Published in final edited form as:

J Comput Graph Stat. 2017 ; 26(4): 803–813. doi:10.1080/10618600.2017.1328364.

A Robust Model-Free Feature Screening Method for Ultrahigh-Dimensional Data

Jingnan Xue* [Graduate Student] and

Department of Statistics, Texas A&M University, College Station, TX 77843

Faming Liang* [professor]

Department of Biostatistics, University of Florida, Gainesville, FL 32611

Abstract

Feature screening plays an important role in dimension reduction for ultrahigh-dimensional data. In this paper, we introduce a new feature screening method and establish its sure independence screening property under the ultrahigh-dimensional setting. The proposed method works based on the nonparanormal transformation and Henze-Zirkler's test; that is, it first transforms the response variable and features to Gaussian random variables using the nonparanormal transformation and then tests the dependence between the response variable and features using the Henze-Zirkler's test. The proposed method enjoys at least two merits. First, it is model-free, which avoids the specification of a particular model structure. Second, it is condition-free, which does not require any extra conditions except for some regularity conditions for high-dimensional feature screening. The numerical results indicate that, compared to the existing methods, the proposed method is more robust to the data generated from heavy-tailed distributions and/or complex models with interaction variables. The proposed method is applied to screening of anticancer drug response genes.

Keywords

Gene Screening; Henze-Zirkler's Test; Nonparanormal Transformation; Sure Independence Screening; Precision Medicine

1 Introduction

Variable selection plays an important role in high-dimensional data analysis. However, under the ultrahigh-dimensional setting, where the number of covariates may grow at an exponential rate of sample size, the current variable selection methods may not work well due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability (Fan et al., 2009). A practical approach is to use a screening procedure to reduce the dimension of feature space to a moderate scale, and then implement a variable selection method on the reduced dataset. To pursue this approach, Fan and Lv (2008) proposed the sure independence screening (SIS) method for linear regression, where the

---

* jnxue@stat.tamu.edu, faliang@ufl.edu.

features are screened based on their respective Pearson correlation coefficients with the response variable. They established the sure screening property, i.e., all active predictors can be selected with probability approaching one as the sample size increases to infinity. Fan and Song (2010) extended SIS to generalized linear models, where the features are screened based on the estimates of the regression coefficients in their respective marginal models. For nonlinear models, Hall and Miller (2009) suggested polynomial transformations of predictors, and Fan et al. (2011) suggested to estimate the nonparametric components using B-splines for each feature and then have the features screened based on their respective magnitudes of the nonparametric components.

All the above methods require the specification of a particular model structure. If the underlying model is correctly specified, these methods can perform well. However, if the underlying model is misspecified, their performance may be corrupted. Under the ultrahigh dimensional setting, specifying a correct model is usually an impossible task, and thus the model-free feature screening methods are appealing. Toward this direction, Zhu et al. (2011) proposed a sure independence ranking and screening (SIRS) method to screen significant features for multi-index models. Li et al. (2012) proposed a distance correlation sure independence screening (DC-SIS) method, where the features are screened based on their distance correlation (Székely et al., 2007) with the response variable. It is known that for two random variables, a zero distance correlation coefficient implies independence. He et al. (2013) proposed a quantile-adaptive sure independence screening (Qa-SIS) method, which employs spline approximation to model the marginal effect of each predictor at a quantile level and then have the predictors screened accordingly. A particular strength of this method is that it can handle the censored data arising in survival analysis. Recently, Cui et al. (2015) proposed a mean variance sure independence screening (MV-SIS) method, where the dependence of two random variables is measured using the mean variance of the conditional distribution function. This method is originally proposed for categorical response variables, but can be extended to the problems for which the response variable is continuous via discretization.

Although the model-free variable screening methods avoid the specification of a particular model structure, they are still based on some assumptions for the predictor and response variables, more or less. For example, DC-SIS requires both the predictors and the response variable to satisfy the sub-exponential tail probability uniformly. That is, practically, the response variable and predictors should be uniformly bounded or follow a multivariate Gaussian distribution. Qa-SIS requires the conditional quantile function's derivative to satisfy a Lipschitz condition and the conditional density function to be uniformly bounded for each feature.

In this article, we propose a new model-free feature screening method and establish its sure independence screening property under very weak conditions. The proposed method works based on the nonparanormal transformation (Liu et al., 2009) and Henze-Zirkler's test (Henze and Zirkler, 1990). It is to first transform the response variable and each of the predictors to Gaussian random variables using the nonparanormal transformation and then test the dependence between the response variable and the predictors using the Henze-Zirkler's test. Compared to the existing methods, the proposed method requires fewer

assumptions to ensure its sure independence screening property and thus performs more robustly. Our numerical studies indicate that the new method can achieve better performance when the covariates follow a heavy-tailed distribution and when the underlying true model is complex with interaction variables.

The rest of this article is organized as follows. In Section 2, we describe the proposed method and establish its sure independence screening property. In Section 3, we conduct simulation studies to evaluate the finite sample performance of the proposed method along with comparisons with the existing methods. In Section 4, we apply the proposed method to screening of anticancer drug response genes. In Section 5, we conclude the article with a brief discussion. Technical proofs are given in the Appendix.

2 Robust Feature Screening

2.1 Henze-Zirkler Sure Independence Screening

Let Y denote a continuous response variable, let $\mathbf{X} = (X_1, \dots, X_p)$ denote p continuous covariates, let n denote the sample size, and let $f(y|\mathbf{x})$ denote the conditional density function of Y given $\mathbf{X}$. Under the ultrahigh-dimensional setting, where $p = O(\exp(n^\tau))$ for some $\tau > 0$, a sparsity condition is generally needed. For example, we may assume that there are only a small number of predictors relevant to the response variable, although p can be much greater than n . Without specifying a parametric form for the regression model, we define the sets of active predictors and inactive predictors as follows:

$$D = \{k: f(y|\mathbf{x}) \text{ functionally depends on } X_k\},$$

$$I = \{k: f(y|\mathbf{x}) \text{ does not functionally depend on } X_k\}.$$

A direct identification of the active predictor set D is usually difficult or even impossible under the ultrahigh-dimensional setting. Therefore, people proposed to first identify a moderate size set of variables which has included all the elements of D , and then apply a variable selection method on this moderate size set to accurately identify D . Note that if $f(y|\mathbf{x})$ functionally depends on X_k , then Y and X_k are usually marginally dependent as well. Hence, the moderate size set of variables can be constructed by selecting only the predictors that are marginally dependent with Y , and this procedure is usually referred to as *independence screening*. In fact, under the partial orthogonality condition (Fan and Song, 2010; Huang et al., 2008), i.e., $\{X_i, i \in D\}$ is independent of $\{X_j, j \in I\}$, it can be shown that $f(y|\mathbf{x})$ functionally depends on X_k if and only if Y and X_k are marginally dependent.

To implement independence screening, we need to find a metric to measure the marginal dependence between each predictor X_k and the response variable Y . Several metrics have already been proposed, see e.g., Zhu et al. (2011), Cui et al. (2015), Li et al. (2012), and He et al. (2013). In this paper, we propose a new one with the basic idea described as follows. Let $F_Y(\cdot)$ denote the CDF of the response variable Y , and let $F_k(\cdot)$ denote the CDF of the predictor X_k . Consider the nonparanormal transformation (Liu et al., 2009)

$$T_y(Y) = \Phi^{-1}(F_y(Y)), \quad T_k(X_k) = \Phi^{-1}(F_k(X_k)), \quad k = 1, \dots, p, \quad (1)$$

where $\Phi(\cdot)$ denotes the CDF of the standard Gaussian distribution. Liu et al. (2009) imposed the nonparanormal transformation on nonparanormal random variables, but here the transformation was applied to general continuous random variables. Then it is easy to see that Y is independent of X_k if and only if $(T_y(Y), T_k(X_k))$ follows the bivariate normal distribution $N_2(\mathbf{0}, \mathbf{I}_2)$, where $\mathbf{I}_2$ denotes the 2×2 identity matrix. The latter can be tested using a multivariate normality test, e.g., Henze-Zirkler's test (Henze and Zirkler, 1990), with the known covariance structure. If $(T_y(Y), T_k(X_k))$ does not follow the distribution $N_2(\mathbf{0}, \mathbf{I}_2)$, then the Henze-Zirkler's test statistic tends to take a larger value. In practice, since F_y and F_k 's are usually unknown, we can use the estimated nonparanormal transformation in Liu et al. (2009). The estimated nonparanormal transformation has been implemented in the R package *huge*.

In summary, the proposed method consists of the following steps:

1. Transform all variables, including the response variable and predictors, to standard Gaussian random variables by the estimated nonparanormal transformation. Let's take the response variable as an example: Let

$$\tilde{T}_y(y_i) = \Phi^{-1}(\tilde{F}_y(y_i)), \quad i = 1, \dots, n,$$

where y_i denotes the i th observation of Y , $\tilde{F}_y$ is the truncated empirical distribution of Y given by

$$\tilde{F}_y(t) = \begin{cases} \delta_n & : \hat{F}_y(t) < \delta_n, \\ \hat{F}_y(t) & : \delta_n \leq \hat{F}_y(t) \leq 1 - \delta_n, \\ 1 - \delta_n & : \hat{F}_y(t) > 1 - \delta_n, \end{cases}$$

$\hat{F}_y(t) \equiv \frac{1}{n} \sum_{i=1}^n 1_{\{y_i \leq t\}}$ is the empirical distribution of Y , and

$\delta_n = \frac{1}{4}n^{-1/4}(\pi \log n)^{-1/2}$ is the default truncation parameter.

2. For each predictor X_k , calculate the Henze-Zirkler test statistic

$$\tilde{\omega}_k^* = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n e^{-\frac{\beta^2}{2} d_{ij}} - \frac{2}{n(1+\beta^2)} \sum_{i=1}^n e^{-\frac{\beta^2}{2(1+\beta^2)} d_i} + \frac{1}{1+2\beta^2}, \quad (2)$$

where β is the smoothing parameter and its optimal value is $(1.25n)^{1/6}/\sqrt{2}$, which corresponds to the optimal bandwidth for a nonparametric kernel density

estimator with Gaussian kernel (Henze and Zirkler, 1990);

$d_{ij} = (\tilde{T}_k(x_{ki}) - \tilde{T}_k(x_{kj}))^2 + (\tilde{T}_y(y_i) - \tilde{T}_y(y_j))^2$; $d_i = \tilde{T}_k^2(x_{ki}) + \tilde{T}_y^2(y_i)$; and $\tilde{T}_k(x_{ki})$ and $\tilde{T}_y(y_i)$ denote the i th realization of $\tilde{T}_k(X_k)$ and $\tilde{T}_y(Y)$, respectively.

3. Select a set of predictors with large values of $\tilde{\omega}_k^*$'s, i.e., set

$$\hat{D} = \left\{ k: \tilde{\omega}_k^* > cn^{-k}, \text{ for } 1 \leq k \leq p \right\},$$

where c and κ are predetermined threshold values.

Since c and κ are usually difficult to determine, we follow other feature screening methods to set the size of $\hat{D}$ to be $[n/\log(n)]$, where $[z]$ denotes the integer part of z . Since the proposed method employs the Henze-Zirkler test statistic to measure the dependence between the transformed response variable and predictors, we call it the Henze-Zirkler sure independence screening method or HZ-SIS for short.

2.2 Theoretical properties

To study the theoretical properties of the HZ-SIS method, we first describe how the HZ-test statistic $\tilde{\omega}_k^*$ in (2) is derived. Define

$$\omega_k = \int_{\mathbb{R}^2} \left| \phi_k(\mathbf{t}) - \exp\left(-\frac{1}{2}\mathbf{t}'\mathbf{t}\right) \right|^2 \varphi_{\beta}(\mathbf{t}) d\mathbf{t}, \quad k = 1, 2, \dots, p,$$

where $\phi_k(\mathbf{t})$ is the characteristic function of $\Phi^{-1}(F_k(X_k))$, $\Phi^{-1}(F_y(Y))$, and $\varphi_{\beta}(\mathbf{t})$ is the density function of $N(0, \beta^2 I_2)$. Recall that $\exp\left(-\frac{1}{2}\mathbf{t}'\mathbf{t}\right)$ is the characteristic function of $N(0, I_2)$. Therefore, ω_k can be viewed as the averaged difference between the characteristic function of the transformed variables and the characteristic function of $N(0, I_2)$. It is easy to verify that ω_k equals zero if and only if X_k and Y are marginally independent.

Given observations $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{X}_i^T = \{x_{1i}, \dots, x_{pi}\}$ denotes the predictor variables in the i th observation, we first use the truncated empirical distribution to estimate the CDF for each variable. In order to estimate ω_k , we re-express it in the following form (Henze and Zirkler, 1990) by some algebra:

$$\omega_k = E \left[e^{-\frac{\beta^2}{2} \left[\left(\Phi^{-1}(F_k(X_k)) - \Phi^{-1}(F_k(X'_k)) \right)^2 + \left(\Phi^{-1}(F_y(Y)) - \Phi^{-1}(F_y(Y')) \right)^2 \right]} - \frac{\beta^2}{2(1+\beta^2)} \left(\Phi^{-1}(F_k(X_k))^2 + \Phi^{-1}(F_y(Y))^2 \right) + \frac{1}{1+2\beta^2} \right],$$

where (X'_k, Y') is an independent copy of (X_k, Y) . With this representation, ω_k can be estimated using a V -statistic, which leads to the HZ-test statistic used in (2).

Next, we study the sure screening property of the HZ-SIS method. As mentioned previously, compared to the existing methods, HZ-SIS requires fewer assumptions for establishing its sure screening property. The assumptions are given as follows.

(C1) There exist positive constants $c > 0$ and $0 < \kappa < 1/4$ such that $\min_{k \in D} \omega_k \geq 2cn^{-\kappa}$.

(C2) The dimension $p = O(\exp(n^\tau))$ for some constant $0 < \tau < \frac{1-4\kappa}{3}$.

Assumption (C1) can be viewed as a regularity condition for sure screening methods, which assumes that the minimum true signal cannot be too weak to be detectable for a given sample size, although it can gradually diminish to zero as the sample size increases to infinity. A similar assumption has been used in other methods, see e.g., Li et al. (2012) and Cui et al. (2015). Assumption (C2) allows an exponential growth of the dimension p as a function of the sample size. It is also regular for ultrahigh dimensional methods. To establish the sure screening property for HZ-SIS, the key step is to establish an exponential probability bound for $|\omega_k^* - \omega_k|$. The following lemma presents such an exponential probability bound with the proof given in the Appendix.

Lemma 1.

If the truncation parameter $\delta_n = \left(\frac{m}{4n^2} \sqrt{2\pi m \log n} \right)^{-1}$, where $m = \frac{2}{3} - \frac{2\kappa}{3}$, then there exist positive constants $c_1 > 0$ such that

$$P\left(\max_{1 \leq k \leq p} |\tilde{\omega}_k^* - \omega_k| > cn^{-k}\right) \leq O\left(\exp\left(-c_1 n^{\frac{1-4\kappa}{3}}\right)\right),$$

for $1 < k \leq p$.

Here we set $m = 1/2$ as the default value for the HZ-SIS method and this default value has been used in all examples of this paper. Based on this lemma, we establish the sure screening property for HZ-SIS in the following theorem.

Theorem 1.

Under conditions (C1) and (C2), we have

$$P(D \subseteq \hat{D}) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Proof. If $D \not\subseteq \hat{D}$, there must exist $k' \in D$ such that $\tilde{\omega}_{k'}^* - cn^{-k}$. Recall that $\omega_k > 2cn^{-k}$ for all $k \in D$. Therefore, we have $|\tilde{\omega}_{k'}^* - \omega_{k'}| > cn^{-k}$. This indicates that $\{D \not\subseteq \hat{D}\} \subseteq \{\text{there exist } k' \in D \text{ such that } |\tilde{\omega}_{k'}^* - \omega_{k'}| > cn^{-k}\}$. Consequently,

$$\begin{aligned}
 P(D \subseteq \hat{D}) &= 1 - P(D \not\subseteq \hat{D}) \\
 &\geq 1 - P(\text{there exist } k' \in D \text{ such that } |\tilde{\omega}_{k'}^* - \omega_{k'}| > cn^{-k}) \\
 &\geq 1 - P\left(\max_{1 \leq k \leq p} |\tilde{\omega}_k^* - \omega_k| > cn^{-k}\right) \geq 1 - O\left\{p \exp\left[-c_1 n^{\frac{1-4k}{3}}\right]\right\} \\
 &= 1 - O\left\{\exp\left[c_1 n^{\frac{1-4k}{3}} + \log(p)\right]\right\} \\
 &= 1 - O\left\{\exp\left[-c_1 n^{\frac{1-4k}{3}} + c_2 n^\tau\right]\right\} = 1 - o(1),
 \end{aligned}$$

which concludes the proof.

3 Simulation Studies

In this section, we used three simulated examples to assess the finite sample performance of HZ-SIS along with comparisons with SIS (Fan and Lv, 2008), DC-SIS (Li et al., 2012), Qa-SIS (He et al., 2013) and MV-SIS (Cui et al., 2015). In addition, NIS (Fan et al., 2011) was implemented for an additive model (Example 3.1), and the sliced inverse regression for interaction detection (SIRI) method (Jiang and Liu, 2014) was implemented for a model with interaction terms (Example 3.2 & 3.3). For each example, we generated 100 independent datasets and summarized the performance of the methods in a few statistics. These statistics include the minimum size of $\hat{D}$ needed to cover all active variables, which is denoted by MSD for short; and for the given size $\nu_n = \lfloor n/\log(n) \rfloor$ of $\hat{D}$, the proportion of $\hat{D}$ covering a single active predictor X_k (denoted by P_k), and the proportion of $\hat{D}$ covering all active variables (denoted by P_a). The reason for choosing the above statistics is that in practice, we usually specify the size ν_n of $\hat{D}$ instead of the thresholding value cn^{-k} for feature screening.

3.1 An Additive Model Example

This example is adopted from Cui et al. (2015). Let

$$f_1(x) = -\sin(2x), \quad f_2(x) = x^2 - \frac{12}{25}, \quad f_3(x) = x, \quad f_4(x) = e^{-x} - \frac{2}{5}\sinh\left(\frac{5}{2}\right),$$

and consider the additive model

$$Y = 3f_1(X_1) + f_2(X_2) - 1.5f_3(X_3) + f_4(X_4) + \varepsilon,$$

where the error term ε follows the $t(1)$ distribution, i.e., the Cauchy distribution. For the predictors, we consider two different distributions:

1. X_k 's, $k = 1, \dots, p$, are generated independently from the distribution $t(4)$;
2. X_k 's, $k = 1, \dots, p$, are generated independently from the Uniform $[-2.5, 2.5]$.

We set $(n, p) = (200, 2000)$ and repeated each case for 100 times. In Qa-SIS, we set $\tau = 0.5$ and the number of basis $d_n = \lceil n^{\frac{1}{5}} \rceil = 3$. In NIS, we took the number of basis $d_n = \lceil n^{\frac{1}{5}} \rceil + 2 = 5$. In MV-SIS, we discretized each predictor into a four-categorical variable using the first, second and third quartiles as knots. For MV-SIS, the same discretization method has been used in all examples of this paper. The results are summarized in Table 1.

Table 1 shows that when the predictors are generated from $t(4)$, a heavy-tailed distribution, HZ-SIS performs best, followed by MV-SIS, DC-SIS and Qa-SIS. This result, combined with the fact that HZ-SIS requires fewer assumptions for the sure screening property, indicates that HZ-SIS is a more robust feature screening method than the existing ones. When the predictors are generated from the uniform distribution, for which the support is bounded, HZ-SIS still performs reasonably well. In this case, it is comparable with MV-SIS and Qa-SIS, but much better than DC-SIS, NIS and SIS.

For the case where the predictors are generated from $t(4)$ distribution, we also plotted histograms of the screening indices for each method. Specifically, for each method, we first combined the screening indices calculated for the 100 datasets. Then we drew a histogram using all 400 indices of active predictors and a histogram using 600 indices of inactive predictors. The latter were randomly selected from a total of 199,600 (100×1996) indices. Finally, the two histograms were merged into the same plot and differentiated by color. As shown in Figure 1, the two histograms produced by HZ-SIS have the smallest overlapping area. This again indicates the superiority of HZ-SIS in identifying active features.

3.2 A Model with Interaction Variables

This example illustrates the performance of HZ-SIS for the models with interaction variables. Let

$$Y = 0.5 + \frac{10X_1}{1 + X_{50}^2} + \varepsilon,$$

The covariates $\mathbf{X} = (X_1, \dots, X_p)^T$ are generated from a multivariate normal distribution with mean $\mathbf{0}$ and the covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. For the error term ε , we considered two cases: (i) ε follows $N(0, 1)$ distribution. (ii) ε follows the Cauchy distribution. We set $(n, p) = (200, 1000)$ and generated 100 datasets for each case.

Jiang and Liu (2014) recently proposed a procedure, sliced inverse regression for interaction detection (SIRI), to conduct high dimensional variable selection for the models with interaction terms. Instead of building a predictive model for the response given combinations of predictors, this procedure works by modeling the conditional distribution of predictors

given the response. Since this procedure includes a screening step, we also implemented this procedure and denoted it as SIRI-SIS. In SIRI-SIS, we used a fixed slicing scheme with 10 slices of equal size ($H=10$). In Qa-SIS procedure, we set $\tau = 0.4$ and the number of basis $d_n = \lceil n^{\frac{1}{5}} \rceil = 3$. The results are summarized in Table 2.

Table 2 indicates that in the case where the random error is generated from normal, all methods can detect X_1 with ease. However, when it comes to detecting X_{50} , HZ-SIS and SIRI-SIS substantially outperforms other methods. For the case where the random error follows the Cauchy distribution, we have a similar conclusion. In addition, our method performs slightly better than SIRI-SIS in this case.

To understand the performance of these methods, we show in Figure 2 the scatter plots of the transformed predictors $\tilde{T}_1(X_1)$, $\tilde{T}_{50}(X_{50})$ and $\tilde{T}_{100}(X_{100})$ versus the transformed response variable $(\tilde{T}_y(Y))$ for a dataset generated in case 1. Given the reference scatter plot of $(\tilde{T}_{100}(X_{100}), \tilde{T}_y(Y))$ for which the theoretical joint distribution is $N(0, I_2)$, we can see that the joint distributions of $(\tilde{T}_1(X_1), \tilde{T}_y(Y))$ and $(\tilde{T}_{50}(X_{50}), \tilde{T}_y(Y))$ substantially deviate from $N(0, I_2)$, and thereby HZ-test is powerful in detecting the dependence of Y on X_1 and X_{50} . However, not all other methods work well for this example. As indicated by the values of P_2 reported in Table 2, SIS and Qa-SIS essentially fail to detect the dependence of Y on X_{50} , and DC-SIS and MV-SIS have only a limited success probability in detecting this dependence.

3.3 A Complex Model with More Interaction Variables

This example illustrates the performance of HZ-SIS for more complex models. Let

$$Y = 1 + A[10X_1 + \exp(X_2^2 + 3X_3)] + 10\frac{X_5}{2+X_6} + 3(X_7 + X_8)^2 + \varepsilon,$$

where A is generated from the set $\{-1, 1\}$ with equal probability, X_k 's are independently generated from $t(4)$ distribution. For the error term ε , we considered two cases: (i) ε follows $N(0, 1)$ distribution. (ii) ε follows the Cauchy distribution. This model is complex, containing more interaction variables than the previous model. We set $(n, p) = (400, 1000)$ and repeated each case 100 times.

In SIRI-SIS, we used a fixed slicing scheme with 10 slices of equal size ($H=10$). In Qa-SIS procedure, we set $\tau = 0.4$ and the number of basis $d_n = \lceil n^{\frac{1}{5}} \rceil = 3$. The results are summarized in Table 3. In both case, HZ-SIS has an overall better performance against any other methods.

4 Screening of Anticancer Drug Response Genes

Recent advances in high-throughput biotechnologies, such as microarray, sequencing technologies and mass spectrometry, have provided an unprecedented opportunity for biomarker discovery. Molecular biomarkers can not only facilitate disease diagnosis, but

also reveal underlying, biologically distinct, patient subgroups with different sensitivities to a specific therapy. The latter is known as disease heterogeneity, which is often observed in complex diseases such as cancer. For example, molecularly targeted cancer drugs are only effective for patients with tumors expressing targets (Gru'nwald and Hidalgo, 2003; Buzdar, 2009). The disease heterogeneity has directly motivated the development of precision medicine, which aims to improve patient care by tailoring optimal therapies to an individual patient according to his/her molecular profile and other clinical characteristics.

Toward the ultimate goal of precision medicine, i.e., selecting right drugs for individual patients, a recent large-scale pharmacogenomics study, namely, cancer cell line encyclopedia (CCLE), has screened multiple anticancer drugs over hundreds of cell lines in order to elucidate the response mechanism of anticancer drugs. The dataset consists of the dose-response data for 24 chemical compounds across over 479 cell lines. For each cell line, it consists of the expression data of 18,988 genes. The dataset is publicly available at www.broadinstitute.org/ccle. Our goal is to screen the genes that respond to each chemical compounds, which will facilitate the followed analysis for identification of anticancer drug response genes. In our analysis, we used the area under the dose-response curve, which is termed as activity area in Barretina et al. (2012), to measure the sensitivity of drug to a given cell line. Compared to other measurements, such as IC_{50} and EC_{50} , the activity area could capture the efficacy and potency of a drug simultaneously.

The drug topotecan (trade name Hycamtin) is a chemotherapeutic agent that is a topoisomerase inhibitor. It is a synthetic, water-soluble analog of the natural chemical compound camptothecin and has been used to treat ovarian cancer, lung cancer and other cancer types. After GlaxoSmithKline received final FDA approval for Hycamtin Capsules in 2007, topotecan became the first topoisomerase I inhibitor for oral use. Table 4 lists the top 10 important genes selected for topotecan by HZ-SIS. For comparison, the table also includes the top 10 genes selected by SIS, DC-SIS, MV-SIS and Qa-SIS.

In the procedure, we set $\tau = 0.5$ and the number of basis $d_n = \lceil n^{\frac{1}{5}} \rceil = 3$. For topotecan, the gene SLFN11 has been recognized as a very important predictor for the sensitivity of topotecan (Barretina et al., 2012; Zoppoli et al., 2012). HZ-SIS ranks it No. 10. In addition to SLFN11, Wang et al. (2014) found the strong relevance of HMGB2 and BCLAF1 to topotecan. HZ-SIS ranks these two genes No. 2 and No. 4, respectively. DC-SIS has a similar performance to HZ-SIS for the drug topotecan, while the other methods do not.

The drug 17-AAG is a derivative of the antibiotic geldanamycin that is being studied in the treatment of cancer, specific young patients with certain types of leukemia or solid tumors, especially kidney tumors. 17-AAG works by inhibiting the gene HSP90, which is expressed in those tumors, and belongs to the family of drugs called antitumor antibiotics. Table 5 reports the top 10 genes ranked by different methods for 17-AAG. According to Hadley and Hendricks (2014) and Barretina et al. (2012), the gene NQO1 is the top predictive biomarker for 17-AAG. HZ-SIS ranks it first among all genes. DC-SIS and MV-SIS also rank it first.

In the supplementary material, we gave the top 10 genes selected by HZ-SIS for each of the rest 22 drugs. The results are promising. For example, Barretina et al. (2012) and Zoppoli et

al. (2012) reported that the gene SLFN11 is predictive of treatment response for both the drug topotecan and irinotecan. As shown in the supplementary material, SLFN11 is also among the top 10 response genes for the drug irinotecan.

5 Discussion

This paper has proposed HZ-SIS as a new model-free feature screening method, and established its sure screening property under very weak conditions. The HZ-SIS method contains two components, nonparanormal transformation and HZ-test. The numerical examples indicate that, compared to the existing methods, HZ-SIS can achieve better performance when the covariates follow a heavy-tailed distribution and when the underlying true model is complex with interaction variables. The reason why HZ-SIS can achieve such a robust performance can be understood from two aspects. First, HZ-SIS does not require any extra conditions except for two regularity conditions that are generally required for high-dimensional feature screening. Second, the truncated empirical CDF estimator used in the estimated nonparanormal transformation helps to reduce the effect of extreme observations.

In HZ-SIS, the HZ-test is employed to test the normality of the nonparanormally transformed data. Other than the HZ-test, other multivariate normality tests, such as Szekeley-Rizzo's goodness-of-fit test (Székely and Rizzo, 2005) and Mardia's skewness and kurtosis tests (Mardia, 1970), can also be applied here. Since none of the tests are universally superior, a combination of different tests might produce a higher power. For example, if we view the feature screening problem as a clustering problem (i.e., separating the active and inactive predictors), the ensemble averaging clustering method by Liang (2008) can be used here to aggregate the results from different tests.

Henze and Zirkler (1990) showed that under the null hypothesis that the testing data are drawn from a multivariate Gaussian distribution, the HZ-test statistic follows a log-normal distribution. This implies that $\tilde{\omega}_k^*$ approximately follows a log-normal distribution if Y and X_k are independent, although a rigorous theoretical justification is still needed to account for the effect caused by the estimation error of the nonparanormal transformation. Hence, compared to the existing variable screening methods, HZ-SIS will have an added advantage that the relevance of an individual predictor to the response variable can be measured with p -value, and thus many of the existing multiple hypothesis tests can be applied to assist feature screening. In particular, we can use the multiple hypothesis testing approach to determine the size of the selected set $\hat{D}$ by controlling the false discovery rate.

Finally, we note that in practice we often encounter into the case where there are strong correlations among predictors. To handle this problem, Fan and Lv (2008) and Zhu et al. (2011) proposed iterative versions for the SIS and SIRS algorithms, respectively. SIS works on the linear regression, so the residual of the response can be used to rule out certain strongly correlated predictors via iterations. SIRS works on multi-index models, where the predictors function through a linear combination, so the residual of predictors, i.e., the projection of the remaining predictors onto the orthogonal complement space of the already selected predictors, can be used for iterations. Similarly, for HZ-SIS, the correlation among predictors can be accounted for in the following procedure:

- (1) Apply HZ-SIS to the data and select $p_{(1)}$ predictors, where $p_{(1)} < [n/\log(n)]$. Denote the set of selected predictors by $A_{(1)}$.
- (2) Treat each predictor $X_k \in A_{(1)}$ as a response variable and conduct regression analysis on the predictors included in the set $A_{(1)}$ using a regularization method, e.g., Lasso(Tibshirani, 1996), SCAD(Fan and Li, 2001) or MCP(Zhang, 2010). Calculate the residual of the regression.
- (3) Apply HZ-SIS to the residuals calculated in step (2). Suppose that $p_{(2)}$ residuals are selected, and the corresponding set of variables is denoted by $A_{(2)}$. Update the totally selected predictor set by $A_{(1)} \cup A_{(2)}$.
- (4) Repeat steps (2) and (3) $m - 1$ times until the totally selected number of predictors $p_{(1)} + \dots + p_{(m)}$ exceeds the pre-specified number $[n/\log(n)]$. The finally selected predictor set is $A_{(1)} \cup \dots \cup A_{(m)}$.

It is obvious that if a predictor is strongly correlated with some already selected predictors, it will not be selected by this procedure. This increases the chance to select the true predictors that are only weakly dependent on the response. In general, we suggest to work on the original predictors in step (2). However, for some problems, e.g., the predictors are bounded, we might choose to work on the nonparanormally transformed predictors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Liang's research was partially supported by grants DMS- 1545202, DMS-1612924 and R01-GM117597. The authors thank the editor, associate editor and two referees for their constructive comments which have led to significant improvement of this paper.

Appendix A Proof of Lemma 1

Define

$$\omega_k^* = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n e^{-\frac{\beta^2}{2} E_{ij}} - \frac{2}{n(1+\beta^2)} \sum_{i=1}^n e^{-\frac{\beta^2}{2(1+\beta^2)} E_i} + \frac{1}{1+2\beta^2},$$

where

$$E_{ij} = \left(\Phi^{-1}(F_k(x_{ki})) - \Phi^{-1}(F_k(x_{kj})) \right)^2 + \left(\Phi^{-1}(F_y(y_i)) - \Phi^{-1}(F_y(y_j)) \right)^2,$$

$$E_i = \Phi^{-1}(F_k(x_{ki}))^2 + \Phi^{-1}(F_y(y_i))^2.$$

For any $\varepsilon > 0$, we have

$$\begin{aligned}
P(|\tilde{\omega}_k^* - \omega_k| > \varepsilon) &= P(|\tilde{\omega}_k^* - \omega_k^* + \omega_k^* - \omega_k| > \varepsilon) \\
&\leq P(|\tilde{\omega}_k^* - \omega_k^*| > \frac{\varepsilon}{2}) + P(|\omega_k^* - \omega_k| > \frac{\varepsilon}{2}).
\end{aligned}$$

For simplicity, in what follows we let $\tilde{T}_k(x) \equiv \Phi^{-1}(\tilde{F}_k(x))$, $T_k(x) \equiv \Phi^{-1}(F_k(x))$ and $g_j \equiv T_j^{-1}$.

For the first term, we have

$$\begin{aligned}
P(|\tilde{\omega}_k^* - \omega_k^*| > \frac{\varepsilon}{2}) &\leq P\left(\frac{1}{n^2} \sum_{i,j} e^{\left| -\frac{\beta^2}{2} D_{ij} - e^{-\frac{\beta^2}{2} E_{ij}} \right|} > \frac{\varepsilon}{4}\right) \\
&\quad + P\left(\frac{1}{n} \sum_i e^{\left| -\frac{\beta^2}{2(1+\beta^2)} D_i - e^{-\frac{\beta^2}{2(1+\beta^2)} E_i} \right|} > \frac{1+\beta^2}{2} \frac{\varepsilon}{4}\right) \\
&\leq P\left(\frac{1}{n^2} \sum_{i,j} \frac{\beta^2}{2} |D_{ij} - E_{ij}| > \frac{\varepsilon}{4}\right) + P\left(\frac{1}{n} \sum_i \frac{1}{2} |D_i - E_i| > \frac{1+\beta^2}{2} \frac{\varepsilon}{4}\right) \\
&= P\left(\frac{1}{n^2} \sum_{i,j} |D_{ij} - E_{ij}| > \frac{\varepsilon}{2\beta^2}\right) + P\left(\frac{1}{n} \sum_i |D_i - E_i| > \frac{(1+\beta^2)\varepsilon}{4}\right).
\end{aligned}$$

Note that we only deal with $P(\frac{1}{n^2} \sum_{i,j} |D_{ij} - E_{ij}| > \frac{\varepsilon}{2\beta^2})$ because $P(\frac{1}{n} \sum_i |D_i - E_i| > \frac{(1+\beta^2)\varepsilon}{4})$

can be calculated in a similar way. First, we calculate

$$\begin{aligned}
D_{ij} - E_{ij} &= (\tilde{T}_k(x_{ki}) - \tilde{T}_k(x_{kj}))^2 + (\tilde{T}_y(y_i) - \tilde{T}_y(y_j))^2 \\
&\quad - (T_k(x_{ki}) - T_k(x_{kj}))^2 - (T_y(y_i) - T_y(y_j))^2 \\
&= (\tilde{T}_k(x_{ki}) - T_k(x_{ki}))^2 + (\tilde{T}_k(x_{kj}) - T_k(x_{kj}))^2 - 2(\tilde{T}_k(x_{ki})(\tilde{T}_k(x_{kj}) - (T_k(x_{ki})(T_k(x_{kj})) \\
&\quad + (\tilde{T}_y(y_i) - T_y(y_i))^2 + (\tilde{T}_y(y_j) - T_y(y_j))^2 - 2(\tilde{T}_y(y_i)(\tilde{T}_y(y_j) - T_y(y_i)(T_y(y_j)) \\
&= (\tilde{T}_k(x_{ki}) - (T_k(x_{ki}))(\tilde{T}_k(x_{ki}) + (T_k(x_{ki})) + (\tilde{T}_k(x_{kj}) - (T_k(x_{kj}))(\tilde{T}_k(x_{kj}) + (T_k(x_{kj})) \\
&\quad - 2(\tilde{T}_k(x_{ki}) - (T_k(x_{ki}))(\tilde{T}_k(x_{kj}) - (T_k(x_{kj})) - 2(\tilde{T}_k(x_{ki}) - (T_k(x_{ki}))T_k(x_{kj})) \\
&\quad - 2(\tilde{T}_k(x_{kj}) - T_k(x_{kj}))T_k(x_{ki}) + (\tilde{T}_y(y_i) - T_y(y_i))(\tilde{T}_y(y_i) + T_y(y_i)) \\
&\quad + (\tilde{T}_y(y_j) - T_y(y_j))(\tilde{T}_y(y_j) + T_y(y_j)) - 2(\tilde{T}_y(y_i) - T_y(y_i))(\tilde{T}_y(y_j) - T_y(y_j)) \\
&\quad - 2(\tilde{T}_y(y_i) - T_y(y_i))T_y(y_j) - 2(\tilde{T}_y(y_j) - T_y(y_j))T_y(y_i).
\end{aligned}$$

Among the ten terms, $(\tilde{T}_k(x_{ki}) - (T_k(x_{ki}))(\tilde{T}_k(x_{kj}) - (T_k(x_{kj}))$ and

$(\tilde{T}_y(y_i) - T_y(y_i))(\tilde{T}_y(y_j) - T_y(y_j))$ are of a higher order, and the other terms share the same

order. Hence, we only consider the probability $P(\frac{1}{n^2} \sum_{i,j} |(\tilde{T}_y(y_i) - T_y(y_i))(T_y(y_j))| > \frac{\varepsilon}{20\beta^2})$.

Define the event $\mathbb{A}_n$ as

$$\mathbb{A}_n \equiv \left\{ g_y(-n^d) \leq y_1, \dots, y_n \leq g_y(n^d) \right\}.$$

Since for the standard Gaussian random variable Z ,

$$P(Z > t) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}t^2}}{t}, \quad \text{if } t > 1, \quad (3)$$

we have

$$P(\mathbb{A}_n^c) \leq \sum_{i=1}^n 2P(y_i > g_y(-n^d)) \leq 2n \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}n^{2d}}}{n^d} = \sqrt{\frac{2}{\pi}} n^{1-d} \exp\left\{-\frac{1}{2}n^{2d}\right\}.$$

Therefore,

$$\begin{aligned} & P\left(\frac{1}{n^2} \sum_{i,j} \left| (\tilde{T}_y(y_i) - T_y(y_i)) T_y(y_j) \right| > \frac{\varepsilon}{20\beta^2}\right) \\ & \leq P\left(\frac{1}{n^2} \sum_{i,j} \left| (\tilde{T}_y(y_i) - T_y(y_i)) T_y(y_j) \right| > \frac{\varepsilon}{20\beta^2}, \mathbb{A}_n\right) + P(\mathbb{A}_n^c) \\ & \leq P\left(\frac{1}{n^2} \sum_{i,j} \left| (\tilde{T}_y(y_i) - T_y(y_i)) T_y(y_j) \right| > \frac{\varepsilon}{20\beta^2}, \mathbb{A}_n\right) + \sqrt{\frac{2}{\pi}} n^{1-d} \exp\left\{-\frac{1}{2}n^{2d}\right\}. \end{aligned}$$

For simplicity, henceforth, we let $\Delta_{ij} = (\tilde{T}_y(y_i) - T_y(y_i)) T_y(y_j)$.

Set the truncation parameter $\delta_n = \frac{1}{4n^{\frac{2}{d}} \sqrt{2\pi m \log n}}$, $m < 1$ and split the interval $[g_y(-n^d), g_h(n^d)]$

into

$$\mathbb{M}_n = (g_y(-\sqrt{m \log n}), g_h(\sqrt{m \log n}))$$

and

$$\mathbb{E}_n = [g_y(-n^d), g_y(-\sqrt{m \log n})] \cup [(g_y(\sqrt{m \log n}), g_y(n^d)].$$

Therefore,

$$P\left(\frac{1}{n^2} \sum_{i,j} |\Delta_{ij}| > \frac{\varepsilon}{20\beta^2}, \mathbb{A}_n\right) \leq P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{E}_n \cup y_j \in \mathbb{E}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}\right) + P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}\right).$$

We now analyze these two terms separately.

From Lemma 12.3 of Abramovich et al. (2006), if we let $\Phi^{-1}(\eta)$ denote the upper η th percentile of the standard Gaussian distribution, for $\eta \in [0, 0.99]$ we have

$$\Phi^{-1}(\eta) = \sqrt{2 \log \frac{1}{1-\eta}} - r(\eta), \quad r(\eta) \in [0, 1.5].$$

Based on this lemma, we can show

$$|\tilde{T}_y(t)| < \Phi^{-1}(1 - \delta_n) = \sqrt{2 \log \frac{1}{\delta_n}} - r(1 - \delta) \leq \sqrt{2 \left[\frac{m}{2} \log(n) + \log(4\sqrt{2\pi m \log n}) \right]} < \sqrt{\log n},$$

for any $t \in \mathbb{R}$, provided that n is sufficiently large.

Then we can bound Δ_{ij} under $\mathbb{A}_n$:

$$|\Delta_{ij}| = |\tilde{T}_{y(y_i)} - T_{y(y_i)}| |T_{y(y_j)}| \leq (|\tilde{T}_{y(y_i)}| + |T_{y(y_i)}|) |T_{y(y_j)}| \leq (\sqrt{\log(n)} + n^d) n^d < 2n^{2d},$$

if n is sufficiently large. Therefore,

$$\begin{aligned} P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{E}_n \cup y_j \in \mathbb{E}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}\right) &\leq P\left(\frac{1}{n^2} \sum_{i,j} 1_{\{y_i \in \mathbb{E}_n \cup y_j \in \mathbb{E}_n\}} > \frac{\varepsilon}{80n^{2d}\beta^2}\right) \\ &\leq P\left(\frac{1}{n^2} \sum_{i,j} 1_{\{y_i \in \mathbb{E}_n\}} + 1_{\{y_j \in \mathbb{E}_n\}} > \frac{\varepsilon}{80n^{2d}\beta^2}\right) \\ &= P\left(\frac{1}{n} \sum_i 1_{\{y_i \in \mathbb{E}_n\}} > \frac{\varepsilon}{160n^{2d}\beta^2}\right) \\ &\leq P\left(\frac{1}{n} \sum_i 1_{\{y_i \in \mathbb{E}_n\}} - P(y_i \in \mathbb{E}_n) > \frac{\varepsilon}{160n^{2d}\beta^2} - n^{-\frac{m}{2}}\right), \end{aligned}$$

where the last inequality follows from the fact that

$$P(y_i \in \mathbb{E}_n) \leq 2P(y_i > g_h(\sqrt{m \log n})) \leq 2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} m \log n} \leq n^{-\frac{m}{2}}.$$

Recall $\varepsilon = cn^{-k}$ and assume $k + 2d < \frac{m}{2}$ we have

$$\frac{\varepsilon}{160n^{2d}\beta^2} - n^{-\frac{m}{2}} = \frac{cn^{-(k+2d)}}{160\beta^2} - n^{-\frac{m}{2}} \geq \frac{cn^{-(k+2d)}}{200\beta^2},$$

if n is sufficiently large. Further, we have

$$\begin{aligned}
P\left(\frac{1}{n} \sum_i 1_{\{y_i \in \mathbb{E}_n\}} - P(y_i \in \mathbb{E}_n) > \frac{\varepsilon}{160n^{2d}\beta^2} - n^{-\frac{m}{2}}\right) \\
\leq P\left(\frac{1}{n} \sum_i 1_{\{y_i \in \mathbb{E}_n\}} - P(y_i \in \mathbb{E}_n) > \frac{cn^{-(k+2d)}}{200\beta^2}\right) \\
\leq \exp\left\{-2n \frac{c^2 n^{-2(k+2d)}}{40000\beta^4}\right\} = \exp\left\{-\frac{c^2}{20000\beta^4} n^{1-2(k+2d)}\right\},
\end{aligned}$$

where the last inequality follows from Hoeffding's inequality.

Now we turn to $P(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |A_{ij}| > \frac{\varepsilon}{40\beta^2})$. Define the event $\mathbb{B}_n$ as

$$\mathbb{B}_n \equiv \left\{ \delta_n \leq \hat{F}_y(g_f(-\sqrt{\beta \log n})) \cup \hat{F}_y(g_f(\sqrt{\beta \log n})) \leq 1 - \delta_n \right\}.$$

Following from (3), we have

$$\begin{aligned}
P(\mathbb{B}_n^c) &\leq 2P(\hat{F}_y(g_f(\sqrt{\beta \log n})) \geq 1 - \delta_n) \\
&= 2P(\hat{F}_y(g_f(\sqrt{\beta \log n})) - F_y(\sqrt{\beta \log n}) \geq 1 - F_y(\sqrt{\beta \log n}) - \delta_n) \\
&\leq 2P(\hat{F}_y(g_f(\sqrt{\beta \log n})) - F_y(\sqrt{\beta \log n}) \geq \frac{1}{2n^{\frac{m}{2}}\sqrt{2\pi m \log n}} - \delta_n) \\
&\leq 2P(\hat{F}_y(g_f(\sqrt{\beta \log n})) - F_y(\sqrt{\beta \log n}) \geq \frac{1}{2n^{\frac{m}{2}}\sqrt{2\pi m \log n}} - \frac{1}{4n^{\frac{m}{2}}\sqrt{2\pi m \log n}}) \\
&= 2P(\hat{F}_y(g_f(\sqrt{\beta \log n})) - F_y(\sqrt{\beta \log n}) \geq \frac{1}{4n^{\frac{m}{2}}\sqrt{2\pi m \log n}}) \\
&\leq 2\exp\left\{-2n \frac{1}{16n^{\frac{m}{2}}\sqrt{2\pi m \log n}}\right\} = 2\exp\left\{-\frac{n^{1-m}}{16\pi m \log n}\right\},
\end{aligned}$$

where last inequality follows from Hoeffding's inequality. Therefore

$$\begin{aligned}
& P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}\right) \\
&= P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}, \mathbb{B}_n\right) + P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}, \mathbb{B}_n^c\right) \\
&\leq P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}, \mathbb{B}_n\right) + P(\mathbb{B}_n^c) \\
&\leq P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}, \mathbb{B}_n\right) + 2\exp\left[-\frac{n^{1-m}}{16\pi m \log n}\right] \\
&\leq P\left(\frac{1}{n^2} \sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n} |\tilde{T}_{y(y_i)} - T_{y(y_i)}| > \frac{\varepsilon}{40\beta^2 \sqrt{m \log n}}, \mathbb{B}_n\right) + 2\exp\left[-\frac{n^{1-m}}{16\pi m \log n}\right] \\
&\leq P\left(\sup_{t \in \mathbb{M}_n} |\tilde{T}_y(t) - T_y(t)| > \frac{\varepsilon}{40\beta^2 \sqrt{m \log n}}, \mathbb{B}_n\right) + 2\exp\left[-\frac{n^{1-m}}{16\pi m \log n}\right].
\end{aligned}$$

Recall that under $\mathbb{B}_n$, for $t \in \mathbb{M}_n$, we have $\tilde{F}_y(t) = \hat{F}_y(t)$. So we can rewrite $\tilde{T}_y(t) - T_y(t)$ as $\Phi^{-1}(\tilde{F}_y(t)) - \Phi^{-1}(F_y(t))$. By the mean value theorem, we further have

$$\Phi^{-1}(\tilde{F}_y(t)) - \Phi^{-1}(F_y(t)) = (\Phi^{-1})'(s)(\hat{F}_y(t) - F_y(t)),$$

where s is between $\hat{F}_y(t)$ and $F_y(t)$. From Lemma 12.3 of Abramovich et al. (2006), we know $(\Phi^{-1})'(s) = \frac{1}{\phi(\Phi^{-1}(s))}$. Also, recall that under $\mathbb{B}_n$, for $t \in \mathbb{M}_n$, both $\hat{F}_y(t)$ and $F_y(t)$ are bounded by $[\delta_n, 1 - \delta_n]$. Therefore,

$$\sup_{s \in [\delta_n, 1 - \delta_n]} (\Phi^{-1})'(s) = \sup_{s \in [\delta_n, 1 - \delta_n]} \frac{1}{\phi(\Phi^{-1}(s))} = \frac{1}{\phi(\Phi^{-1}(1 - \delta_n))} \leq \frac{1}{\phi(\sqrt{2 \log(\frac{1}{\delta_n})})} = \frac{\sqrt{2\pi}}{\delta_n}.$$

Combining them together, we are able to show

$$\begin{aligned}
& P\left(\sup_{t \in \mathbb{M}_n} \left| \tilde{T}_y(t) - T_y(t) \right| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n\right) \\
& \leq P\left(\sup_{t \in \mathbb{M}_n} \left| \Phi^{-1}(\hat{F}_y(t)) - \Phi^{-1}(F_y(t)) \right| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n\right) \\
& \leq P\left(\sup_{s \in [\delta n, 1 - \delta n]} \frac{1}{\phi(\Phi^{-1}(s))} \sup_{t \in \mathbb{M}_n} \left| \hat{F}_y(t) - F_y(t) \right| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n\right) \\
& = P\left(\frac{\sqrt{2\pi}}{\delta n} \sup_{t \in \mathbb{M}_n} \left| \hat{F}_y(t) - F_y(t) \right| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n\right) \\
& = P\left(\sup_{t \in \mathbb{M}_n} \left| \hat{F}_y(t) - F_y(t) \right| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}} \frac{1}{\frac{m}{\sqrt{2\pi}4n^2\sqrt{2\pi m\log n}}}, \mathbb{B}_n\right) \\
& = P\left(\sup_{t \in \mathbb{M}_n} \left| \hat{F}_y(t) - F_y(t) \right| > \frac{\varepsilon n}{320\beta^2\pi m\log n} \frac{-\frac{m}{2}}{2}, \mathbb{B}_n\right).
\end{aligned}$$

Using the Dvoretzky-Kiefer-Wolfowitz inequality, we have

$$\begin{aligned}
P\left(\sup_{t \in \mathbb{M}_n} \left| \hat{F}_y(t) - F_y(t) \right| > \frac{\varepsilon n}{320\beta^2\pi m\log n} \frac{-\frac{m}{2}}{2}, \mathbb{B}_n\right) & \leq \exp\left\{-2n\left[\frac{\varepsilon n}{320\beta^2\pi m\log n} \frac{-\frac{m}{2}}{2}\right]^2\right\} \\
& = \exp\left\{-2\frac{\varepsilon^2 n^2}{102400\beta^4 m^2 \pi^2 \log^2 n}\right\}.
\end{aligned}$$

Now it remains to deal with $P(|\omega_k^* - \omega_k| > \frac{\varepsilon}{2})$. Recall that ω_k^* is a V-statistic, and the corresponding U-statistic ω_k^{**} is given by

$$\omega_k^{**} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n e^{-\frac{\beta^2}{2} E_{ij}} - \frac{2}{n(1+\beta^2)} \sum_{i=1}^n e^{-\frac{\beta^2}{2(1+\beta^2)} E_i} + \frac{1}{1+2\beta^2}.$$

By noting that $e^{-\frac{\beta^2}{2} E_{ij}}$ when $i=j$, and $e^{-\frac{\beta^2}{2} E_{ij}} < 1$ for other cases, it is easy to show

$$\begin{aligned}
|\omega_k^* - \omega_k^{**}| & = \left| \frac{1}{n^2} \left(n + \sum_{i=1}^n \sum_{j \neq i}^n e^{-\frac{\beta^2}{2} E_{ij}} \right) - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n e^{-\frac{\beta^2}{2} E_{ij}} \right| \\
& = \left| \frac{1}{n} - \frac{1}{n^2(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n e^{-\frac{\beta^2}{2} E_{ij}} \right| \leq \frac{1}{n} + \frac{1}{n} = \frac{2}{n}.
\end{aligned}$$

Recall that ε is of a lower order of $\frac{1}{n}$ therefore, we can consider $P(|\omega_k^{**} - \omega_k| > \frac{\varepsilon}{2})$ instead.

The kernel $h(x_{ki}; x_{kj}; y_i; y_j)$ of ω_k^{**} is bounded,

$$\begin{aligned} \left| h(x_{ki}; x_{kj}; y_i; y_j) \right| &= \left| e^{-\frac{\beta^2}{2} E_{ij}} - \frac{1}{1+\beta^2} e^{-\frac{\beta^2}{2(1+\beta^2)} E_i} - \frac{1}{1+\beta^2} e^{-\frac{\beta^2}{2(1+\beta^2)} E_j} + \frac{1}{1+2\beta^2} \right| \\ &\leq 1 + \frac{2}{1+\beta^2} + \frac{1}{1+2\beta^2} \leq 4. \end{aligned}$$

Therefore, it follows from Theorem 5.6.1.A of Sering (1980) that

$$P(|\omega_k^{**} - \omega_k| > \frac{\varepsilon}{2}) \leq \exp \left\{ -2 \left\lfloor \frac{n}{2} \right\rfloor \frac{\varepsilon^2}{2^2 8^2} \right\} = \exp \left\{ - \left\lfloor \frac{n}{2} \right\rfloor \frac{\varepsilon^2}{128} \right\},$$

where $\left\lfloor \frac{n}{2} \right\rfloor$ denotes the integer part of $\frac{n}{2}$.

In summary, by letting $\varepsilon = cn^{-k}$, we have

$$P(|\tilde{\omega}_k^* - \omega_k| > cn^{-k}) \leq O \left\{ \exp \left\{ -c_1 n^{2d} \right\} + \exp \left\{ -c_2 n^{1-2(k+2d)} \right\} + \exp \left\{ -c_3 n^{1-m-2k} \right\} \right\},$$

with the additional constraint $k+2d < \frac{m}{2}$. To optimize the convergence rate, we should let

$m = 2(k+2d)$ and $d = \frac{1}{6} - \frac{2}{3}k$, then we can obtain

$$P(|\tilde{\omega}_k^* - \omega_k| > cn^{-k}) \leq O \left\{ \exp \left\{ -c_1 n^{\frac{1-4k}{3}} \right\} \right\}.$$

Hence,

$$P \left(\max_{1 \leq k \leq p} |\tilde{\omega}_k^* - \omega_k| > cn^{-k} \right) \leq p \max_{1 \leq k \leq p} P(|\tilde{\omega}_k^* - \omega_k| > cn^{-k}) \leq O \left\{ p \exp \left\{ -c_1 n^{\frac{1-4k}{3}} \right\} \right\},$$

which completes the proof.

References

- Abramovich F, Benjamini Y, Donoho DL, and Johnstone IM (2006), ‘‘Adapting to unknown sparsity by controlling the false discovery rate,’’ *The Annals of Statistics*, 34, 584–653.
- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin A, Kim S, Wilson C, Lehár J, Kryukov G, Sonkin D, Reddy A, Liu M, Murray L, Berger M, Monahan J, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J and Mapa F, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels I, and et al. (2012), ‘‘The Cancer cell line encyclopedia enables predictive modeling of anticancer drug sensitivity,’’ *Nature*, 483, 603–607. [PubMed: 22460905]

- Buzdar A (2009), "Role of biologic therapy and chemotherapy in hormone receptor and HER2-positive breast cancer," *The Annals of Oncology*, 20, 993–999. [PubMed: 19150946]
- Cui H, Li R, and Zhong W (2015), "Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis," *Journal of the American Statistical Association*, 110, 630–641. [PubMed: 26392643]
- 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, 544–557. [PubMed: 22279246]
- Fan J and Li R (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.
- 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. [PubMed: 19603084]
- Fan J, Samworth R, and Wu Y (2009), "Ultrahigh Dimensional Feature Selection: Beyond The Linear Model," *J. Mach. Learn. Res.*, 10, 2013–2038. [PubMed: 21603590]
- Fan J and Song R (2010), "Sure independence screening in generalized linear models with NP-dimensionality," *Ann. Statist.*, 38, 3567–3604.
- Grunwald V and Hidalgo M (2003), "Developing inhibitors of the epidermal growth factor receptor for cancer treatment," *Journal of the National Cancer Institute*, 95, 851–867. [PubMed: 12813169]
- Hadley KE and Hendricks DT (2014), "Use of NQO1 status as a selective biomarker for oesophageal squamous cell carcinomas with greater sensitivity to 17-AAG," *BMC Cancer*, 14, 1–8. [PubMed: 24383403]
- Hall P and Miller H (2009), "Using generalized correlation to effect variable selection in very high dimensional problems," *Journal of Computational and Graphical Statistics*, 18, 533–550.
- He X, Wang L, and Hong HG (2013), "Correction: Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data," *The Annals of Statistics*, 41, 2699.
- Henze N and Zirkler B (1990), "A class of invariant consistent tests for multivariate normality," *Communications in statistics - Theory and Methods*, 10, 3595–3617.
- Huang J, Horowitz JL, and Ma S (2008), "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *Ann. Statist.*, 36, 587–613.
- Jiang B and Liu J (2014), "Sliced Inverse Regression with Variable Selection and Interaction Detection," *ArXiv e-prints*
- Li R, Zhong W, and Zhu L (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [PubMed: 25249709]
- Liang F (2008), "Clustering gene expression profiles using mixture model ensemble averaging approach," *JP Journal of Biostatistics*, 2, 57–80.
- Liu H, Lafferty J, and Wasserman L (2009), "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *J. Mach. Learn. Res.*, 10, 2295–2328.
- Mardia K (1970), "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, 57, 519–530.
- Serfling R (1980), *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series, Wiley.
- Székely G and Rizzo M (2005), "A new test for multivariate normality," *Journal of Multi-variate Analysis*, 93, 58–80.
- Székely G, Rizzo M, and Bakirov N (2007), "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, 35, 2769–2794.
- Tibshirani R (1996), "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
- Wang K, Shrestha R, Wyatt AW, Reddy A, Lehr J, Wang Y, Lapuk A, and Collins CC (2014), "A Meta-Analysis Approach for Characterizing Pan-Cancer Mechanisms of Drug Sensitivity in Cell Lines," *PLoS One*, 9, 1–16.
- Zhang C-H (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38, 894–942.

- 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, 1464–1475. [PubMed: 22754050]
- Zoppoli G, Regairaz M, Leo E, Reinhold W, Varma S, Ballestrero A, Doroshow J, and Pommier Y (2012), “Putative DNA/RNA helicase schlafen11 (SLFN11) sensitizes cancer cells to DNA-damaging agents,” *Proceedings of the National Academy of Sciences USA*, 109, 15030–15035.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1:
Histograms of screening indices produced by different methods for the additive model example with the predictors generated from the distribution $\mathcal{U}(4)$.

Figure 2:
Scatter plots of the transformed response variable $T_Y(Y)$ versus the transformed predictors $T_1(X_1)$, $T_{50}(X_{50})$ and $T_{100}(X_{100})$

Table 1:

Results for the additive model example. For MSD, we report the median with the associated interquartile range (IQR) in the parentheses. Cases 1 and 2 refer to that the predictors are generated from $t(4)$ and Uniform $[-2.5, 2.5]$, respectively.

	Method	MSD	P_1	P_2	P_3	P_4	P_a
Case 1	SIS	976.50(1023.00)	0.07	0.08	0.96	0.98	0.01
	NIS	1342.50(704.75)	0.01	0.20	0.05	0.76	0.00
	DC-SIS	279.50(656.75)	0.27	0.46	0.57	0.95	0.16
	MV-SIS	24.00(118.00)	0.83	0.73	0.97	0.95	0.58
	Qa-SIS	347.50(653.50)	0.02	0.81	0.22	0.98	0.00
	HZ-SIS	11.50(22.00)	0.98	0.90	0.97	0.94	0.80
Case 2	SIS	1216.50(964.75)	0.10	0.02	1.00	1.00	0.00
	NIS	924.00(1257.25)	0.16	0.17	0.30	0.33	0.06
	DC-SIS	197.00(339.00)	0.20	0.31	0.98	0.98	0.06
	MV-SIS	11.00(28.00)	0.94	0.83	1.00	1.00	0.78
	Qa-SIS	8.00(15.50)	0.91	0.91	1.00	1.00	0.82
	HZ-SIS	24.50(52.25)	0.71	0.88	0.99	1.00	0.64

Table 2:

Results for the model with interaction variables. For MSD, we report the median with the associated interquartile range (IQR) given in the parentheses. Cases 1 and 2 refer to that the random errors are generated from the $N(0,1)$ and Cauchy distributions, respectively.

	Method	MSD	P_1	P_{50}	P_a
Case 1	SIS	686.00(321.25)	1.00	0.00	0.00
	SIRI-SIS	3.00(1.00)	1.00	0.99	0.99
	DC-SIS	50.50(52.25)	1.00	0.39	0.39
	MV-SIS	34.00(73.75)	1.00	0.52	0.52
	Qa-SIS	426.00(478.00)	1.00	0.02	0.02
	HZ-SIS	3.00(1.00)	1.00	1.00	1.00
Case 2	SIS	575.50(387.00)	0.78	0.03	0.01
	SIRI-SIS	11.00(52.00)	1.00	0.69	0.69
	DC-SIS	167.00(251.25)	1.00	0.12	0.12
	MV-SIS	97.50(158.75)	1.00	0.23	0.23
	Qa-SIS	414.50(403.75)	1.00	0.02	0.02
	HZ-SIS	9.50(22.00)	1.00	0.86	0.86

Table 3:

Results for the complex model with more interaction variables. For MSD, we report the median with the associated interquartile range (IQR) given in the parentheses. Cases 1 and 2 refer to that the random errors are generated from $N(0,1)$ and the Cauchy distribution, respectively.

	Method	MSD	P_1	P_2	P_3	P_5	P_6	P_7	P_8	P_a
Case1	SIS	841.50(196.25)	0.07	0.30	0.21	0.78	0.06	0.08	0.06	0.00
	SIRI-SIS	470.00(377.25)	0.73	0.99	1.00	0.34	0.05	0.48	0.53	0.00
	DC-SIS	911.00(117.00)	0.04	0.94	0.13	0.04	0.07	0.04	0.07	0.00
	MV-SIS	511.50(322.00)	0.28	0.63	1.00	0.88	0.09	0.38	0.33	0.00
	Qa-SIS	420.50(305.25)	0.02	0.97	0.98	0.09	0.05	0.05	0.08	0.00
	HZ-SIS	245.50(341.75)	0.90	1.00	1.00	0.93	0.30	0.76	0.53	0.12
Case2	SIS	882.50(183.50)	0.10	0.29	0.16	0.84	0.07	0.09	0.11	0.00
	SIRI-SIS	432.00(378.25)	0.56	0.99	1.00	0.17	0.06	0.45	0.40	0.00
	DC-SIS	899.50(134.25)	0.10	0.97	0.11	0.10	0.09	0.03	0.08	0.00
	MV-SIS	484.00(325.00)	0.17	0.63	1.00	0.85	0.11	0.39	0.36	0.00
	Qa-SIS	416.00(453.25)	0.04	0.97	0.99	0.06	0.08	0.03	0.08	0.00
	HZ-SIS	200.50(309.25)	0.79	1.00	1.00	0.89	0.26	0.71	0.40	0.10

Table 4:

Top 10 genes selected for the drug topotecan by different methods.

Rank	SIS	DC-SIS	MV-SIS	Qa-SIS	HZ-SIS
1	CADM2	ITGB5	HMGB2	FLJ35816	RFXAP
2	MMP27	PPIC	KIF15	GATS	HMGB2
3	WNT5B	HMGB2	RFXAP	HS3ST3A1	ITGB5
4	CDX4	ARHGAP19	ARHGAP19	ASIC4	BCLAF1
5	ELF4	RFXAP	CD63	TRAV26-2	CPSF6
6	ECI2	CPSF6	TAF5	LOC100128239	HAUS1
7	GLIPR1	CD63	CPSF6	LOC100507300	ILF3
8	ABCC9	TAF5	ELAVL1	VPS72	ELAVL1
9	ADCY5	CNTRL	ILF3	PIK3IP1	TAF5
10	PPIC	SLFN11	S100A10	CRSF6	SLFN11

Table 5:

Top 10 genes selected for the drug 17-AAG by different methods.

Rank	SIS	DC-SIS	MV-SIS	Qa-SIS	HZ-SIS
1	UXT	NQO1	NQO1	MMP24	NQO1
2	IGFN1	MMP24	INO80	ATP6V0E1	OGDHL
3	MSH2	ZNF610	MMP24	ZFP30	TMEM198
4	ROCK1	ZFP30	ZNF610	GPR35	ZBTB7A
5	DDA1	NFKB1	ZFP30	SLC1A5	GYG2
6	SCEL	CDH6	PRPUSD4	GPX2	CDH6
7	ST5	OGDHL	LOC1005007372	CNTRL	ZNF610
8	THUMPD3	LOC100507372	PCSK1N	VPS72	RPUSD4
9	ITGA9	PRUSD4	NFKB1	LOC100507373	CSK
10	C20orf141	INO80	ZBTB7A	ZNF610	CTCF