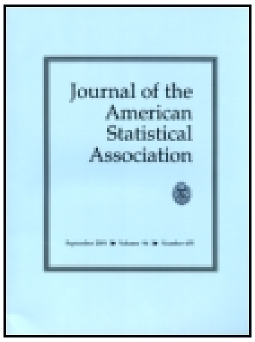

Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

A Generic Sure Independence Screening Procedure

Wenliang Pan , Xueqin Wang , Weinan Xiao & Hongtu Zhu

A Generic Sure Independence Screening Procedure

WENLIANG PAN, XUEQIN WANG, WEINAN XIAO AND HONGTU ZHU[*]

Abstract

Extracting important features from ultra-high dimensional data is one of the primary tasks in statistical learning, information theory, precision medicine and biological discovery. Many of the sure independent screening methods developed to meet these needs are suitable for special models under some assumptions. With the availability of more data types and possible models, a model-free generic screening procedure with fewer and less restrictive assumptions is desirable. In this paper, we propose a generic nonparametric sure independence screening procedure, called BCor-SIS, on the basis of a recently developed

[*] Wenliang Pan and Weinan Xiao are Research Associates, Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, P. R. China. Xueqin Wang is Professor, Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, P. R. China; and Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, 510080, China; and Xinhua College, Sun Yat-Sen University, Guangzhou, 510520, China. Hongtu Zhu is the corresponding author and the Bao-Shan Jing Professor, Department of Biostatistics, The University of Texas, MD Anderson Cancer Center, Houston, Texas, 77030. Email: hzhu5@mdanderson.org. Dr. Wang's research is partially supported by the National Science Foundation of China (NSFC) for Excellent Young Scholars (11322108), NCET(12-0559), and NSFC(11001280). Dr. Zhu's work was partially supported by the US National Institutes of Health (grant MH086633), the National Science Foundation (grants SES-1357666 and DMS-1407655), a senior investigator grant from the Cancer Prevention Research Institute of Texas, and the endowed Bao-Shan Jing Professorship in Diagnostic Imaging. All authors contributed equally to this paper; the authors are listed in alphabetic order.

universal dependence measure: Ball correlation. We show that the proposed procedure has strong screening consistency even when the dimensionality is an exponential order of the sample size without imposing sub-exponential moment assumptions on the data. We investigate the flexibility of this procedure by considering three commonly encountered challenging settings in biological discovery or precision medicine: iterative BCor-SIS, interaction pursuit, and survival outcomes. We use simulation studies and real data analyses to illustrate the versatility and practicability of our BCor-SIS method.

Keywords: Ball Correlation; Sure Independence; Rank; Variable Screening.

1. INTRODUCTION

Ultra-high dimensional data arise from a variety of applications; analyzing such data poses major computational and statistical challenges to modern statistical inference. For instance, many studies in precision medicine search for risk factors among various types of data such as clinical, genomic and protein data for complex diseases. Many penalized variable selection techniques have been introduced to identify a small set of 'significant' factors related to disease status. Such penalized variable selection methods include the least absolute shrinkage and selection operator (LASSO, (Tibshirani, 1996)) and smoothly clipped absolute deviation (SCAD, Fan and Li (2001)), among many others. However, most of these methods suffer from the curse of dimensionality due to diverging spectra and noise accumulation in the ultra-high dimensional feature space (Fan, Feng and Tong, 2010). High variance and overfitting have been major concerns in this setting.

To overcome the issues associated with ultra-high dimensionality, many marginal screening techniques, such as the sure independence screening (SIS) procedure, have been shown to filter out many uninformative variables in many scenarios (Fan and Lv, 2008). Subsequently, standard penalized variable selection methods can be applied to the remaining variables. A desired marginal screening procedure possesses the sure screening property; that is, with probability close to 1, the procedure retains all of the important variables. The key idea of the SIS procedure is to rank all predictors by using a utility measure between the response and each predictor and then to retain the top variables for further investigation. The SIS procedure has been rapidly extended to various models and data types (Fan, Song et al., 2010; Fan et al., 2009; Zhao and Li, 2012; Gorst-Rasmussen and Scheike, 2013). Further extensions to complex cases have been proposed along the same lines. Zhu et al. (2011) used the expectation of the square of the correlation between the predictor and an indicator function of the response for an ultra-high-dimensional multi-index model (SIRS), and Li et al. (2012) used distance correlation to carry out marginal screening (DC-SIS). These two methods are model-free but not robust to the predictors whose distributions are heavy tail. Similar to the work of Li et al. (2012), Shao and Zhang (2014) proposed a martingale difference correlation for high-dimensional variable screening (MDC-SIS). For variable interaction, Fan et al. (2017) proposed a sure independent screening procedure based on Pearson correlation (P-IT), and Kong et al. (2017) developed one based on distance correlation (DC-IT).

Table 1 summarizes the strengths and weaknesses of a set of representative SIS procedures. Specifically, we consider six important aspects, including multivariate response, group predictor, survival response, collinear predicts, predictor interaction, nonlinear model, and robustness. The aim of this paper is to develop a screening procedure that can work for all these six aspects, while detecting complex associations under less restrictive assumptions. Our screening procedure is based on a recently developed universal dependence measure: Ball correlation (BCor, Anonymity (2017)). BCor efficiently measures the dependence between two random vectors, which is between 0 and 1, and 0 if and only if these two random vectors are independent under some mild conditions. This property enables us to use the empirical BCor between the response of interest and each predictor vector to rank the predictors.

Our proposed SIS procedure, called BCor-SIS, is a generic method that is model-free and has fewer and less restrictive data assumptions. Four defining characteristics

Table 1: The comparison of sure independent screening methods, including SIS: sure independence screening; I-SIS: iterative SIS; SIRS: sure independent ranking and screening; DC-SIS: distance correlation SIS; I-DC-SIS: iterative DC-SIS; MDC-SIS: martingale difference correlation; P-IT: Pearson correlation for variable interaction; DC-IT: distance correlation for variable interaction; CRIS: censored rank independence screening; BCor-SIS: Ball correlation SIS; and I-BCor-SIS: iterative BCor-SIS. We consider six aspects, including A: multivariate response; B: group predictor; C: survival response; D: collinear predictors; E: predictor interaction; F: nonlinear model; and G: robustness. Moreover, "×" and "√", respectively, denote "not working" and "working".

Methods	A	B	C	D	E	F	G
SIS	×	×	×	×	×	×	×
I-SIS	×	×	×	√	×	×	×
SIRS	×	×	×	×	×	×	×
DC-SIS	√	√	×	×	√	√	×
I-DC-SIS	√	√	×	√	√	√	×
MDC-SIS	√	√	×	×	√	√	×
P-IT	×	×	×	×	√	×	×
DC-IT	√	√	×	×	√	√	×
CRIS	×	×	√	×	×	×	√
BCor-SIS	√	√	√	×	√	√	√
I-BCor-SIS	√	√	×	√	√	√	√

of BCor-SIS are given here. i) It has a strong screening consistency property (Huang et al., 2013) without finite sub-exponential moments of the data even when the dimensionality is an exponential order of the sample size since empirical BCor is a function of some indictors. ii) It is nonparametric and has the property of robustness, which is also due to the boundedness of the empirical BCor. iii) It works well for complex responses and/or predictors from the definition of BCor. iv) It can extract important features even when the underlying model is complicated. We also evaluate BCor-SIS and investigate its flexibility in three commonly encountered challenging settings in biological discovery or precision medicine: iterative BCor-SIS, interaction pursuit, and survival outcomes. We show that the modified procedures also theoretically exhibit strong or sure screening properties, and illustrate their versatility and practicability by simulation studies and real data analyses. We have developed the BCor-SIS package using Rcpp and released it through the website https://github.com/BIG-S2/SBISIS.

The rest of this paper is organized as follows. We introduce our BCor-SIS procedure and establish its strong screening property in Section 2. We consider three challenging settings of BCor-SIS and the sure screening properties in Section 3. We carry out simulations to examine the finite sample performance of BCor-SIS in Section 4. We apply BCor-SIS to a real data set in Section 5. We make some concluding remarks in Section 6. We defer the technical details to the Appendix.

2. METHODOLOGIES

2.1 A Review of Ball Correlation

We first introduce the definition of Ball covariance and then review some of its theoretical properties. Ball covariance is an important tool for measuring the dependence between two random vectors. Specifically, denote $W = (X, Y)$ and let X and Y be random vectors in two separable Banach spaces $(\mathscr{X}, \zeta_X)$ and $(\mathscr{Y}, \zeta_Y)$, where ζ_X and ζ_Y are norms. Let θ be a Borel probability measure on $\mathscr{X} \times \mathscr{Y}$ and (X, Y) be a B-valued random variable defined on a probability space $(\Omega, \mathfrak{A}, P)$ such that $(X, Y) \sim \theta$, $X \sim \mu$, and $Y \sim \nu$, where μ and ν are, respectively, a Borel probability measure on $\mathscr{X}$ and $\mathscr{Y}$. Moreover, we use $\bar{B}(x_1, \zeta_X(x_1, x_2))$ (or $\bar{B}_{\zeta_X}(x_1, x_2)$) to denote the closed ball with the center x_1 and the radius $\zeta_X(x_1, x_2)$ in $\mathscr{X}$ and $\bar{B}(y_1, \zeta_Y(y_1, y_2))$ (or $\bar{B}_{\zeta_Y}(y_1, y_2)$) to denote the closed ball with the center y_1 and the radius $\zeta_Y(y_1, y_2)$ in $\mathscr{Y}$.

Definition 1. *The Ball covariance* $\mathbf{BCov}(X, Y)$ *is defined as the square root of*

$$\mathbf{BCov}^2(X, Y) = \int\int_{U \times V} [\theta - \mu \otimes \nu]^2(\bar{B}_{\zeta_X}(x_1, x_2) \times \bar{B}_{\zeta_Y}(y_1, y_2))$$
$$\theta(dx_1, dy_1)\theta(dx_2, dy_2),$$

where $\mu \otimes \nu$ is a product measure on $\mathscr{X} \times \mathscr{Y}$.

Definition 2. *The Ball correlation* $\mathbf{BCor}(X, Y)$ *is defined as the square root of*

$$\mathbf{BCor}^2(X, Y) = \mathbf{BCov}^2(X, Y)/\sqrt{\mathbf{BCov}^2(X, X) \times \mathbf{BCov}^2(Y, Y)},$$

if $\mathbf{BCov}(X, X) \times \mathbf{BCov}(Y, Y) > 0$, *or* 0 *otherwise.*

We then calculate the sample forms of $\mathbf{BCov}(X, Y)$ and $\mathbf{BCor}(X, Y)$ based on an observed random sample $\mathbf{W} = (\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ generated from the joint distribution of random vectors X and Y. Let $\delta^X_{ij,k} := I(X_k \in \bar{B}_{\zeta_X}(X_i, X_j))$, where $I(A)$ denotes an indicator function of event A. Therefore, $\delta^X_{ij,k}$ indicates whether X_k is located in the closed ball $\bar{B}_{\zeta_X}(X_i, X_j)$, and $\delta^X_{ij,kl} = \delta^X_{ij,k}\delta^X_{ij,l}$ denotes whether both X_k and X_l fall into the closed ball $\bar{B}_{\zeta_X}(X_i, X_j)$. We also define $\xi^X_{ij,klst} = (\delta^X_{ij,kl} + \delta^X_{ij,st} - \delta^X_{ij,ks} - \delta^X_{ij,lt})/2$. Similarly, we define $\delta^Y_{ij,k}$, $\delta^Y_{ij,kl}$, and $\xi^Y_{ij,klst}$ for Y. Next, we introduce the following definitions.

Definition 3. *Empirical Ball covariance* $\mathbf{BCov}_n(\mathbf{X}, \mathbf{Y})$ *is defined as the square root of* $\mathbf{BCov}^2_n(\mathbf{X}, \mathbf{Y}) = n^{-6} \sum_{i,j,k,l,s,t=1}^{n} \xi^X_{ij,klst}\xi^Y_{ij,klst}$.

According to Definition 3, if we replace $\delta^X_{ij,kl}$ and $\delta^Y_{ij,kl}$ by symmetric positive definite kernel functions $k_X(X_k, X_l)$ and $k_Y(Y_k, Y_l)$, then Ball covariance is equivalent to Hilbert-Schmidt independence criterion (HSIC, Gretton et al. (2008)).

Definition 4. *Empirical Ball correlation* $\mathbf{BCor}_n(\mathbf{X},\mathbf{Y})$ *is defined as the square root of*

$$\mathbf{BCor}_n^2(\mathbf{X},\mathbf{Y}) = \mathbf{BCov}_n^2(\mathbf{X},\mathbf{Y})/\sqrt{\mathbf{BCov}_n^2(\mathbf{X},\mathbf{X}) \times \mathbf{BCov}_n^2(\mathbf{Y},\mathbf{Y})},$$

if $\mathbf{BCov}_n^2(\mathbf{X},\mathbf{X}) \times \mathbf{BCov}_n^2(\mathbf{Y},\mathbf{Y}) > 0$*, or* 0 *otherwise.*

We review some theoretical properties of Ball correlation as follows.

Lemma 1. *Let* S_θ*,* S_μ *and* S_ν *be the support sets of* θ*,* μ*, and* ν*, respectively. Consider three conditions as follows:*

- ***Condition 1:*** θ *is a discrete distribution or continuous distribution with continuous density function, or more generally a mixture of these two distributions;*

- ***Condition 2:*** $S_\mu \times S_\nu = S_\theta$ *and* $\mathscr{X} \times \mathscr{Y}$ *is a finite dimensional Banach space;*

- ***Condition 3:*** $\mathscr{X} \times \mathscr{Y} = S_\theta$ *and* $\mathscr{X} \times \mathscr{Y}$ *is a separable Banach space with a normalized basis.*

If one of Conditions 1-3 holds, then $\mathbf{BCor}(X,Y) = 0$ *is equivalent to* $\theta = \mu \otimes \nu$*.*

Lemma 1 shows that Ball correlation can capture various dependence relationships between X and Y in Banach space under some mild conditions. Although distance correlation and HSIC are also powerful dependence measures for detecting multivariate and nonlinear relationship, they require strong negative type condition (Lyons, 2013) or positive type condition (Sejdinovic et al., 2013). Specifically, distance correlation (or HSIC) is suitable for the type of spaces that can be embed to Hilbert space or reproducing kernel Hilbert space, whereas many standard spaces are neither negative type nor positive type. In contrast, Ball correlation can detect multivariate, nonlinear and non-Euclidean (i.e., infinite dimension and manifold) dependences without assuming the negative and positive type conditions.

Lemma 2. *(Properties of Ball correlation)*

(i) $\mathbf{BCor}(X,Y)$ *is an nondecreasing function of* $|\rho|$ *when* X *and* Y *are standard normal variables with* $Cov(X,Y) = \rho$*.*

(ii) $0 \leq \mathbf{BCor}_n(\mathbf{X}, \mathbf{Y}), \mathbf{BCor}(X, Y) \leq 1$.

(iii) *If there exist a vector* $\mathbf{a}$*, a nonzero real number* b *and an orthonormal matrix* C *such that* $\mathbf{Y} = \mathbf{a} + bC\mathbf{X}$ *or* $Y = \mathbf{a} + bX$*, then* $\mathbf{BCor}_n(\mathbf{X}, \mathbf{Y}) = \mathbf{BCor}(X, Y) = 1$ *holds.*

Lemma 2 (i) shows that BCor is order preserving with respect to the correlation strength under the Gaussian assumption. We will use this result to establish a relationship between I-BCor-SIS and I-SIS (Fan and Lv, 2008) in Theorem 2. The second item explains why BCor is standardized based on the Cauchy-type inequality. The third item ensures that $\mathbf{BCor}_n(\mathbf{X}, \mathbf{Y})$ reaches the maximum when $\mathbf{X}$ and $\mathbf{Y}$ are linear relationships.

Lemma 3. $\mathbf{BCov}_n(\mathbf{X}, \mathbf{Y})$ *and* $\mathbf{BCor}_n(\mathbf{X}, \mathbf{Y})$ *almost surely converges to* $\mathbf{BCov}(X, Y)$ *and* $\mathbf{BCor}(X, Y)$*, respectively.*

Lemma 4. *If* X *and* Y *are independent, then* $n\mathbf{BCov}_n^2(\mathbf{X}, \mathbf{Y})$ *converges to a mixture of* χ^2 *distribution.*

$$n\mathbf{BCov}_n^2(\mathbf{X}, \mathbf{Y}) \xrightarrow[n \longrightarrow \infty]{d} \sum_{v=1}^{\infty} \lambda_v Z_v^2,$$

where Z_v*s' are independent standard normal random variables and* λ_v*s' are nonnegative constants that depend on the distribution of* (X, Y)*.*

2.2 BCor-based Sure Independence Screening Procedure

In this section, we propose the BCor-based SIS procedure (BCor-SIS) as follows. The BCor-SIS procedure shares the same model-free property as DC-SIS (Li et al., 2012), but it has the distinctive feature of not requiring finite moments, since BCor is the rank function of distance. This leads to robustness for data with a heavy-tailed distribution.

Let Y be a response vector and $X = (X_1^T, \ldots, X_p^T)^T$ be a vector of predictors, where each X_r is a $q_r \times 1$ vector for either grouped or categorical data for $r = 1, \ldots, p$. The goal of feature screening is to identify a set of predictors in X that is relevant to at least one component of Y. We define the index sets of active and inactive predictors without specifying a statistical model as follows:

$$\begin{aligned}\mathcal{A} &= \{r:\ P(B|X_r) \text{ is non-constant in } X_r \text{ for some } B \in \sigma(Y)\}, \\ \mathcal{M} &= \{r:\ P(B|X_r) \text{ is a constant in } X_r \text{ for any } B \in \sigma(Y)\}.\end{aligned} \tag{1}$$

Under the conditions of Lemma 1, $\mathcal{A}$ and $\mathcal{M}$ are equivalent to $\{r:\ \rho_r > 0\}$ and $\{r:\ \rho_r = 0\}$, respectively. A good screening approach is to identify an index set that includes all indexes in $\mathcal{A}$, while excluding as many indexes as possible in $\mathcal{M}$ as the sample size tends to infinity.

Our BCor-SIS is based on the assumption that the predictors with larger BCor are more strongly correlated with the response vector. Specifically, BCor-SIS consists of two steps:

(i) Calculate $\widehat{\rho}_r = \mathbf{BCor}_n^2(\mathbf{X}_r, \mathbf{Y})$, which is an estimate of $\rho_r = \mathbf{BCor}^2(X_r, Y)$, and use it as a marginal utility of X_r for $r = 1, \ldots, p$;

(ii) Select the X_rs that fall into $\widehat{\mathcal{A}}_n^* = \{r:\ \widehat{\rho}_r \geq \tau_n,\ r = 1, \ldots, p\}$, where τ_n is a pre-specified constant. We discuss the selection of τ_n in the following section.

2.3 Theoretical Properties

In this section, we study the screening property of BCor-SIS. We need two conditions as follows:

(C1) There exist a constant $c > 0$ and $0 \leq \kappa < 1/2$ such that $\min_{r\in\mathcal{A}} \rho_r \geq 2cn^{-\kappa}$.

(C2) Assume $\log(p) = o(n^{1-2\kappa})$, where κ is defined in condition (C1).

Condition (C1) is critical for performing ultra high-dimensional feature screening. It requires that the values of BCor are not too small between each active predictor vector and the response vector. It is similar to condition 3 of Fan and Lv (2008) and condition (C2) of Li et al. (2012). Condition (C2) assumes that p diverges at an exponential rate of n. SIS, SIRS, DC-SIS, and MDC-SIS also require the common sub-exponential moment assumption on Y and each X_r to establish the sure screening property. Our BCor-SIS procedure does not require this condition, which makes it suitable for more relaxed model assumptions.

With only conditions (C1) and (C2), we prove that BCor-SIS possesses strong screening consistency (Huang et al., 2013). We defer the proof to the Appendix.

Theorem 1. (*Strong screening consistency of BCor-SIS*) *There exists a positive constant $c_1 > 0$ such that*

$$\mathbb{P}(\max_{1\leq r\leq p} |\widehat{\rho}_r - \rho_r| \geq cn^{-\kappa}) \leq O(p \times \exp(-c_1 n^{1-2\kappa})).$$

If condition (C1) holds and (X_r, Y) satisfies the conditions of Lemma 1, then for any $\tau_n \in (0, 2cn^{-\kappa})$, there exists a constant $c_2 > 0$ such that

$$\mathbb{P}(\mathcal{A} \subset \widehat{\mathcal{A}}_n^*) \geq 1 - O(\gamma \exp(-c_2 n^{1-2\kappa})),\ \mathbb{P}(\widehat{\mathcal{A}}_n^* \subset \mathcal{A}) \geq 1 - O(\gamma^* \exp(-c_2 n^{1-2\kappa})),$$

where γ and γ^* are the cardinality of $\mathcal{A}$ and $\widehat{\mathcal{A}}_n^*$. Thus, if condition (C2) also holds, then the property of strong screening consistency holds,

$$\mathbb{P}(\widehat{\mathcal{A}}_n^* = \mathcal{A}) \xrightarrow[n\to\infty]{a.s.} 1.$$

Remark 1. *The property of strong screening consistency says that the selected set $\widehat{\mathcal{A}}_n^*$ is exactly equal to the active set $\mathcal{A}$ with probability 1. This implies the SIS property instead of vice versa. Huang et al. (2013) proved this strong screening consistency of feature screening for ultrahigh-dimensional categorical data, but that work requires the assumption that the measure between the response and the predictors in an inactive set is equal to 0. Wang et al. (2015) also proved the strong screening consistency property for SIS under the restricted diagonally dominant (RDD) condition.*

2.4 Tuning Parameter Selection

We discuss the choice of the tuning parameter τ_n for $\widehat{\mathcal{A}}_n^*$. Theorem 1 shows that the ideal value of τ_n lies in the interval $(0, 2cn^{-\kappa})$, but the true values of c and κ are unknown. We introduce two rules to determine the value of τ_n below.

The first rule is a soft cutoff rule, which introduces some auxiliary variables for the choice of threshold (Zhu et al., 2011). First, we generate m auxiliary variables, denoted as $z_1, \ldots, z_m$, from some pre-specified distribution, such as a standard normal distribution. In practice, we usually set $m = p$. Second, we calculate the BCor between each z_j and the response Y for $j = 1, \ldots, m$. Third, we set τ_n as the largest value of all BCor values between z_j and Y for all j. Fourth, we select all the predictors for which the BCor is larger than τ_n. The rationale for the soft cutoff rule is based on the theoretical results in Theorem 1. Since all $z_1, \ldots, z_m$ are independent of Y, the true BCor between z_j and Y is equal to zero for all j. It follows from Theorem 1 that τ_n lies between 0 and $2cn^{-\kappa}$ almost surely. Therefore, it provides a reasonable estimated value of τ_n.

Furthermore, there is another predictor selection soft rule, which is based on controlling the false positive rate, which is proposed by Zhao and Li (2012). Similar to Zhu et al. (2011), the method based on controlling the false positive rate is also a data-driven method. It takes advantage of the statistical distribution of the active predictor size and its false discovery rate (FDR). The active predictor size is the one that minimize its expectation of FDR.

The second rule is a hard cutoff rule proposed by Fan and Lv (2008). This method is based on a sparsity assumption that only $o(n)$ predictors are truly associated with the response variable. This phenomenon is common in gene selection or risk control problems. Therefore, the hard cutoff rule suggests selecting d variables with the largest BCor values, where d is usually chosen to be $[n/\log n]$ or $n-1$, in which $[a]$ denotes the integer part of a. According to Theorem 1, we have the following corollary.

Corollary 1. *If conditions (C1) and (C2) hold, then we have*

$$\mathbb{P}(\max_{r\in\mathcal{M}} \widehat{\rho}_r < \min_{r\in\mathcal{A}} \widehat{\rho}_r) \xrightarrow[n\to\infty]{a.s.} 1.$$

Corollary 1 has several important implications. The BCor values of all active predictors are larger than those of all inactive predictors asymptotically. It is reasonable to choose the predictors for which the BCor is among the largest d ones. In the next section, we adopt the hard cutoff method in our simulations and examine the effects of different ds on the convergence rate of the sure screening property.

3. EXTENSIONS

We consider three extensions of BCor-SIS, including iterative BCor-SIS, linear interaction models, and censored responses.

3.1 Iterative BCor-SIS

Since the BCor-SIS approach picks out the important predictors based on their marginal correlation with Y, it may suffer from two possible issues. The first issue is that BCor-SIS may miss some important predictors that are marginally uncorrelated, but jointly correlated with the response. The second issue is that BCor-SIS may mistakenly select an irrelevant variable that is highly correlated with some of the truly active predictors.

We consider an extension of BCor-SIS by accounting for the joint distribution information. Similar to the work of Zhong and Zhu (2015), we use an iterative BCor-SIS approach to enhance its power. Let d be the pre-specified number of total selected predictors. The key steps of the iterative BCor-SIS are as follows. First, we use BCor-SIS to select a small subset of k ($k < d$) predictors. Second, we regress the response and the remaining predictors over this subset of predictors in order to remove the influence of the subset of selected predictors. The residuals can be explained as the projection of the response and all remaining predictors onto the orthogonal complement space of the selected predictors. Third, we treat the residuals as new predictors and repeat the previous two steps until we pick d predictors. More specifically, the procedure is shown in algorithm 1.

A key advantage of the I-BCor-SIS procedure is that it takes advantage of the information of selected predictors. When an active variable is marginally independent of the response due to its correlation with other active variables, Step 2 aims to break down its plausible marginal independence with the response and make it marginally detectable. Moreover, when many irrelevant variables are highly correlated with the active variables that have strong signals, Step 2 can dramatically reduce the effects of these irrelevant variables on correct selection once their correlated active variables are

Algorithm 1 I-BCor-SIS

Step 1: We apply the BCor-SIS method to the response $\boldsymbol{Y}$ and all predictors $\boldsymbol{X}$. Suppose that $d^{(1)}$ predictors are selected. We denote this subset of predictors as $\widehat{\mathcal{A}}_1$.

Step 2: Denote the design matrix of $\widehat{\mathcal{A}}_1$ and $\{X_1, \ldots, X_p\} \backslash \widehat{\mathcal{A}}_1$ as $\boldsymbol{X}_1$ and X_1^c, respectively. Then we define the predictor residual matrix as

$$\boldsymbol{X}^* = \boldsymbol{X}_1^c - E(\boldsymbol{X}_1^c | \boldsymbol{X}_1), \boldsymbol{Y}^* = \boldsymbol{Y} - E(\boldsymbol{Y} | \boldsymbol{X}_1),$$

where $E(\boldsymbol{X}_1^c | \boldsymbol{X}_1)$ and $E(\boldsymbol{Y} | \boldsymbol{X}_1)$ are the projections on $\boldsymbol{X}_1$. Next we apply the BCor-SIS procedure to $\boldsymbol{Y}^*$ and the predictors in $\boldsymbol{X}^*$, and then select $d^{(2)}$ predictors. We denote this subset of predictors as $\widehat{\mathcal{A}}_2$.

Step 3: We update $\widehat{\mathcal{A}}_1$ with $\widehat{\mathcal{A}}_1 \bigcup \widehat{\mathcal{A}}_2$. We repeat step 2 until the total number of selected predictors exceeds the prespecified number d. The final selected predictor set is $\widehat{\mathcal{A}}_1$.

selected in Step 1. This can enhance the detection of the rest of the active variables. Although the I-BCor-SIS procedure uses the hard cutoff rule to determine the active variable set, the soft cutoff rule is also applicable here.

The following theorem shows that the iterative BCor-SIS is asymptotically equivalent to forward-stepwise selection (Hastie et al., 2004) when both $\mathbf{X}$ and Y are normally distributed.

Theorem 2. *Suppose that $\mathbf{X} \sim N(\eta_x, \Sigma)$ and $Y \sim N(\eta_y, \sigma)$, the I-BCor-SIS is asymptotically equivalent to the forward-stepwise selection. Furthermore, the components of $\boldsymbol{X}^*$ at every step are independent predictors and their indexes are the same as the predictors selected asymptotically by the forward-stepwise selection.*

As shown in Hastie et al. (2004), the forward-stepwise selection is sub-optimal compared to best subset selection. Thus, Theorem 2 implies that I-BCor-SIS is no worse than I-SIS (Fan and Lv, 2008) when n is large, whereas I-BCor-SIS can be used to detect a more complex nonlinear relationship, which is shown in Tables 2 and 3 below.

3.2 Linear Interaction Models

Although BCor-SIS is a model-free procedure, its performance can be competitive even under some specific modeling settings. As an illustration, we consider a linear interaction model. This type of interaction models has wide applications, such as in the analysis of gene-gene interactions and gene-environment interactions in genome-wide association studies, and treatment and covariate interactions in personalized medicine. Specifically, we consider a response variable Y and a $p \times 1$ vector of covariates $(X_1, \ldots, X_p)$. A linear interaction model is given by

$$\begin{aligned} Y =& \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{k=1}^{p-1} \sum_{j=k+1}^{p} \gamma_{kj} X_i X_j + \sum_{k=1}^{p-2} \sum_{j=k+1}^{p-1} \sum_{l=j+1}^{p} \gamma_{kjl} X_l X_j X_k \\ &+ \sum_{m=1}^{p-3} \sum_{j=m+1}^{p-2} \sum_{k=j+1}^{p-1} \sum_{l=k+1}^{p} \gamma_{mjkl} X_m X_j X_k X_l + \varepsilon, \end{aligned}$$

where β_0, β_j, γ_{kj}, γ_{kjl}, and γ_{mjkl} are, respectively, the intercept and regression coefficients for the main effects, and those for the second-, third-, and fourth-order interactions, and ε is a measurement error independent of $(X_1, \ldots, X_p)$ with mean zero and finite variance. Since the interactions often have heavy-tailed distributions and our BCor-SIS method can efficiently deal with such heavy-tailed distribution data, BCor-SIS should be suitable for the interaction pursuit. The interaction screening procedure based on Ball correlation is summarised as follows.

We write $X_r^* = f(X_r)$ and $Y^* = g(Y)$, where $f(\cdot)$ and $g(\cdot)$ are some specific transformations that may improve performance in interaction variable screening. Fan et al. (2017) suggested that the interaction predictor X_r can be retained by ranking the marginal correlations between the squared response Y^2 and the squared predictor X_r^2, which implies that the squared transformation can strength the dependence between Y and X_r. Indeed, absolute transformation maybe an better alternative because of the moment conditions. Here we still choose the squared transformation since Ball correlation does not require moment condition. Define two quantities of Ball correlation as $\rho_{r_1} = \mathbf{BCor}^2(X_{r_1}, Y)$ and $\rho^*_{r_2} = \mathbf{BCor}^2(X^*_{r_2}, Y^*)$ for $1 \leq r_1, r_2 \leq p$.

Similarly, let $\mathcal{A}$, $\mathcal{S}$, and $\mathcal{M}$, respectively, denote the index set of active main variables, active interaction variables and inactive variables as follows:

$$\begin{aligned}
\mathcal{A} &= \{r: \ \text{ some } X_r \text{ are main variables which affect } Y\},\\
\mathcal{S} &= \{r: \ \text{ some } X_r \text{ are interaction variables which affect } Y\},\\
\mathcal{M} &= \{r: \ \text{ any } Y \text{ does not depend on } X_r\}.
\end{aligned}$$

Similarly, we can define the estimators as $\widehat{\mathcal{A}}_n^* = \{r_1 : \widehat{\rho}_{r_1} \geq \tau_1,\ r_1 = 1, \dots, p\}$ and $\widehat{\mathcal{S}}_n^* = \{r_2 : \widehat{\rho}_{r_2}^* \geq \tau_2,\ r_2 = 1, \dots, p\}$.

(I1) There exist some constants $c_1, c_2 > 0$ and $0 \leq \kappa_1, \kappa_2 < 1/2$ such that $\min_{r_1 \in \mathcal{A}} \rho_{r_1} \geq 2c_1 n^{-\kappa_1}$ and $\min_{r_2 \in \mathcal{S}} \rho_{r_2}^* \geq 2c_2 n^{-\kappa_2}$.

(I2) Assume $\log(p) = o(n^{1-2\max\{\kappa_1,\kappa_2\}})$, where κ_1, κ_2 is defined in condition (I1).

Theorem 3 implies that new screening procedure based on BCor-SIS for the linear interaction model satisfies strong screening consistency property.

Theorem 3. (*Strong screening consistency for linear interaction models*) *There exist some positive constants c_3 and c_4 such that*

$$\begin{aligned}
&\mathbb{P}(\max_{1 \leq r_1 \leq p} |\widehat{\rho}_{r_1} - \rho_{r_1}| \geq c_1 n^{-\kappa_1}) \leq O(p \times \exp(-c_3 n^{1-2\kappa_1})),\\
&\mathbb{P}(\max_{1 \leq r_2 \leq p} |\widehat{\rho}_{r_2}^* - \rho_{r_2}^*| \geq c_2 n^{-\kappa_2}) \leq O(p \times \exp(-c_4 n^{1-2\kappa_2})),
\end{aligned}$$

when $n > \max\{(320/c_1)^{\frac{1}{1-\kappa_1}}, (320/c_2)^{\frac{1}{1-\kappa_2}}\}$. If condition (I1) holds and (X_r, Y) satisfies the conditions of Lemma 1, then for any $\tau_1 \in (0, 2c_1 n^{-\kappa_1})$ and $\tau_2 \in (0, 2c_2 n^{-\kappa_2})$, there exist some positive constants c_5 and c_6 such that

$$\mathbb{P}(\mathcal{A} = \widehat{\mathcal{A}}_n^*, \mathcal{S} = \widehat{\mathcal{S}}_n^*) \geq 1 - O(p\exp(-c_5 n^{1-2\kappa_1}) + p\exp(-c_6 n^{1-2\kappa_2})).$$

Furthermore, if condition (I2) also holds, then the property of strong screening consistency holds, that is,

$$\mathbb{P}(\mathcal{A} = \widehat{\mathcal{A}}_n^*, \mathcal{S} = \widehat{\mathcal{S}}_n^*) \xrightarrow[n\to\infty]{a.s.} 1.$$

3.3 Censored Survival Data

In this subsection, we extend BCor-SIS to deal with censored survival data and high-dimensional predictors. Let T be the failure time variable, C be the censoring time variable, and $X = (X_1, \ldots, X_p)^T$ be a $p \times 1$ vector of the predictors, respectively. Suppose that we consider a random sample of n subjects and observe data $\{(X_i, V_i, \Delta_i) : i = 1, \ldots, n\}$, where $X_i = (X_{i1}, \ldots, X_{pi})^T$, $V_i = \min(T_i, C_i)$, and $\Delta_i = I(T_i \leq C_i)$, in which $I(\cdot)$ is an indicator function of an event. It is assumed that the censoring time C is independent of failure time T and covariates X. Similarly, let $\mathcal{A}$ and $\mathcal{M}$, respectively, denote the index set of active variables and inactive variables:

$$\mathcal{A} = \{r : \quad \text{some } T \text{ depends on } X_r\},$$
$$\mathcal{M} = \{r : \quad \text{any } T \text{ does not depend on } X_r\}.$$

Our goal is to select the set of active variables $X_{\mathcal{A}}$. We propose a new empirical BCov for the survival response as follows.

$$\begin{aligned}
&\mathcal{D}_n(T, X_r)\\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{S}^3(V_i)}\{\frac{1}{n}\sum_{k=1}^{n} I(V_k > V_i, X_{rk} > X_{ri}) - \frac{1}{n}\sum_{k=1}^{n} I(V_k > V_i)\frac{1}{n}\sum_{k=1}^{n} I(X_{rk} > X_{ri})\}^2\\
=&\frac{1}{n^5}\sum_{i=1}^{n}\sum_{k,l,u,v=1}^{n}\frac{\Delta_i}{\widehat{S}^3(V_i)}(\delta^V_{i,kl} + \delta^V_{i,uv} - \delta^V_{i,ku} - \delta^V_{i,lv})(\delta^{X_r}_{i,kl} + \delta^{X_r}_{i,uv} - \delta^{X_r}_{i,ku} - \delta^{X_r}_{i,lv}),
\end{aligned}$$

where $\delta^V_{i,kl} = I(V_k > V_i, V_l > V_i)$, $\delta^{X_r}_{i,kl} = I(X_{rk} > X_{ri}, X_{rl} > X_{ri})$, and $\widehat{S}(\cdot)$ is the Kaplan-Meier (KM) estimator of $S(t) = P(C \geq t)$. We show below that $\mathcal{D}_n(T, X_r)$ is a consistent estimator of $\mathcal{D}(T, X_r)$.

Proposition 1. *We have*

$$\begin{aligned}
\mathcal{D}(T, X_r) =&E\{\frac{\Delta_i}{S^3(V_i)}(\delta^V_{i,kl} + \delta^V_{i,uv} - \delta^V_{i,ku} - \delta^V_{i,lv})(\delta^{X_r}_{i,kl} + \delta^{X_r}_{i,uv} - \delta^{X_r}_{i,ku} - \delta^{X_r}_{i,lv})\}\\
=&E[\{\mathbb{P}(T > T', X_r > X'_r | T', X'_r) - \mathbb{P}(T > T' | T')\mathbb{P}(X_r > X'_r | X'_r)\}^2].
\end{aligned}$$

To derive the sure screening property of our new statistic, we further impose an additional condition on the distribution of censoring time C as follows. For simplicity, we denote $\mathcal{D}(T, X_r)$ as $\mathcal{D}_r$.

(S1) There exist a constant $c > 0$ and $0 \leq \kappa < 1/2$ such that $\min_{r \in \mathcal{A}} \rho_r \geq 2cn^{-\kappa}$.

(S2) $\mathbb{P}(C = \nu) > 0$ and $\mathbb{P}(C > \nu) = 0$ for some $\nu > 0$.

Condition (S2) has been widely used in the literature (Peng and Fine, 2009; Song et al., 2014).

Theorem 4. *Under condition* (S2)*, for any positive constants* $c_5 \leq c_6$*, when* $n > \max\{D^2(1-\delta)^{-2}[1.5^{(1/3)} - 1]^{-2}\|S\|_\infty^{-2}, 49D^2c_5^{-2}n^{2\kappa}(1-\delta)^{-2}\|S\|_\infty^{-2}, (c_5/1.01)^{\frac{1}{\kappa}}\}$*, there exist positive constants* c_1, c_2*, and* c_4 *such that*

$$\mathbb{P}(\max_{1 \leq r \leq p} |\widehat{\mathcal{D}}_r - \mathcal{D}_r| > c_6 n^{-\kappa}) \leq \gamma[2.5n \exp(-c_1 n) + 2\exp(-c_4 n^{1-2\kappa}) \\ + 2.5n \exp(-c_2 n^{1-2\kappa})],$$

where $\| \cdot \|_\infty$ *is the* L_∞ *norm, and* D *is a constant that is defined in Lemma 2 of the Appendix. If condition* (S1) *holds, then for* $\tau_n = c_7 n^{-\kappa}$ *with* $0 < c_7 \leq c$*, we have*

$$\mathbb{P}(\mathcal{A} \subset \widehat{\mathcal{A}}_n^*) \geq 1 - \gamma[2.5n \exp(-c_1 n) + 2\exp(-c_4 n^{1-2\kappa}) + 2.5n \exp(-c_2 n^{1-2\kappa})],$$

where γ *is the cardinality of* $\mathcal{A}$*.*

Theorem 4 ensures the SIS property of our method for survival data. Moreover, γ can reach an exponential rate as n goes to infinity. Similar to the conditions in the work of Song et al. (2014), our method does not require the tail probability condition for the covariates. Thus, it is suitable for survival data with a heavy-tailed distribution. However, compared with the method of Song et al. (2014), our method can deal with more complex relationships between the covariates and survival time, such as the interaction and square relationships. We will elaborate more on this point in the next section.

4. SIMULATION STUDIES

In this section, we conduct Monte Carlo simulation studies to numerically compare BCor-SIS with SIS (Fan and Lv, 2008), SIRS (Zhu et al., 2011), DC-SIS (Li et al., 2012), and MDC-SIS (Shao and Zhang, 2014). For a fair comparison, we consider different scenarios, including linear and nonlinear models, a normal distributed error structure, a heavy-tailed distributed error structure, a group predictor scenario, and a multi-response scenario, among which three examples are given in the Appendix for the sake of space. When the true model is a standard linear model with normal noise levels, the BCor-SIS approach can perform as well as SIS, SIRS, DC-SIS and MDC-SIS. In all other scenarios, BCor-SIS outperforms all competing screening methods.

Following Li et al. (2012), we generated $X = (X_1, \ldots, X_p)^T$ from a multivariate normal distribution with a zero mean vector and covariance matrix $\mathbf{\Sigma} = (\sigma_{jk})_{p\times p}$, where $\sigma_{jk} = 0.8^{|j-k|}$ for $1 \leq j, k \leq p$. We set p to be 1000 and the sample size n to be 150. We repeated each experiment 500 times. We consider the following two criteria:

- P_m, the likelihood that an individual active predictor is selected for a given size d in the 500 replications; and
- P_a, the likelihood that all active predictors are selected for a given size d in the 500 replications.

$(P_m,\ P_a)$ is used to verify the sure screening property. Ideally, the sure screening property ensures that both P_m and P_a are close to one when the model size d is sufficiently large. We set d to be $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$, and $d_3 = 3[n/\log n]$, respectively.

4.1 Simulation Results for I-BCor-SIS

Example 1: We compare BCor-SIS with I-BCor-SIS in the following four models:

$$\begin{aligned}
&(1.a) : Y = X_1 + 1.25X_2 + 0.75X_8 - 2.4X_{16} + \epsilon \\
&(1.b) : Y = 3X_1^2 + 5X_2 + 5X_8 - 8X_{16} + \epsilon, \\
&(1.c) : Y = 2I(\omega > 0)\omega + 1 \ \text{ with } \ \omega = 5X_1^2 - 5X_2^2 + 3X_8 + 2X_{16} + \epsilon, \\
&(1.d) : Y = 2I(\omega > 0)\omega + 1 \ \text{ with } \ \omega = 5X_1X_2 + 3X_8 + 3X_{16} + \epsilon,
\end{aligned}$$

where $\epsilon \sim N(0, 1)$.

Here, we choose the additive models and set p to be 2,000, sample size n to be 200, and $d^{(1)}$ and $d^{(2)}$ to be 5 in Algorithm 1. Moreover, we set the correlation parameter

σ_0 to be 0.2, 0.5, and 0.8, respectively. In model (1.a), the dependency between X_1 and Y is linear, but it is nonlinear in models (1.b)-(1.d). Tables 2 and 3 present the simulation results of I-SIS (Fan and Lv, 2008), I-DC-SIS (Zhong and Zhu, 2015), and I-BCor-SIS. In all cases, I-BCor-SIS outperforms I-SIS and I-DC-SIS for the nonlinear models.

Table 2: The proportions of P_m and P_a in example 1. The user-specified model sizes $d_1 = [n/\log n], d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$.

			P_m				P_a	P_m				P_a
		selection rate	X_1	X_2	X_8	X_{16}	All	X_1	X_2	X_8	X_{16}	All
σ_0	sizes		(1.a)					(1.b)				
0.2	d_1	I-SIS$_1$	0.11	0.11	0.29	0.18	0.01	0.06	0.10	0.97	1.00	0.01
		I-DC-SIS$_1$	0.89	0.94	0.69	0.28	0.13	0.28	0.27	0.98	1.00	0.08
		I-BCor-SIS$_1$	1.00	1.00	0.97	0.69	0.66	0.97	1.00	0.98	1.00	0.95
	d_2	I-SIS$_2$	0.13	0.12	0.29	0.21	0.01	0.09	0.10	0.97	1.00	0.02
		I-DC-SIS$_2$	0.9	0.94	0.69	0.28	0.14	0.30	0.28	0.98	1.00	0.09
		I-BCor-SIS$_2$	1.00	1.00	0.97	0.69	0.66	0.97	1.00	0.98	1.00	0.95
	d_3	I-SIS$_3$	0.11	0.11	0.29	0.18	0.01	0.11	0.14	0.97	1.00	0.03
		I-DC-SIS$_3$	0.89	0.94	0.69	0.28	0.13	0.30	0.29	0.98	1.00	0.09
		I-BCor-SIS$_3$	1.00	1.00	0.97	0.69	0.66	0.97	1.00	0.98	1.00	0.95
0.5	d_1	I-SIS$_1$	0.01	0.02	0.2	0.07	0.00	0.04	0.14	0.71	0.99	0.00
		I-DC-SIS$_1$	0.97	1.00	0.25	0.13	0.03	0.10	0.15	0.91	1.00	0.02
		I-BCor-SIS$_1$	1.00	1.00	0.96	0.60	0.59	0.90	0.91	0.97	1.00	0.79
	d_2	I-SIS$_2$	0.01	0.05	0.22	0.08	0.00	0.06	0.17	0.71	0.99	0.01
		I-DC-SIS$_2$	0.97	1.00	0.25	0.13	0.03	0.11	0.15	0.92	1.00	0.02
		I-BCor-SIS$_2$	1.00	1.00	0.96	0.60	0.59	0.9	0.91	0.97	1.00	0.79
	d_3	I-SIS$_3$	0.02	0.05	0.27	0.09	0.00	0.08	0.18	0.71	0.99	0.01
		I-DC-SIS$_3$	0.97	1.00	0.27	0.14	0.03	0.12	0.16	0.92	1.00	0.03
		I-BCor-SIS$_3$	1.00	1.00	0.96	0.60	0.59	0.9	0.91	0.97	1.00	0.79
0.8	d_1	I-SIS$_1$	0.02	0.04	0.06	0.04	0.00	0.02	0.02	0.17	0.50	0.00
		I-DC-SIS$_1$	0.98	0.96	0.08	0.05	0.00	0.26	0.24	0.39	0.77	0.01
		I-BCor-SIS$_1$	1.00	1.00	0.94	0.53	0.50	0.8	0.76	0.85	1.00	0.60
	d_2	I-SIS$_2$	0.04	0.04	0.06	0.06	0.00	0.05	0.05	0.19	0.52	0.00
		I-DC-SIS$_2$	0.98	0.96	0.14	0.08	0.00	0.27	0.25	0.40	0.77	0.02
		I-BCor-SIS$_2$	1.00	1.00	0.94	0.58	0.54	0.81	0.79	0.86	1.00	0.62
	d_3	I-SIS$_3$	0.05	0.06	0.08	0.07	0.00	0.05	0.10	0.21	0.52	0.00
		I-DC-SIS$_3$	0.99	0.97	0.15	0.10	0.00	0.3	0.25	0.40	0.77	0.02
		I-BCor-SIS$_3$	1.00	1.00	0.94	0.58	0.54	0.81	0.8	0.87	1.00	0.62

Table 3: The proportions of P_m and P_a in example 1. The user-specified model sizes $d_1 = [n/\log n], d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$.

			P_m				P_a	P_m				P_a
		selection rate	X_1	X_2	X_8	X_{16}	All	X_1	X_2	X_8	X_{16}	All
σ_0	sizes		(1.c)					(1.d)				
0.2	d_1	I-SIS$_1$	0.09	0.13	0.96	0.98	0.01	0.15	0.09	0.90	0.50	0.00
		I-DC-SIS$_1$	0.15	0.22	1.00	1.00	0.05	1.00	0.95	0.98	0.62	0.57
		I-BCor-SIS$_1$	0.61	0.66	1.00	0.98	0.41	1.00	1.00	1.00	1.00	1.00
	d_2	I-SIS$_2$	0.10	0.15	0.96	0.98	0.02	0.17	0.11	0.9	0.51	0.00
		I-DC-SIS$_2$	0.16	0.25	1.00	1.00	0.06	1.00	0.95	0.99	0.63	0.58
		I-BCor-SIS$_2$	0.62	0.67	1.00	0.98	0.43	1.00	1.00	1.00	1.00	1.00
	d_3	I-SIS$_3$	0.12	0.17	0.96	0.99	0.03	0.2	0.14	0.9	0.52	0.00
		I-DC-SIS$_3$	0.19	0.28	1.00	1.00	0.06	1.00	0.95	0.99	0.64	0.59
		I-BCor-SIS$_3$	0.64	0.68	1.00	0.98	0.45	1.00	1.00	1.00	1.00	1.00
0.5	d_1	I-SIS$_1$	0.09	0.07	0.71	0.74	0.00	0.04	0.04	0.54	0.28	0.00
		I-DC-SIS$_1$	0.15	0.17	0.96	0.96	0.01	0.95	0.72	0.79	0.39	0.19
		I-BCor-SIS$_1$	0.53	0.64	1.00	0.98	0.34	1.00	1.00	1.00	1.00	1.00
	d_2	I-SIS$_2$	0.12	0.12	0.74	0.74	0.00	0.07	0.04	0.57	0.29	0.00
		I-DC-SIS$_2$	0.2	0.22	0.96	0.96	0.02	0.95	0.74	0.79	0.41	0.21
		I-BCor-SIS$_2$	0.54	0.64	1.00	0.98	0.35	1.00	1.00	1.00	1.00	1.00
	d_3	I-SIS$_3$	0.15	0.13	0.74	0.75	0.00	0.09	0.06	0.59	0.31	0.00
		I-DC-SIS$_3$	0.2	0.24	0.96	0.96	0.02	0.95	0.74	0.79	0.43	0.21
		I-BCor-SIS$_3$	0.58	0.66	1.00	0.98	0.39	1.00	1.00	1.00	1.00	1.00
0.8	d_1	I-SIS$_1$	0.05	0.06	0.22	0.18	0.00	0.04	0.01	0.16	0.03	0.00
		I-DC-SIS$_1$	0.17	0.09	0.63	0.47	0.00	0.9	0.57	0.29	0.10	0.01
		I-BCor-SIS$_1$	0.49	0.46	0.96	0.94	0.31	0.99	1.00	1.00	0.91	0.91
	d_2	I-SIS$_2$	0.07	0.07	0.24	0.21	0.00	0.07	0.02	0.18	0.05	0.00
		I-DC-SIS$_2$	0.2	0.12	0.65	0.48	0.02	0.91	0.57	0.30	0.10	0.01
		I-BCor-SIS$_2$	0.52	0.47	0.96	0.94	0.33	0.99	1.00	1.00	0.91	0.91
	d_3	I-SIS$_3$	0.09	0.07	0.24	0.21	0.00	0.1	0.05	0.18	0.07	0.00
		I-DC-SIS$_3$	0.22	0.12	0.65	0.51	0.02	0.91	0.58	0.32	0.10	0.01
		I-BCor-SIS$_3$	0.52	0.48	0.96	0.95	0.33	0.99	1.00	1.00	0.92	0.92

4.2 Simulation Results for Interaction Selection

We compare our screening methods, BCor-SIS and I-BCor-SIS, with four screening methods, including SIS (Fan and Lv, 2008), SIRS (Zhu et al., 2011), DC-SIS (Li et al., 2012), and P-IT (Fan et al., 2017). We generate $X = (X_1, \ldots, X_p)^T$ from a multivariate normal distribution with zero mean vector and covariance matrix $\Sigma = (\sigma_{jk})$, where $\sigma_{jk} = 0.5^{|k-j|}$ for $1 \leq k, j \leq p$. Here we set p to be 2000, the sample size n to be 200, and $d = 2[n/\log n]$. We repeat each experiment 500 times.

Example 2: We consider the following four models

$$\begin{aligned}
(2.a) : Y =& 3X_1X_5 + 2X_{10} + 2X_{15} + \epsilon, \\
(2.b) : Y =& 3X_1X_5 + 3X_{10}X_{15} + \epsilon, \\
(2.c) : Y =& 3X_1X_5X_{10} + 3X_{15} + \epsilon, \\
(2.d) : Y =& 3X_1X_5X_{10}X_{15} + \epsilon,
\end{aligned}$$

where $\epsilon \sim N(0, 1)$.

Table 4 reveals that the interaction variables screening procedure based on Ball correlation outperforms other methods in all models. Specifically, the results shows that the new screening procedure based on Ball correlation is suitable to the linear interaction model especially for high order interaction variables.

4.3 Simulation Results for Censored Survival Data

In this subsection, we compare our method with four screening methods, including correlation screening (CS), log-rank statistics screening (LRSS, Gorst-Rasmussen and Scheike (2013)), partial likelihood ratio screening (PLRS), and censored rank independent screening (CRIS, Song et al. (2014)). For correlation screening, we used an inverse probability of censoring weighted method to compute the Pearson correlation between the survival time and predictors, which is the generation of Fan and Lv (2008) for survival data. For partial likelihood ratio screening, a marginal Cox model was fitted for each predictor and the partial likelihood ratio statistic was constructed by

Table 4: The proportion of P_m and P_a in example 2. The user-specified model sizes $d_2 = 2[n/\log n]$.

	P_m				P_a	P_m				P_a
selection rate	X_1	X_5	X_{10}	X_{16}	All	X_1	X_5	X_{10}	X_{16}	All
			(2.a)					(2.b)		
SIS	0.098	0.086	1.00	1.00	0.008	0.150	0.132	0.130	0.118	0.002
SIRS	0.058	0.030	1.00	1.00	0.004	0.064	0.058	0.058	0.072	0.001
DC-SIS	0.114	0.096	1.00	1.00	0.016	0.438	0.430	0.362	0.392	0.050
MDC-SIS	0.102	0.104	1.00	1.00	0.014	0.126	0.100	0.102	0.100	0.000
P-IT	0.812	0.778	1.00	1.00	0.646	0.868	0.876	0.870	0.854	0.570
BCor-SIS	0.968	0.974	1.00	1.00	0.946	0.974	0.970	0.994	0.998	0.938
			(2.c)					(2.d)		
SIS	0.118	0.106	0.190	1.00	0.002	0.112	0.156	0.132	0.128	0.002
SIRS	0.050	0.036	0.080	1.00	0.001	0.058	0.090	0.080	0.090	0.001
DC-SIS	0.140	0.126	0.200	1.00	0.004	0.616	0.656	0.612	0.612	0.128
MDC-SIS	0.106	0.076	0.166	1.00	0.000	0.174	0.206	0.204	0.172	0.000
P-IT	0.722	0.750	0.754	1.00	0.396	0.576	0.590	0.584	0.560	0.088
BCor-SIS	0.896	0.898	0.948	1.00	0.766	1.000	1.000	0.996	0.996	0.992

the corresponding model, which is asymptotically equivalent to the method proposed by Zhao and Li (2012).

Example 3: We generate T_i from the following three transformation models:

(3.a) $H(T) = -3X_1 - 2X_2 + 0.8X_9 + X_{10} + \varepsilon,$

(3.b) $H(T) = -3X_1X_2 + 0.8X_9 + X_{10} + \varepsilon,$

(3.c) $H(T) = -3X_1^2 - 2I(X_2 < 0) + 0.8X_9 + X_{10} + \varepsilon,$

where $H(T) = \log\{2(e^{4t} - 1)\}$ and $\varepsilon \sim N(0, 1)$. We set $n = 300$ and $p = 2000$. The covariates, X_i, are generated from a multivariate normal distribution with a mean of zero and a first-order autoregressive structure $\Sigma = (\sigma_{jk})$ with $\sigma_{jk} = 0.5^{|j-k|}$ for $j, k = 1, \ldots, p$. The censoring time is generated from a uniform distribution on $[0, c]$, where c is chosen to achieve censoring proportions of 15% and 40%.

Table 5 presents all the simulation results. For model (3.a), BCor-SIS performs slightly worse than CRIS, but much better than CS, LRSS and PLRS. In contrast, for models (3.b) and (3.c), BCor-SIS significantly outperforms all four competing methods under the nonlinear model settings.

Table 5: The proportion of P_m and P_a in example 3. The user-specified model sizes $d_2 = 2[n/\log n]$.

	(c =15%)					(c =40%)				
	P_m				P_a	P_m				P_a
selection rate	X_1	X_2	X_9	X_{10}	All	X_1	X_2	X_9	X_{10}	All
					(3.a)					
CS	0.827	0.800	0.447	0.487	0.273	0.693	0.547	0.127	0.233	0.033
LRSS	0.807	0.760	0.413	0.473	0.093	0.813	0.787	0.440	0.467	0.100
PLRS	0.033	0.027	0.080	0.047	0.000	0.053	0.060	0.053	0.020	0.000
CRIS	0.980	0.980	0.973	0.987	0.940	0.993	0.993	0.960	0.973	0.933
BCor-SIS	0.993	0.993	0.953	0.980	0.920	0.987	0.993	0.913	0.960	0.873
					(3.b)					
CS	0.180	0.220	0.347	0.353	0.033	0.107	0.153	0.367	0.433	0.000
LRSS	0.107	0.153	0.573	0.553	0.007	0.093	0.107	0.613	0.573	0.000
PLRS	0.060	0.060	0.073	0.047	0.000	0.047	0.033	0.073	0.047	0.000
CRIS	0.133	0.133	1.000	0.993	0.040	0.113	0.100	1.000	1.000	0.020
BCor-SIS	1.000	0.980	1.000	1.000	0.980	0.967	0.947	1.000	1.000	0.913
					(3.c)					
CS	0.200	0.140	0.347	0.313	0.007	0.073	0.200	0.367	0.407	0.000
LRSS	0.173	0.307	0.500	0.520	0.013	0.100	0.347	0.573	0.573	0.013
PLRS	0.047	0.033	0.040	0.073	0.000	0.027	0.020	0.067	0.047	0.000
CRIS	0.313	0.947	1.000	1.000	0.313	0.280	0.927	1.000	1.000	0.273
BCor-SIS	1.000	0.987	1.000	1.000	0.987	1.000	0.813	0.933	0.980	0.760

5. ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

To motivate the proposed methodology, we consider a large database with imaging, genetic, and clinical data collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (adni.loni.usc.edu). We consider the joint analysis of hippocampus surface and genetic data collected through ADNI-1. To reduce the population stratification effect, we included 708 Caucasians (421 men and 287 women) from healthy controls and individuals with Alzheimer's disease (AD) and mild cognitive impairment (MCI) (163 AD, 347 MCI, and 198 healthy controls). The scans were performed on a variety of 1.5 T MRI scanners with protocols individualized for each scanner and include standard T1-weighted images obtained using volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. We applied a hippocampal subregional analysis package based on surface fluid registration to all segmented hippocampus surfaces extracted from structural MRI images. Then, we calculated the radial distance map of the hippocampal surface, which retains information on the deformation along the surface normal direction. More details can be found in Wang et al. (2011).

We considered the 708 subjects' genotype variables acquired by using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) in the ADNI-1 database, which includes 620,901 single nucleotide polymorphisms (SNPs). By following Wang et al. (2012), we focus on SNPs belonging to the top 40 AD candidate genes. After the quality control procedures, 1072 SNPs remained in the final data analysis.

The objective of this data analysis was to examine the genetic effect of each of 1,072 candidate SNPs on either the left or right hippocampus. To achieve this objective, we applied the DC-SIS, BCor-SIS and I-BCor-SIS ($d^{(1)} = d^{(2)} = 3$) procedures to screen the candidate genes and compared the results obtained from these three approaches. Moreover, we chose the 15000 radial distances of either the left or right hippocampus as a functional phenotype after regressing out age and gender and then computed the standard L_2 norm.

Tables 6 and 7 list the top 10 potential genes selected by the above three approaches. The data analysis results have confirmed the important role of well-known genes such as APOE-ϵ4 and SORCS1, where APOE-ϵ4 is the most influential gene of the left or right hippocampus. A flood of literature (Lescai et al., 2011; Hao et al., 2016) suggested that APOE-ϵ4 is the top genetic risk factor of Alzheimer's disease, thus it can potentially affect the shape of the hippocampus. In contract to DC-SIS, BCor-SIS and I-BCor-SIS are inclined to screen more types of useful genes. Take left hippocampus for example, I-BCor-SIS ranked the gene LOC651924 as the top 10 genes, which was ignored by BCor-SIS and DC-SIS. Belbin et al. (2011) provided support for LOC651924 as risk modifiers of late-onset Alzheimer's disease by meta-analyses of all published follow-up case-control association studies. Besides, I-BCor-SIS procedure also detected gene CH25H, which was taken as an Alzheimer's disease risk factor in Shibata et al. (2006).

Table 6: ADNI data analysis results: top 10 selected SNPs based on left hippocampus data

DC-SIS		BCor-SIS		I-BCor-SIS	
SNP	gene	SNP	gene	SNP	gene
rs429358	APOE-ϵ4	rs429358	APOE-ϵ4	rs429358	APOE-ϵ4
rs2418828	SORCS1	rs2152676	SORCS1	rs2152676	SORCS1
rs2152676	SORCS1	rs3798729	NEDD9	rs3798729	NEDD9
rs10786978	SORCS1	rs16871247	NEDD9	rs1785469	CLU
rs6584766	SORCS1	rs4287912	TF	rs2784945	LOC651924
rs11218301	SORL1	rs2418828	SORCS1	rs10887927	CH25H
rs2486154	SORCS1	rs12989701	BIN1	rs2784940	LOC651924
rs12625444	PRNP	rs1262099	SORCS1	rs7091546	SORCS1
rs2276346	SORL1	rs8177184	TF	rs2745251	ECE1
rs6139494	PRNP	rs17195022	SORCS1	rs11141914	DAPK1

6. DISCUSSION

We propose a novel nonparametric feature screening procedure based on the Ball correlation. Without finite sub-exponential moments, we proved its strong property of sure screening when the dimensionality is an exponential order of the sample size. We used Monte Carlo simulations to demonstrate its screening accuracy compared to that of several popular methods in some important scenarios. Compared with the existing methods, our proposed method is a generic procedure that is model-free and has fewer and less restrictive assumptions of the data.

Some issues deserve further study. The computational complexity of BCor-SIS is $O(pn^2 \log n)$ for multivariate responses and group predictors, but can be reduced to be $O(pn^2)$ for univariate responses. The threshold used in the proposed method is

Table 7: ADNI data analysis: top 10 selected SNPs based on right hippocampus data

DC-SIS		BCor-SIS		I-BCor-SIS	
SNP	gene	SNP	gene	SNP	gene
rs429358	APOE-ϵ4	rs429358	APOE-ϵ4	rs429358	APOE-ϵ4
rs4311	ACE	rs7216307	GRN	rs7216307	GRN
rs12415086	SORCS1	rs913778	DAPK1	rs913778	DAPK1
rs4344419	SORCS1	rs10512188	DAPK1	rs4353	ACE
rs376382	IL33	rs6701713	CR1	rs12415086	SORCS1
rs386880	IL33	rs4818921	CLU	rs4344419	SORCS1
rs7908795	SORCS1	rs7919814	SORCS1	rs11601726	GAB2
rs2900784	SORCS1	rs10868558	DAPK1	rs7071961	SORCS1
rs2250938	SORCS1	rs4311	ACE	rs9380116	NEDD9
rs1056719	DAPK1	rs1385741	CD2AP	rs1930056	DAPK1

adopted from those of Fan and Lv (2008) and Zhu et al. (2011). It is also of interest to develop a new criterion to determine the threshold for finite samples; however, we leave this topic for future research.

REFERENCES

Anonymity (2017), "Ball Covariance: a generic measure of dependence in Banach Space," *Manuscript*, .

Belbin, O., Carrasquillo, M. M., Crump, M., Culley, O. J., Hunter, T. A., Ma, L., Bisceglio, G., Zou, F., Allen, M., Dickson, D. W. et al. (2011), "Investigation of 15 of the top candidate genes for late-onset Alzheimer's disease," *Human Genetics*, 129(3), 273–282.

Fan, J., Feng, Y., and Song, R. (2009), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557.

Fan, J., Feng, Y., and Tong, X. (2010), "A ROAD to Classification in High Dimensional Space.," *Journal of The Royal Statistical Society Series B*, 74, 745–770.

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*, 70, 849–911.

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

Fan, Y., Kong, Y., Li, D., and Lv, J. (2017), Interaction pursuit with feature screening and selection,, Technical report, University of South California.

Gorst-Rasmussen, A., and Scheike, T. (2013), “Independent screening for single-index hazard rate models with ultrahigh dimensional features,” *Journal of the Royal Statistical Society: Series B*, 75, 217–245.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Scholkopf, B., and Smola, A. J. (2008), “A Kernel Statistical Test of Independence,” , pp. 585–592.

Hao, X., Yan, J., Yao, X., Risacher, S. L., Saykin, A. J., Zhang, D., and Shen, L. (2016), “Diagnosis-guided method for identifying multi-modality neuroimaging biomarkers associated with genetic risk factors in Alzheimer’s disease,” *Pac Symp Biocomput.*, 21, 108–119.

Hastie, T., Tibshirani, R., and Friedman, J. (2004), *The Elements of Statistical Learning.* Springer.

Huang, D., Li, R., and Wang, H. (2013), “Feature screening for ultrahigh dimensional categorical data with applications,” *Journal of Business & Economic Statistics*, 32, 237–244.

Kong, Y., Li, D., Fan, Y., and Lv, J. (2017), "Interaction pursuit in high-dimensional multi-response regression via distance correlation.," *The Annals of Statistics*, 45, 897–922.

Lescai, F., Chiamenti, A. M., Codemo, A., Pirazzini, C., D'Agostino, G., Ruaro, C., Ghidoni, R., Benussi, L., Galimberti, D., Esposito, F. et al. (2011), "An APOE haplotype associated with decreased $\varepsilon 4$ expression increases the risk of late onset Alzheimer's disease," *Journal of Alzheimer's Disease*, 24, 235–245.

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

Lyons, R. (2013), "Distance covariance in metric spaces," *The Annals of Probability*, 41, 3284–3305.

Peng, L., and Fine, J. P. (2009), "Competing risks quantile regression," *Journal of the American Statistical Association*, 104, 1440–1453.

Sejdinovic, D., Sriperumbudur, B. K., Gretton, A., and Fukumizu, K. (2013), "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *Annals of Statistics*, 41(5), 2263–2291.

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.

Shibata, N., Kawarai, T., Lee, J. H., Lee, H.-S., Shibata, E., Sato, C., Liang, Y., Duara, R., Mayeux, R. P., St George-Hyslop, P. H. et al. (2006), "Association studies of cholesterol metabolism genes (CH25H, ABCA1 and CH24H) in Alzheimer's disease," *Neuroscience Letters*, 391(3), 142–146.

Song, R., Lu, W., Ma, S., and Jeng, X. J. (2014), "Censored Rank Independence Screening for High-dimensional Survival Data.," *Biometrika*, 101, 799–814.

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

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., and Shen, L. (2012), “Identifying quantitative trait loci via group-sparse multitask regression and feature selection,” *Bioinformatics*, 28(2), 229–237.

Wang, X., Leng, C., and Dunson, D. B. (2015), “On the consistency theory of high dimensional variable screening,” *Neural Information Processing Systems*,

Wang, Y., Song, Y., Rajagopalan, P., An, T., Liu, K., Chou, Y., Gutman, B. A., Toga, A. W., and Thompson, P. M. (2011), “Surface-based TBM boosts power to detect disease effects on the brain: an N=804 ADNI study,” *NeuroImage*, 56, 1993–2010.

Zhao, S. D., and Li, Y. (2012), “Principled sure independence screening for Cox models with ultra-high-dimensional covariates,” *Journal of Multivariate Analysis*, 105, 397–411.

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., Li, L., Li, R., and Zhu, L. (2011), “Model-Free Feature Screening for Ultrahigh-Dimensional Data,” *Journal of the American Statistical Association*, 106, 1464–1475.