

Statistica Sinica Preprint No: SS-2016-0264R1

Title	Model-Free Feature Screening for Ultrahigh Dimensional Data through a Modified Blum-Kiefer-Rosenblatt Correlation
Manuscript ID	SS-2016.0264.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0264
Complete List of Authors	Liping ZHU and Yeqing Zhou
Corresponding Author	Liping ZHU
E-mail	zhulp1@hotmail.com
Notice: Accepted version subject to English editing.	

Model-Free Feature Screening for Ultrahigh Dimensional Data Through a Modified Blum-Kiefer-Rosenblatt Correlation

Yeqing Zhou and Liping Zhu

Shanghai University of Finance and Economics

Renmin University of China

Abstract: In this paper we introduce a modified Blum-Kiefer-Rosenblatt correlation (MBKR for short) to rank the relative importance of each predictor in ultrahigh dimensional regressions. We advocate using the MBKR for two reasons. First, the MBKR is nonnegative and equals zero if and only if two random variables are independent, indicating that the MBKR can detect nonlinear dependence. We illustrate that the sure independence screening procedure based on the MBKR (MBKR-SIS for short) is effective to detect nonlinear effects including interactions and heterogeneity, particularly when both continuous and discrete predictors are involved simultaneously. Second, the MBKR is conceptually simple, easy to implement and affine-invariant. The MBKR is free of tuning parameters and no iteration is required in estimation. It remains unchanged when order-preserving transformations are applied to the response or predictors, indicating that the MBKR-SIS is robust to the presence of extreme values and outliers in the observations. We also show that, under mild conditions, the MBKR-SIS procedure has the desirable sure screening and ranking consistency properties,

which guarantee that all important predictors can be retained after screening with probability approaching one. We also propose an iterative screening procedure to detect the important predictors which are marginally independent of the response variable. We demonstrate the merits of the MBKR-SIS procedure through simulations and an application to a real-world dataset.

Key words and phrases: Blum-Kiefer-Rosenblatt correlation, Feature screening, Independence test, Ranking consistency property, Sure screening property.

1. Introduction

With the advance of science and technology, ultrahigh dimensional data arise in many scientific fields. For instance, in order to identify gene mutations which probably cause a disease, medical scientists typically collect thousands of gene expression levels or genetic markers from a relatively small number of subjects, borrowing the strength of microarray technology. To analyze such datasets effectively, it is often assumed that the disease depends upon only a few among thousands of gene expression levels or genetic markers. Such an assumption, usually referred to as sparsity, illuminates extensive researches on feature selection in statistical area in the past two decades.

To identify important features in regressions, many penalized least squares and penalized likelihood algorithms such as bridge regression (Frank and Friedman, 1993), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001),

adaptive Lasso (Zou, 2006), Dantzig selector (Candes and Tao, 2007), non-negative garrote (Yuan and Lin, 2006), are proposed. These penalized algorithms are effective for mean regressions and parametric models, yet suffer from simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability when the predictors are ultrahigh dimensional and the sample size is relatively small (Fan, Samworth and Wu, 2009).

To analyze ultrahigh dimensional data, marginal screening procedures are typically regarded as acceptable preludes to penalized regressions. Marginal screening breaks an ultrahigh dimensional regression into many low-dimensional problems, hence reduces the overall computational complexity dramatically.

There are two classes of marginal screening procedures in the literature. One is concerned with the conditional mean regression, while the other is concerned with the conditional distribution function. To be precise, we let $Y \in \mathbb{R}^1$ be the response variable and $\mathbf{x} = (X_1, \dots, X_p)^T \in \mathbb{R}^p$ be the predictor vector. The first class of screening procedures aims to identify important features indexed by

$$\mathcal{I} \stackrel{\text{def}}{=} \{k : E(Y | \mathbf{x}) \text{ varies with } X_k\}. \quad (1.1)$$

In this area, Fan and Lv (2008) proposed a sure independence screening (SIS for short) procedure based on Pearson correlation coefficient, assuming $E(Y | \mathbf{x})$ is a linear function of \mathbf{x} . The idea of Pearson correlation

based screening was later generalized from many different perspectives. For example, Hall and Miller (2009) suggested using polynomial transformations of the predictors in SIS. Li et al. (2012) recommended Kendall's rank correlation and Shao and Zhang (2014) proposed martingale difference correlation in place of Pearson correlation to perform marginal screening. Fan and Song (2010) and Fan, Feng and Song (2011) extended the linear model assumption to generalized linear model and nonparametric additive model, respectively. The second class of screening procedures aims to identify the important features indexed by

$$\mathcal{A} \stackrel{\text{def}}{=} \{k : F(y | \mathbf{x}) \text{ varies with } X_k \text{ for some } y \in \mathbb{R}^1\}, \quad (1.2)$$

where $F(y | \mathbf{x})$ stands for the conditional distribution function of Y given \mathbf{x} . In this area, Zhu et al. (2011) proposed a sure independent ranking and screening procedure (SIRS for short) and Li, Zhong and Zhu (2012) suggested a distance correlation based sure independence screening procedure (DC-SIS for short). Mai and Zou (2013) and Mai and Zou (2015) suggested a nonparametric Kolmogorov filter when the response variable is binary. We remark here that $\mathcal{A} = \mathcal{I}$ when Y is binary. Though all the aforementioned marginal screening procedures assumed the marginal distribution functions of either \mathbf{x} or Y , or both, have exponential tails, they all have the desirable sure screening property, a terminology introduced by Fan and Lv (2008).

That is, all important predictors, possibly together with a few unimportant ones, are retained after screening with an overwhelming probability.

In addition to main effects which are of our interests, interactions and heterogeneity are quite common phenomena in ultrahigh dimensional data. Detecting interactions through mean regressions indexed by \mathcal{I} hinges upon either a utility which can measure possibly nonlinear relation between X_k and Y or a correctly specified mean model for $E(Y | \mathbf{x})$. In analysis of ultrahigh dimensional data, we are quite often lack of prior information on the regression structure (Zhu et al., 2011). Modelling interactions in $E(Y | \mathbf{x})$ will increase the model size from $O(p)$ to $O(p^2)$ if two-way interactions are concerned and to $O(p^3)$ if three-way interactions are concerned, consequently will bring in an unbearable computation even for marginal screening. The Pearson correlation based screening procedures, such as Fan and Lv (2008) and Li et al. (2012), may fail to detect interactions even in linear models. Heterogeneity is another important issue and is characterized by $\text{var}(Y | \mathbf{x})$. Those mean regression based screening procedures cannot detect important features which merely describe the heterogeneity of the data. We will demonstrate these issues through simulated examples in Sections 2 and 3.

The overcoming goal of regression analysis is to characterize how the

conditional distribution function of Y varies with the realizations of \mathbf{x} . As a prelude to subsequent regression analysis, an ideal screening procedure is expected to retain the important predictors indexed by \mathcal{A} rather \mathcal{I} . In effect, identifying \mathcal{A} rather \mathcal{I} is sufficient to capture both the interactions and heterogeneity, in addition to main effects. Important covariates involved in interactions and heterogeneity are generally nonlinear effects. A prerequisite to identify \mathcal{A} is to design a utility which can measure the possibly nonlinear relations between X_k and Y without requiring any information of the underlying regression structure. Our goal in this paper is to design a utility which can detect the main effects, the interactions and the heterogeneity simultaneously, or more precisely, the important predictors indexed by \mathcal{A} . Following the idea of Blum, Kiefer and Rosenblatt (1961), we introduce a modified Blum-Kiefer-Rosenblatt correlation (MBKR for short) to rank the relative importance of the predictors. The MBKR correlation measures the relation between X_k and Y through

$$MBKR(X_k, Y) \stackrel{\text{def}}{=} \int_{\mathbb{R}^1} \int_{\mathbb{R}^1} \frac{\{F_{X_k, Y}(x_k, y) - F_{X_k}(x_k)F_Y(y)\}^2}{F_{X_k}(x_k)\{1 - F_{X_k}(x_k)\}F_Y(y)\{1 - F_Y(y)\}} dF_{X_k}(x_k)dF_Y(y), (1.3)$$

where F_{X_k} and F_Y are the respective marginal distribution functions of X_k and Y , and $F_{X_k, Y}(x_k, y)$ is the joint distribution of (X_k, Y) . We advocate using the MBKR for at least two reasons.

1. The MBKR is nonnegative and equals zero if and only if two random variables are independent, indicating that the MBKR can detect nonlinear dependence. We illustrate through simulations that the sure independence screening procedure based on the MBKR (MBKR-SIS for short) is effective to detect nonlinear effects including interactions and heterogeneity, particularly when both continuous and discrete predictors are involved simultaneously. Interactions and heterogeneity are quite common phenomena in regression analysis and yet detecting them appears to be a challenging task for many existing screening procedures.
2. The MBKR is conceptually simple, easy to implement and affine invariant. The MBKR is free of tuning parameters and no iteration is required in estimation. It remains unchanged when order-preserving transformations are applied to either the response or the predictors. This property also indicates that the MBKR-SIS is robust to the presence of extreme values and outliers in the observations.

We study the asymptotic properties of the sample MBKR. We show that the sample MBKR is n consistent if X_k and Y are independent and root n consistent otherwise. If X_k and Y are independent and both are continuous random variables, the asymptotic distribution of the sample MBKR does

not depend on the marginal distribution of X_k or that of Y . This property is appealing in that the critical value can be easily determined when applying the sample MBKR to test independence between X_k and Y . We also show that, under mild conditions, the MBKR-SIS procedure has the desirable sure screening and ranking consistency properties, which guarantee that all important predictors can be retained after screening with an overwhelming probability.

The rest of this paper is organized as follows. In Section 2, we introduce the notion of the MBKR correlation and consider two applications. One is to test independence between two random variables and the other is to screen out irrelevant features for ultrahigh dimensional data. We also investigate the theoretical properties of our proposals thoroughly. In Section 3, we evaluate the finite sample performance of our proposals through Monte Carlo simulations and an application to a rat eye dataset consisting of gene expression levels and genetic markers. We also propose an iterative approach to detect the important predictors which are marginally independent of the response. We conclude this paper with a brief discussion in Section 4. The technical details are relegated to the supplement.

2. A Modified Blum-Kiefer-Rosenblatt Correlation

2.1. Two Relevant Utilities

The MBKR defined in (1.3) is originated from the Blum-Kiefer-Rosenblatt correlation (BKR for short) introduced by Blum, Kiefer and Rosenblatt (1961). The latter is defined as

$$BKR(X_k, Y) = \int_{\mathbb{R}^1} \int_{\mathbb{R}^1} \{F_{X_k, Y}(x_k, y) - F_{X_k}(x_k)F_Y(y)\}^2 dF_{X_k}(x_k) dF_Y(y). \quad (2.1)$$

By definition, $F_{X_k, Y}(x_k, y) - F_{X_k}(x_k)F_Y(y) = \text{cov}\{I(X_k \leq x_k), I(Y \leq y)\}$.

Throughout we use $I(A)$ to denote an indicator function, which equals one if the event A is true and zero otherwise. The difference between the MBKR and the BKR is that the integrand of the former is $\text{corr}^2\{I(X_k \leq x_k), I(Y \leq y)\}$ while that of the latter is $\text{cov}^2\{I(X_k \leq x_k), I(Y \leq y)\}$. Here the notation $\text{corr}(\cdot, \cdot)$ stands for the Pearson correlation coefficient. If both Y and X_k are continuous, using correlation or covariance does not make significant difference because $F_Y(Y)$ and all $F_{X_k}(X_k)$ s, for $k = 1, \dots, p$, follow uniform distributions coincidentally. If some of the predictors are continuous while others are discrete or categorical, then the difference between using correlation and covariance to rank the relative importance of X_k may not be negligible. We will illustrate this issue through a toy example shortly.

Another relevant measure is the Hoeffding's index (Hoeffding, 1948), which is defined as

$$H(X_k, Y) = \int_{\mathbb{R}^1} \int_{\mathbb{R}^1} \{F_{X_k, Y}(x_k, y) - F_{X_k}(x_k)F_Y(y)\}^2 dF_{X_k, Y}(x_k, y). \quad (2.2)$$

The Hoeffding's index, the BKR and the MBKR correlations have quite similar forms, and all can be seen to be nonnegative with equality to zero under independence between X_k and Y . However, the Hoeffding's index has a serious drawback in that it may be zero if there exists an association between X_k and Y . In other words, it does not lead to a consistent test of independence. An example is the case that $\text{pr}(X_k = 0, Y = 1) = \text{pr}(X_k = 1, Y = 0) = 1/2$ (Hoeffding, 1948, page 548).

We first use a toy example to explain why we prefer using the MBKR to rank the relative importance of each predictor.

Example 1. We draw $\mathbf{z} = (Z_1, Z_2, Z_3)^T$ from multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{kl})_{3 \times 3}$, where $\sigma_{kl} = 0.9^{|k-l|}$, and draw ε independently from standard normal distribution. Let the predictors $\mathbf{x} = (X_1, X_2, X_3)^T = \{Z_1, Z_2, I(Z_3 \geq 0)\}^T$ and the response $Y = \kappa X_1 + X_3 + \varepsilon$, for $\kappa = 1, \sqrt{3}$ and 3. In this example, X_1 and X_2 are continuous and X_3 is binary. X_1 and X_3 are truly important predictors while X_2 is not predictive for Y when X_1 and X_3 are given. In other words, $\mathcal{A} = \{1, 3\}$ and $\mathcal{A}^c = \{2\}$. The superscript c in \mathcal{A}^c stands for complement of \mathcal{A} . The sample size $n = 100$ and 200. We apply the MBKR, the BKR and the Hoeffding's index to rank the relative importance of the predictors. Apparently, the larger a measure between each predictor and the response

is, the more important the predictor is regarded as by this measure. To compare their performance, we report the probability of ranking important predictors above the unimportant ones, namely, $\text{pr}(\min_{k \in \mathcal{A}} \omega_k \geq \max_{k \in \mathcal{A}^c} \omega_k)$, where ω_k can be $MBKR(X_k, Y)$, $BKR(X_k, Y)$ and $H(X_k, Y)$, corresponding to three different correlation measures. This criterion is very stringent. It quantifies the capability of a measure to rank the important predictors prior to those unimportant ones. The closer to one this probability is, the better a measure is.

We repeat each scenario 1,000 times and chart the simulation results in Table 1.

Table 1: The empirical probabilities of $\text{pr}(\min_{k \in \mathcal{A}} \omega_k \geq \max_{k \in \mathcal{A}^c} \omega_k)$ based on 1,000 repetitions for Example 1.

measure	n=100			n=200		
	$\kappa = 1$	$\kappa = \sqrt{3}$	$\kappa = 3$	$\kappa = 1$	$\kappa = \sqrt{3}$	$\kappa = 3$
Hoeffding's index	0.513	0.380	0.142	0.645	0.349	0.074
BKR correlation	0.513	0.380	0.142	0.645	0.349	0.074
MBKR correlation	0.845	0.984	0.960	0.911	1.000	0.995

It can be clearly seen from Table 1 that the MBKR ranks X_1 and X_3 above X_2 with high probabilities across all scenarios, while the BKR and the Hoeffding's index yield exactly identical results. We remark here that

our limited experience indicates that the BKR and the Hoeffding's index usually have very similar performances in many scenarios. For the MBKR method, $\text{pr}(\min_{k \in \mathcal{A}} \omega_k \geq \max_{k \in \mathcal{A}^c} \omega_k) \geq 0.845$ when $n = 100$ and $\text{pr}(\min_{k \in \mathcal{A}} \omega_k \geq \max_{k \in \mathcal{A}^c} \omega_k) \geq 0.911$ when $n = 200$, and its performance is relatively stable for different κ values. However, both the BKR and the Hoeffding's index deteriorate sharply as κ , the coefficient of X_1 in this example, increases. Surprisingly, for the BKR correlation and the Hoeffding's index, the probability $\text{pr}(\min_{k \in \mathcal{A}} \omega_k \geq \max_{k \in \mathcal{A}^c} \omega_k)$ decreases from 0.645 to 0.074 when $n = 200$.

This example indicates that applying the MBKR to quantify the relative importance of each predictor is probably more effective than using the BKR or the Hoeffding's index, particularly when some predictors are discrete or categorical while others are continuous. The MBKR has some other apparent appealing properties. For example, $MBKR(X_k, Y)$ is always nonnegative and equals zero if and only if X_k and Y are independent; $MBKR(X_k, Y) = MBKR\{m_k(X_k), m(Y)\}$ for monotone functions m_k and m and $MBKR(X_k, Y) = MBKR(Y, X_k)$, indicating that the MBKR correlation is affine invariant. Considering these nice properties, we will work with the MBKR correlation only in what follows unless stated otherwise.

2.2. An Estimation

Suppose $\{(\mathbf{x}_i, Y_i), i = 1, \dots, n\}$ is a random sample from the population

(\mathbf{x}, Y) . In this section, we propose an estimation for the MBKR correlation.

Let F_{n,X_k} , $F_{n,Y}$ and $F_{n,X_k,Y}$ be the respective empirical versions of F_{X_k} , F_Y and $F_{X_k,Y}$. To be precise,

$$F_{n,X_k}(x_k) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n I(X_{i,k} \leq x_k), \quad F_{n,Y}(y) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n I(Y_i \leq y) \text{ and}$$

$$F_{n,X_k,Y}(x_k, y) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n I(X_{i,k} \leq x_k, Y_i \leq y).$$

Define $0/0 = 0$. A natural estimator of $MBKR(X_k, Y)$ is

$$\widehat{MBKR}(X_k, Y) \stackrel{\text{def}}{=} n^{-2} \sum_{i=1}^n \sum_{j=1}^n \frac{\{F_{n,X_k,Y}(X_{i,k}, Y_j) - F_{n,X_k}(X_{i,k})F_{n,Y}(Y_j)\}^2}{F_{n,X_k}(X_{i,k})\{1 - F_{n,X_k}(X_{i,k})\}F_{n,Y}(Y_j)\{1 - F_{n,Y}(Y_j)\}}. \quad (2.3)$$

A remarkable property of this estimator is that it depends on the ranks of X_k and Y only. Thus, the MBKR remains unchanged if order-preserving transformations are applied to either X_k or Y . This also indicates that the MBKR is resilient to the presence of outliers and extreme values.

Theorem 1 states the asymptotic properties of $\widehat{MBKR}(X_k, Y)$.

Theorem 1. (Convergence in Distribution)

1. If X_k and Y are not independent, then $MBKR(X_k, Y) > 0$ and

$$n^{1/2} \left\{ \widehat{MBKR}(X_k, Y) - MBKR(X_k, Y) \right\} \xrightarrow{d} \mathcal{N}(0, \sigma^2), \text{ as } n \rightarrow \infty,$$

where $\sigma^2 \stackrel{\text{def}}{=} \text{var}(Z)$ and Z is defined in (S1.1) in the supplement.

2. If X_k and Y are independent, then $MBKR(X_k, Y) = 0$ and

$$n \widehat{MBKR}(X_k, Y) \xrightarrow{d} \sum_{l=1}^{\infty} \sum_{m=1}^{\infty} \lambda_{lm} \chi_{lm}^2(1), \text{ as } n \rightarrow \infty,$$

where $\chi_{lm}^2(1)$ are independent chi-square random variables with one degree of freedom and λ_{lm} s may depend on the distributions of X_k and Y . In particular, if both X_k and Y are continuous random variables, then $\lambda_{lm} = 1/\{l(l+1)m(m+1)\}$.

Theorem 1 indicates that, if both X_k and Y are continuous, the MBKR offers a distribution-free test of independence. To be precise, we can reject the null hypothesis $H_0 : X_k$ and Y are independent, if $n \widehat{MBKR}(X_k, Y) \geq c_\alpha$ where c_α is the upper $\alpha \times 100\%$ quantile of the limiting distribution $\sum_{l=1}^{\infty} \sum_{m=1}^{\infty} \chi_{lm}^2(1)/\{l(l+1)m(m+1)\}$ and α is the user-specified significance level. The power of this test is

$$\begin{aligned} \beta_n(\varphi) &\stackrel{\text{def}}{=} \text{pr}(n \widehat{MBKR}(X_k, Y) > c_\alpha \mid MBKR(X_k, Y) = \varphi > 0) \\ &= \text{pr}\left(\frac{n^{1/2}\{\widehat{MBKR}(X_k, Y) - \varphi\}}{\sigma} > \frac{c_\alpha - n\varphi}{n^{1/2}\sigma} \mid MBKR(X_k, Y) = \varphi > 0\right) \\ &\approx 1 - \Phi\left(\frac{c_\alpha - n\varphi}{n^{1/2}\sigma}\right) \rightarrow 1, \text{ as } n \rightarrow \infty, \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0, 1)$.

Next we give another toy example to show that the MBKR is effective to detect the interactions and the heterogeneity of data.

Example 2. We consider two simulated models in this toy example:

$$Y = (5\kappa)(X_1X_2) + \varepsilon; \quad (2.4)$$

$$Y = \exp\{\kappa(X_1 + X_2)\}\varepsilon. \quad (2.5)$$

In the interaction model (2.4) and the heterogeneity model (2.5), we draw $\mathbf{x} = (X_1, X_2)^T$ from multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{kl})_{2 \times 2}$, where $\sigma_{kl} = 0.8^{|k-l|}$. We draw ε from standard cauchy distribution. We set $\kappa = 0, 0.2, 0.4, \dots, 1$ in both models (2.4) and (2.5). We test the following hypothesis $H_0 : X_1$ and Y are independent, versus $H_1 : \text{otherwise}$. Apparently, the null hypothesis H_0 is true when $\kappa = 0$ and false otherwise.

We compare the performances of six tests of independence. The first four are based on Pearson correlation, Kendall's rank correlation, the distance correlation (Székely et al., 2007) and the rank-based distance correlation (Székely and Rizzo, 2009), respectively. The distance correlation and rank-based distance correlation are also nonnegative with equality to zero if and only if two random variables are independent. The Pearson correlation and Kendall's rank correlation are often used to detect linear dependence between two random variables. The fifth is based on the utility used in the SIRS procedure (Zhu et al., 2011), which is defined as $\omega_k = E\{\Omega_k^2(Y)\}/\text{var}(X_k)$ and $\Omega_k(y) = \text{cov}\{X_k, I(Y \leq y)\}$. The last is

based on the MBKR correlation $\omega_k = MBKR(X_k, Y)$. The aforementioned correlation measures are used in several existing screening procedures in the literature. In particular, the Pearson correlation was used in the SIS procedure (Fan and Lv, 2008); the Kendall's rank correlation was suggested by Li et al. (2012) for screening out irrelevant predictors; the distance correlation was used in the DC-SIS procedure (Li, Zhong and Zhu, 2012). For fair comparison, we include rank-based distance correlation to test independence between X_k and Y . We will compare their performances in detecting interactions and heterogeneity in this example.

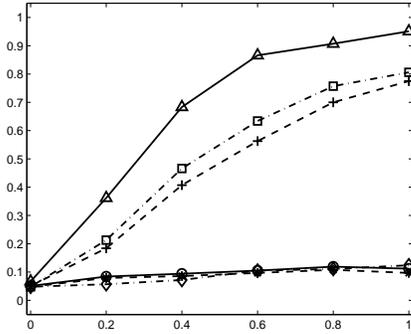
We fix the sample size $n = 50$ and the significance level $\alpha = 0.05$. We repeat the simulations 1,000 times, and report the sizes and the power curves in Figure 1(A) for model (2.4) and Figure 1(B) for model (2.5).

It can be clearly seen from Figure 1 that, when $\kappa = 0$, the sizes of all six tests are very close to the significance level 0.05. It is thus fair to compare their power curves. The power curves of the distance correlation, the rank-based distance correlation and the MBKR increase gradually as the κ values increase in both models. The MBKR is apparently the most powerful to detect interactions and heterogeneity, followed by distance correlation, rank-based distance correlation in model (2.4) and SIRS, rank-based distance correlation and distance correlation in model (2.5). In both models,

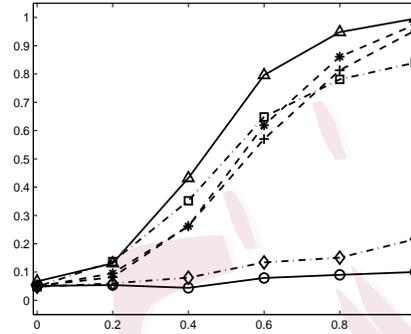
the performances of distance correlation are slightly influenced by the heavily tailed distribution of the unobservable error term. Though the distance correlation performs satisfactorily in this example, the asymptotic distribution of its sample estimate is not tractable, making the determination of critical values very time-consuming. The Pearson correlation and the Kendall's rank correlation completely lose power to detect interactions and heterogeneity as their power curves are almost flat as the κ value increases. This toy example indicates that the MBKR, among the six aforementioned correlation measures, is probably the most powerful utility to detect the interactions and heterogeneity, even when all the predictors are continuous.

2.3. A Screening Procedure

In this section we design a sure independence screening procedure based on the MBKR correlation (MBKR-SIS for short) for ultrahigh dimensional data. Suppose Y is the response variable and $\mathbf{x} = (X_1, \dots, X_p)^T$ is the associated predictor vector. Here the predictor dimension p is extremely large and the available sample size n is relatively small. With a sample of size n , we aim to screen out as many unimportant predictors as possible. The unimportant predictors are indexed by \mathcal{A}^c , upon which the response variable Y does not depend when the important predictors are given. In mathematical symbols, $Y \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}^c} \mid \mathbf{x}_{\mathcal{A}}$. The index set of our primary interest



(A): model (2.4)



(B): model (2.5)

Figure 1: The empirical power curves based on the Pearson correlation test (dashdot line marked with diamond), the Kendall's rank correlation test (solid line marked with circle), the distance covariance test (dashdot line marked with square), the rank-based distance covariance test (dashed line marked with plus), the marginal utility in SIRS test (dashed line marked with star) and the MBKR test (solid line marked with up triangle), respectively. The horizontal axis: the κ value varies from 0 to 1 in models (2.4) and (2.5). The vertical axis: the size and power curves increase from 0 to 1.

is \mathcal{A} defined in (1.2).

Because the conditional distribution of Y given \mathbf{x} varies with $\mathbf{x}_{\mathcal{A}}$ only, it is natural to expect that Y depends on X_k , for $k \in \mathcal{A}$, though such dependence can be nonlinear. If we are willing to use the $MBKR(X_k, Y)$ to measure the possibly nonlinear relation between X_k and Y , we can naturally expect $MBKR(X_k, Y)$, for $k \in \mathcal{A}$, cannot be very small. We spell out this assumption below.

(A1) The truly important predictors satisfy

$$\min_{k \in \mathcal{A}} MBKR(X_k, Y) \geq 2dn^{-\gamma}, \text{ for some constants } d > 0, 0 \leq \gamma < 1/4.$$

Assumption (A1) requires the signals of the important predictors be strong enough to be detectable by the MBKR correlation. Similar conditions are widely assumed in the marginal screening literature. See, for example, Condition 3 in Fan and Lv (2008) and Condition (C2) in Li, Zhong and Zhu (2012). For $\gamma > 0$, the MBKR correlation may shrink to zero. In other words, we allow for the weak signal case as well.

If the signal level is not too small, i.e., (A1) is true, we suggest the following MBKR-SIS procedure which retains the predictors indexed by

$$\widehat{\mathcal{A}} = \{k : \widehat{MBKR}(X_k, Y) \geq dn^{-\gamma}, k = 1, \dots, p\}. \quad (2.6)$$

With Assumption (A1), we can easily establish the desirable sure screening property for the MBKR-SIS procedure.

Theorem 2. (Sure Screening Property) *If p satisfies $np \exp(-d_1 n^{1-4\gamma}) \rightarrow 0$ for a positive constant d_1 , then under Assumptions (A1), we have*

$$pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \geq 1 - O\{n|\mathcal{A}| \exp(-d_1 n^{1-4\gamma})\},$$

where $|\mathcal{A}|$ denotes the cardinality of the index set \mathcal{A} .

The following theorem characterizes the size of reduced model after screening. Let $|\widehat{\mathcal{A}}|$ stand for the cardinality of $\widehat{\mathcal{A}}$.

Theorem 3. (*Minimum Model Size*) *Under the conditions of Theorem 2, there exists a positive constant d_2 , such that,*

$$pr \left\{ |\widehat{\mathcal{A}}| \leq O(n^\gamma \sum_{k=1}^p |MBKR(X_k, Y)|) \right\} \geq 1 - O \{np \exp(-d_2 n^{1-4\gamma})\}.$$

We further notice that, the relation of conditional independence, that is $Y \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}^c} \mid \mathbf{x}_{\mathcal{A}}$, indicates that the conditional distribution of Y give $\mathbf{x}_{\mathcal{A}}$ is independent of $\mathbf{x}_{\mathcal{A}^c}$. This motivates us to expect that Y depends more upon $\mathbf{x}_{\mathcal{A}}$ than upon $\mathbf{x}_{\mathcal{A}^c}$, though such dependence can be nonlinear. Intuitively speaking, $MBKR(X_k, Y)$, for $k \in \mathcal{A}$, is larger than $MBKR(X_k, Y)$, for $k \in \mathcal{A}^c$, if we use the MBKR correlation to measure nonlinear dependence. Such an intuition is formulated in the following assumption.

(A2) $\liminf_{p \rightarrow \infty} \{ \min_{k \in \mathcal{A}} MBKR(X_k, Y) - \max_{k \in \mathcal{A}^c} MBKR(X_k, Y) \} \geq d_3$, where d_3 is a positive constant.

With Assumption (A2), we can easily establish the ranking consistency property for the MBKR-SIS procedure.

Theorem 4. (*Rank Consistency Property*) *We assume Condition (A2) in addition to the conditions of Theorem 2. Then*

$$\liminf_{n \rightarrow \infty} \left\{ \min_{k \in \mathcal{A}} \widehat{MBKR}(X_k, Y) - \max_{k \in \mathcal{A}^c} \widehat{MBKR}(X_k, Y) \right\} > 0 \text{ almost surely.}$$

Theorem 4 ensures that the important predictors will be ranked prior to the unimportant ones with an overwhelming probability, if the signals between the important and the unimportant predictors are distinguishable.

We remark here that, to ensure the above desirable sure screening and ranking consistency properties, we do not impose any moment conditions on either \mathbf{x} or Y . In effect, we merely use the ranks of the observations to achieve the consistency. This also indicates that the MBKR will be resistant to the presence of extreme values and outliers in the observations.

3. Numerical Studies

3.1. Some Additional Simulations

In this section, we conduct some additional Monte Carlo simulations to evaluate the finite sample performance of our proposed MBKR-SIS method in comparison to five existing screening methods such as the SIS (Fan and Lv, 2008), SIRS (Zhu et al., 2011), Kendall's rank correlation based sure independent screening procedure (Li et al., 2012, RRCS for short), distance correlation based on SIS (Li, Zhong and Zhu, 2012, DC-SIS for short) and the sure independent screening procedure using rank-based distance correlation (Székely and Rizzo, 2009, RDC-SIS for short).

Following Li, Zhong and Zhu (2012), we use the following three criteria to assess their performance.

1. \mathcal{S} : The minimum model size to ensure that all important predictors are retained after screening. We report the 5%, 25%, 50%, 75% and 95% quantiles of \mathcal{S} out of 1,000 replications.
2. \mathcal{P}_a : The proportion of including all the important predictors for a given model size $[n/\log n] \approx 37$ out of 1,000 replications.
3. \mathcal{P}_s : The proportion of including each single important predictor for a given model size $[n/\log n] \approx 37$ out of 1,000 replications.

Example 3. We set the sample size $n = 200$ and the predictor dimension $p = 2,000$ throughout. We generate the predictor vector $\mathbf{x} = (X_1, X_2, \dots, X_p)^\top$ from multivariate t distribution $T(\mathbf{0}, \mathbf{\Sigma}, v)$ with $v = 1$ (one degree of freedom) and $\mathbf{\Sigma} = (\sigma_{kl})_{p \times p}$ for $\sigma_{kl} = 0.9^{|k-l|}$. The error term is independently drawn from standard cauchy distribution. We consider the following five models.

$$Y = 0.5X_1 + 0.4X_2 + 0.3X_3 + 0.2X_4 + 0.1X_5 + \varepsilon; \quad (3.1)$$

$$Y = 0.5X_1 + 0.4X_2 + 0.3X_3 + 0.2X_4 + 0.1X_5 + \exp\{3I(X_{20} \leq 4)X_{20}\}\varepsilon; \quad (3.2)$$

$$Y = 2X_1X_2 + 2X_{20}X_{21} + \varepsilon; \quad (3.3)$$

$$Y = 2X_1X_2X_3X_4 + 6X_{21}X_{22} + \varepsilon; \quad (3.4)$$

$$Y = 4X_1X_2 + 3X_3^2 + \exp\{5I(X_{20} \leq 3)X_{20}\}\varepsilon. \quad (3.5)$$

Model (3.1) is a homoscedastic linear model, and models (3.2)-(3.5) contain either interactions or heterogeneity or both. Model (3.4) contains a four-way interaction term. Such complicated interactions are rarely considered in literature.

Each experiment is repeated 1,000 times. The simulation results are charted in Table 2 for \mathcal{S} and in Table 3 for both \mathcal{P}_a and \mathcal{P}_s .

It can be clearly seen that the MBKR-SIS performs the best in most scenarios, following by the RDC-SIS method. The MBKR-SIS retains all interactions and heterogeneity terms with a very high probability. The medians of the minimum model size \mathcal{S} are very close to the number of truly important predictors across all scenarios, and the interquartiles of the minimum model size \mathcal{S} are also very small, indicating that the MBKR ranks the important predictors above the unimportant ones with an overwhelming probability and the performances of the MBKR-SIS are very stable. In terms of the 95% quantile of the minimum model size \mathcal{S} , the RDC-SIS method is slightly worse than the MBKR-SIS method.

The SIS method is designed for homoscedastic linear models. In model (3.1), the median of the minimum model size \mathcal{S} is however as large as 155, which is close to the sample size $n = 200$. This is probably because the SIS procedure is not very effective to detect some weak signals, especially when

the error distribution is heavily tailed. To be precise, given the retained model size $S = \lceil n/\log n \rceil$, the SIS has only a chance of 50.6% to detect X_5 . The SIRS, the RRCS and the DC-SIS perform satisfactorily in model (3.1) while none of them is capable of detecting either heterogeneity terms or interactions in the other four models. In particular, the SIRS detects X_{20} with negligible probabilities 0.036 in model (3.2), 0.024 in model (3.5) and identifies X_{21} with probabilities 0.021 in model (3.3), 0.019 in model (3.4). These observations are in line with what we observed from Example 2. The performance of DC-SIS is not satisfactory, partly because the distribution functions of \mathbf{x} and Y are heavily tailed in these examples.

3.2. An Application

In this section we illustrate the performance of MBKR-SIS through an empirical analysis of rat eye microarray expression dataset collected by Scheetz et al. (2006). They experimented on 120 twelve-week-old male rats, obtained 31,042 different probe sets, and conducted genome-wide linkage analysis with 399 genetic markers. The dataset is available at Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) with GEO accession number GSE5680. In order to gain insight into genetic variation involved in human eye disease, Scheetz et al. (2006) applied expression quantitative trait locus (eQTL) mapping method to 18,976 probes that are considered

Table 2: Simulation results for Example 3. The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size \mathcal{S} such that all the truly important predictors are retained after screening, out of 1,000 replications.

model	quantiles	SIS	SIRS	RRCS	DC-SIS	RDC-SIS	MBKR-SIS
(3.1)	5%	5	5	5	5	5	5
	25%	12	6	5	5	5	5
	50%	155	16	5	10	5	5
	75%	930	137	5	352	5	5
	95%	1896	1638	6	1660	6	6
(3.2)	5%	678	73	26	376	6	6
	25%	1325	392	187	879	7	6
	50%	1699	931	603	1337	9	7
	75%	1889	1538	1282	1674	12	9
	95%	1982	1926	1822	1950	17	13
(3.3)	5%	444	625	557	271	5	4
	25%	978	1252	1234	744	9	5
	50%	1458	1655	1632	1201	15	7
	75%	1791	1908	1840	1627	27	12
	95%	1959	1992	1971	1923	70	31
(3.4)	5%	547	725	730	477	8	6
	25%	1129	1422	1445	1012	15	8
	50%	1569	1759	1744	1467	26	13
	75%	1831	1909	1898	1775	47	20
	95%	1971	1991	1984	1964	122	49
(3.5)	5%	423	597	437	206	5	4
	25%	1020	1178	1043	646	9	5
	50%	1466	1596	1533	1088	16	8
	75%	1791	1871	1817	1509	30	13
	95%	1966	1987	1967	1888	83	41

Table 3: Simulation results for Example 3. For a given model size $S = \lceil n/\log(n) \rceil$, the proportion \mathcal{P}_s of each single important predictor is retained after screening and the proportion \mathcal{P}_a of all truly important predictors are retained after screening.

model			SIS	SIRS	RRCS	DC-SIS	RDC-SIS	MBKR-SIS
(3.1)	\mathcal{P}_s	X_1	0.738	0.788	1.000	0.879	1.000	1.000
		X_2	0.845	0.825	1.000	0.952	1.000	1.000
		X_3	0.778	0.816	1.000	0.902	1.000	1.000
		X_4	0.640	0.774	1.000	0.809	1.000	1.000
		X_5	0.506	0.727	1.000	0.685	1.000	1.000
	\mathcal{P}_a	ALL	0.348	0.607	1.000	0.597	1.000	1.000
(3.2)	\mathcal{P}_s	X_1	0.021	0.758	1.000	0.026	1.000	1.000
		X_2	0.011	0.806	1.000	0.024	1.000	1.000
		X_3	0.029	0.778	1.000	0.039	1.000	1.000
		X_4	0.016	0.740	0.999	0.025	1.000	1.000
		X_5	0.021	0.660	0.993	0.022	0.998	0.999
	X_{20}	0.026	0.036	0.077	0.327	1.000	1.000	
\mathcal{P}_a	ALL	0.000	0.024	0.077	0.002	0.998	0.999	
(3.3)	\mathcal{P}_s	X_1	0.068	0.027	0.028	0.104	0.957	0.996
		X_2	0.073	0.023	0.025	0.122	0.955	0.992
		X_{20}	0.060	0.028	0.033	0.095	0.957	0.991
		X_{21}	0.061	0.021	0.038	0.090	0.948	0.987
	\mathcal{P}_a	ALL	0.001	0.000	0.000	0.006	0.841	0.969
(3.4)	\mathcal{P}_s	X_1	0.089	0.030	0.036	0.089	0.861	0.982
		X_2	0.087	0.030	0.040	0.104	0.918	0.991
		X_3	0.083	0.029	0.046	0.105	0.938	0.998
		X_4	0.079	0.031	0.032	0.097	0.860	0.979
		X_{21}	0.020	0.019	0.021	0.019	0.985	0.979
	X_{22}	0.017	0.026	0.025	0.018	0.983	0.980	
\mathcal{P}_a	ALL	0.001	0.000	0.000	0.000	0.676	0.922	
(3.5)	\mathcal{P}_s	X_1	0.056	0.029	0.039	0.090	0.912	0.962
		X_2	0.056	0.031	0.044	0.091	0.974	0.989
		X_3	0.049	0.029	0.040	0.095	0.988	0.998
		X_{20}	0.015	0.024	0.082	0.145	0.927	0.991
	\mathcal{P}_a	ALL	0.000	0.000	0.002	0.008	0.829	0.946

sufficiently expressed and exhibit at least two-fold variation. Following Scheetz et al. (2006) and Huang, Ma and Zhang (2008), our analysis only focuses on these 18,976 probes and 399 genetic markers, where genes expression levels are continuous and genetic markers are categorical with 3 classes. Chiang et al. (2006) found that the gene TRIM32 at probe 1389163_at is a critical gene to the Bardet-Biedl syndrome, which is a genetic human disease concerning about retina. Our goal is to find out which gene expression levels and genetic markers are the most predictive for the expression level of the gene TRIM32. Because the number of predictors is extremely large while the sample size is relatively small, a screening approach seems to be a natural choice to screen out most of irrelevant genes before an elaborative second-stage analysis. During our analysis, all 18,976 probes are scaled to have zero mean and unit variance.

We first implement the aforementioned six screening procedures, the SIS, the SIRS, the DC-SIS, the RDC-SIS, the RRCS and the MBKR-SIS, on the whole dataset to reduce the dimensionality of the involved genes to $[n/\log(n)] = 25$. We further conduct a second-stage selection based on the retained predictors to obtain a more interpretable model. Following Huang, Horowitz and Wei (2010), we apply the elaborative group SCAD (GS) to nonparametric additive model. The additive components are approximated

by cubic splines. The tuning parameters are selected via generalized cross-validation. To evaluate the performances of different methods, we report the number of probes selected by each method together with the median of the absolute values of residuals, denoted by “Size” and “MAR”, respectively, in the the second and third columns of Table 4.

Table 4: Results for analyzing the rat eye data.

Method	All Data		Test Data	
	Size	MAR	Med. Size	Med. PE
SIS-GS	10	0.0463	8(2.42)	0.2131(0.1128)
SIRS-GS	7	0.0526	7(2.72)	0.1829(0.1008)
RRCS-GS	9	0.0507	7(3.27)	0.1851(0.0991)
DC-SIS-GS	11	0.0425	8(2.72)	0.2095(0.1110)
RDC-SIS-GS	8	0.0530	7(3.35)	0.1854(0.1073)
MBKR-SIS-GS	8	0.0406	7(3.29)	0.1824(0.0981)

Our proposed screening procedure followed by group SCAD selection, the MBKR-SIS-GS, appears superior to other methods in terms of MAR. The boxplot and the histogram of gene TRIM32 are given in Figure 2 (A) and (B). It can be seen that the responses contain some obvious outliers and its distribution is negative skewed. This may explain why the SIS-GS and DC-SIS-GS appear conservative in terms of model size.

We further study the eight genes selected by the MBKR-SIS-GS procedure, which are at probes 1373534_at, 1372453_at, 1372710_at, 1399134_at,

1393510_at, 1378590_at 1380583_s_at and 1373165_at respectively. Six of them are also detected by DC-SIS-GS while five are identified by RRCS-GS and RDC-SIS-GS separately. The genes at probes 1399134_at, 1372453_at, 1373534_at and 1393510_at are also regarded as important by SIRS-GS. SIS-GS only selects two same genes as MBKR-SIS-GS. We also find that the gene at probe 1380583_s_at is missed by all other five competitors.

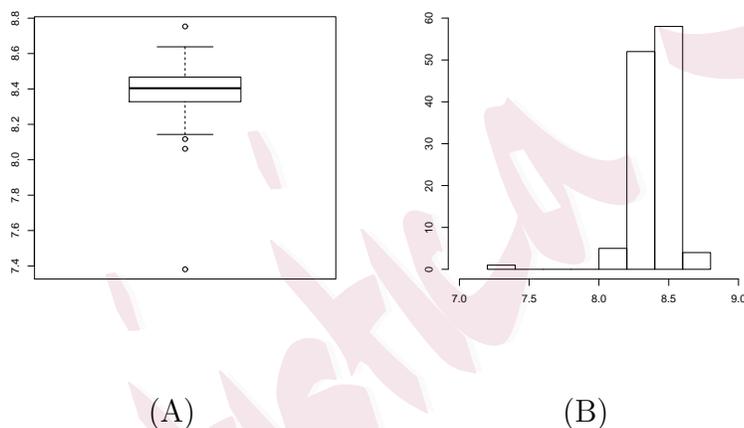


Figure 2: (A) is the boxplot of gene TRIM32 at probe 1389163_at, (B) is the histogram of gene TRIM32 at probe 1389163_at.

Next we randomly split the data into a training set of size 100 and a test set of size 20. Subsequent analyses on the training set are identical to our previous analyses on the whole dataset. We select the tuning parameter for training set again through generalized cross-validation and evaluate the performance of different methods. We repeat this random partition step

200 times. The median of the model size (Med. Size) selected by each method with its standard deviation is summarized in the fourth column of Table 4. In the fifth column of Table 4, we also report the medians of the absolute values of prediction errors (Med. PE) based on the test data set. Their standard deviations are also reported in the parentheses of the fifth column. The smaller the Med. Size and Med. PE are, the better the procedure performs. Our analyses indicate that, in contrast to the other five methods, our proposed approach selects the relatively small number of genes with smallest prediction errors.

3.3. An Iterative Procedure

In Section 3.1 we demonstrate through simulations that the MBKR-SIS procedure is effective in retaining the important predictors which satisfy Assumption (A1). In some situations, however, some important predictors may be marginally independent of the response variable and hence their signals are very weak. To retain such weak signals in the screening stage, we follow the idea of Zhu et al. (2011) and introduce an iterative procedure in this section.

We describe the iterative MBKR-SIS procedure (MBKR-ISIS for short) in what follows. For notational clarity, denote $\mathbf{y} = (Y_1, \dots, Y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$. By definition, \mathbf{y} is an $n \times 1$ vector and \mathbf{X} is an $n \times p$ matrix.

1. We apply the MBKR-SIS procedure to the observations (\mathbf{X}, \mathbf{y}) . Denote by $\mathbf{x}_{\mathcal{A}_1} = (X_1^{(1)}, \dots, X_{p_1}^{(1)})^T$ the predictors selected in this step, where p_1 is a user-specified number. Throughout our simulations, we follow Zhu et al. (2011) and set $p_1 = 5$.
2. Denote $\mathbf{X}_1 = (\mathbf{x}_{1, \mathcal{A}_1}, \dots, \mathbf{x}_{n, \mathcal{A}_1})^T$, $\mathbf{X}_1^c = (\mathbf{x}_{1, \mathcal{A}_1^c}, \dots, \mathbf{x}_{n, \mathcal{A}_1^c})^T$ and $\mathbf{X}_{\text{new}} = \{\mathbf{I}_{n \times n} - \mathbf{X}_1(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T\} \mathbf{X}_1^c$. By definition, \mathbf{X}_1 is an $n \times |\mathcal{A}_1|$ matrix and \mathbf{X}_1^c is an $n \times (p - |\mathcal{A}_1|)$ matrix. We further apply the MBKR-SIS procedure to $(\mathbf{X}_{\text{new}}, \mathbf{y})$ to select p_2 additional predictors. Denote by $\mathbf{x}_{\mathcal{A}_2} = (X_1^{(2)}, \dots, X_{p_2}^{(2)})^T$ the predictors selected in this step.
3. Update \mathcal{A}_1 with $\mathcal{A}_1 \cup \mathcal{A}_2$ and p_1 with $p_1 + p_2$. Repeat the second step until the total number of selected predictors reaches a pre-specified number. The final model selected by MBKR-ISIS is indexed by \mathcal{A}_1 .

The above MBKR-ISIS procedure is quite different from the ISIS procedure proposed by Fan and Lv (2008). To be precise, instead of working with $(\mathbf{X}_{\text{new}}, \mathbf{y})$ in the second step, the ISIS procedure works with $(\mathbf{X}_1^c, \mathbf{y}_{\text{new}})$, where $\mathbf{y}_{\text{new}} = \{\mathbf{I}_{n \times n} - \mathbf{X}_1(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T\} \mathbf{y}$. Another difference is that ISIS uses Pearson correlation while MBKR-ISIS uses the modified BKR correlation defined in (1.3). More importantly, ISIS assumes implicitly that Y depends on \mathbf{x} linearly. In the present context, we hope to retain the model-

free favor of the MBKR-SIS procedure and hence are undesired to impose any dependence structure between Y and \mathbf{x} . The idea of regressing \mathbf{X}_1^c onto \mathbf{X}_1 was first proposed by Zhu et al. (2011) and later elaborated by Zhong and Zhu (2015). The merit of this idea is that it does not assume the dependence of Y onto \mathbf{x} , though it imposes additional distributional assumptions on \mathbf{x} to ensure its validity.

Next we demonstrate the performance of our proposed MBKR-ISIS approach through simulations. We compare both the non-iterative and the iterative versions of our MBKR-SIS and SIS. The simulation studies of SIS/ISIS are conducted by the R package SIS. The sample size n is fixed to be 200 and the predictor dimension p is 2,000.

Example 4. We consider three models:

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon; \quad (3.6)$$

$$Y = I(5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon > 0); \quad (3.7)$$

$$Y = \exp\{(5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4)/2 + \varepsilon\}. \quad (3.8)$$

In the above three models, the predictors $\mathbf{x} = (X_1, X_2, \dots, X_p)^T$ are generated from multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p \times p}$, $\sigma_{ii} = 1$ for $i = 1, \dots, p$, $\sigma_{4i} = \sigma_{i4} = \sqrt{\rho}$ for $i = 1, \dots, p$ and $i \neq 4$, $\sigma_{ij} = \rho$ for $i \neq j$, $j \neq 4$ and $i \neq 4$. We set $\rho = 0.5$. For model (3.6), we generate ε from two distributions: standard normal and standard Cauchy distributions. For

model (3.7) and (3.8), we draw ε from standard normal distribution only. In all three models, the important predictor X_4 is marginally independent of Y . We repeat each experiment 1,000 times.

Table 5 charts the probabilities that every single active predictor and all four truly important predictors rank at top $[n/\log(n)]$ in 1,000 repetitions. The simulation results indicate that both SIS and MBKR-SIS fail to detect X_4 with very high probabilities. The ISIS procedure performs satisfactorily when ε is drawn from normal distribution in model (3.6). However, its performance deteriorates dramatically if the error distribution is heavily-tailed or if the underlying true model is nonlinear. By contract, our proposed MBKR-ISIS is robust to heavily tailed distribution and performs effectively with respective $\mathcal{P}_a = 0.992$ and $\mathcal{P}_a = 0.962$ corresponding to two different distributions of error terms in model (3.6), $\mathcal{P}_a = 0.949$ in model (3.7) and $\mathcal{P}_a = 0.989$ in model (3.8).

4. Discussion

In this paper we suggest a modified Blum-Kiefer-Rosenblatt correlation to measure the possibly nonlinear relation between two random variables. Both the original and the modified Blum-Kiefer-Rosenblatt correlations are nonnegative and equal zero if and only if two random variables are independent, indicating that these two measures lead to a consistent test of

Table 5: Simulation results for Example 4. For a given model size $S = \lceil n/\log(n) \rceil$, the proportion \mathcal{P}_s of each single important predictor is retained after screening and the proportion \mathcal{P}_a of all truly important predictors are retained after screening.

model	ε			SIS	ISIS	MBKR-SIS	MBKR-ISIS
(3.6)	$N(0, 1)$	\mathcal{P}_s	X_1	0.996	1.000	0.994	1.000
			X_2	0.996	1.000	0.997	1.000
			X_3	0.996	1.000	0.995	0.999
			X_4	0.000	1.000	0.000	0.993
		\mathcal{P}_a	ALL	0.000	1.000	0.000	0.992
(3.6)	$t(1)$	\mathcal{P}_s	X_1	0.491	0.557	0.981	0.996
			X_2	0.462	0.551	0.982	0.993
			X_3	0.494	0.567	0.986	0.999
			X_4	0.000	0.818	0.000	0.974
		\mathcal{P}_a	ALL	0.000	0.311	0.000	0.962
(3.7)	$N(0, 1)$	\mathcal{P}_s	X_1	0.975	0.655	0.981	0.998
			X_2	0.981	0.652	0.980	0.993
			X_3	0.971	0.654	0.978	0.997
			X_4	0.000	0.656	0.000	0.961
		\mathcal{P}_a	ALL	0.000	0.648	0.000	0.949
(3.8)	$N(0, 1)$	\mathcal{P}_s	X_1	0.542	0.641	0.993	0.999
			X_2	0.547	0.636	0.995	1.000
			X_3	0.534	0.660	0.994	1.000
			X_4	0.000	0.863	0.000	0.990
		\mathcal{P}_a	ALL	0.000	0.247	0.000	0.989

independence. Compared with the original Blum-Kiefer-Rosenblatt correlation, this modified correlation appears more useful when some of the pre-

dictors are categorical or discrete while others are continuous. Our limited numerical experience indicates that, if all the predictors are continuous, the modified measure behaves very similarly to its original version. We develop a model-free sure independence screening procedure based on a modified Blum-Kiefer-Rosenblatt correlation. Our proposed screening procedure is robust against heavy-tailed distributions, outliers and extreme values. We also propose an iterative approach to detect the important predictors which are marginally independent of the response. In contrast to existing screening approaches, our proposals are effective to detect important predictors upon which the response variable depends nonlinearly. Numerical analysis also shows that our proposed screening approaches are superior to alternative ones in many scenarios. However, we also notice that our proposals are not very effective when the underlying true model is not very sparse or contains many weak signals. Research along this line is warranted.

Supplementary Materials: The technical proofs of Theorems 1-4 are included in supplemental materials.

Acknowledgements: Zhou's research is supported by Innovation Fund of Graduate Students (CXJJ-2016-416), Shanghai University of Finance and Economics, Shanghai, China. Zhu is the corresponding author and

his research is supported by National Natural Science Foundation of China (11371236 and 11422107), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002) and National Youth Top-notch Talent Support Program, China.

References

- Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, **32**, 485–498.
- Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics*, **35**, 2313–2351.
- Chiang, A. P., Beck, J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. (2006). Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*, **103**, 6287–6292.
- 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.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle

REFERENCES37

- 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*, **70**, 849–911.
- Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, **10**, 2013–2038.
- Fan, J., Samworth, R., and Wu, Y. SIS: sure independence screening. R package version 0.6; 2010. Available from: <http://CRAN.R-project.org/package=SIS>.
- Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, **38**, 3567–3604.
- Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–135.
- 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.
- Hoeffding, W. (1948). A non-parametric test of independence. *Annals of Mathematical Statistics*, **19**, 546–557.
- Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, **38**, 2282–2313.
- Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regres-

REFERENCES38

- sion models. *Statistica Sinica*, **18**, 1603–1618.
- Li, G., Peng, H., Zhang, J., and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, **40**, 1846–1877.
- Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107**, 1129–1139.
- Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, **100**, 229–234.
- Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: a nonparametric model-free screening method. *The Annals of Statistics*, **43**, 1471–1497.
- Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, **103**, 14429–14434.
- Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, **109**, 1302–1318.
- Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, **3**, 1236–1265.
- Székely, G. J., Rizzo, M. L., Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **35**, 2769–2794.

REFERENCES39

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, **58**, 267–288.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68**, 49–67.

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

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.

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

School of Statistics and Management, Shanghai University of Finance and Economics,
Shanghai 200433, China.

E-mail: yqzhou1991@hotmail.com

Research Center for Applied Statistical Science and Institute of Statistics and Big Data, Renmin
University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn.