379.50	651.00	834.75	957.15
DC_CLS_0.9	476.75	698.00	834.00	925.00	986.00	256.85	525.25	710.00	872.00	977.00

Table 12: Model 2: $n = 200, p = 3000$

	$\rho = 0$					$\rho = 0.5$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	813.60	1681.50	2212.50	2667.25	2932.10	1034.35	1728.50	2364.50	2696.00	2944.05
DC_SIS	359.80	951.50	1537.50	2250.50	2845.05	229.95	847.00	1435.50	2116.75	2796.00
DC_CLS_0.5	1382.75	2144.00	2522.50	2784.50	2965.10	433.75	1182.25	1998.50	2502.25	2906.15
DC_CLS_0.9	1370.75	2157.00	2564.50	2803.00	2966.05	1548.30	2211.25	2514.00	2782.25	2937.05

From the density plot above in Figure 8, we can see that although DC-CLS is not performing the best, DC-CLS can perform better than SIS under certain circumstances, especially when the correlation within the predictors is nonzero. This can also be verified through the result percentiles table below. From the table below, we can find that for all cases, DC-SIS performs the best through all three procedures. DC-CLS with $\tau = 0.5$ can possibly perform better than SIS but not as well as DC-SIS.

In all, based on the simulation studies above, although DC-CLS shows promising results in the proof of concept, it is not competitive for general uses.

4 Discussion

In this paper, we proposed a new screening procedure that is based on distance correlation and complete least squares. However, our simulation studies reveal that this procedure is not effective for general use. But since the proof of concept shows promising results, we consider a couple reasons that can vary the results of the simulation. In this section, we will demonstrate how two different factors could possibly lead to different results, including selection of τ and whether to use the magnitude of CLS estimator.

4.1 Selection of τ

To assess how different τ produces different result, we chose to use Model(2) with a correlation of 0.8 within the design matrix and computed the CLS estimators for different τ value so that $\tau = 0, 0.1, 0.2, ..., 0.9$.

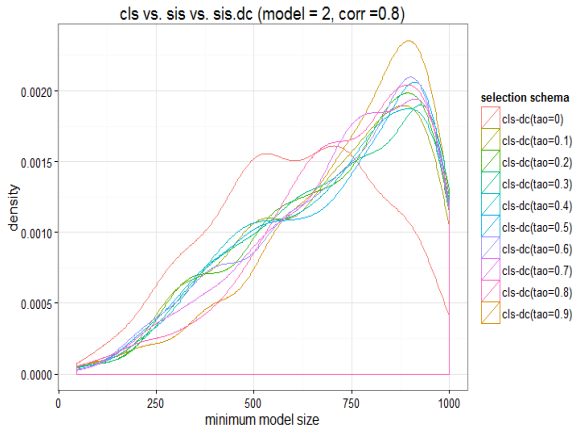

Figure 9: *density of minimum model size for Model(2) with $\rho = 0.8$ and various τ values*

As shown in the plot above, we can see that with different τ values, the curve for the density of minimum model size shifts from left to right and back and forth, which indicates that the selection of τ can be critical for the CLS estimator so that it directly changes the ranking of each predictor that is in the pool.

Reyes (2012) defined the τ in the CLS estimator to be

$$\tau = \lambda_2/\lambda_1 = \sum_{k=1}^{p} \omega_k \binom{p-2}{k-2} / \sum_{k=1}^{p} \omega_k \binom{p-1}{k-1} \text{ [8]}$$

where k is an integer such that $1 \leq k \leq p$, p is the size of the predictors in the data set and ω_k is a pre-specified model weight for the k-th model. However, in our simulation study, we arbitrarily selected the τ values to be 0.5 or 0.9 for the CLS estimators, which do not exactly follow the given definition of CLS estimator. Hence, it is probable that using arbitrary τ values in our simulation studies can lead to CLS estimators that do not actually minimize the corresponding CLS objective function.

Moreover, for each different model and correlation structure of design matrices, τ values might vary to minimize the corresponding CLS objective functions. As mentioned by Reyes (2012), with different τ values, the CLS estimator can be tending toward either univariate marginal estimates or full OLS estimates. Hence, τ values can vary the CLS estimators largely [8]. However, in our simulation study, τ values are selected to be the same for all models with all possible correlation structures in the design matrix, which can possibly cause CLS not to estimate the best results.

According to the discussion regarding to the selection of τ above, to fully assess the CLS estimator, future work may need to be done on selecting the appropriate τ value for its corresponding model.

4.2 Magnitude of CLS Estimator

The process of converting the estimator to a ranking may also influence results. Consider Model(1) with a correlation of 0.5 within the design matrix. We used the case used in the previous section. Then, we generated the density plots of minimum model size for the two cases.

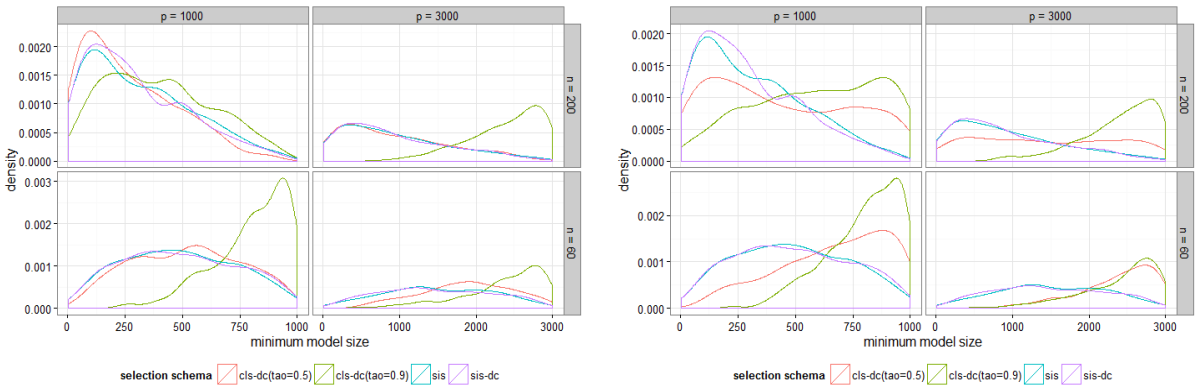

Figure 10: *density of minimum model size for Model(1) with $\rho = 0.5$ ranking by **left:** raw estimators, **right:** magnitude of estimators*

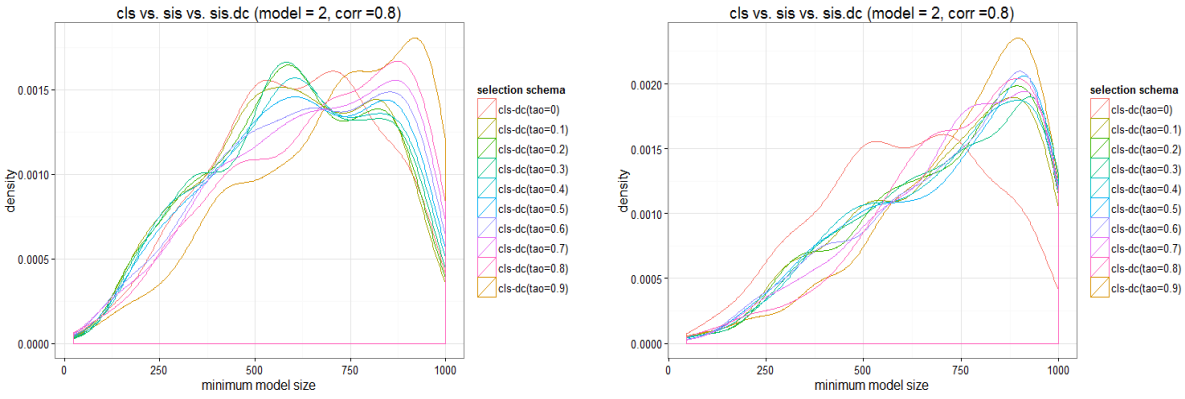

Figure 11: *density of minimum model size for case in previous section ranking by **left:** raw estimators, **right:** magnitude of estimators*

From the plot above, we can see that using the magnitude of the CLS estimators, the curves for the density plot are shifted from the ones that use the raw CLS estimators. And it can be seen that the selection can be more effective sometime when using the raw estimators for the ranking. Although we were able to learn about the properties of distance correlation and complete least squares separately, we were not able to investigate how distance correlation has affected CLS estimator for the screening procedure.

To be specific, it is possible that the estimators are not distributed symmetrically about zero so that taking the magnitude of the estimators can be inappropriate.

According to the discussion regarding the magnitude of CLS estimators, to better assess the CLS estimator with distance correlation, future work is required investigating the properties of the CLS estimator that uses distance correlation matrix inside.

5 Conclusion

In this paper, we proposed a new feature screening procedure using distance correlation and complete least squares. We used a new estimator based on the original CLS estimator to rank the importance of each predictor within the design matrix. In the proof of concept, we used small simulations to show how this screening procedure can potentially perform better and capture the predictors that are marginally independent of the response under certain given circumstances. However, through the simulation studies, it can be shown that our proposed screening is not as effective as the other two screening procedures, SIS and DC-SIS. To reveal the full potential of this new feature screening procedure, DC-CLS, further study on theoretical analysis of the method and estimators needs to be conducted.

References

[1] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849911, 2008.

[2] Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):11291139, 2012.

[3] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267288, 1996.

[4] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) J Royal Statistical Soc B*, 67(2):301320, 2005.

[5] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):14181429, 2006.

[6] Robert Tibshirani, Iain Johnstone, Trevor Hastie, and Bradley Efron. Least angle regression. *Ann. Statist. The Annals of Statistics*, 32(2):407499, 2004.

[7] Gbor J. Szkely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Statist. The Annals of Statistics*, 35(6):27692794, 2007.

[8] Eric M. Reyes. Complete least squares: A new variable screening and selection method., 2012.