MODEL-FREE FORWARD SCREENING VIA CUMULATIVE DIVERGENCE

Tingyou Zhou[a], Liping Zhu[b], Chen Xu[c] and Runze Li[d]

[a] School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, P. R. China. [b] Institute of Statistics and Big Data and Center for Applied Statistics, Renmin University of China, Beijing, P. R. China. [c] Department of Mathematics and Statistics University of Ottawa, Ottawa, Canada. [d] Department of Statistics and The Methodology Center, The Pennsylvania State University at University Park, U.S.A.

Abstract

Feature screening plays an important role in the analysis of ultrahigh dimensional data. Due to complicated model structure and high noise level, existing screening methods often suffer from model misspecification and the presence of outliers. To address these issues, we introduce a new metric named cumulative divergence (CD), and develop a CD-based forward screening procedure. This forward screening method is model-free and resistant to the presence of outliers in the response. It also incorporates the joint effects among covariates into the screening process. With a data-driven threshold, the new method can automatically determine the number of features that should be retained after screening. These merits make the CD-based screening very appealing in practice. Under certain regularity conditions, we show that the proposed method possesses sure screening property. The performance of our proposal is illustrated through simulations and a real data example.

KEY WORDS: Cumulative divergence; feature screening; forward screening; high dimensionality; sure screening property; variable selection.

*Liping Zhu is the corresponding author. Email: zhu.liping@ruc.edu.cn. Institute of Statistics and Big Data and Center for Applied Statistics, Renmin University of China, 59 Zhongguancun Avenue, Haidian District, Beijing 100872, P. R. China.

1. INTRODUCTION

Regression analysis with ultrahigh dimensional covariates arises in many scientific fields such as agriculture, biomedicine, economics, finance, and genetics. It is desirable to identify the important covariates that are truly influential to the response. Traditional best subset selection methods are computationally infeasible in the presence of ultrahigh dimensional covariates. In the past two decades, many regularization methods, such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006), and Dantzig selector (Candes and Tao, 2007), have been proposed for variable selection. However, when the covariates are ultrahigh dimensional, Fan et al. (2009) stated that these regularization methods suffer from the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability.

To deal with ultrahigh dimensionality, Fan and Lv (2008) suggested screening out most unimportant covariates before implementing an elaborative variable selection. They proposed a sure independent screening procedure (SIS) for linear models using marginal Pearson correlation between each covariate and the response. Since the seminal work of Fan and Lv (2008), feature screening has received extensive attention in the past decade. In particular,Wang (2009) proposed a forward regression and Chang et al. (2013) suggested a marginal likelihood ratio test to screen out unimportant covariates in linear models. Li et al., (2012) suggested replacing Pearson correlation with Kendall's rank correlation in the presence of outliers. Ma, Li and Tsai (2017) proposed quantile partial correlation for feature screening in linear quantile regression. Fan and Song (2010) and Xu and Chen (2014) suggested maximum likelihood estimate and Mai and Zou (2013) proposed Kolmogorov-Smirnov statistic to screen out unimportant features in generalized linear models. Fan et al. (2011) and He, Wang and Hong (2013) suggested nonparametric screening procedures for additive models. Song et al. (2014)

proposed an independent screening procedure for varying coefficient models. These model-based screening procedures are effective if the working model is close to the underlying true model, and may be very ineffective otherwise.

To minimize the impact of model misspecification, several model-free screening methods have been developed. For instance, Zhu et al., (2011) proposed a sure independent ranking and screening procedure for a general class of index models. Li, Zhong and Zhu (2012) suggested distance correlation for feature screening, which can simultaneously deal with grouped covariates and multivariate response. Shao and Zhang (2014) introduced martingale difference correlation to perform screening as long as the mean function of the response is concerned. These model-free methods are favored when we are lack of prior information on the regression structure. However, most of them are based on marginal correlations and are vulnerable in the presence of outliers.

In the present work, we develop a model-free forward screening procedure for ultra-high dimensional data. Forward screening is related to but much more challenging than conditional screening. For conditional screening, the conditioning set is fixed. However, for our proposed forward screening procedure, the conditioning set is iteratively updated in a data-driven fashion. Moreover, existing conditional screening procedures are model-based and there is little literature on model-free conditional screening (Wang, 2009; Xu and Chen, 2014; Barut, Fan and Verhasselt, 2016). To the best of our knowledge, how to design a model-free forward screening has not been studied yet. We aim to fill in this gap in this paper. To this end, we first introduce the concept of cumulative divergence (CD), a new correlation metric to characterize functional dependence. We show that the CD is robust to the presence of outliers in the conditioning variable. We further propose a CD-based forward screening procedure. At each step of the forward screening, a new covariate will be added to an active index set based on its conditional

CD with the response. This procedure stops when the conditional CD of all remaining covariates is less than a certain threshold. Compared with marginal screening methods, the forward screening incorporates the joint correlation among the covariates. With a data-driven threshold, our proposal can adaptively determine the number of features that should be retained after screening. Therefore, it is convenient for implementation without ad hoc tuning steps. Due to its robust property, our proposal performs well even when the underlying true model is misspecified. It is also robust in the presence of outliers. This appealing property makes the CD-based forward screening attractive for handling ultrahigh dimensional noisy data. Under some regularity conditions, we show that our forward screening method possesses the sure screening property in the terminology of Fan and Lv (2008). We further demonstrate the finite sample performance of the proposed procedures through simulations and a real data example.

We summarize the major contributions of this paper as follows. (1) The proposed forward screening approach is distinguished from marginal screening approaches in that the joint correlations among the covariates are taken into account by the proposed forward screening procedure and yet are ignored by the marginal screening methods (Zhu et al.,, 2011; Li, Zhong and Zhu, 2012). (2) The proposed forward screening procedure is model-free, and hence robust to model misspecification. Thus, the proposed procedure is different from existing model-based forward regression and conditional screening methods (Wang, 2009; Xu and Chen, 2014; Barut, Fan and Verhasselt, 2016). This model-free property is very appealing in ultrahigh dimensional data analysis, especially when we are often lack of information on the underlying regression structure. (3) We propose the CD to quantify deviation from mean independence. The CD is robust to the presence of outliers in the conditioning variable, is thus different from the martingale difference correlation (Shao and Zhang, 2014). Our proposed CD-based forward screening approach inherits this robustness property and is robust to the presence of

outliers in the response.

This paper is organized as follows. In Section 2, we introduce the notion of cumulative divergence and study its properties. In Section 3, we propose a model-free forward screening procedure and establish its sure screening property. In Section 4, we assess the finite sample performance of our proposed forward screening procedure through comprehensive numerical studies. Some concluding remarks are given in Section 5. All technical details are relegated to the Appendix and a supplementary document.

2. THE CUMULATIVE DIVERGENCE

In each step of the forward screening procedure to be developed, we have to determine whether a covariate should be selected through testing whether the conditional mean function of the response variable is independent of this covariate. This motivates us to start with a simplified problem by testing *mean independence* that

$$H_0 : E(Y \mid X) = E(Y) \text{ almost surely} \quad \text{versus} \quad H_1 : \text{otherwise}. \tag{2.1}$$

Let $(\widetilde{X}, \widetilde{Y})$ be an independent copy of (X, Y). We assume $\text{var}(X) > 0$ and $0 < \text{var}(Y) < \infty$ throughout. We do not require $\text{var}(X) < \infty$. Let " $\Leftrightarrow$ " stand for "the statements on both the left and the right hand sides are equivalent", and $\text{supp}(X)$ stand for the support of the conditioning variable X. We first note that

$$\begin{aligned} & E(Y \mid X) = E(Y) \ \text{ almost surely} \\ \Leftrightarrow \ & E(Y \mid X < x_0) = E(Y), \ \ \text{for all } x_0 \in \text{supp}(X) \\ \Leftrightarrow \ & \text{cov}\{Y, \mathbf{1}(X < x_0)\} = 0, \ \ \text{for all } x_0 \in \text{supp}(X) \\ \Leftrightarrow \ & E\left[\text{cov}^2\{Y, \mathbf{1}(X < \widetilde{X}) \mid \widetilde{X}\}\right] = 0. \end{aligned} \tag{2.2}$$

This motivates us to define the *cumulative covariance* (CCov) and the CD as follows.

DEFINITION 2.1. *Assume* $\text{var}(X) > 0$ *and* $0 < \text{var}(Y) < \infty$. *The cumulative covariance, denoted* $\text{CCov}(Y \mid X)$, *and the cumulative divergence, denoted* $\text{CD}(Y \mid X)$, *between random variables* X *and* Y *are defined, respectively, by*

$$\text{CCov}(Y \mid X) \stackrel{\text{def}}{=} E\left[\text{cov}^2\{Y, \mathbf{1}(X < \widetilde{X}) \mid \widetilde{X}\}\right] \text{ and} \tag{2.3}$$

$$\text{CD}(Y \mid X) \stackrel{\text{def}}{=} \text{CCov}(Y \mid X)/\text{var}(Y). \tag{2.4}$$

The definition of CD allows for $\text{var}(X) = \infty$, indicating that the distribution of X can be heavy-tailed. Since the rank of X is used in the definition of $\text{CCov}(Y \mid X)$, this also indicates that $\text{CD}(Y \mid X)$ is robust to outliers in the conditioning variable X. The following theorem states that the CD possesses several other appealing properties.

Theorem 1. *The CD has the following properties.*

1. *Assume* $\text{var}(X) > 0$ *and* $0 < \text{var}(Y) < \infty$, *then* $0 \leq \text{CD}(Y \mid X) \leq 1/4$ *and* $\text{CD}(Y \mid X) = 0$ *if and only if* $E(Y \mid X) = E(Y)$ *almost surely. In addition,* $\text{CD}(X \mid Y) = \text{CD}(Y \mid X) = 0$ *if* $F(y \mid X) = F(y)$ *for all* $y \in \mathbb{R}$, *where* $F(y \mid X) \stackrel{def}{=} pr(Y < y \mid X)$ *and* $F(y) = pr(Y < y)$, *for* $y \in \mathbb{R}$.

2. *For* $a, b \in \mathbb{R}$ *with* $a \neq 0$, *and an arbitrary strictly monotone transformation M(X),* $\text{CD}(Y \mid X) = \text{CD}\{aY + b \mid M(X)\}$.

3. *If* X *and* Y *are jointly normal with Pearson correlation* ρ, *then* $\text{CD}(Y \mid X) = \text{CD}(X \mid Y) = \rho^2/\left(2\sqrt{3}\pi\right)$. *In particular,* $\text{CD}(X \mid X) = 1/\left(2\sqrt{3}\pi\right)$.

4. *Let* $\widetilde{X}$ *be an independent copy of* X. *If* Y *is normal and all involved moments exist,* $\text{CD}(Y \mid X)/\text{var}(Y) = E\left[E^2\left\{\partial F(\widetilde{X} \mid Y)/\partial Y \mid \widetilde{X}\right\}\right]$.

The first assertion of Theorem 1 indicates that $\mathrm{CD}(Y \mid X)$ is a useful measure to detect whether the conditional mean function of Y depends on X functionally. In particular, $\mathrm{CD}(Y \mid X) = 0$ if and only if $E(Y \mid X) = E(Y)$. This ensures that the CD is a useful tool to test (2.1). In general, $\mathrm{CD}(Y \mid X) \neq \mathrm{CD}(X \mid Y)$ even if $\mathrm{var}(X) = \mathrm{var}(Y)$. If X and Y are independent, then $\mathrm{CD}(Y \mid X) = \mathrm{CD}(X \mid Y) = 0$; and if X and Y are jointly normal, $\mathrm{CD}(Y \mid X) = \mathrm{CD}(X \mid Y)$.

The second assertion of Theorem 1 indicates that the CD is invariant with respect to strictly monotone transformation of X. This invariant property matches the fact that $E(Y \mid X) = E\{Y \mid M(X)\}$ and is however not shared by other popular correlation measures, such as Pearson correlation, martingale difference (Shao and Zhang, 2014), or distance correlation (Székely, Rizzo and Bakirov, 2007; Székely and Rizzo, 2009). This property implies that the CD is robust against model misspecification and the presence of outliers, because it merely uses the rank rather than the observed values of X. The virtue of robustness makes the associated forward screening procedure to be developed in Section 3 potentially attractive for ultrahigh dimensional noisy data.

The third assertion of Theorem 1 implies that, when X and Y are jointly normal with Pearson correlation ρ and unit variance, our proposed CD is closely related to other popular correlation measures through ρ. In particular, Kendall's rank correlation (Huber and Ronchetti, 2009) equals to $2\arcsin(\rho)/\pi$, the squared martingale difference correlation equals to $\rho^2\{4(1-\sqrt{3}+\pi/3)\}^{-1/2}$, and the squared distance correlation is $\big\{\rho\arcsin(\rho) + (1-\rho^2)^{1/2} - \rho\arcsin(\rho/2) - (4-\rho^2)^{1/2} + 1\big\}/(1+\pi/3-\sqrt{3})$.

A sample version of the CD can be conveniently constructed. Specifically, let $\{(X_i, Y_i), i = 1, \ldots, n\}$ be a random sample from the joint distribution of (X, Y).

We estimate $\mathrm{CCov}(Y \mid X)$ and $\mathrm{CD}(Y \mid X)$ respectively by

$$\begin{aligned} \widehat{\mathrm{CCov}}(Y \mid X) &\stackrel{\text{def}}{=} n^{-3} \sum_{j=1}^{n} \left[\sum_{i=1}^{n} (Y_i - \overline{Y}) \left\{ \mathbf{1}(X_i < X_j) - F_n(X_j) \right\} \right]^2 \text{ and} \\ \widehat{\mathrm{CD}}(Y \mid X) &\stackrel{\text{def}}{=} \widehat{\mathrm{CCov}}(Y \mid X) / \widehat{\mathrm{var}}(Y), \end{aligned} \tag{2.5}$$

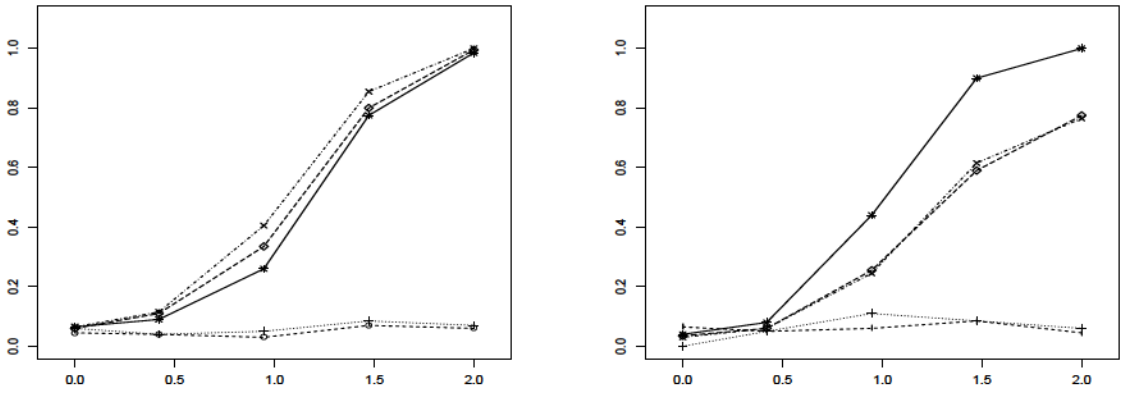
where

$$\overline{Y} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} Y_i, \quad F_n(X_j) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \mathbf{1}(X_i < X_j) \quad \text{and} \quad \widehat{\mathrm{var}}(Y) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

To decide critical values in the test for the hypothesis (2.1), we propose a wild bootstrap procedure as follows. Define $\varepsilon_i = Y_i - \overline{Y}$ and $Y_i^* = \overline{Y} + a_i \varepsilon_i$, where a_i satisfies $\mathrm{pr}(a_i = 1) = \mathrm{pr}(a_i = -1) = 1/2$. The wild bootstrap sample is $\{(X_i, Y_i^*), i = 1 \ldots, n\}$. We repeat the wild bootstrap procedure m times to obtain $\widehat{\mathrm{CD}}^{(1)}(Y^* \mid X), \ldots, \widehat{\mathrm{CD}}^{(m)}(Y^* \mid X)$. Denote τ the $(1-\alpha)$-th quantile of $\{\widehat{\mathrm{CD}}^{(1)}(Y^* \mid X), \ldots, \widehat{\mathrm{CD}}^{(m)}(Y^* \mid X)\}$. We reject H_0 at the significance level α if $\widehat{\mathrm{CD}}(Y \mid X)$ calculated from the original sample $\{(X_i, Y_i), i = 1, \ldots, n\}$ is greater than τ and accept H_0 otherwise.

We conduct a simulation study to compare the finite-sample performance of the CD with that of four commonly-used correlation: Pearson correlation, rank correlation, distance correlation and martingale difference correlation. We consider two scenarios for generating the conditioning variable X. In the first scenario X is standard normal and in the second scenario X follows Cauchy distribution. Let $Y = c \exp(-X^2) + \varepsilon$, where $\varepsilon \sim N(0, 1)$. We set $c = 0.0$, 0.5, 1.0, 1.5 and 2.0. The null hypothesis H_0 in (2.1) holds true when $c = 0$. We set the sample size $n = 100$ and summarize the simulation results in Figure 1 when the significance level $\alpha = 0.05$.

It can be clearly seen from Figure 1 that the sizes of all tests are close to the significance level $\alpha = 0.05$. Both the Pearson correlation and the rank correlation test



(A) (B)

Figure 1: The power curves of the Pearson correlation test (dashed line marked with circles), the Kendall's rank correlation test (dotted line marked with plus signs), the martingale difference correlation test (dotdash line marked with cross signs), the distance correlation test (longdash line marked with diamond signs) and the cumulative divergence test (solid line marked with star signs), respectively. In Figure 1 (A), both X and ε are standard normal. In Figure 1 (B), X follows Cauchy distribution and ε is standard normal.

fail to detect the non-monotone mean dependence. The CD test is much more powerful than both the martingale difference correlation test and the distance correlation test when X follows Cauchy distribution. This simulated example empirically confirms that the robustness property of the CD test.

3. A FORWARD SCREENING PROCEDURE

In this section we propose a model-free forward screening procedure based on the CD. This new forward screening procedure inherits the appealing properties of the CD.

To ease subsequent presentation, we introduce the following notations. Let Y be the response and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be the p-dimensional covariate vector. Let $\mathcal{F}$ be a working index set and $\mathcal{F}^c$ be its complement. Both $\mathcal{F}$ and $\mathcal{F}^c$ are subsets of $\{1, 2, \ldots, p\}$. We define $\mathbf{x}_{\mathcal{F}} \stackrel{\text{def}}{=} \{X_k, k \in \mathcal{F}\}$ the covariate vector indexed by $\mathcal{F}$ and $\mathbf{\Sigma}_{\mathcal{F}} \stackrel{\text{def}}{=} \mathrm{var}(\mathbf{x}_{\mathcal{F}})$. Let $|\mathcal{F}|$ stand for the cardinality of $\mathcal{F}$. We assume throughout that

$E(\mathbf{x}) = \mathbf{0}$ for simplicity.

The goal of feature selection is to identify the smallest index set $\mathcal{A}$ such that

$$Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}, \tag{3.1}$$

where $\perp\!\!\!\perp$ stands for statistical independence. Model (3.1) implies immediately that $F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}_{\mathcal{A}})$, for $y \in \mathbb{R}$. Therefore, identifying $\mathbf{x}_{\mathcal{A}}$ which satisfies model (3.1) is equivalent to seeking for the smallest index set

$$\mathcal{A} \stackrel{\text{def}}{=} \{k : F(y \mid \mathbf{x}) \text{ depends functionally on } X_k \text{ for } y \in \mathbb{R}, \quad k = 1, \ldots, p\}.$$

Model (3.1) covers a wide variety of existing models. Interested readers can refer to Section 2.1 of Zhu et al., (2011) for details.

We first note that model (3.1) ensures that $Y \perp\!\!\!\perp X_k \mid \mathbf{x}_{\mathcal{F}}$ for all $k \in \mathcal{F}^c$ and all $\mathcal{A} \subseteq \mathcal{F}$. Therefore, given a working index set $\mathcal{F}$, assessing whether X_k, $k \in \mathcal{F}^c$, is truly important for the response variable Y amounts to testing the hypothesis that

$$H_0 : Y \perp\!\!\!\perp X_k \mid \mathbf{x}_{\mathcal{F}} \quad \text{versus} \quad H_1 : \text{otherwise}. \tag{3.2}$$

The law of iterated expectations implies immediately that $E\{X_k - E(X_k \mid \mathbf{x}_{\mathcal{F}}) \mid Y\} = E\{E(X_k \mid \mathbf{x}_{\mathcal{F}}, Y) - E(X_k \mid \mathbf{x}_{\mathcal{F}}) \mid Y\}$. Under H_0 in (3.2), $E(X_k \mid \mathbf{x}_{\mathcal{F}}, Y) = E(X_k \mid \mathbf{x}_{\mathcal{F}})$, and hence $E\{X_k - E(X_k \mid \mathbf{x}_{\mathcal{F}}) \mid Y\} = 0$. Under H_1 in (3.2), X_k is dependent upon Y even when $\mathbf{x}_{\mathcal{F}}$ is given. Thus it is reasonable to expect that $E(X_k \mid \mathbf{x}_{\mathcal{F}}, Y) \neq E(X_k \mid \mathbf{x}_{\mathcal{F}})$ and accordingly $E\{X_k - E(X_k \mid \mathbf{x}_{\mathcal{F}}) \mid Y\} \neq 0$. These, together with Theorem 1, motivate us to use $\omega_{k|\mathcal{F}} \stackrel{\text{def}}{=} \text{CD}\{X_k - E(X_k \mid \mathbf{x}_{\mathcal{F}}) \mid Y\}$ to test (3.2).

To ensure that $\omega_{k|\mathcal{F}}$ has nontrivial power in test for (3.2), we further assume that

A1. $E\{\partial F(y \mid \mathbf{x})/\partial X_k\} \neq 0$ for some $y \in \mathbb{R}$, for all $k \in \mathcal{A}$.

It is remarkable that (3.1) ensures that $E\{\partial F(y \mid \mathbf{x})/\partial X_k\} = 0$ for all $y \in \mathbb{R}$, and all $k \in \mathcal{A}^c$. This fact, together with Assumption A1, ensures that the important and the unimportant covariates are separable, which is stated in Theorem 2.

Theorem 2. *Under H_0 in (3.2), we have $\omega_{k|\mathcal{F}} = 0$. If we further assume that $\mathbf{x}$ is normal and Assumption A1 holds, then* $\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing} \max_{k\in\mathcal{F}^c\cap\mathcal{A}} \omega_{k|\mathcal{F}} > 0$.

Theorem 2 guarantees that, if all the truly important covariates have been selected into $\mathcal{F}$ already, then for any $k \in \mathcal{F}^c$, we have $\omega_{k|\mathcal{F}} = 0$. However, if there are a few important covariates that have not been found yet, that is, $\mathcal{F}^c \cap \mathcal{A} \neq \varnothing$, then there must exist $k \in \mathcal{F}^c \cap \mathcal{A}$ such that $\max_{k\in\mathcal{F}^c\cap\mathcal{A}} \omega_{k|\mathcal{F}} > 0$. This motivates us to reject H_0 in (3.2) when the sample version of $\max_{k\in\mathcal{F}^c\cap\mathcal{A}} \omega_{k|\mathcal{F}}$ is sufficiently large.

How to estimate $\omega_{k|\mathcal{F}}$ is a nontrivial task because it involves estimating $E(X_k \mid \mathbf{x}_\mathcal{F})$. A fully nonparametric estimate of $E(X_k \mid \mathbf{x}_\mathcal{F})$ is apparently undesirable, especially when $\mathbf{x}_\mathcal{F}$ is high dimensional. In the present context, we assume that

A2. $E(X_k \mid \mathbf{x}_\mathcal{F}) = g_{k|\mathcal{F}}(\mathbf{x}_\mathcal{F}, \boldsymbol{\beta}_{k|\mathcal{F}})$, where $g_{k|\mathcal{F}}$ is known and $\boldsymbol{\beta}_{k|\mathcal{F}}$ is unknown.

We allow $E(X_k \mid \mathbf{x}_\mathcal{F})$ to be a general parametric function. When $\mathbf{x}$ follows elliptically contoured distribution, $E(X_k \mid \mathbf{x}_\mathcal{F})$ is indeed a linear function of $\mathbf{x}_\mathcal{F}$, for all k and $\mathcal{F} \subseteq \{1, \ldots, p\}$. Examples of elliptically contoured distribution include multivariate normal distribution, multivariate t-distribution, symmetric multivariate Laplace distribution and multivariate logistic distribution, etc.

Let $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ be a random sample from $(\mathbf{x}, Y)$, where each covariate, for notational clarity, is assumed to be marginally standardized to have zero mean and

unit variance in advance. To carry out the CD test for (3.2), we estimate $\omega_{k|\mathcal{F}}$ by

$$\widehat{\omega}_{k|\mathcal{F}} = n^{-2}\sum_{j=1}^{n}\left[\sum_{i=1}^{n}\mathbf{1}(Y_i < Y_j)\{X_{ik} - g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}}, \widehat{\boldsymbol{\beta}}_{k|\mathcal{F}})\}\right]^2 \Bigg/ \sum_{i=1}^{n}\left\{X_{ik} - g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}}, \widehat{\boldsymbol{\beta}}_{k|\mathcal{F}})\right\}^2,$$

where $\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}}$ is obtained through the nonlinear least squares. That is,

$$\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}} \stackrel{\text{def}}{=} \underset{\boldsymbol{\beta}_{k|\mathcal{F}}}{\operatorname{argmin}} \sum_{i=1}^{n}\{X_{ik} - g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})\}^2 \tag{3.3}$$

We reject H_0 in (3.2) when $\widehat{\omega}_{k|\mathcal{F}}$ is sufficiently large. Deciding the critical value for the CD test amounts to studying the asymptotic distribution of $\widehat{\omega}_{k|\mathcal{F}}$. Let $g'_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$, $g''_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ and$g'''_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ be the first, the second and the third derivatives of $g_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ with respect to $\boldsymbol{\beta}_{k|\mathcal{F}}$, respectively. We denote $g'_{l_1,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ the l_1-th component of $g'_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$, $g''_{l_1l_2,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ the (l_1, l_2)-th component of $g''_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ and $g'''_{l_1l_2l_3,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ the (l_1, l_2, l_3)-th component of $g'''_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$. Let $\delta_{k|\mathcal{F}} \stackrel{\text{def}}{=} X_k - E(X_k \mid \mathbf{x}_{\mathcal{F}})$ and C be a generic constant. We assume the following conditions.

(B1) There exists $\vartheta > 0$ such that $p = o\{\exp(an^{\vartheta})\}$ for any $a > 0$.

(B2) For any working index set $\mathcal{F} \subseteq \{1, 2, \ldots, p\}$ and $k \in \mathcal{F}^c$, $E(X_k^4) \le C, E(\delta_{k|\mathcal{F}}^8) \le C$, $E\{|g'_{l_1,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})|^8\} \le C$; $|g_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})| \le G_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})$ with $E[\{G_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})\}^4] \le C$; $|g'_{l_1,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})| \le G_{l_1,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})$ with $E[\{G_{l_1,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})\}^4] \le C$; $|g''_{l_1l_2,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})| \le G_{l_1l_2,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})$ with $E[\{G_{l_1l_2,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})\}^4] \le C$; $|g'''_{l_1l_2l_3,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})| \le G_{l_1l_2l_3,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})$ with $E[\{G_{l_1l_2l_3,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}})\}^4] \le C$, for all l_1, l_2, l_3 and $\boldsymbol{\beta}_{k|\mathcal{F}}$.

(B3) There exists c_0 such that $\|\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\|_\infty < c_0$ for all p, where $\|A\|_\infty \stackrel{\text{def}}{=} \max_l \sum_m |a_{lm}|$ stands for the infinity norm of the matrix $A = (a_{lm})$.

Condition (B1) allows p to diverge exponentially faster than n. Condition (B2) is widely used to study the asymptotic behavior of nonlinear least squares estimation. See, e.g.,

Jennrich (1969) and White (1981). This condition can be simplified dramatically when $g_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})$ is linear. Theorem 3 requires condition (B3) holds true for $|\mathcal{F}| = o(n^{1/5})$. Many precision matrices satisfy Condition (B3). In particular, if we denote $\boldsymbol{\Sigma}_{\mathcal{F}}^{-1} = (\sigma_{-1,lm})_{|\mathcal{F}|\times|\mathcal{F}|}$ and let $\sigma_{-1,lm}$ equal 1 if $l = m$ and r_n otherwise, then this condition is satisfied as long as $|r_n| \leq (c_0 - 1)/(|\mathcal{F}| - 1)$. If $\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}$ is a banded or block-diagonal matrix and each row has d nonzero entries, for example, $\sigma_{-1,lm}$ equals 1 if $l = m$, r if $1 \leq |l-m| < d$ and 0 if $|l-m| \geq d$, condition (B3) simply requires $|r| \leq (c_0-1)/(d-1)$. Condition (B3) can also be satisfied by many other sparse precision matrices. If $\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}$ is a power-decay matrix, say, $\sigma_{-1,lm} = \rho_n^{|l-m|}$, for $|\rho_n| < 1$, condition (B3) is satisfied as long as $(1-\rho_n^{|\mathcal{F}|}) \leq c_0(1-\rho_n)$. Condition (B3) is also implied by $\|\boldsymbol{\Sigma}^{-1}\|_\infty < c_0$. Similar conditions are also assumed in the literature. See, for example, Mai et al. (2012, page 34-35) and Bickel and Levina (2008, page 2580).

Theorem 3. *In addition to Conditions (B1)-(B3), we further assume* $|\mathcal{F}| = o(n^{1/5})$.

1. *Under* H_0 *in (3.2), we have,* $\omega_{k|\mathcal{F}} = 0$ *and* $pr\big(n\widehat{\omega}_{k|\mathcal{F}} < q\ \kappa_{k|\mathcal{F}}\big) - pr\left(Q_{k|\mathcal{F}} < q\right) \to 0$, *for any* $q \in \mathbb{R}^+$, *where* $Q_{k|\mathcal{F}} \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \lambda_{j,k|\mathcal{F}} \chi_j^2(1)$, $\kappa_{k|\mathcal{F}}$ *is defined in (B.1),* $\chi_j^2(1)s$ *are independent* $\chi^2(1)$ *random variables,* $\lambda_{j,k|\mathcal{F}}s$ *are nonnegative constants that depend on the joint distribution of* $(X_k, \mathbf{x}_{\mathcal{F}}, Y)$ *and* $E\left(Q_{k|\mathcal{F}}\right) = 1$.

2. *Under* H_1 *in (3.2) and if* $\omega_{k|\mathcal{F}} > 0$ *for* $k \in \mathcal{F}^c$, *we have* $pr\left\{n^{1/2}(\widehat{\omega}_{k|\mathcal{F}} - \omega_{k|\mathcal{F}}) < t\right\} - pr\left(T_{k|\mathcal{F}} < t\right) \to 0$, *for any* $t \in \mathbb{R}$, *where* $T_{k|\mathcal{F}}$ *is a normal random variable with mean zero and variance* $\Delta_{k|\mathcal{F}}$ *and* $\Delta_{k|\mathcal{F}}$ *is defined in (B.3).*

The condition $|\mathcal{F}| = o(n^{1/5})$ seems somewhat stringent. By refining Assumption A2, such as $E(X_k \mid \mathbf{x}_{\mathcal{F}}) = g_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}\boldsymbol{\beta}_{k|\mathcal{F}})$, this condition can be weakened to $|\mathcal{F}| = o(n^{1/3})$. This condition is in line with that of Huber (1973), Fan and Peng (2004) and Tan and Zhu (2018). We impose this condition because $\boldsymbol{\beta}_{k|\mathcal{F}}$ is unknown and has to be

estimated from data. We do not impose sparsity assumption on $\boldsymbol{\beta}_{k|\mathcal{F}}$ but we do require the convergence rate of $\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}}$ be fast enough to ensure the weak convergence of $\widehat{\omega}_{k|\mathcal{F}}$. The requirement on $\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}}$ can be met under the condition that $|\mathcal{F}| = o(n^{1/5})$.

Theorem 3 shows that $\widehat{\omega}_{k|\mathcal{F}}$ is root-n consistent under H_0 and n-consistent under H_1, indicating that the CD test has nontrivial power in test for (3.2). We adopt the wild bootstrap procedure introduced in Section 2 to determine critical values.

Next we adapt our proposed CD test for (3.2) with a working index set $\mathcal{F}$ to a forward screening procedure for ultrahigh dimensional feature selection in model (3.1). The rationale of our proposed forward screening procedure is as follows. If H_0 in (3.2) is rejected, we update $\mathcal{F}$ with $\mathcal{F} \cup \{k\}$, because X_k is possibly influential for Y. With the updated $\mathcal{F}$, we further consider testing (3.2) until H_0 is accepted for all $k \in \mathcal{F}^c$. It is reasonable to expect $\mathcal{A} \subseteq \mathcal{F}$ when the forward screening procedure stops. To provide theoretical justification for our proposal, we assume the following condition.

A3. There exist a positive constant C and $\varpi \in [0, 1/2)$ such that

$$\min_{\mathcal{F}:\mathcal{F}^c \cap \mathcal{A} \neq \emptyset} \max_{k \in \mathcal{F}^c \cap \mathcal{A}} \omega_{k|\mathcal{F}} > C n^{-\varpi}. \tag{3.4}$$

Assumption A3 requires that the signal strength of the truly important covariates, conditional on the covariates $\mathbf{x}_{\mathcal{F}}$ that have already been selected, is strong enough to be detectable. It is also justified in Theorem 2. This assumption is different from the marginal signal assumptions used in the screening literature in that the marginal signal strength is quantified through setting the working index set $\mathcal{F}$ to be a null set. See, for example, condition 3 in Fan and Lv (2008), condition E in Fan and Song (2010), condition C in Fan et al. (2011), condition (C1) in Zhu et al., (2011), and condition (C2) in Li, Zhong and Zhu (2012). It is generally required that the marginal signal strength

of all truly important covariates must be greater than a certain threshold in the existing screening literature. By contrast, Assumption A3 quantifies the signal strength of the truly important covariates conditional on the selected covariates $\mathbf{x}_{\mathcal{F}}$, which ensures that $\omega_{k|\mathcal{F}}$ plays a similar role as the regression coefficients in linear models. Similar assumptions are also made in the literature. See, for example, condition (C3) in Wang (2009, page 1513) and condition 1 in Barut, Fan and Verhasselt (2016, page 1270). These assumptions are generally regarded as mild and reasonable.

To establish the sure screening property for the proposed screening procedure, we further assume the following conditions.

(B4) The cardinality of $\mathcal{A}$ satisfies $|\mathcal{A}| = O(n^{1/5-\gamma})$ for $\gamma \in (0, 1/5]$.

(B5) Let M and υ be two generic positive constants. Assume that $E|X_k|^m \leq m!M^{m-2}\upsilon/2$ for all $m \geq 2$, $k = 1, 2, \ldots, p$. Assume in addition that $E|g_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{0,k|\mathcal{F}})|^m \leq m!M^{m-2}\upsilon/2$ and $E|g'_{l_1,k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{0,k|\mathcal{F}})|^m \leq m!M^{m-2}\upsilon/2$ for all $m \geq 2$, $\mathcal{F} \subseteq \{1, 2, \ldots, p\}$ and $k \in \mathcal{F}^c$.

Assumption (B4) is also a technical condition and is closely related to the assumption that $|\mathcal{F}| = o(n^{1/5})$ used in Theorem 3. Condition (B5) is milder than the sub-Gaussian assumption (Buldygin and Kozachenko, 1980, Lemma 1).

Theorem 4. *Suppose that Conditions (B1)-(B5) and Assumption A3 are satisfied. If we further assume* $|\mathcal{F}| = o(n^{1/5})$, $3/5 - 2\varpi - \vartheta > 0$ *and set* $\nu < Cn^{-\varpi}/2$ *in the forward screening procedure, then* $pr\left(\min_{\mathcal{F}:\mathcal{F}^c\cap\mathcal{A}\neq\varnothing} \max_{k\in\mathcal{F}^c\cap\mathcal{A}} \widehat{\omega}_{k|\mathcal{F}} > \nu\right) \to 1$ *as* $n \to \infty$.

Theorem 4 ensures that the proposed procedure can retain all important covariates with an overwhelming probability if ν is chosen properly. Such a desirable property is referred to as the sure screening property. The CD is a robust correlation metric, and our forward screening procedure is also robust to model misspecification. Such

merits are particularly appealing for analyzing ultrahigh dimensional data in the absence of prior knowledge of model structure and data quality. Unlike existing model-free marginal screening methods, the proposed method is a stepwise procedure, which incorporates joint correlation among ultrahigh dimensional features in the forward screening process. It thus provides more reliable results in practice. With a data-driven choice of ν, the procedure adaptively determines the number of features to be retained after selection. This makes the implementation of our proposed forward screening method practically convenient, since our proposal does not require additional ad hoc tuning steps.

We describe the algorithm for our proposed forward screening procedure as follows.

Step 1 Start with an initial index set $\mathcal{F} = \varnothing$.

Step 2 For all $k \in \mathcal{F}^c$, calculate $\widehat{\omega}_{k|\mathcal{F}}$. Denote $k^* = \arg\max_{k \in \mathcal{F}^c} \widehat{\omega}_{k|\mathcal{F}}$. If $\widehat{\omega}_{k^*|\mathcal{F}} > \nu$, update $\mathcal{F}$ with $\mathcal{F} \cup \{k^*\}$. The data-driven ν will be determined as follows.

(a) Generate $\widetilde{X}_{ik^*} = g_{k^*|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}}, \widehat{\boldsymbol{\beta}}_{k^*|\mathcal{F}}) + a_i \widehat{\delta}_{i,k^*|\mathcal{F}}$, $i = 1, 2, \ldots, n$, where $\widehat{\delta}_{i,k^*|\mathcal{F}} = X_{ik^*} - g_{k^*|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}}, \widehat{\boldsymbol{\beta}}_{k^*|\mathcal{F}})$, and a_i are independent and identically distributed random weights satisfying $\mathrm{pr}(a_i = 1) = \mathrm{pr}(a_i = -1) = 1/2$. We calculate $\widehat{\widetilde{\omega}}_{k^*|\mathcal{F}} \stackrel{\text{def}}{=} \widehat{\mathrm{CD}}\big\{\widetilde{X}_{k^*} - E\big(\widetilde{X}_{k^*} \mid \mathbf{x}_{\mathcal{F}}\big) \mid Y\big\}$ using $\{(\widetilde{X}_{ik^*}, \mathbf{x}_{i\mathcal{F}}, Y_i),\ i = 1, 2, \ldots, n\}$.

(b) Repeat the above wild boostrap procedure for B times to obtain $\widehat{\widetilde{\omega}}^{(1)}_{k^*|\mathcal{F}}$, $\widehat{\widetilde{\omega}}^{(2)}_{k^*|\mathcal{F}}$, $\ldots$, $\widehat{\widetilde{\omega}}^{(B)}_{k^*|\mathcal{F}}$. Set ν to be the $(1-\alpha)$-th upper quantile of $\big\{\widehat{\widetilde{\omega}}^{(1)}_{k^*|\mathcal{F}}, \widehat{\widetilde{\omega}}^{(2)}_{k^*|\mathcal{F}}, \ldots, \widehat{\widetilde{\omega}}^{(B)}_{k^*|\mathcal{F}}\big\}$. We update the working index set $\mathcal{F}$ with $\mathcal{F} \cup \{k^*\}$ if $\widehat{\omega}_{k^*|\mathcal{F}} > \nu$.

Step 3 Repeat **Step 2** until no covariate can be added into the working index set $\mathcal{F}$.

Assumption A2 requires that the minimal signal strength be greater than $Cn^{-\varpi}$, and Theorem 4 requires the cutoff ν to be smaller than one half of the minimal signal

strength. These requirements ensure that our proposal possesses the desirable sure screening property. In practice, however, the magnitude of minimal signal strength is generally unknown. Consequently, how to choose an optimal cutoff ν is not straightforward. To put our proposed procedure into practice, at each step and for each covariate, we choose $\alpha = 0.01$ and set the cutoff to be the 99-th percentile of asymptotic null distribution of $\widehat{\omega}_{k|\mathcal{F}}$ in our algorithm. This works satisfactorily in our numerical studies.

4. NUMERICAL STUDIES

4.1. Simulations

In this section, we conduct Monte Carlo simulations to assess the finite sample performance of the CD-based forward screening procedure. For convenience of presentation, we refer to our proposed forward screening method as C-FS. We compare C-FS with the following five competitors: the forward regression designed for linear model by Wang (2009, FR), the least absolute shrinkage and selection operator proposed by Tibshirani (1996, LASSO), the sure independent ranking and screening procedure proposed by Zhu et al., (2011, SIRS), the distance correlation based sure independence screening procedure proposed by Li, Zhong and Zhu (2012, DC-SIS), and the Pearson correlation based sure independence screening procedure proposed by Fan and Lv (2008, SIS) .

To determine the number of features to be retained after screening, we use a BIC-type criterion for FR, as suggested by Wang (2009). The model size (tuning parameter) of the LASSO was chosen by 10-fold cross validation. For SIRS, DC-SIS and SIS, we follow the convention by retaining $[n/\log(n)]$ top ranked covariates into the screened model. It should be noted that our C-FS algorithm automatically determines the screening size with a wild bootstrap procedure.

We adopt the following criteria to evaluate the performance of above methods.

1. $\mathcal{P}_{\text{ind}}$: With a given size, $\mathcal{P}_{\text{ind}}$ is the empirical probability that an influential covariate is retained after screening.

2. $\mathcal{P}_{\text{all}}$: With a given size, $\mathcal{P}_{\text{all}}$ is the empirical probability that all the influential covariates are retained after screening.

3. FPR: Let $\widehat{\mathcal{A}}$ be the index set of the retained covariates and $\mathcal{A}$ be the index set of truly influential covariates. The false positive rate (FPR) is defined as $|\widehat{\mathcal{A}}\backslash\mathcal{A}|/|\mathcal{A}^c|$, where $\widehat{\mathcal{A}}\backslash\mathcal{A}$ is the index set of irrelevant covariates that are retained after screening and $|\mathcal{M}|$ denotes the cardinality of the set $\mathcal{M}$.

4. TPR: The true positive rate (TPR) is defined as $|\widehat{\mathcal{A}}\cap\mathcal{A}|/|\mathcal{A}|$, where $\widehat{\mathcal{A}}\cap\mathcal{A}$ denotes the set of influential covariates that are correctly retained after screening

We report both the mean and the standard errors of the FPR and TPR values based on 500 repetitions. We set the sample size $n = 200$, the covariate dimension $p = 3,000$ and the bootstrap times $B = 1,000$.

Example 1. We generate data from a linear model $Y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + c_0\varepsilon$, where $\boldsymbol{\beta} = (5, 5, 5, -15\rho^{1/2}, 0, \ldots, 0)^{\mathrm{T}}$, $c_0 = 1$ if $\varepsilon \sim N(0, 1)$ and $c_0 = 0.1$ if $\varepsilon \sim t(1)$. We consider the following two scenarios to generate the covariate $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$.

(1) The elliptical case: The covariate $\mathbf{x}$ is drawn from multivariate normal population with mean zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$, where $\sigma_{ii} = 1$, $i = 1, \ldots, p$, $\sigma_{i4} = \sigma_{4i} = \rho^{1/2}$ for $i \neq 4$, and $\sigma_{ij} = \rho$, for $i \neq j$, $i \neq 4$ and $j \neq 4$.

(2) The non-elliptical case: Set $\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\{\widehat{\mathrm{var}}(\mathbf{z})\}^{-1/2}\{\mathbf{z} - \widehat{E}(\mathbf{z})\}$, where $\boldsymbol{\Sigma}$ is defined in the first scenario, $\mathbf{z} \overset{\text{def}}{=} (Z_1, \ldots, Z_p)^{\mathrm{T}}$, Z_ks are independent of each other and follow $\chi^2(2)$ distribution.

In the above two scenarios, we set ρ to be $0.1, 0.5$ and 0.9, respectively, to stand for small, moderate and high correlation. This example was also used by Fan and Lv (2008) and Zhu et al., (2011). The simulation results are summarized in Tables 1-2.

Table 1: The mean and the standard errors of both the FPR and the TPR values based on 500 repetitions for Example 1.

ε	method	$\rho = 0.1$				$\rho = 0.5$				$\rho = 0.9$			
		FPR		TPR		FPR		TPR		FPR		TPR	
		mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
When $\mathbf{x}$ follows elliptical distribution													
$\mathcal{N}(0,1)$	C-FS	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
	FR	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
	LASSO	0.01	0.00	1.00	0.00	0.06	0.00	0.75	0.00	0.04	0.00	0.75	0.00
	SIRS	0.00	0.00	0.75	0.00	0.01	0.00	0.75	0.01	0.01	0.00	0.59	0.29
	DC-SIS	0.00	0.00	0.75	0.00	0.01	0.00	0.75	0.02	0.01	0.00	0.58	0.29
	SIS	0.00	0.00	0.75	0.00	0.00	0.00	0.75	0.00	0.01	0.00	0.60	0.28
$t(1)$	C-FS	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
	FR	0.00	0.00	0.90	0.29	0.00	0.00	0.90	0.29	0.00	0.00	0.77	0.42
	LASSO	0.00	0.00	0.84	0.35	0.03	0.02	0.63	0.27	0.02	0.02	0.52	0.34
	SIRS	0.01	0.00	0.75	0.01	0.01	0.00	0.75	0.02	0.01	0.00	0.62	0.27
	DC-SIS	0.01	0.00	0.75	0.04	0.01	0.00	0.74	0.06	0.01	0.00	0.59	0.28
	SIS	0.01	0.00	0.72	0.14	0.01	0.00	0.71	0.16	0.01	0.00	0.51	0.33
When $\mathbf{x}$ follows non-elliptical distribution													
$\mathcal{N}(0,1)$	C-FS	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
	FR	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
	LASSO	0.01	0.01	1.00	0.00	0.06	0.00	0.75	0.00	0.04	0.00	0.75	0.00
	SIRS	0.01	0.00	0.75	0.02	0.01	0.00	0.74	0.05	0.01	0.00	0.59	0.28
	DC-SIS	0.01	0.00	0.75	0.02	0.01	0.00	0.74	0.05	0.01	0.00	0.59	0.27
	SIS	0.01	0.00	0.75	0.02	0.01	0.00	0.74	0.04	0.01	0.00	0.62	0.25
$t(1)$	C-FS	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
	FR	0.00	0.00	0.94	0.21	0.00	0.00	0.93	0.21	0.00	0.00	0.85	0.30
	LASSO	0.00	0.00	0.86	0.32	0.03	0.02	0.62	0.28	0.02	0.02	0.49	0.35
	SIRS	0.00	0.00	0.75	0.00	0.01	0.00	0.74	0.05	0.01	0.00	0.57	0.29
	DC-SIS	0.01	0.00	0.75	0.04	0.01	0.00	0.74	0.06	0.01	0.00	0.56	0.30
	SIS	0.01	0.00	0.71	0.15	0.01	0.00	0.69	0.18	0.01	0.00	0.50	0.32

In this example, X_4 is marginally independent of Y. It is thus not surprising to observe from Table 2 that SIRS, DC-SIS and SIS fail to retain X_4, as they consider only the marginal effects. The performance of LASSO is decent for $\rho = 0.1$ and

Table 2: The empirical probabilities $\mathcal{P}_{\text{ind}}$ and $\mathcal{P}_{\text{all}}$ based on 500 repetitions for Example 1.

ε	method	$\rho = 0.1$					$\rho = 0.5$					$\rho = 0.9$				
		$\mathcal{P}_{\text{ind}}$				$\mathcal{P}_{\text{all}}$	$\mathcal{P}_{\text{ind}}$				$\mathcal{P}_{\text{all}}$	$\mathcal{P}_{\text{ind}}$				$\mathcal{P}_{\text{all}}$
		X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
When $\mathbf{x}$ follows elliptical distribution																
$\mathcal{N}(0,1)$	C-FS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	FR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LASSO	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
	SIRS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.79	0.79	0.79	0.00	0.00
	DC-SIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.76	0.77	0.78	0.00	0.00
	SIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.81	0.80	0.81	0.00	0.00
$t(1)$	C-FS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	FR	0.90	0.90	0.90	0.89	0.89	0.90	0.90	0.91	0.90	0.89	0.77	0.77	0.77	0.77	0.76
	LASSO	0.85	0.85	0.86	0.79	0.79	0.84	0.83	0.83	0.00	0.00	0.70	0.70	0.70	0.00	0.00
	SIRS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.83	0.82	0.83	0.00	0.00
	DC-SIS	0.99	1.00	0.99	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.80	0.78	0.78	0.00	0.00
	SIS	0.95	0.96	0.96	0.00	0.00	0.95	0.95	0.94	0.00	0.00	0.69	0.68	0.68	0.00	0.00
When $\mathbf{x}$ follows non-elliptical distribution																
$\mathcal{N}(0,1)$	C-FS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	FR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LASSO	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
	SIRS	1.00	1.00	1.00	0.01	0.01	0.99	0.99	0.99	0.00	0.00	0.78	0.80	0.80	0.00	0.00
	DC-SIS	1.00	1.00	1.00	0.01	0.01	0.99	0.99	0.99	0.00	0.00	0.77	0.80	0.80	0.00	0.00
	SIS	1.00	1.00	1.00	0.01	0.01	0.99	1.00	0.99	0.00	0.00	0.83	0.83	0.84	0.00	0.00
$t(1)$	C-FS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	FR	0.94	0.94	0.95	0.93	0.91	0.93	0.93	0.93	0.95	0.90	0.83	0.82	0.83	0.91	0.78
	LASSO	0.88	0.87	0.88	0.80	0.80	0.83	0.83	0.82	0.00	0.00	0.65	0.66	0.65	0.00	0.00
	SIRS	1.00	1.00	1.00	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.76	0.76	0.76	0.00	0.00
	DC-SIS	1.00	1.00	1.00	0.00	0.00	0.99	0.99	0.99	0.00	0.00	0.75	0.74	0.74	0.00	0.00
	SIS	0.94	0.94	0.96	0.00	0.00	0.92	0.93	0.92	0.00	0.00	0.65	0.66	0.69	0.00	0.00

$\varepsilon \sim \mathcal{N}(0,1)$, but it deteriorates sharply as ρ increases. Both FR and C-FS perform well when ε is normal. However, when $\varepsilon \sim t(1)$ and $\mathbf{x}$ is elliptical, FR has an average TPR as low as 0.77 for $\rho = 0.9$. In this scenario, FR is also quite unstable in terms of the large standard deviations. By contrast, the proposed C-FS attains stable and satisfactory performance in all scenarios. The simulation results when $\mathbf{x}$ follows non-elliptical distribution are quite similar to those when $\mathbf{x}$ follows elliptical distribution.

Example 2. We consider three models where Y depends on $\mathbf{x}_{\mathcal{A}}$ nonlinearly.

(a) $Y = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2\exp(X_{20} + c_0\varepsilon)$.

(b) $Y = X_1 + 0.8X_2 + 0.6(X_5 + 1)^2 + 0.4X_{10}^3 + 0.2\exp(|X_{20} + 1| + c_0\varepsilon)$.

(c) $Y = \beta_1(X_1)X_2 + \beta_2(X_1)X_3 + \beta_3(X_1)X_4 + \beta_4(X_1)X_5 + c_0\varepsilon$.

In all three models, ε and c_0 are generated in the same way as in Example 1. In Example 2(a) and 2(b), we consider two scenarios for generating $\mathbf{x}$.

(1) The elliptical case: The covariate $\mathbf{x}$ is drawn from multivariate normal population with mean zero and covariance matrix $\mathbf{\Sigma} = (0.5^{|i-j|})_{p\times p}$.

(2) The non-elliptical case: Set $\mathbf{x} = \mathbf{\Sigma}^{1/2}\{\widehat{\mathrm{var}}(\mathbf{z})\}^{-1/2}\{\mathbf{z} - \widehat{E}(\mathbf{z})\}$, where $\mathbf{\Sigma} = (0.5^{|i-j|})_{p\times p}$, $\mathbf{z} \stackrel{\text{def}}{=} (Z_1, \ldots, Z_p)^{\mathrm{T}}$, Z_ks are independent and follow $\chi^2(2)$ distribution.

In Example 2(c), we generate U_1 and U_2 independently from uniform distribution on $[0,1]$, and set $X_1 = (U_1 + U_2)/2$, $\beta_1(X_1) = 4(1 - X_1^2)$, $\beta_2(X_1) = 3\{1 + \sin(2\pi X_1)\}$, $\beta_3(X_1) = 2\{1 + (1 - X_1)^3/2\}$, and $\beta_4(X_1) = \exp(|X_1|)$. Define $X_k = (Z_k + 3U_1)/4$, $k = 2, 3, \ldots, p$, where Z_ks are independently drawn from (1) the standard normal distribution in the elliptical case and (2) the $\chi^2(2)$ distribution in the non-elliptical case.

The simulation results for Example 2 are charted in Tables 3-4. In Example 2(a), none of SIRS, DC-SIS or SIS is able to identify X_{20} as an important covariate. This

is because these methods are relatively sensitive to the transformation of variables. When $\varepsilon \sim t(1)$, the new C-FS is the only method that has satisfactory performance. This confirms the robustness behavior of C-FS. In Example 2(b), those model based methods, such as FR, LASSO and SIS, fail to retain all the important covariates, because the linear model assumption is violated. In comparison, the model free methods (C-FS, SIRS, DC-SIS) perform relatively better. The performance of C-FS is the best, due to its robustness property against the outliers in the response. In Example 2(c), the marginal screening methods, such as SIRS, DC-SIS and SIS, fail to detect X_1 in all scenarios. Both C-FS and LASSO outperform FR when ε is normal, our C-FS is the only method that remains satisfactory performance when $\varepsilon \sim t(1)$.

Table 3: The mean and the standard errors of both the FPR and the TPR values based on 500 repetitions for Example 2.

	method	$\mathbf{x}$ follows elliptical distribution				$\mathbf{x}$ follows non-elliptical distribution			
		$\varepsilon \sim \mathcal{N}(0,1)$		$\varepsilon \sim t(1)$		$\varepsilon \sim \mathcal{N}(0,1)$		$\varepsilon \sim t(1)$	
		FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR
		mean std	mean std	mean std	mean std	mean std	mean std	mean std	mean std
(a)	C-FS	0.01 0.00	0.98 0.06	0.01 0.00	0.98 0.07	0.01 0.00	0.97 0.07	0.01 0.00	0.98 0.07
	FR	0.00 0.00	0.88 0.18	0.00 0.00	0.25 0.40	0.00 0.00	0.63 0.28	0.00 0.00	0.21 0.34
	LASSO	0.00 0.00	0.83 0.28	0.00 0.00	0.22 0.38	0.00 0.00	0.45 0.44	0.00 0.00	0.14 0.32
	SIRS	0.01 0.00	0.90 0.10	0.01 0.00	0.86 0.09	0.01 0.00	0.98 0.07	0.01 0.00	0.97 0.08
	DC-SIS	0.01 0.00	0.91 0.10	0.01 0.00	0.44 0.42	0.01 0.00	0.98 0.08	0.01 0.00	0.48 0.46
	SIS	0.01 0.00	0.94 0.10	0.01 0.00	0.32 0.41	0.01 0.00	0.84 0.27	0.01 0.00	0.31 0.42
(b)	C-FS	0.00 0.00	0.98 0.06	0.00 0.00	1.00 0.03	0.00 0.00	0.99 0.04	0.00 0.00	1.00 0.02
	FR	0.00 0.00	0.60 0.28	0.00 0.00	0.19 0.35	0.00 0.00	0.44 0.23	0.00 0.00	0.14 0.24
	LASSO	0.00 0.00	0.54 0.42	0.00 0.00	0.17 0.35	0.00 0.00	0.15 0.29	0.00 0.00	0.04 0.17
	SIRS	0.01 0.00	0.96 0.08	0.01 0.00	0.96 0.08	0.01 0.00	0.99 0.04	0.01 0.00	1.00 0.03
	DC-SIS	0.01 0.00	0.98 0.06	0.01 0.00	0.41 0.46	0.01 0.00	0.96 0.14	0.01 0.00	0.39 0.46
	SIS	0.01 0.00	0.93 0.15	0.01 0.00	0.29 0.41	0.01 0.00	0.60 0.30	0.01 0.00	0.22 0.33
(c)	C-FS	0.01 0.00	0.95 0.09	0.01 0.00	0.99 0.05	0.01 0.00	0.94 0.11	0.01 0.00	0.98 0.06
	FR	0.00 0.00	0.89 0.12	0.00 0.00	0.48 0.40	0.00 0.00	0.88 0.14	0.00 0.00	0.49 0.39
	LASSO	0.00 0.00	0.96 0.09	0.00 0.00	0.43 0.42	0.00 0.00	0.93 0.12	0.00 0.00	0.43 0.42
	SIRS	0.01 0.00	0.73 0.10	0.01 0.00	0.76 0.09	0.01 0.00	0.70 0.13	0.01 0.00	0.73 0.10
	DC-SIS	0.01 0.00	0.72 0.11	0.01 0.00	0.74 0.12	0.01 0.00	0.69 0.12	0.01 0.00	0.70 0.13
	SIS	0.01 0.00	0.73 0.10	0.01 0.00	0.54 0.29	0.01 0.00	0.72 0.11	0.01 0.00	0.53 0.28

4.2. An Application

Table 4: The empirical probabilities $\mathcal{P}_{\text{ind}}$ and $\mathcal{P}_{\text{all}}$ based on 500 repetitions for Example 2.

	method	$\varepsilon \sim \mathcal{N}(0,1)$						$\varepsilon \sim t(1)$					
		$\mathcal{P}_{\text{ind}}$					$\mathcal{P}_{\text{all}}$	$\mathcal{P}_{\text{ind}}$					$\mathcal{P}_{\text{all}}$
When $\mathbf{x}$ follows elliptical distribution													
		X_1	X_2	X_3	X_4	X_{20}	ALL	X_1	X_2	X_3	X_4	X_{20}	ALL
(a)	C-FS	1.00	1.00	0.99	0.93	0.97	0.90	1.00	1.00	1.00	0.97	0.91	0.90
	FR	0.96	0.97	0.91	0.62	0.96	0.58	0.28	0.29	0.27	0.20	0.20	0.19
	LASSO	0.92	0.94	0.90	0.70	0.71	0.62	0.25	0.25	0.24	0.19	0.15	0.15
	SIRS	1.00	1.00	1.00	1.00	0.48	0.48	1.00	1.00	1.00	1.00	0.32	0.32
	DC-SIS	1.00	1.00	1.00	1.00	0.56	0.56	0.53	0.53	0.52	0.48	0.14	0.14
	SIS	1.00	1.00	1.00	0.99	0.71	0.71	0.37	0.37	0.37	0.33	0.14	0.13
(b)	C-FS	1.00	0.96	1.00	1.00	0.96	0.91	1.00	0.99	1.00	1.00	0.99	0.98
	FR	0.59	0.38	0.67	0.64	0.73	0.10	0.21	0.15	0.22	0.22	0.16	0.11
	LASSO	0.59	0.55	0.57	0.53	0.48	0.31	0.18	0.18	0.18	0.17	0.12	0.11
	SIRS	1.00	1.00	1.00	0.98	0.84	0.83	1.00	1.00	1.00	0.99	0.82	0.82
	DC-SIS	1.00	1.00	1.00	0.97	0.95	0.91	0.44	0.43	0.42	0.39	0.34	0.32
	SIS	0.96	0.95	0.93	0.88	0.95	0.76	0.31	0.31	0.29	0.29	0.23	0.19
		X_1	X_2	X_3	X_4	X_5	ALL	X_1	X_2	X_3	X_4	X_5	ALL
(c)	C-FS	1.00	1.00	1.00	0.96	0.78	0.75	1.00	1.00	1.00	1.00	0.94	0.94
	FR	0.99	1.00	1.00	0.93	0.56	0.51	0.45	0.59	0.64	0.45	0.28	0.24
	LASSO	0.89	1.00	1.00	0.99	0.90	0.80	0.24	0.53	0.54	0.46	0.36	0.22
	SIRS	0.00	0.99	0.99	0.93	0.73	0.00	0.00	1.00	1.00	0.97	0.81	0.00
	DC-SIS	0.00	0.99	1.00	0.92	0.69	0.00	0.00	0.99	0.99	0.94	0.76	0.00
	SIS	0.00	1.00	1.00	0.94	0.72	0.00	0.00	0.78	0.79	0.64	0.49	0.00
When $\mathbf{x}$ follows non-elliptical distribution													
		X_1	X_2	X_3	X_4	X_{20}	ALL	X_1	X_2	X_3	X_4	X_{20}	ALL
(a)	C-FS	1.00	1.00	0.98	0.89	1.00	0.86	1.00	1.00	0.98	0.93	0.99	0.89
	FR	0.67	0.69	0.54	0.27	0.97	0.21	0.23	0.26	0.17	0.09	0.32	0.08
	LASSO	0.52	0.54	0.46	0.28	0.48	0.26	0.16	0.17	0.14	0.08	0.14	0.08
	SIRS	1.00	1.00	1.00	1.00	0.88	0.88	1.00	1.00	1.00	1.00	0.83	0.83
	DC-SIS	0.99	1.00	0.99	0.98	0.93	0.91	0.52	0.53	0.51	0.46	0.40	0.37
	SIS	0.84	0.87	0.81	0.71	0.98	0.67	0.32	0.32	0.31	0.25	0.35	0.22
(b)	C-FS	0.99	0.96	1.00	1.00	1.00	0.95	1.00	0.99	1.00	1.00	1.00	0.99
	FR	0.16	0.08	0.50	0.53	0.90	0.00	0.06	0.03	0.16	0.20	0.27	0.00
	LASSO	0.09	0.08	0.21	0.18	0.22	0.04	0.03	0.02	0.06	0.05	0.06	0.01
	SIRS	1.00	1.00	1.00	0.98	0.99	0.97	1.00	1.00	1.00	0.99	1.00	0.98
	DC-SIS	0.96	0.96	0.98	0.88	0.99	0.86	0.39	0.39	0.42	0.36	0.41	0.33
	SIS	0.37	0.37	0.63	0.66	0.96	0.24	0.15	0.14	0.24	0.26	0.32	0.08
		X_1	X_2	X_3	X_4	X_5	ALL	X_1	X_2	X_3	X_4	X_5	ALL
(c)	C-FS	1.00	0.99	0.97	0.95	0.78	0.73	1.00	1.00	1.00	0.98	0.91	0.89
	FR	0.96	1.00	0.98	0.92	0.55	0.49	0.48	0.62	0.63	0.45	0.29	0.24
	LASSO	0.82	1.00	0.99	0.99	0.87	0.73	0.26	0.53	0.54	0.45	0.35	0.22
	SIRS	0.00	0.99	0.95	0.89	0.66	0.00	0.00	1.00	0.99	0.95	0.70	0.00
	DC-SIS	0.00	0.99	0.99	0.88	0.59	0.00	0.00	0.99	0.99	0.92	0.60	0.00
	SIS	0.00	1.00	0.99	0.92	0.69	0.00	0.00	0.78	0.77	0.66	0.44	0.00

We further illustrate the performance of the proposed C-FS method through a rat eye expression dataset, which was previously studied by Scheetz et al. (2006) and Huang et al. (2008). This dataset consists of 31,042 probe sets of 120 twelve-week-old male rats, yet only 18,976 probes were sufficiently expressed. The response variable TRIM32 is among these 18,976 probes. This probe was found to cause Bardet-Biedl syndrome (Chiang et al., 2006). We rank the remaining 18,975 probes according to their variances and retain only 3,000 probes with the largest variances. Our analysis is based on the selected 3,000 probes, in addition to the probe TRIM32. The goal is to identify the probes that affect the expression level of TRIM32 considerably.

The sample size $n = 120$ is small compared with the covariate dimension $p = 3,000$. We apply the aforementioned six feature selection/screening methods to this dataset and denote the retained covariates as $\mathbf{x}_{\widehat{\mathcal{A}}}$. We order the entries of $\widehat{\mathcal{A}}$ according to the relative importance of each retained covariate. Specifically, for SIS, DC-SIS, and SIRS, $\widehat{\mathcal{A}}$ is the index set of the covariates with s largest marginal effects; for C-FS, FR, and LASSO, $\widehat{\mathcal{A}}$ is the index set of the first s covariates that enter the active set.

We assess the performance of these methods as follows. Given a model size s, we fit an additive model

$$Y = \sum_{j=1}^{s} f_{kj}(X_{kj}) + \varepsilon_k, \tag{4.1}$$

where $k = 1, \ldots, 6$ represents C-FS, FR, LASSO, SIRS, DC-SIS , and SIS respectively. The subscript kj denotes the jth element in $\widehat{\mathcal{A}}_k$ and s is set from 1 to 10. We stop our comparison at the 10-th step because the C-FS algorithm with critical values decided by bootstrap stops at this step. In other words, all remaining null hypotheses are accepted at the significance level 0.01 and there is no need to add additional covariates.

We estimate the unknown functions f_{kj} by the R package `mgcv`, where the adjusted

R^2 and the explained deviance are summarized in Table 5. Since the deviance explained

Table 5: The adjusted R^2 and the explained deviance of the six methods.

model	adjusted R^2						dev.explained					
size	C-FS	FR	LASSO	SIRS	DC-SIS	SIS	C-FS	FR	LASSO	SIRS	DC-SIS	SIS
1	0.33	0.62	0.62	0.33	0.33	0.62	0.35	0.62	0.62	0.35	0.35	0.62
2	0.63	0.68	0.68	0.68	0.68	0.68	0.66	0.69	0.69	0.71	0.69	0.69
3	0.66	0.70	0.68	0.69	0.69	0.69	0.71	0.71	0.70	0.71	0.71	0.71
4	0.72	0.72	0.70	0.70	0.70	0.70	0.74	0.73	0.72	0.73	0.71	0.72
5	0.72	0.74	0.70	0.74	0.70	0.70	0.75	0.75	0.73	0.77	0.72	0.72
6	0.79	0.75	0.73	0.76	0.71	0.73	0.83	0.77	0.77	0.80	0.73	0.77
7	0.81	0.76	0.74	0.76	0.72	0.74	0.84	0.78	0.77	0.80	0.74	0.77
8	0.81	0.77	0.77	0.75	0.72	0.73	0.84	0.79	0.80	0.79	0.74	0.77
9	0.82	0.77	0.77	0.75	0.71	0.74	0.86	0.79	0.81	0.80	0.74	0.78
10	0.84	0.78	0.77	0.74	0.73	0.74	0.88	0.80	0.81	0.78	0.76	0.78

is defined as the proportion of the null deviance explained by the fitted model, the method with larger deviance has a better performance. From Table 5, we observe that, as s increases, the proposed C-FS tends to outperform all other marginal effect based methods. This is partly because these procedures may fail to identify some truly important covariates. In comparison, the performances of C-FS, FR and LASSO are relatively satisfactory as they consider the joint effects. Among these methods, the proposed C-FS has the highest R^2 and explained deviance.

We use the five-fold cross-validation to further compare the prediction performance. Specifically, we randomly partition the dataset into five equal sized subsamples, denoted $\mathcal{D}_1, \ldots, \mathcal{D}_5$. For each subsample $\mathcal{D}_k$, we use the remaining four subsamples to fit model (4.1) with $s = 10$, then calculate the mean squared prediction error on the subsample $\mathcal{D}_k$. We repeat this procedure such that the prediction is performed on each subsample exactly once. The mean squared prediction error of C-FS, FR, LASSO, SIRS, DC-SIS and SIS are 0.43, 0.74, 0.49, 0.78, 0.45 and 0.60, respectively. In this example, the C-FS gives the best prediction, followed by the DC-SIS and the LASSO.

5. CONCLUDING REMARKS

In this article, we proposed a CD-based forward screening procedure, which is model-free and robust to the presence of outliers in the response. By using a stepwise searching framework, the proposed procedure incorporates joint correlations among features in the screening process and thus provides more reliable results in applications.

In general, this forward screening procedure shares similar spirit to the iterative screening approaches (Zhu et al.,, 2011; Zhong and Zhu, 2015). However, how to decide model sizes for the iterative screening approaches and how to study their theoretical properties are rarely discussed in existing literature. Equipped with our proposed model-free forward screening procedure, a data-driven method is proposed to determine which covariates should be retained. We also show that our proposed forward screening procedure possesses the desirable sure screening property.

A model-free and robust method is often computationally intensive. Our experiences shows that the proposed C-FS procedure is more computationally demanding than other procedures. This is possibly caused by stepwise searching and adaptive thresholding, which make the proposed method more reliable and fully automatic. We conjecture that a simplified procedure may lead to less computational cost with a small sacrifice of numerical stability. In our simulation setup, each run takes no more than 4 minutes on average for $n = 200$ and $p = 3000$ on PC Intel Core2 Duo T9600 2.8GHz 4GB RAM server. This numerical cost is often acceptable in practice. It would be interesting to further develop a more efficient algorithm for the CD-based method.

APPENDIX: LEMMAS AND PROOFS OF THEOREMS

Appendix A: Some Lemmas

Lemma 1. *(Bernstein's Inequality, Van and Wellner (1996, Lemma 2.2.11)) Let* $X_1, X_2, \ldots, X_n$ *be independent random variables with mean* 0 *and* $E|X_i|^m \leq m! M^{m-2} \upsilon_i/2$ *for every* $m \geq 2$ *and* $i = 1, 2, \ldots, n$*, where* M *and* υ_i *are positive constants. Then*

$$pr\{|X_1 + X_2 + \ldots + X_n| > \varepsilon\} \leq 2\exp\left\{-\frac{\varepsilon^2}{2(\upsilon + M\varepsilon)}\right\}, \text{ for } \upsilon \geq \sum_{i=1}^{n} \upsilon_i.$$

For notational clarity, we denote $s = |\mathcal{F}|$ in what follows.

Lemma 2. *If* $s = o(n^{1/5})$ *and conditions (B1)-(B5) hold true, then for any* $k \in \mathcal{F}^c$*, and* $\varepsilon_n = Cn^{-\kappa}$*,* C *and* κ *are positive constants,*

$$pr\left\{(\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}} - \boldsymbol{\beta}_{0,k|\mathcal{F}})^{\tau}(\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}} - \boldsymbol{\beta}_{0,k|\mathcal{F}}) > \varepsilon_n\right\} < 2s\exp(-c_1 n s^{-1}\varepsilon_n),$$

where c_1 is a positive constant and $\boldsymbol{\beta}_{0,k|\mathcal{F}}$ is defined by $\underset{\boldsymbol{\beta}_{k|\mathcal{F}}}{\operatorname{argmin}} E\{X_k - g_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}}, \boldsymbol{\beta}_{k|\mathcal{F}})\}^2$.

Proof of Lemma 2: The proof is given in the supplementary document. □

Appendix B: Proof of Theorems

Proof of Theorem 2: We prove the first part. Under H_0, for any $y \in \mathbb{R}$, we have

$$\begin{aligned} E\left\{\mathbf{1}(Y < y)E(X_k \mid \mathbf{x}_{\mathcal{F}})\right\} &= E\left[E\left\{\mathbf{1}(Y < y) \mid \mathbf{x}_{\mathcal{F}}\right\} X_k\right] = E[E\{\mathbf{1}(Y < y) \mid \mathbf{x}_{\mathcal{F}}, X_k\}X_k] \\ &= E\{\mathbf{1}(Y < y)E(X_k \mid \mathbf{x}_{\mathcal{F}}, X_k)\} = E\{\mathbf{1}(Y < y)X_k\}. \end{aligned}$$

This completes the proof of the first part. Next we prove the second part. Define $\boldsymbol{\xi}(y) = \{\xi_1(y), \ldots, \xi_p(y)\}^{\tau} \overset{\text{def}}{=} E\left\{\partial F(y \mid \mathbf{x})/\partial \mathbf{x}\right\}$. Stein's lemma yields that $\boldsymbol{\xi}(y) = \boldsymbol{\Sigma}^{-1}E\{\mathbf{1}(Y < y)\mathbf{x}\}$. Assumption A1 ensures that, for any $k \in \mathcal{A}$, there exists some $y \in \mathbb{R}$ such that $\xi_k(y) \neq 0$. Next, we show that for any $\mathcal{F}$ satisfying $\mathcal{F}^c \cap \mathcal{A} \neq \varnothing$,

$\max_{k\in\mathcal{F}^c\cap\mathcal{A}} \omega_{k|\mathcal{F}} > 0$ holds. Define $\Omega_{k|\mathcal{F}}(y) \stackrel{\text{def}}{=} E[\mathbf{1}(Y < y)\{X_k - E(X_k \mid \mathbf{x}_\mathcal{F})\}]$. The normality of $\mathbf{x}$ indicates $E(X_k \mid \mathbf{x}_\mathcal{F}) = \boldsymbol{\beta}_{k|\mathcal{F}}^{\mathrm{T}}\mathbf{x}_\mathcal{F}$, where $\boldsymbol{\beta}_{k|\mathcal{F}} = \boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},k}$. Thus $\Omega_{k|\mathcal{F}}(y) = (1, -\boldsymbol{\beta}_{k|\mathcal{F}}^{\mathrm{T}})\left[E\{\mathbf{1}(Y < y)X_k\}, E\{\mathbf{1}(Y < y)\mathbf{x}_\mathcal{F}^{\mathrm{T}}\}\right]^{\mathrm{T}}$. Without loss of generality, we assume $(\mathbf{x}_\mathcal{F}, X_k)$ be the first $|\mathcal{F}| + 1$ elements of $\mathbf{x}$. It follows that

$$\Omega_{k|\mathcal{F}}(y) = (-\boldsymbol{\beta}_{k|\mathcal{F}}^{\mathrm{T}}, 1)(\mathbf{I}_{|\mathcal{F}|+1}, \mathbf{0}_{(|\mathcal{F}|+1)\times(p-|\mathcal{F}|-1)})\boldsymbol{\Sigma}\boldsymbol{\xi}(y) = (\boldsymbol{\Sigma}_{k,\mathcal{S}} - \boldsymbol{\Sigma}_{k,\mathcal{F}}\boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{S}})\boldsymbol{\xi}(y),$$

where $\boldsymbol{\Sigma}_{\mathcal{F}_1,\mathcal{F}_2} = E(\mathbf{x}_{\mathcal{F}_1}^{\mathrm{T}}\mathbf{x}_{\mathcal{F}_2})$, for $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{S}$. Under model (3.1), $\Omega_{k|\mathcal{F}}(y) = (\boldsymbol{\Sigma}_{k,\mathcal{F}^c\cap\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{F}}\boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}})\boldsymbol{\xi}_{\mathcal{F}^c\cap\mathcal{A}}(y)$. This yields that

$$\begin{aligned}\sum_{k\in\mathcal{F}^c\cap\mathcal{A}} \Omega_{k|\mathcal{F}}^2(y) &= \sum_{k\in\mathcal{F}^c\cap\mathcal{A}} \boldsymbol{\xi}_{\mathcal{F}^c\cap\mathcal{A}}(y)^{\mathrm{T}}(\boldsymbol{\Sigma}_{k,\mathcal{F}^c\cap\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{F}}\boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}})^2\boldsymbol{\xi}_{\mathcal{F}^c\cap\mathcal{A}}(y) \\ &= \boldsymbol{\xi}_{\mathcal{F}^c\cap\mathcal{A}}(y)^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}^c\cap\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}}\boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}})^2\boldsymbol{\xi}_{\mathcal{F}^c\cap\mathcal{A}}(y).\end{aligned}$$

Define

$$\boldsymbol{\Sigma}_{\mathcal{F}\cup(\mathcal{F}^c\cap\mathcal{A})} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{\Sigma}_\mathcal{F} & \boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}} \\ \boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}} & \boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}^c\cap\mathcal{A}} \end{pmatrix}.$$

Because $(\boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}^c\cap\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}}\boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}})^{-1}$ is sub-matrix of $\boldsymbol{\Sigma}_{\mathcal{F}\cup(\mathcal{F}^c\cap\mathcal{A})}^{-1}$, we have

$$\rho_{\max}((\boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}^c\cap\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}}\boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}})^{-1}) < \rho_{\max}(\boldsymbol{\Sigma}_{\mathcal{F}\cup(\mathcal{F}^c\cap\mathcal{A})}^{-1}).$$

Accordingly,

$$\rho_{\min}(\boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}^c\cap\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}}\boldsymbol{\Sigma}_\mathcal{F}^{-1}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}}) > \rho_{\min}(\boldsymbol{\Sigma}_{\mathcal{F}\cup(\mathcal{F}^c\cap\mathcal{A})}) > \rho_{\min}(\boldsymbol{\Sigma}),$$

where $\rho_{\min}(\mathbf{M})$ and $\rho_{\max}(\mathbf{M})$ represent the maximum and minimum eigenvalue of ma-

trix $\mathbf{M}$, respectively. This leads to that

$$\begin{aligned}\max_{k\in\mathcal{F}^c\cap\mathcal{A}}\Omega^2_{k|\mathcal{F}}(y) &\geq |\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{k\in\mathcal{F}^c\cap\mathcal{A}}\Omega^2_{k|\mathcal{F}}(y)\\ &\geq |\mathcal{F}^c\cap\mathcal{A}|^{-1}\rho^2_{\min}(\boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}^c\cap\mathcal{A}}-\boldsymbol{\Sigma}_{\mathcal{F}^c\cap\mathcal{A},\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F},\mathcal{F}^c\cap\mathcal{A}})\boldsymbol{\xi}^{\mathrm{T}}_{\mathcal{F}^c\cap\mathcal{A}}(y)\boldsymbol{\xi}_{\mathcal{F}^c\cap\mathcal{A}}(y)\\ &\geq |\mathcal{F}^c\cap\mathcal{A}|^{-1}\rho^2_{\min}(\boldsymbol{\Sigma})\boldsymbol{\xi}^{\mathrm{T}}_{\mathcal{F}^c\cap\mathcal{A}}(y)\boldsymbol{\xi}_{\mathcal{F}^c\cap\mathcal{A}}(y)>0,\end{aligned}$$

for some $y\in\mathbb{R}$. This completes the proof of the second part of Theorem 2. □

For notational clarity, we denote $\mu_{k|\mathcal{F}}\stackrel{\text{def}}{=}E(X_k\mid\mathbf{x}_{\mathcal{F}})$ in what follows.

Proof of Theorem 3: Define

$$\zeta_{n,k|\mathcal{F}}(y)\stackrel{\text{def}}{=}n^{-1/2}\sum_{i=1}^n\mathbf{1}(Y_i<y)\left\{X_{ik}-g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}},\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}})\right\},\ \text{for } y\in\mathbb{R}.$$

By Taylor's expansion, it follows that

$$\begin{aligned}\zeta_{n,k|\mathcal{F}}(y) &= n^{-1/2}\sum_{i=1}^n\mathbf{1}(Y_i<y)\{X_{ik}-g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}})\}\\ &+ E\{\mathbf{1}(Y<y)g'_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}})\}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{k|\mathcal{F}}-\boldsymbol{\beta}_{0,k|\mathcal{F}})+o_p(1).\end{aligned}$$

(S.1) in the supplement gives that

$$\begin{aligned}\zeta_{n,k|\mathcal{F}}(y) &= n^{-1/2}\sum_{i=1}^n V_{k|\mathcal{F}}(X_{ik},\mathbf{x}_{i\mathcal{F}},Y_i;y)+o_p(1),\quad\text{where}\\ V_{k|\mathcal{F}}(X_{ik},\mathbf{x}_{i\mathcal{F}},Y_i;y) &= \mathbf{1}(Y_i<y)\{X_{ik}-g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}})\}+E\{\mathbf{1}(Y<y)g'_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}})^{\mathrm{T}}\}\\ &\quad\boldsymbol{\Sigma}^{-1}_{k|\mathcal{F}}\{X_{ik}-g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}})\}g'_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}}),\end{aligned}$$

and $\boldsymbol{\Sigma}_{k|\mathcal{F}}=E\left[\left\{g'_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}})\right\}\left\{g'_{k|\mathcal{F}}(\mathbf{x}_{\mathcal{F}},\boldsymbol{\beta}_{0,k|\mathcal{F}})\right\}^{\mathrm{T}}\right]$.

Suppose H_0 in (3.2) holds true. Theorem 2 ensures that $\omega_{k|\mathcal{F}}=0$. Let $\zeta_{k|\mathcal{F}}(\cdot)$ be a mean zero Gaussian process with covariance function $\mathrm{cov}\{\zeta_{k|\mathcal{F}}(y_1),\zeta_{k|\mathcal{F}}(y_2)\}=$

$E\left\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y_1)V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y_2)\right\}$. It is easy to verify that $E\{\zeta_{n,k|\mathcal{F}}(y)\} = o(1)$ and $E\{\zeta^2_{n,k|\mathcal{F}}(y)\} = \text{cov}\{\zeta_{k|\mathcal{F}}(y), \zeta_{k|\mathcal{F}}(y)\} + o(1)$. Thus we have $\zeta_{n,k|\mathcal{F}}(\cdot) \xrightarrow{d} \zeta_{k|\mathcal{F}}(\cdot)$, and consequently $\int_{-\infty}^{+\infty} \zeta^2_{n,k|\mathcal{F}}(y)dF_n(y) \xrightarrow{d} \int_{-\infty}^{+\infty} \zeta^2_{k|\mathcal{F}}(y)dF(y)$. This, together with the fact that $n\,\widehat{\text{CCov}}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} = \int_{-\infty}^{+\infty} \zeta^2_{n,k|\mathcal{F}}(y)dF_n(y)$, yields that $n\,\widehat{\text{CCov}}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} \xrightarrow{d} \int_{-\infty}^{+\infty} \zeta^2_{k|\mathcal{F}}(y)dF(y)$ (Kuo, 1975). Therefore,

$$n\left[E\left\{\int_{-\infty}^{+\infty} \zeta^2_{k|\mathcal{F}}(y)dF(y)\right\}\right]^{-1} \widehat{\text{CCov}}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_{j,k|\mathcal{F}}\chi^2_j(1).$$

By Slutsky's theorem, it follows that $n\kappa^{-1}_{k|\mathcal{F}}\widehat{\omega}_{k|\mathcal{F}} \longrightarrow \sum_{j=1}^{\infty} \lambda_{j,k|\mathcal{F}}\chi^2_j(1)$, where

$$\begin{aligned}\kappa_{k|\mathcal{F}} \stackrel{\text{def}}{=} & E\Big[\mathbf{1}(Y < \widetilde{Y})\{X_k - g_{k|\mathcal{F}}(\mathbf{x}_\mathcal{F}, \boldsymbol{\beta}_{0,k|\mathcal{F}})\} + E\{\mathbf{1}(Y < \widetilde{Y})g'_{k|\mathcal{F}}(\mathbf{x}_\mathcal{F}, \boldsymbol{\beta}_{0,k|\mathcal{F}})^{\mathrm{T}}\}\boldsymbol{\Sigma}^{-1}_{k|\mathcal{F}} \\ & \{X_k - g_{k|\mathcal{F}}(\mathbf{x}_\mathcal{F}, \boldsymbol{\beta}_{0,k|\mathcal{F}})\}g'_{k|\mathcal{F}}(\mathbf{x}_\mathcal{F}, \boldsymbol{\beta}_{0,k|\mathcal{F}})\Big]^2 \Big/ \text{var}\{X_k - g_{k|\mathcal{F}}(\mathbf{x}_{i\mathcal{F}}, \boldsymbol{\beta}_{0,k|\mathcal{F}})\}. \quad \text{(B.1)}\end{aligned}$$

Suppose H_1 in (3.2) holds true. Recall that

$$\begin{aligned}& \widehat{\text{CCov}}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} = \int \{n^{-1/2}\zeta_{n,k|\mathcal{F}}(y)\}^2 dF_n(y) \\ = & \int [n^{-1/2}\zeta_{n,k|\mathcal{F}}(y) - E\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y)\} + E\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y)\}]^2 dF_n(y) \\ = & \int 2E\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y)\}[n^{-1/2}\zeta_{n,k|\mathcal{F}}(y) - E\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y)\}]dF_n(y) \\ & \qquad + \int E^2\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y)\}dF_n(y) + o_p(n^{-1/2}) \\ = & \int 2E\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y)\}n^{-1/2}\zeta_{n,k|\mathcal{F}}(y)dF(y) - 2\text{CCov}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} \\ & \qquad + \int E^2\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_\mathcal{F}, Y; y)\}dF_n(y) + o_p(n^{-1/2}).\end{aligned}$$

Thus

$$\widehat{\text{CCov}}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} - \text{CCov}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} = n^{-1}\sum_{i=1}^{n} Z_{i,k|\mathcal{F}} + o_p(n^{-1/2}),$$

with

$$Z_{i,k|\mathcal{F}} \stackrel{\text{def}}{=} 2\left[\int E\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_{\mathcal{F}}, Y; y)\}V_{k|\mathcal{F}}(X_{ik}, \mathbf{x}_{i\mathcal{F}}, Y_i; y)dF(y) - \text{CCov}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\}\right]$$
$$+ \quad E^2\{V_{k|\mathcal{F}}(X_k, \mathbf{x}_{\mathcal{F}}, Y; Y_i)\} - \text{CCov}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\}, \tag{B.2}$$

where the expectation is taken with respect to $(X_k, \mathbf{x}_{\mathcal{F}}, Y)$. By the central limit theorem, $n^{1/2}\Big[\widehat{\text{CCov}}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\} - \text{CCov}\{(X_k - \mu_{k|\mathcal{F}}) \mid Y\}\Big]$ converges in distribution to $\mathcal{N}(0, \varsigma^2_{k|\mathcal{F}})$, where $\varsigma^2_{k|\mathcal{F}} \stackrel{\text{def}}{=} \text{var}(Z_{i,k|\mathcal{F}})$. By Slutsky's theorem, we have $n^{1/2}(\widehat{\omega}_{k|\mathcal{F}} - \omega_{k|\mathcal{F}}) \xrightarrow{d} \mathcal{N}\left(0, \Delta_{k|\mathcal{F}}\right)$, where

$$\Delta_{k|\mathcal{F}} \stackrel{\text{def}}{=} \varsigma^2_{k|\mathcal{F}} / \left\{\text{var}(X_k - \mu_{k|\mathcal{F}})\right\}^2. \tag{B.3}$$

This completes the proof of Theorem 3.

The uniform consistency of $\widehat{\omega}_{k|\mathcal{F}}$ paves the road for proving Theorem 4.

PROPOSITION 1. *Under conditions (B1)-(B5), for any $\varepsilon_n > 0$, there exists positive constants c_1, c_2,c_3, c_4 and sufficiently small $s_{\varepsilon_n} \in (0, 2/\varepsilon_n)$ such that*

$$pr\Big\{\max_{k\in\mathcal{F}^c}|\widehat{\omega}_{k|\mathcal{F}} - \omega_{k|\mathcal{F}}| > \varepsilon_n\Big\} \leq O\Big[p\exp\{n\log(1 - \varepsilon_n s_{\varepsilon_n}/2)/3\} + p\exp(-c_1 n\varepsilon_n^2)$$
$$+pn\exp(-c_2 n\varepsilon_n^2) + ps\exp(-c_3 n s^{-2}\varepsilon_n^2) + ps\exp(-c_4 n s^{-2}\varepsilon_n)\Big].$$

Set $\varepsilon_n = Cn^{-\kappa}$, for some constants $C > 0$ and $\kappa > 0$. If there exists $\vartheta > 0$ such that $p = o\{\exp(an^{\vartheta})\}$ for any $a > 0$ and $3/5 - 2\kappa - \vartheta > 0$, then

$$\max_{\mathcal{F}:|\mathcal{F}|=o(n^{1/5})} pr\left\{\max_{k\in\mathcal{F}^c}|\widehat{\omega}_{k|\mathcal{F}} - \omega_{k|\mathcal{F}}| > Cn^{-\kappa}\right\} \to 0 \text{ as } n \to \infty.$$

PROOF OF PROPOSITION 1: The proof is given in the supplementary document. □

PROOF OF THEOREM 4: For notational clarity, we define the random event $E_1 = \{\text{There exists an index set } \mathcal{F}, \text{ such that} \mathcal{F}^c \cap \mathcal{A} \neq \emptyset \text{ and } \max_{k \in \mathcal{F}^c \cap \mathcal{A}} \widehat{\omega}_{k|\mathcal{F}} \leq \nu\}$. For such an $\mathcal{F}$, we have, by Assumption A3, $\max_{k \in \mathcal{F}^c \cap \mathcal{A}} \omega_{k|\mathcal{F}} - \max_{k \in \mathcal{F}^c \cap \mathcal{A}} \widehat{\omega}_{k|\mathcal{F}} > Cn^{-\varpi} - \nu$. Consequently,

$$\max_{k \in \mathcal{F}^c \cap \mathcal{A}} (\omega_{k|\mathcal{F}} - \widehat{\omega}_{k|\mathcal{F}}) \geq \max_{k \in \mathcal{F}^c \cap \mathcal{A}} \omega_{k|\mathcal{F}} - \max_{k \in \mathcal{F}^c \cap \mathcal{A}} \widehat{\omega}_{k|\mathcal{F}} \geq Cn^{-\varpi}/2.$$

Define the random event $E_2 = \left\{ \max_{k \in \mathcal{F}^c \cap \mathcal{A}} (\omega_{k|\mathcal{F}} - \widehat{\omega}_{k|\mathcal{F}}) > Cn^{-\varpi}/2, \text{ for } \mathcal{F} \text{ in } E_1 \right\}$. The above discussions imply that $E_1 \subseteq E_2$. It follows that

$$\text{pr}\left(\min_{\mathcal{F}: \mathcal{F}^c \cap \mathcal{A} \neq \varnothing} \max_{k \in \mathcal{F}^c \cap \mathcal{A}} \widehat{\omega}_{k|\mathcal{F}} > \nu \right) = 1 - \text{pr}(E_1) \geq 1 - \text{pr}(E_2).$$

Proposition 1 implies that $\text{pr}(E_2) \to 0$, which completes the proof of Theorem 4. □

Acknowledgment

The authors thank the AE, and the reviewers for their constructive comments, which have led to a significant improvement of the earlier version of this article. Liping Zhu is the corresponding author.

Funding

This work was supported by National Natural Science Foundation of China (NNSFC) grants 11731011, 11690014, 11690015, 11731011 and 11801501, NSERC RGPIN-2016-05024, NSF grant DMS 1820702 and NIDA, NIH grant P50 DA039838, the Ministry of Education Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002) and National Youth Top-notch Talent Support Program, P. R. China. The content is solely the responsibility of the authors and does not necessarily represent the official views of NNSFC, MEC, NSF, NIH or NIDA.

REFERENCES

Barut, E., Fan, J. and Verhasselt, A. (2016). "Conditional sure independence screening" *Journal of the American Statistical Association*. **111**, 1266–1277.

Bickel, P. and Levina, E. (2008). "Covariance regularization by thresholding" *The Annals of Statistics*. **36**, 2577–2604.

Buldygin, V. V. and Kozachenko, Y. V. (1980). Sub-Gaussian random variables. *Ukrainian Mathematical Journal*, **32**, 483–489.

Candes, E. and Tao, T. (2007). "The Dantzig selector: statistical estimation when p is much larger than n (with discussion)." *The Annals of Statistics*. **35**, 2313–2404.

Chang, J., Tang, C. Y., and Wu, Y. (2013). "Marginal empirical likelihood and sure independence feature screening." *The Annals of Statistics*, **41**, 2123–2148.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M., and Shefield, V. C. (2006). "Homozygosity mapping with snp arrays identifies a novel gene for bardet-biedl syndrome." *Proceeding 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 it 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 (with discussion)." *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

Fan, J. and Peng, H. (2004). "Nonconcave penalized likelihood with a diverging number of parameters." *Annals of Statistics*, **32**, 928–961.

Fan, J., Samworth, R. and Wu, Y. (2009). "Ultrahigh dimensional feature selection: beyond the linear model." *Journal of Machine Learning Research*, **10**, 1829–1853.

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

He, X., Wang, L., and Hong, H. (2013). "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data." *The Annals of Statistics* **41**, 342–369.

Huang, J., Ma, S. G., and Zhang, C. H. (2008). "Adaptive lasso for sparse high-dimensional regression models." *Statistica Sinica*, **18**, 1603–1618.

Huber, P. J. (1973). "Robust regression: asymptotics, conjectures and Monte Carlo". *The Annals of Statistics*, **1**: 799–821.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. The Second Edition. Wiley: New York.

Jennrich, R. I. (1969). "Asymptotic properties of non-linear least squares estimators." *The Annals of Mathematical Statistics*, **40**: 633–643.

Kuo, H. H. (1975) *Gaussian Measures in Banach Spaces.* Lecture Notes in Mathematics. Springer: Berlin.

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.

Ma, S., Li, R. and Tsai, C.-L. (2017). "Variable screening via quantile partial correlation". *Journal of American Statistical Association*, **112**, 650–663.

Mai, Q. and Zou, H. (2013). "The Kolmogorov filter for variable screening in high-dimensional binary classification." *Biometrika*, **100**, 229–234.

Mai, Q., Zou, H., and Yuan, M. (2012). "A direct approach to sparse discriminant analysis in ultra-high dimensions." *Biometrika*, **99**, 29–42.

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., Shefield, V. C., and Stone, E. M. (2006). "Regulation of gene expression in the mammalian eye and its relevance to eye disease." *Proceeding of the National Academy of Sciences*, **103**, 14429–14434.

Shao, X. F. and Zhang, J. S. (2014) "Martingale difference correlation and its use in high dimensional variable screening." *Journal of the American Statistical Association*. **109**, 1302–1318.

Song, R., Yi, F., and Zou, H. (2014). "On varying-coefficient independence screening for high-dimensional varying-coefficient models." *Statistica Sinica*, **24**, 1735–1752.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) "Measuring and testing dependence by correlation of distances." *The Annals of Statistics*, **35**, 2769–2794.

Székely, G. J. and Rizzo, M. L. (2009) "Brownian distance covariances." *The Annals of Statistics*, **3**, 1236–1265.

Tan, F. L. and Zhu, L. X. (2018). "Adaptive-to-model checking for regressions with diverging number of predictors." *The Annals of Statistics*, to appear.

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

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes,* New York: Springer.

Wang, H. (2009). "Forward regression for ultra-high dimensional variable screening." *Journal of the American Statistical Association*, **104**, 1512–1524.

White, H. (1981)."Consequences and detection of misspecified nonlinear regression models. " *Journal of the American Statistical Association*, **76**, 419–433.

Xu, C. and Chen, J. (2014). "The Sparse MLE for ultra-high-dimensional feature screening." *Journal of the American Statistical Association*, **109**, 1257–1269.

Zhong, W. and Zhu, L. P. (2015). "An iterative approach to distance correlation-based sure independence screening." *Journal of Statistical Computation and Simulation*, **85**, 1-15.

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.