Statistics and Its Interface Volume 10 (2017) 399–406

Adaptive model-free sure independence screening

Canhong Wen, Shan Zhu, Xin Chen, and Xueqin Wang*

Variable screening procedure is popularly used in ultrahigh-dimensional data analysis. It ranks the importance of the predictor variables by marginal correlations and then screens out the variables that are weakly correlated or uncorrelated with the response variables. Though demonstrated their effectiveness, the performance of most variable screening approaches depend on the pre-determined threshold of the size of selected predictor variables, which is some integer multiples of $\lceil n/\log(n)\rceil$ with n being the sample size. To circumvent this issue, we propose a novel data-driven variable screening procedure that can automatically determine the threshold. In our proposal, we rank the importance of the predictor variables by the p-values using some modified independent tests, with the smaller p-values indicating higher correlation. Compared with the existing counterpart, extensive simulation studies and a real genetic data indicate the preference of our procedure.

Keywords and phrases: Adaptive threshold, Distance correlation, False discovery rate, Sure independence screening, Ultrahigh dimensional data.

1. INTRODUCTION

Owing to rapid advances of technologies and science, ultrahigh-dimensional data emerge increasingly in contemporary scientific research areas including biological science, social science and so on. The major challenge in dealing with such kind of data lies in the ultrahigh-dimensionality, which means that the number of predictors p is much larger than the sample size n. This limitation leads to the rank deficiency of the design matrix and thus traditional statistical methods cannot be applied directly.

Variable screening procedure has been proposed to deal with ultrahigh-dimensional data and received increasing attention in recent literature. It works by ranking the importance of predictors through marginal utility measures (for example, correlation) between response and predictors and selecting the top few variables as the most important ones. The framework of variable screening started with the seminal work of Fan and Lv [5], in which they proposed the sure independence screening (SIS) procedure with the Pearson correlation as the marginal measure. When the data come from a linear model with Gaussian errors, Fan and Lv [5] have shown that the SIS procedure possesses the desirable sure screening property, that is, with probability very close to one, the procedure retains all of the important variables in the selected model. Soon afterward, SIS has been extended to generalized linear models [6], additive models [4], Cox models [17], compress sensing [15] and so on.

Since specifying a proper/useful model for ultrahigh-dimensional data is challenging, model-free sure screening procedures are more appealing in practice as a first step of analysis. Compare to the SIS technique and its extension, the model-free sure screening procedures measure the importance of variables by utilities that don't rely on any model. Thus these model-free methods tend to have robust performance and could be quite flexible in practical implementation. For example, Zhu et al. [18] developed sure screening procedure with their newly proposed marginal utility measure, which is concerned with the entire conditional distribution of the response given the predictors. Li, Zhong and Zhu [8] introduced model-free sure screening procedure called DC-SIS with utility measure being the distance correlation, a recently proposed measurement of independence by Székely et al. [13, 10].

All the aforementioned variable screening procedures involve the choice of an appropriate threshold of the size of selected predictors. The selection of the threshold could influence the performance of variable screening in ultrahigh-dimensional data substantially and statistical accuracy in the follow-up analysis. Therefore, selecting an appropriate threshold becomes an important question of interest, both theoretically and practically.

For SIS, Fan and Lv (2008) recommended the threshold value being $\lceil n/\log(n)\rceil$ and shown its consistency with the sure screening property in numerical studies. For DC-SIS, Li, Zhong and Zhu [8] extended the criterion and chose the threshold as the multiples of $\lceil n/\log(n)\rceil$. However, these thresholds are solely depended on sample sizes, which might influence the screening results significantly as said in Li, Zhong and Zhu [8]. Rather than just fixing the threshold to be a function of sample size, Zhu et al. [18] set a prespecified value $\lceil n/\log(n)\rceil$ as the maximum for their threshold in their proposal and a data-driven cutoff value obtained by adding artificial auxiliary variables to the data. Yet the auxiliary variables added may not reflect the true relation between response and predictors. Furthermore, this might increase computation burden and makes it infeasible in ultrahigh-dimensional data analysis.

In this work, we propose an adaptive model-free sure independence screening procedure, in which the threshold is totally determined by the data. The marginal measure in the

*Corresponding author.

proposal is the p-values from the modified distance correlation test, a modified version which was introduced to correct the bias in the squared distance covariance [11]. Rather than simply choosing the top few variables with smallest p-values, we adopt the multiple testing correction to adjust the p-values and select the predictor variables under some criterions. Compared with existing methods, the most distinguishable feature is that the threshold here is totally determined by the data itself but not the sample size only. This might increase the power in detection of important variables and thus increase the statistical accuracy in further analysis. Besides, the proposed procedure is model-free and has robust performance in practical implementation.

The rest of this paper is organized as follows. In Section 2, we give some preliminaries and then demonstrate the motivation by a simple example. In Section 3, we develop a novel model-free feature screening approach with adaptive threshold. Section 4 demonstrates its finite performance by Monte Carlo simulations and Section 5 presents its application to a GAW17 dataset. A brief discussion in included in Section 6. Some extra numerical study can be found in the Appendix.

2. PRELIMINARIES AND MOTIVATION

2.1 Preliminaries

Let $Y \in R^q$ be the response variable, $X = (X_1, ..., X_p)^T$ be the predictor variables. $X_k (k = 1, \ldots, p)$ is r_k dimensional and can be grouped or categorical data here. The goal of variable screening is to identify all the variables in the predictors X that are relevant to the response variable Y. To be more formal, define the index sets of active and inactive predictors without specifying a regression model by

$$\begin{aligned}(1)\\ \mathcal{A} =& \{r : \ F(y|x) \text{ functionally depends on } X_r \text{ for some } y\},\\ \mathcal{I} =& \{r : \ F(y|x) \text{ does not functionally depends on } X_r \text{ for any } y\}.\end{aligned}$$

where $F(y|x)$ denotes the conditional distribution function of y given x. In the framework of variable screening, the target is to identify an index set that includes all indexes in $\mathcal{A}$ and includes as less as possible indexes in $\mathcal{I}$ as the sample size tends to infinity.

Suppose $\mathbf{W} = (\mathbf{X}_1, \ldots, \mathbf{X}_p, \mathbf{Y}) = \{(X_{i1}, \ldots, X_{ip}, Y_i) : i = 1, \ldots, n\}$ is a random sample from the joint distribution of X and Y. Denote the Euclidean distance between X_{ik} and X_{jk} by $a^k_{ij} = |X_{ik} - X_{jk}|_{r_k}$, and those between Y_i and Y_j by $b_{ij} = |Y_i - Y_j|_q$. Furthermore, define the doubly centered distance matrix A^k_{ij} of $\mathbf{X}_k$ by

$$(2) \qquad A^k_{ij} = a^k_{ij} - \bar{a}^k_{i\cdot} - \bar{a}^k_{\cdot j} + \bar{a}^k_{\cdot\cdot}, \quad i, j = 1, \ldots, n,$$

where

$$\bar{a}^k_{i\cdot} = \frac{1}{n}\sum_{l=1}^{n} a^k_{il}, \quad \bar{a}^k_{\cdot j} = \frac{1}{n}\sum_{s=1}^{n} a^k_{sj}, \quad \bar{a}^k_{\cdot\cdot} = \frac{1}{n^2}\sum_{s,l=1}^{n} a^k_{sl}.$$

The doubly centered distance matrix of B_{ij} of $\mathbf{Y}$ is defined similarly. Then the distance covariance and distance correlation between $\mathbf{X}_k$ and $\mathbf{Y}$ by

$$(3) \qquad \mathcal{V}^2(\mathbf{X}_k, \mathbf{Y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} A^k_{ij} B_{ij},$$

and

$$(4) \qquad \mathcal{R}(\mathbf{X}_k, \mathbf{Y}) = \frac{\mathcal{V}(\mathbf{X}_k, \mathbf{Y})}{\sqrt{\mathcal{V}(\mathbf{X}_k, \mathbf{X}_k)\mathcal{V}(\mathbf{Y}, \mathbf{Y})}},$$

if $\mathcal{V}(\mathbf{X}_k, \mathbf{X}_k)\mathcal{V}(\mathbf{Y}, \mathbf{Y}) > 0$, otherwise $\mathcal{R}^2(\mathbf{X}_k, \mathbf{Y}) = 0$.

The DC-SIS marginally ranks the importance of each predictor $\mathbf{X}_k$ by the distance correlation with $\mathbf{Y}$, namely $\mathcal{R}(\mathbf{X}_k, \mathbf{Y})$. Then the top d predictors are identified as selected variables, where d is chosen to be an integer multiple of $\lceil n/\log(n) \rceil$ as in [8].

2.2 Motivation

To demonstrate the need of an adaptive threshold selection criterion, we begin with a simple example where a linear model $Y = X\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ and X is generated from multivariate Gaussian distribution with zero mean and autoregressive covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$. More specifically, the covariance matrix Σ has entries $\sigma_{ij} = \rho^{|i-j|}, ij, = 1, \ldots, p$. In this example, we fix the sample size $n = 200$, the dimensionality $p = 100$ and $\rho = 0.5$. The sizes of the true models p_1, i.e., the numbers of non-zero coefficients, were chosen from 5 to 50 and the non-zero components of the p-vectors β were randomly chosen. We set $a = 4\log(n)/n^{1/2}$ and picked non-zero coefficients of the form $(-1)^U(a + |Z|)$, where U was drawn from a Bernoulli distribution with parameter 0.4 and Z was drawn from the standard Gaussian distribution. We repeated the above procedures for 100 times.

Figure 1 gives boxplots of the selected model size and the true positive rate for DC-SIS. As seen in the upper panel, the model size of DC-SIS remains the same since the sample size is fixed and the threshold equals to $\lceil n/\log(n) \rceil$ as recommended in [8]. However, as p_1 increases, this threshold cannot guarantee most of the true active predictors are included. Actually, it leads to an sharply decrease of the true positive rate as one can see from the lower panel of Figure 1.

3. ADAPTIVE MODEL-FREE SURE INDEPENDENCE SCREENING

In this section, we propose an adaptive model-free sure screening procedure with data-driven threshold. Unlike DC-SIS, the marginal measure we use here is the p-values of the modified distance correlation [11]. These marginal p-values are sorted and transformed into a series of order q-values $\{q_{(k)} : k = 1, \ldots, p\}$ by controlling the false discovery rate (FDR, Benjamini and Hochberg [2], Benjamini and Yekutieli

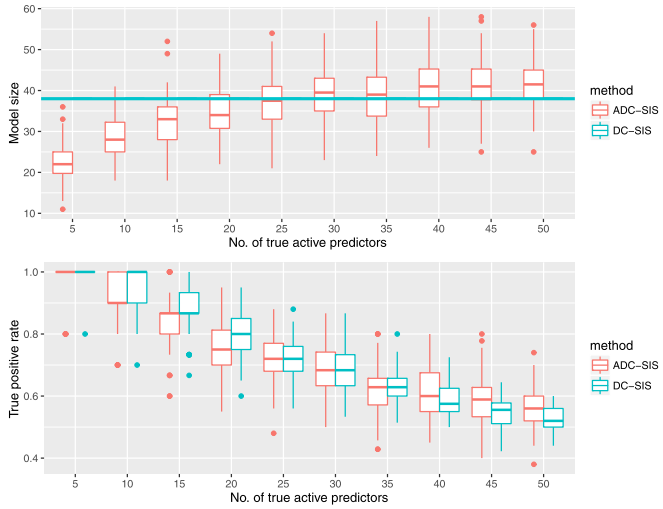

Figure 1. Motivation example: Boxplot of model size and true positive rate for DC-SIS and ADC-SIS.

[3]). Finally, the threshold is determined by the largest k such that $q_{(k)} \leq \alpha$, where α is the significant level.

The modified distance covariance was introduced to correct the bias in the sample squared distance covariance, for details see [11] and [12]. To be precise, we modify the doubly centered distance matrices A_{ij}^k, B_{ij} as

$$\tilde{A}_{ij}^k = \begin{cases} \frac{n}{n-1}(A_{ij}^k - \frac{a_{ij}^k}{n}), & i \neq j; \\ \frac{n}{n-1}(\bar{a}_{i\cdot}^k - \bar{a}_{\cdot\cdot}^k), & i = j, \end{cases}$$

$$\tilde{B}_{ij} = \begin{cases} \frac{n}{n-1}(B_{ij} - \frac{b_{ij}}{n}), & i \neq j; \\ \frac{n}{n-1}(\bar{b}_{i\cdot} - \bar{b}_{\cdot\cdot}), & i = j. \end{cases}$$

Then the modified distance covariance and modified distance correlation are given by

(5)
$$\tilde{\mathcal{V}}^2(\mathbf{X}_k, \mathbf{Y}) = \frac{1}{n(n-3)} \left(\sum_{i \neq j} \tilde{A}_{ij}^k \tilde{B}_{ij} - \frac{2}{n-2} \sum_{i=1}^{n} \tilde{A}_{ii}^k \tilde{B}_{ii} \right)$$

and

(6)
$$\tilde{\mathcal{R}}(\mathbf{X}_k, \mathbf{Y}) = \frac{\tilde{\mathcal{V}}(\mathbf{X}_k, \mathbf{Y})}{\sqrt{\tilde{\mathcal{V}}(\mathbf{X}_k, \mathbf{X}_k)\tilde{\mathcal{V}}(\mathbf{Y}, \mathbf{Y})}},$$

if $\tilde{\mathcal{V}}(\mathbf{X}_k, \mathbf{X}_k)\tilde{\mathcal{V}}(\mathbf{Y}, \mathbf{Y}) > 0$, otherwise $\tilde{\mathcal{R}}^2(\mathbf{X}_k, \mathbf{Y}) = 0$.

To obtain the p-values of $\tilde{\mathcal{R}}_n(\mathbf{X}_k, \mathbf{Y})$, we could use the nonparametric permutation procedure just like the distance covariance test [13]. However, it involves at least hundreds of resamples and thus the computation burden is heavy especially in ultrahigh-dimensional data. Alternatively, one can use the asymptotic distribution of the modified distance correlation statistic $\tilde{\mathcal{R}}(\mathbf{X}_k, \mathbf{Y})$. It has been shown that under independence the distribution of a transformation, i.e.,

(7)
$$\tau_n^k = \sqrt{\frac{n(n-3)}{2} - 1} \cdot \frac{\tilde{\mathcal{R}}(\mathbf{X}_k, \mathbf{Y})}{\sqrt{1 - \tilde{\mathcal{R}}^2(\mathbf{X}_k, \mathbf{Y})}},$$

converges to Student t distribution with degrees of freedom being $n(n-3)/2-1$, as dimensions of $\mathbf{X}_k$ and $\mathbf{Y}$ tend to infinity [11]. Although the dimensions of $\mathbf{X}_k$ and $\mathbf{Y}$ are fixed and small in this paper, the approach of using Student t asymptotic distribution is still appropriate, as shown in Appendix. Thereby, we prefer to the asymptotic t test for its computational flexibility and comparable performance with the permutation bootstrap test. That is, the p-values is given by $\Pr(T > \tau_n^k)$, where T is a Student t random variable with degrees of freedom being $n(n-3)/2-1$.

Next we adjust the p-values in multiple testing by controlling the false discovery rate (FDR). FDR control offers a way to increase power while maintaining some principled bound on error. We adopt the Benjamini-Hochberg-Yekutieli (BH) procedure [2, 3] since it performs best in very sparse cases, which matches the assumption in ultrahigh-dimensional variable screening methods. As shown in Appendix, the performance of the asymptotic t statistic is comparable with the bootstrap test when α is around 0.1. Thus the significant level used in the BH procedure is set to be 0.1 throughout this paper. This proposed procedure is thus referred as adaptive distance correlation sure independence screening (ADC-SIS) and the algorithm is summarized as follows:

Adaptive Distance Correlation Sure Independence Screening (ADC-SIS)

---

1. For each predictor $X_k (k = 1, \ldots, p)$, calculate the modified distance correlation and transform it into τ_n^k as in Equation (7);
2. Derive the p-values $p_k = Pr(T > \tau_n^k)$ for each predictors and sort them in an increasing order;
3. Derive the q-values q_r using the BH procedure, that is, $q_{(k)} = p/k \sum_{j=1}^{p} 1/j \cdot p_{(k)}$, where $p_{(k)}$ is the k-th smallest p-value;
4. For a given α, determine the threshold K by the largest k such that $q_{(k)} \leq \alpha$;
5. Select the predictors corresponding to the q-values $\{q_{(1)}, q_{(2)}, \ldots, q_{(K)}\}$.

---

4. SIMULATION STUDIES

In this section, we illustrate the variable screening performance of ADC-SIS in various examples in comparison with DC-SIS. Three measures are used to evaluate the relative performance of ADC-SIS. The first measure is the true positive rate (TPR), the proportion of the truly selected predictors among the true active predictors. The second measure is the false positive rate (FPR), which is the proportion of falsely selected predictors among the true inactive predictors. These two measures are commonly used in the

biomedical literature. Ideally, one wishes to have TPR to be close to one and FPR to be close to zero simultaneously. Besides, the size of the selected model is reported to measure the performance of ADC-SIS.

4.1 Motivation example

We evaluate the screening performance of ADC-SIS in the motivation example, which was discussed in Section 2.2. The screening results are given in Figure 1. When the number of true active predictors is small, i.e., $p_1 = 5$ or 10, the average selected model size for ADC-SIS is much smaller than those for DC-SIS, while the proportions of true active predictors among selected variables for both methods are similar. When p_1 is larger, the average proportion of true active predictors among selected variables for ADC-SIS is significantly higher than those for DC-SIS, whereas the average selected model size for ADC-SIS is slightly larger.

4.2 Example I: univariate response

In this section, we consider situations that the response variable is univariate. To imitate the real data in genetic studies, the predictors are mixtures of continuous and binary data. To be specific, the predictor X is generated from a multivariate normal distribution with $\mathrm{cov}(X_i, X_j) = 0.5^{|i-j|}$. We discrete X_{12} using its median, i.e., $X_{12} = I(X_{12} < \mathrm{median}(X_{12}))$, where $I(\cdot)$ is the indicator function. We consider the following five scenarios:

$$(1.a): Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 X_{12} + c_4\beta_4 X_{22} + \varepsilon;$$
$$(1.b): Y = c_1\beta_1 \sin(X_1) + c_2\beta_2 X_2 + c_3\beta_3 X_{12} + c_4\beta_4 X_{22} + \varepsilon;$$
$$(1.c): Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 X_{12} + c_4\beta_4 X_{22}^2 + \varepsilon;$$
$$(1.d): Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 X_{12} + c_4\beta_3 X_{22} + \varepsilon;$$
$$(1.e): Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 X_{12} X_{22} + \varepsilon;$$

where $(c_1, c_2, c_3, c_4) = (2, 0.5, 3, 2)$ means different effect sizes of predictors to the response. The non-zero regression coefficient $\beta_j (j = 1, 2, 12, 22)$ is set to $(-1)^U(a + |Z|)$, where $a = 4\log n/\sqrt{n}$ with n being sample size, $U \sim \mathrm{Bernoulli}(0.4)$ and $Z \sim \mathcal{N}(0, 1)$. The random error ϵ is generated from a standard normal distribution. For each scenarios, four different sample sizes ranging from 100 to 400 have been considered. The dimension of predictors is set to be $p = 1000$. For each sample sizes, a total of 1000 random replications have been conducted. The first three scenarios are all additive models, in which scenario (1.a) depicts linear model and the rest represents non-linear models. The last two scenarios considers interaction terms, which is a common phenomenon in genetic data analysis. While scenario (1.d) considers only interaction in continuous predictors, scenario (1.e) includes interaction between continuous and binary predictors.

Table 1 and Figure 2 present the variable screening results for ADC-SIS with sample size being 100, 200, 300 and 400. For comparison, we also obtained the screening results for DC-SIS with threshold $d = \lceil n/\log(n) \rceil$ as recommend in [8]. Specially, Table 1 lists the mean and standard error

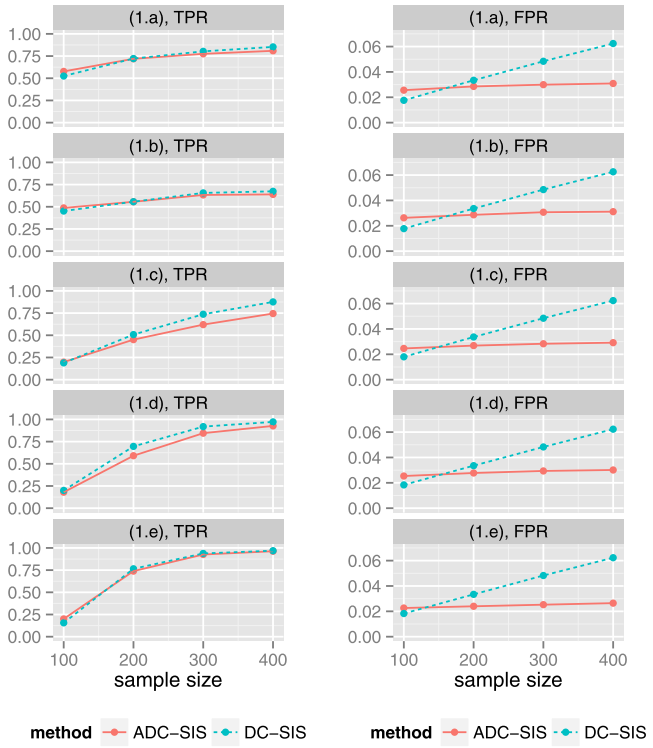

Figure 2. Univariate response example: The average true positive rate (TPR) and false positive rate (FPR) for ADC-SIS and DC-SIS with sample size being $n = 100, 200, 300$ and 400.

Table 1. Univariate response example: the mean size of the selected model for $n = 100, 200, 300$ and 400 observations for ADC-SIS. The numbers in parentheses are the corresponding standard errors. For comparison, the model size for DC-SIS, i.e., $\lceil n/\log(n) \rceil$, is included in the last row

	Sample size n			
Scenario	100	200	300	400
(1.a)	29(7.37)	32(7.46)	34(7.60)	35(7.23)
(1.b)	30(7.25)	32(7.39)	34(7.17)	35(7.44)
(1.c)	28(7.19)	30(7.15)	32(7.36)	33(7.17)
(1.d)	28(7.61)	31(7.03)	33(7.28)	34(7.17)
(1.e)	25(6.97)	28(7.10)	29(6.96)	30(7.22)
DC-SIS	21	37	52	66

of the selected model size and Figure 2 displays the plots of the average FPR and TPR bases on 1000 replications. One can see from the figure that the TPRs for both DC-SIS and ADC-SIS are close to 1 as the sample size n increases, which supports the assertion that the sure screening property is possessed by both procedures. Moreover, ADC-SIS performs competitively with DC-SIS in terms of TPR, which suggests the feasibility of using FDR to select the threshold in variable screening procedures. As sample size n increases, the FPR of DC-SIS increases rapidly as the threshold (and thus the model size) is only determined by the sample size.

At the same time, the average FPR of our proposal almost stay the same as sample size increases. Compared with DC-SIS, ADC-SIS has much smaller mean FPR when the sample size is large and comparable mean FPR when sample size is small. This indicates that ADC-SIS helps reduce false positive when more data are included, while DC-SIS couldn't control the false positive.

Table 1 clearly shows that as expected, ADC-SIS leads to an adaptive threshold with different scenarios. In particular, the average selected model size is different in scenarios (1.a)–(1.c) even if they all reflect additive effect of predictors to the response. Besides, scenario (1.e) is much smaller than the first three scenarios since the coefficient of X_2 for scenario (1.e) is much larger. In comparison, the selected model size for DC-SIS is fixed and non-adaptive to data. Furthermore, when $n = 400$, the average model size for ADC-SIS is only half of those for DC-SIS, whereas they have comparable performance in terms of TPR. Both Figure 2 and Table 1 suggest that ADC-SIS seems to be more adaptive to the data and this adaptive property seems to be helpful in screening variables.

4.3 Example II: multivariate response

In this example, we consider the multivariate response cases. To be precise, we generate $Y = (Y_1, Y_2)^T$ from multivariate normal distribution with mean $\mathbf{0}$ and variance $\Sigma = (\sigma_{ij})_{2\times 2}$, where $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = \sigma(x)$. For $\sigma(x)$, we consider the following two scenarios:

$$(2.a)\ :\sigma(x) = \sin(\beta_1^T x) \text{ where } \beta_1 = (0.8, 0.6, 0, ..., 0)^T;$$

$$(2.b)\ :\sigma(x) = \{\exp(\beta_2^T x) - 1\}/\{\exp(\beta_2^T x) + 1\}, \text{ where } \beta_2 = (2 - U_1, 2 - U_2, 2 - U_3, 2 - U_4, 0, ..., 0)^T, \text{ and } U_i (i = 1, \ldots, 4) \text{ is independently generated from Uniform[0,1].}$$

Figure 3 depicts the average TPR and FPR for ADC-SIS and DC-SIS. Although the mean TPRs for both methods are low when $n = 100$, both of them grow rapidly as sample size increases as one can see from the figure. This numerical example supports the sure screening property for ADC-SIS and DC-SIS. While ADC-SIS has comparable performance with DC-SIS in terms of TPR, the average FPR of DC-SIS is much higher than that of ADC-SIS.

Table 2 summaries the selected model sizes for ADC-SIS and DC-SIS, from which we can see that the mean model sizes of ADC-SIS is much less than those of DC-SIS. Compared with the examples with univariate response, ADC-SIS has quite different selected model size because of its adaptive property. However, the selected model size for DC-SIS is the same since the threshold value depends only on the sample size.

5. REAL DATA ANALYSIS: GAW 17 DATA

To demonstrate the practical efficiency of the ADC-SIS approach, we consider here a GAW 17 mini-exome data described in [1]. The data are a hybrid of real exome sequence data and simulated synthetic quantitative phenotypes. The sequence data are used to provide a realistic pattern of number and frequency of SNPs, whereas the simulated phenotypes provide a way to investigate relative performance as the true causal SNPs are known. For each sample, 200 replicates of the phenotypes were simulated.

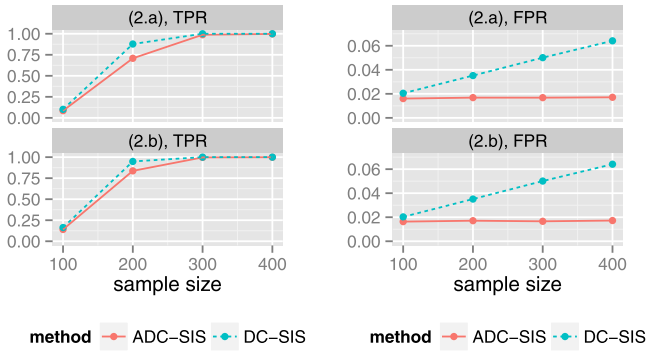

Figure 3. Multivariate response example: The average true positive rate (TPR) and false positive rate (FPR) for ADC-SIS and DC-SIS with sample size being $n = 100, 200, 300$ *and* 400.

Table 2. Multivariate response example: the mean size of the selected model for $n = 100, 200, 300$ *and* 400 *observations for ADC-SIS. The numbers in parentheses are the corresponding standard errors. For comparison, the model size for DC-SIS, i.e.,* $\lceil n/\log(n) \rceil$, *is included in the last row*

	Sample size n			
Scenario	100	200	300	400
(2.a)	17(6.52)	18(6.52)	19(6.47)	19(6.36)
(2.b)	17(6.71)	19(6.05)	19(6.14)	19(6.33)
DC-SIS	21	37	52	66

We focus here on the GAW 17 unrelated data with metric phenotype Q1 and Q2. The corresponding sequence data matrix contains information on 24,487 SNPs for $n = 697$ individuals. By construction, phenotype Q1 is correlated with 39 SNPs and has a residual heritability of 0.44, while phenotype Q2 is correlated with 72 SNPs with a relatively lower residual heritability of 0.29. Following the preprocessing procedure in [19], we have a total of 8,020 SNPs and a reduced true unique SNPs for phenotypes Q1(38) and Q2(71).

For each phenotype, ADC-SIS and DC-SIS were used for screening SNPs. The threshold in DC-SIS was fixed to $\lceil n/\log(n) \rceil = \lceil 697/\log(697) \rceil = 107$, as suggested in [8]. Figure 4 reports the screening accuracy for ADC-SIS and DC-SIS procedures in terms of true positive, which is defined as the number of truly identified active predictors.

As can be seen from Figure 4, ADC-SIS uniformly outperforms DC-SIS in terms of true positives. In particular, for phenotype Q1, the median true positives for ADC-SIS is

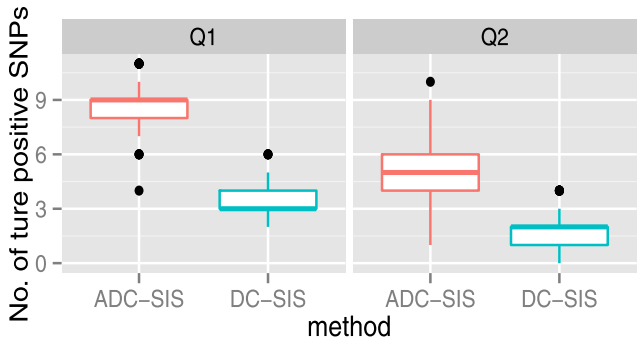

Figure 4. True positive SNPs being selected by ADC-SIS and DC-SIS for phenotypes Q1 and Q2. For DC-SIS, the threshold is fixed to be $\lceil n/\log(n)\rceil = \lceil 697/\log(697)\rceil = 107$.

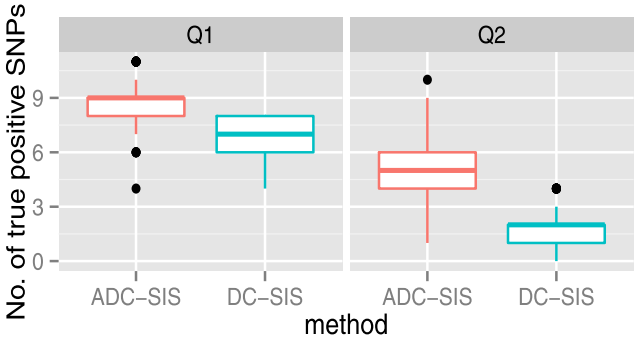

Figure 5. True positive SNPs being selected by ADC-SIS and DC-SIS for phenotypes Q1 and Q2. For DC-SIS, the threshold is chosen to be the same as those in ADC-SIS in each replication.

Table 3. GAW17 data: The mean size of the selected model for ADC-SIS. The numbers in parentheses are the corresponding standard errors. For comparison, the model size for DC-SIS, i.e., $\lceil n/\log(n)\rceil$*, is included in the last column*

phenotype	ADC-SIS	DC-SIS
Q1	651(147)	107
Q2	166(62)	107

9 while those for DC-SIS is only 3. Table 3 lists the mean and standard deviation of model sizes of the selected model for ADC-SIS. Note that for DC-SIS, the threshold is fixed ($d = 107$) and the same for phenotypes Q1 and Q2. However, the mean model sizes for phenotype Q1 is much higher than those for Q2. This coincides with the construction of these two phenotypes, i.e., Q1 is characterized by SNPs with strong effects and moderate minor allele frequencies, the true SNPs for Q2 have a very low minor allele frequencies.

To make a more fair comparison, we compare the performance when DC-SIS has the same threshold value with ADC-SIS in each replicates. The boxplots of the true positives for both two methods are shown in Figure 5. It is clearly that ADC-SIS still leads to a higher true positives even with the same threshold for this particular data.

6. DISCUSSION

In this paper, we examine the problem of the choice of selected predictor variables in ultrahigh-dimensional data analysis and propose a novel sure independence screening procedure with adaptive threshold. The proposal eliminates the predictor variables that are weakly correlated or uncorrelated with the response variable via a modified distance correlation test. Then the threshold is determined by the size that the ascending order of adjusted p-values is no more than the significant level.

The extensive simulation studies suggests the preference of our procedure and the possible consistency of the selection of the active predictors as sample size increases. Theoretical study of the sure screening property and model selection consistency for the proposal is yet to be established.

As the referees suggested, the idea of adaptive threshold could be naturally extended to feature screening procedures using correlation-based independence tests when both X_k and Y are univariate. For example, [9] proposed a feature screening procedure based on Kendall rank correlation and studies its screening properties for linear regression models and transformation regression models. In the nonparanormal graphical models, [16] proposed sparse estimation scheme based on the Spearman's rank correlation and shown it has some desirable theoretical properties. Instead of approximate p-values of distance correlation as in our proposal, it would be possible to obtain p-values directly from the asymptotic distribution. One could use technique such as Fisher's Z transform to obtain the asymptotical normality, i.e., Kendall rank correlation and Spearman's rank correlation [14, 7].

APPENDIX. COMPARISON BETWEEN PERMUTATION TEST AND ASYMPTOTIC T TEST

In this section, we study extensive simulations to compare the relative performance of permutation test based on the modified distance correlation and the asymptotic t test. Throughout this section, we study the correlation between univariate random variables X and Y.

We will start with examination of independent cases. For the distribution of X and Y, we consider the following three scenarios: (a) standard normal distribution; (b) Student t distribution with degree of freedom 1; (c) exponential distribution with mean 1. Next we will consider the cases when X and Y are linear related with correlation $\rho = 1$ or $\rho = 0.5$. Then we will examine the situations with nonlinear relationship between X and Y. To this end, seven patterns are considered: wave, trapezoid, diamond, quadratic, X shade,

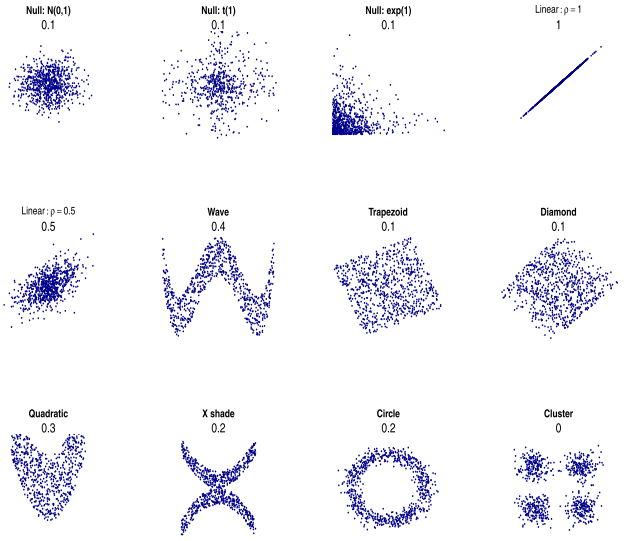

Figure 6. Several sets of (X, Y) points, with the Distance correlation coefficient of X and Y for each set.

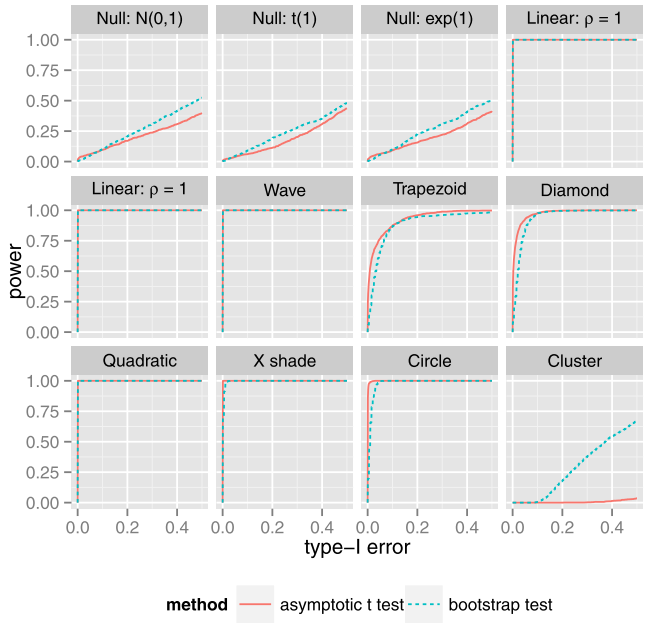

Figure 7. Plot of power for permutation test and asymptotic t test in different cases.

circle and cluster. The scatter plot for these cases with the sample size being $n = 800$ is given in Figure 6.

We compare the performance of permutation test and asymptotic t test in the three independent cases and eight dependent cases. The sample size is fixed to be 200 and the test procedure is repeated for 1000 times. Then for each $\alpha \in (0, 0.5)$, Figure 7 gives a plot of the power, i.e., $Pr(\text{p-value} < \alpha)$, versus α.

It is obviously that permutation test and asymptotic t test have similar performance when X and Y are correlated, especially when they are linear correlated. To look further into Figure 7, the asymptotic t test performs slightly better when $\alpha < 0.1$ for trapezoid and diamond patterns. When X and Y are independent, both permutation test and asymptotic t test can control the type I error. However, the asymptotic t test is more conservative when $\alpha > 0.1$.

ACKNOWLEDGEMENT

Wang's research is supported by National Natural Science Foundation of China for Excellent Young Scholar [11322108], New Century Excellent Talents Supporting Plan [12-0559], National Natural Science Foundation of China [11001280], the Research Fund for the Doctoral Program of Higher Education [20110171110037] and National Natural Science Foundation of China [11301324]. The authors thank the editor and two referees for their constructive comments, which have led to a significant improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NNSFC.

Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH grant R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org). The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.

Received 15 December 2015

REFERENCES

[1] Almasy, L., Dyer, T. D., Peralta, J. M., Kent, J. W., Charlesworth, J. C., Curran, J. E. and Blangero, J. (2011). Genetic Analysis Workshop 17 mini-exome simulation. In *BMC proceedings* **5** S2. BioMed Central Ltd.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300. MR1325392

[3] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 1165–1188. MR1869245

[4] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106**. MR2847969

[5] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911. MR2530322

[6] Fan, J., Song, R. et al. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38** 3567–3604. MR2766861

[7] Hollander, M., Wolfe, D. A. and Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons. MR3221959

[8] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107** 1129–1139. MR3010900

[9] Li, G., Peng, H., Zhang, J., Zhu, L. et al. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40** 1846–1877. MR3015046

[10] Székely, G. J., Rizzo, M. L. et al. (2009). Brownian distance covariance. *The Annals of Applied Statistics* **3** 1236–1265. MR2752127
[11] Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivariate Analysis* **117** 193–213. MR3053543
[12] Szekely, G. J. and Rizzo, M. L. (2013). Partial Distance Correlation with Methods for Dissimilarities. *arXiv preprint arXiv:1310.2926*. MR3269983
[13] Székely, G. J., Rizzo, M. L., Bakirov, N. K. et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769–2794. MR2382665
[14] Van der Vaart, A. W. (2000). *Asymptotic statistics* **3**. Cambridge University Press. MR1652247
[15] Xue, L. and Zou, H. (2011). Sure independence screening and compressed random sensing. *Biometrika* **98** 371–380. MR2806434
[16] Xue, L., Zou, H. et al. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics* **40** 2541–2571. MR3097612
[17] Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* **105** 397–411. MR2877525
[18] Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**. MR2896849
[19] Zuber, V., Silva, A. P. D. and Strimmer, K. (2012). A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies. *BMC Bioinformatics* **13** 284.

Canhong Wen
Southern China Research Center of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: wencanhong@gmail.com

Shan Zhu
Southern China Research Center of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: sarina_66@163.com

Xin Chen
Department of Statistics and Applied Probability
National University of Singapore
Singapore, SG 117546
Singapore
E-mail address: stacx@nus.edu.sg

Xueqin Wang
Southern China Research Center of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, GD 510275
China
E-mail address: wangxq88@mail.sysu.edu.cn
url: http://scrcss.sysu.edu.cn/