Model-Free Feature Screening for Ultrahigh-Dimensional Data

Li-Ping ZHU, Lexin LI, Runze LI, and Li-Xing ZHU

With the recent explosion of scientific data of unprecedented size and complexity, feature ranking and screening are playing an increasingly important role in many scientific studies. In this article, we propose a novel feature screening procedure under a unified model framework, which covers a wide variety of commonly used parametric and semiparametric models. The new method does not require imposing a specific model structure on regression functions, and thus is particularly appealing to ultrahigh-dimensional regressions, where there are a huge number of candidate predictors but little information about the actual model forms. We demonstrate that, with the number of predictors growing at an exponential rate of the sample size, the proposed procedure possesses consistency in ranking, which is both useful in its own right and can lead to consistency in selection. The new procedure is computationally efficient and simple, and exhibits a competent empirical performance in our intensive simulations and real data analysis.

KEY WORDS: Feature ranking; Ultrahigh-dimensional regression; Variable selection.

1. INTRODUCTION

High-dimensional data are frequently collected in a large variety of areas such as biomedical imaging, functional magnetic resonance imaging, tomography, tumor classifications, and finance. In high-dimensional data, the number of variables or parameters p can be much larger than the sample size n. Such a "large p, small n" problem has imposed many challenges for statistical analysis, and calls for new statistical methodologies and theories (Donoho 2000; Fan and Li 2006). The sparsity principle, which assumes that only a small number of predictors contribute to the response, is frequently adopted and deemed useful in the analysis of high-dimensional data. Following this general principle, a large number of variable selection approaches have been developed in the recent literature to estimate a sparse model and select significant variables simultaneously. Examples include Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), nonnegative garrote (Breiman 1995), group Lasso (Yuan and Lin 2006), adaptive Lasso (Zou 2006), and Dantzig selector (Candes and Tao 2007). See the article by Fan and Lv (2010) for an overview.

While those variable selection methods have been successfully applied in many high-dimensional analyses, modern applications in areas such as genomics, proteomics, and high-frequency finance further push the dimensionality of data to an even larger scale, where p may grow exponentially with n. Such ultrahigh-dimensional data present simultaneous challenges of computational expediency, statistical accuracy, and algorithm stability (Fan, Samworth, and Wu 2009). It is difficult to directly apply the aforementioned variable selection methods to those ultrahigh-dimensional statistical learning problems due to the computational complexity inherent in those methods. To address those challenges, Fan and Lv (2008) emphasized the importance of feature screening in ultrahigh-dimensional data analysis, and proposed sure independence screening (SIS) and iterated sure independence screening (ISIS) in the context of linear regression models. Furthermore, Fan, Samworth, and Wu (2009) and Fan and Song (2010) extended SIS and ISIS from a linear model to a generalized linear model. Each of those proposals focuses on a specific model, and its performance is based upon the belief that the imposed working model is close to the true model.

In this article, we propose a model-free feature screening approach for ultrahigh-dimensional data. Compared with the SIS, the most distinguishable feature of our proposal is that we only impose a very general model framework instead of a specific model. It is so general that the newly proposed procedure can be viewed as a model-free screening method, and it covers a wide range of commonly used parametric and semiparametric models. This feature makes our proposed procedure particularly appealing for feature screening when there are a huge number of candidate variables, but little information suggesting that the actual model is linear or follows any other specific parametric form. This flexibility is achieved by using the newly proposed marginal utility measure that is concerned with the entire conditional distribution of the response given the predictors. In addition, our method is robust to outliers and heavy-tailed responses in that it only uses the ranks of the observed response values. Theoretically, we establish that the proposed method possesses a *consistency in ranking* (CIR) property. That is, in probability, our marginal utility measure always ranks an active predictor above an inactive one, and thus guarantees a clear separation between the active and inactive predictors. The CIR property

Li-Ping Zhu is Associate Professor, School of Statistics and Management, Shanghai University of Finance and Economics (E-mail: *zhu.liping@mail.shufe.edu.cn*). Lexin Li is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: *li@stat.ncsu.edu*). Runze Li is the corresponding author and Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111 (E-mail: *rli@stat.psu.edu*). Li-Xing Zhu is Chair Professor of Statistics, Department of Mathematics, Hong Kong Baptist University, Hong Kong (E-mail: *lzhu@hkbu.edu.hk*). Li-Ping Zhu's research was supported by National Natural Science Foundation of China grant 11071077 and National Institute on Drug Abuse (NIDA) grant R21-DA024260. Lexin Li's research was supported by NSF grant DMS 1106668. Runze Li's research was supported by NSF grant DMS 0348869, National Natural Science Foundation of China grant 11028103, and National Institute on Drug Abuse (NIDA) grant P50-DA10075. Li-Xing Zhu's research was supported by Research Grants Council of Hong Kong grant HKBU2034/09P. The authors are grateful to Dr. Yichao Wu for sharing the ideas through personal communication about the iterative screening approach presented in this article. The authors thank the editor, the associate editor, and reviewers for their suggestions, which have helped greatly improve the article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIDA.

Journal of the American Statistical Association
December 2011, Vol. 106, No. 496, Theory and Methods
DOI: 10.1198/jasa.2011.tm10563

can be particularly useful in some genomic studies (Choi et al. 2009) where ranking is more of a concern than selection. Moreover, it leads to *consistency in selection*; that is, it simultaneously selects all active predictors and excludes all inactive predictors in probability, provided an ideal cutoff of the utility measure is available. The proposed procedure is valid provided that the total number of predictors p grows slower than $\exp(an)$ for any fixed $a > 0$. This rate is similar to the exponential rate achieved by the SIS procedures. Given a rank of all candidate features, we further propose a combination of hard and soft thresholding strategies to obtain the cutoff point that separates the active and inactive predictors. The soft threshold is constructed by adding a series of auxiliary variables, motivated by the idea of adding pseudo variables in model selection proposed by Luo, Stefanski, and Boos (2006) and Wu, Boos, and Stefanski (2007). Similarly to the iterative SIS procedures, we also propose an iterative version of our new screening method. This is due to the fact that the marginal utility measure may miss an active predictor that is marginally independent of the response, a phenomenon also observed in the SIS procedures. The iterative procedure is shown to resolve this issue effectively. Computationally, the proposed screening procedure does not require any complicated numerical optimization and is very simple and fast to implement.

The rest of the article is organized as follows. In Section 2, we first present our general model framework, then develop the new feature ranking and screening approach. Section 3 illustrates the finite sample performance by both Monte Carlo simulations and a real data analysis. All technical proofs are given in the Appendix.

2. A UNIFIED FEATURE SCREENING APPROACH

2.1 A General Model Framework

Let Y be the response variable with support Ψ_y, and Y can be both univariate and multivariate. Let $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a covariate vector. Here we adopt the same notation system as used by Fan and Lv (2008) where a boldface lowercase letter denotes a vector and a boldface capital letter denotes a matrix. We first develop the notion of active predictors and inactive predictors *without* specifying a regression model. We consider the conditional distribution function of Y given $\mathbf{x}$, denoted by $F(y \mid \mathbf{x}) = P(Y < y \mid \mathbf{x})$. Define two index sets:

$$\begin{aligned}\mathcal{A} = \{&k : F(y \mid \mathbf{x}) \text{ functionally depends} \\ &\text{on } X_k \text{ for some } y \in \Psi_y\},\\ \mathcal{I} = \{&k : F(y \mid \mathbf{x}) \text{ does not functionally} \\ &\text{depend on } X_k \text{ for any } y \in \Psi_y\}.\end{aligned}$$

If $k \in \mathcal{A}$, X_k is referred to as an active predictor, whereas if $k \in \mathcal{I}$, X_k is referred to as an inactive predictor. Let $\mathbf{x}_{\mathcal{A}}$, a $p_1 \times 1$ vector, consist of all X_k with $k \in \mathcal{A}$. Similarly, let $\mathbf{x}_{\mathcal{I}}$, a $(p - p_1) \times 1$ vector, consist of all inactive predictors X_k with $k \in \mathcal{I}$.

Next we consider a general model framework under which we are to develop our unified screening approach. Specifically, we consider that $F(y \mid \mathbf{x})$ depends on $\mathbf{x}$ only through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$ for some $p_1 \times K$ constant matrix $\boldsymbol{\beta}$. In other words, we assume that

$$F(y \mid \mathbf{x}) = F_0(y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}), \tag{2.1}$$

where $F_0(\cdot \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}})$ is an unknown distribution function for a given $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$. We make the following remarks. First, $\boldsymbol{\beta}$ may not be identifiable; what is identified is the space spanned by the columns of $\boldsymbol{\beta}$. However, the identifiability of $\boldsymbol{\beta}$ is of no concern here because our primary goal is to identify active variables rather than to estimate $\boldsymbol{\beta}$ itself. Actually, our screening procedure does not require an explicit estimation of $\boldsymbol{\beta}$. Second, the form of (2.1) is fairly common in a large variety of parametric and semiparametric models where the response Y depends on the predictors $\mathbf{x}$ through a number of linear combinations $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$. As we will show next, (2.1) covers a wide range of existing models and, in many cases, K is as small as just one, two, or three.

Before we continue the pursuit of feature screening, we examine some special cases of model (2.1) to show its generality. Note that many existing regression models for a continuous response can be written in the following form:

$$h(Y) = f_1(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}) + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} + f_2(\boldsymbol{\alpha}_3^{\mathrm{T}}\mathbf{x}_{\mathcal{A}})\varepsilon, \tag{2.2}$$

where $h(\cdot)$ is a monotone function, $f_2(\cdot)$ is a nonnegative function, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, and $\boldsymbol{\alpha}_3$ are unknown coefficients, and it is assumed that ε is independent of $\mathbf{x}$. Here $h(\cdot), f_1(\cdot)$, and $f_2(\cdot)$ may be either known or unknown. Clearly model (2.2) is a special case of (2.1) if we choose $\boldsymbol{\beta}$ to be a basis of the column space spanned by $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$, and $\boldsymbol{\alpha}_3$. Meanwhile, it is seen that model (2.2) with $h(Y) = Y$ includes the following special cases: the linear regression model, the partially linear model (Härdle, Liang, and Gao 2000), the single-index model (Härdle, Hall, and Ichimura 1993), and the partially linear single-index model (Carroll et al. 1997). Model (2.2) also includes the transformation regression model for a general transformation $h(Y)$.

In survival data analysis, the response Y is the time to event of interest, and a commonly used model for Y is the accelerated failure time model:

$$\log(Y) = \alpha_0 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} + \varepsilon,$$

where ε is independent of $\mathbf{x}$. Different choices for the error distribution of ε lead to models that are frequently seen in survival analysis; that is, the extreme value distribution for ε yields the proportional hazards model (Cox 1972), and the logistic distribution for ε yields the proportional odds model (Pettitt 1982). It can again be easily verified that all those survival models are special cases of model (2.1).

Various existing models for discrete responses such as binary outcomes and count responses can be treated as a generalized partially linear single-index model (Carroll et al. 1997)

$$g_1\{E(Y \mid \mathbf{x})\} = g_2(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}) + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}, \tag{2.3}$$

where the conditional distribution of Y given $\mathbf{x}$ belongs to the exponential family, $g_1(\cdot)$ is a link function, $g_2(\cdot)$ is an unknown function, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are unknown coefficients. While model (2.3) includes the generalized linear model and the generalized single-index model as special cases, (2.3) itself is a special case of (2.1), which allows an unknown link function $g_1(\cdot)$ as well.

In summary, a large variety of existing models with various types of response variables can be cast into the common model framework of (2.1). As a consequence, our feature screening approach developed under (2.1) offers a unified approach that works for a wide range of existing models.

2.2 A New Screening Procedure

To facilitate presentation, we assume throughout this article that $E(X_k) = 0$ and $\text{var}(X_k) = 1$ for $k = 1, \ldots, p$. Define $\boldsymbol{\Omega}(y) = E\{\mathbf{x}F(y \mid \mathbf{x})\}$. It then follows by the law of iterated expectations that $\boldsymbol{\Omega}(y) = E[\mathbf{x}E\{\mathbf{1}(Y < y) \mid \mathbf{x}\}] = \text{cov}\{\mathbf{x}, \mathbf{1}(Y < y)\}$. Let $\Omega_k(y)$ be the kth element of $\boldsymbol{\Omega}(y)$, and define

$$\omega_k = E\{\Omega_k^2(Y)\}, \qquad k = 1, \ldots, p. \tag{2.4}$$

Then ω_k is to serve as the population quantity of our proposed marginal utility measure for predictor ranking. Intuitively, one can see that, if X_k and Y are independent, then X_k and the indicator function $\mathbf{1}(Y < y)$ change independently. Consequently $\Omega_k(y) = 0$ for any $y \in \Psi_y$ and $\omega_k = 0$. On the other hand, if X_k and Y are related, then there exists some $y \in \Psi_y$ such that $\Omega_k(y) \neq 0$, and hence ω_k must be positive. This observation motivates us to employ the sample estimate of ω_k to rank all the predictors. We will summarize this intuitive observation more rigorously in Corollary 1 in the next section.

Given a random sample $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ from $\{\mathbf{x}, Y\}$, we next derive a sample estimator of ω_k. For ease of presentation, we assume that the sample predictors are all standardized; that is, $n^{-1}\sum_{i=1}^n X_{ik} = 0$ and $n^{-1}\sum_{i=1}^n X_{ik}^2 = 1$ for $k = 1, \ldots, p$. A natural estimator for ω_k is

$$\tilde{\omega}_k = \frac{1}{n}\sum_{j=1}^n \left\{ \frac{1}{n}\sum_{i=1}^n X_{ik}\mathbf{1}(Y_i < Y_j) \right\}^2, \qquad k = 1, \ldots, p,$$

where X_{ik} denotes the kth element of $\mathbf{x}_i$. As shown in the proof of Theorem 2,

$$\widehat{\omega}_k = \frac{n^3}{n(n-1)(n-2)}\tilde{\omega}_k$$

is a U-statistic. This enables us to directly use the theory of U-statistics to establish asymptotic property of $\widehat{\omega}_k$. Note that $\widehat{\omega}_k$ is a scaled version of $\tilde{\omega}_k$. They lead to the same result of feature ranking and screening.

In sum, we propose to rank all the candidate predictors $X_k, k = 1, \ldots, p$, according to $\widehat{\omega}_k$ from the largest to smallest. We then select the top ones as the active predictors. Later we will propose a thresholding rule for obtaining the cutoff value that separates the active and inactive predictors.

Before we turn to the theoretical properties of the proposed procedure, we will examine some simple settings to get more insight into our proposal. First, we consider a case where $K = 1$ and $\mathbf{x} \sim \text{N}_p(\mathbf{0}, \sigma^2\mathbf{I}_p)$ with unknown σ^2. Note that the normality assumption on $\mathbf{x}$ is not necessary and will be relaxed later, to derive the measure's properties. For ease of presentation, we write $\mathbf{x} = (\mathbf{x}_{\mathcal{A}}^{\text{T}}, \mathbf{x}_{\mathcal{I}}^{\text{T}})^{\text{T}}$, and define $\mathbf{b} = (b_1, \ldots, b_p)^{\text{T}} = (\boldsymbol{\beta}^{\text{T}}, \mathbf{0}^{\text{T}})^{\text{T}}$. It follows by a direct calculation that

$$\boldsymbol{\Omega}(y) = E\{\mathbf{x}F_0(y \mid \mathbf{b}^{\text{T}}\mathbf{x})\} = c(y)\mathbf{b},$$

where $c(y) = \|\mathbf{b}\|^{-1}\int_{-\infty}^{\infty} vF_0(y \mid v\|\mathbf{b}\|)\phi(v; 0, \sigma^2)\,dv$ with $\phi(v; 0, \sigma^2)$ being the density function of $\text{N}(0, \sigma^2)$ at v. Then $\omega_k = E\{\Omega_k^2(Y)\} = E\{c^2(Y)\}b_k^2$. If $E\{c^2(Y)\} > 0$, then

$$\max_{k\in\mathcal{I}} \omega_k < \min_{k\in\mathcal{A}} \omega_k, \tag{2.5}$$

and $\omega_k = 0$ if and only if $k \in \mathcal{I}$. This implies that the quantity ω_k may be used for feature screening in this setting.

2.3 Theoretical Properties

The property (2.5) allows us to perform feature ranking and feature screening. To ensure this property in general, we impose the following conditions. It is interesting to note that all the conditions are placed on the distribution of $\mathbf{x}$ only.

(C1) The following inequality condition holds uniformly for p:

$$\frac{K^2\lambda_{\max}\{\text{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{I}}^{\text{T}})\,\text{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{A}}^{\text{T}})\}}{\lambda_{\min}^2\{\text{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\text{T}})\}} < \frac{\min_{k\in\mathcal{A}}\omega_k}{\lambda_{\max}\{\boldsymbol{\Omega}_{\mathcal{A}}\}}, \tag{2.6}$$

where $\boldsymbol{\Omega}_{\mathcal{A}} = E\{\boldsymbol{\Omega}_{\mathcal{A}}(Y)\boldsymbol{\Omega}_{\mathcal{A}}^{\text{T}}(Y)\}$, $\boldsymbol{\Omega}_{\mathcal{A}}(y) = \{\Omega_1(y), \ldots, \Omega_{p_1}(y)\}^{\text{T}}$, and $\lambda_{\max}\{\mathbf{B}\}$ and $\lambda_{\min}\{\mathbf{B}\}$ denote the largest and smallest eigenvalues of a matrix $\mathbf{B}$, respectively. Note that $\lambda_{\min}(\mathbf{B})$ and $\lambda_{\max}(\mathbf{B})$ may depend on the dimension of $\mathbf{B}$. Throughout this article, when we say that "$a < b$ holds uniformly for p," it means that $\limsup_{p\to\infty}\{a(p) - b(p)\} < 0$.

(C2) The linearity condition:

$$E\{\mathbf{x} \mid \boldsymbol{\beta}^{\text{T}}\mathbf{x}_{\mathcal{A}}\} = \text{cov}(\mathbf{x}, \mathbf{x}_{\mathcal{A}}^{\text{T}})\boldsymbol{\beta}\{\text{cov}(\boldsymbol{\beta}^{\text{T}}\mathbf{x}_{\mathcal{A}})\}^{-1}\boldsymbol{\beta}^{\text{T}}\mathbf{x}_{\mathcal{A}}. \tag{2.7}$$

(C3) The moment condition: there exists a positive constant t_0 such that

$$\max_{1\le k\le p} E\{\exp(tX_k)\} < \infty \quad \text{for } 0 < t \le t_0.$$

Condition (C1) dictates the correlations among the predictors, and is the key assumption to ensure that the proposed screening procedure works properly. We make the following remarks about this condition. First, as the dimension K of $\boldsymbol{\beta}$ in (2.1) increases, the condition becomes more stringent. Therefore, a model with a small K is favored by our procedure. In many commonly used models, however, K is indeed small, as partially shown in Section 2.1. Second, for the left side of (2.6), the numerator measures the correlation between the active predictors $\mathbf{x}_{\mathcal{A}}$ and the inactive ones $\mathbf{x}_{\mathcal{I}}$, while the denominator measures the correlation among the active predictors themselves. When $\mathbf{x}_{\mathcal{A}}$ and $\mathbf{x}_{\mathcal{I}}$ are uncorrelated, (C1) holds automatically. For the proposed screening method to work well, this condition rules out the case in which there is strong collinearity between the active and inactive predictors, or among the active predictors themselves. This is very similar to condition 4 of the article by Fan and Lv (2008, p. 870). Third, the quantity $\min_{k\in\mathcal{A}}\omega_k$ on the right side of (2.6) reflects the signal strength of individual active predictors, which in turn controls the rate of probability error in selecting the active predictors. This aspect is similar to condition 3 of the article by Fan and Lv (2008, p. 870), which requires the contribution of an active predictor to be sufficiently large. Finally, we note that (2.6) is not scale invariant, since $\boldsymbol{\Sigma} = \text{cov}(\mathbf{x}, \mathbf{x}^{\text{T}})$ is not taken into account. This is similar to the linear SIS procedure of Fan and Lv (2008), which is based upon the covariance vector $\text{cov}(\mathbf{x}, Y)$ alone without the term $\boldsymbol{\Sigma}$. Fan and Lv (2008) imposed the concentration property [Fan and Lv 2008, equation (16) on p. 870] that implicitly requires the marginal variances of all predictors be of the same order. In our setup, we always marginally standardize all the predictors to have sample variance equal to 1.

Condition (C2) holds if $\mathbf{x}$ follows a normal or an elliptical distribution (Fang, Kotz, and Ng 1990). This condition was first proposed by Li (1991) and has been widely used in the dimension-reduction literature. It is remarkable though that Condition (C2) is itself weaker than both the normality and the elliptical symmetry conditions because we only require it to hold for the true value of $\boldsymbol{\beta}$. Furthermore, Hall and Li (1993) showed that the linearity condition holds asymptotically if the number of predictors p diverges while the dimension K remains fixed. For this reason, we view the linearity condition as a mild assumption in ultrahigh-dimensional regressions, where p is essentially very large and grows at a fast rate toward infinity.

Condition (C3) is concerned with the moments of the predictors. This condition holds for a variety of distributions, including the normal distribution and the distributions with bounded support. Compared with the usual conditions imposed in the feature screening literature, (C3) relaxed the normality assumption assumed by Fan and Lv (2008), in which both $\mathbf{x}$ and $Y \mid \mathbf{x}$ are assumed to be normally distributed.

Next we present the theoretical properties of the proposed screening measure. The proof is given in the Appendix. It is the main theoretical foundation for our feature screening procedure.

Theorem 1. Under Conditions (C1)–(C3), the following inequality holds uniformly for p:

$$\max_{k\in\mathcal{I}} \omega_k < \min_{k\in\mathcal{A}} \omega_k. \tag{2.8}$$

The following corollary reveals that the quantity ω_k is in fact a measure of the correlation between the marginal covariate X_k and the linear combinations $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$.

Corollary 1. Under the linearity condition (C2) and for $k = 1, \ldots, p$, $\omega_k = 0$ if and only if $\operatorname{cov}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}, X_k) = \mathbf{0}$.

Theorem 1 and Corollary 1 together offer more insights into the newly proposed utility measure ω_k. First, it is easy to see that, when X_k is independent of Y, $\omega_k = 0$. On the other hand, $k \in \mathcal{I}$ alone does not necessarily imply that $\omega_k = 0$. The quantity is zero only if X_k is uncorrelated with $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$. Theorem 1, however, ensures that ω_k of an inactive predictor is always smaller than ω_k of an active predictor, which is sufficient for the purpose of predictor ranking.

We next present the main theoretical result on feature ranking in terms of the utility measure $\widehat{\omega}_k$.

Theorem 2 (Consistency in ranking). In addition to the conditions in Theorem 1, we further assume that $p = o\{\exp(an)\}$ for any fixed $a > 0$. Then, for any $\varepsilon > 0$, there exists a sufficiently small constant $s_\varepsilon \in (0, 2/\varepsilon)$ such that

$$P\Big(\sup_{k=1,\ldots,p} |\widehat{\omega}_k - \omega_k| > \varepsilon\Big) \le 2p\exp\{n\log(1 - \varepsilon s_\varepsilon/2)/3\}.$$

In addition, if we write $\delta = \min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k$, then there exists a sufficiently small constant $s_{\delta/2} \in (0, 4/\delta)$ such that

$$P\Big(\max_{k\in\mathcal{I}} \widehat{\omega}_k < \min_{k\in\mathcal{A}} \widehat{\omega}_k\Big) \ge 1 - 4p\exp\big\{n\log\big(1 - \delta s_{\delta/2}/4\big)/3\big\}.$$

Note that $p = o\{\exp(an)\}$. Thus, the right side of the above equation approaches 1 with an exponential rate as $n \to \infty$. Theorem 2 justifies using $\widehat{\omega}_k$ to rank the predictors, and it establishes the consistency in ranking. That is, $\widehat{\omega}_k$ always ranks an active predictor above an inactive one in probability, and so guarantees a clear separation between the active and inactive predictors. Provided an ideal cutoff is available, this property would lead to consistency in selection in the ultrahigh-dimensional setup. Next we propose a thresholding rule to obtain a cutoff value to separate the active and inactive predictors.

2.4 Thresholding Rule

The thresholding rule is based upon a combination of a soft cutoff value obtained by adding artificial auxiliary variables to the data, and a hard cutoff that retains a fixed number of predictors after ranking.

The idea of introducing auxiliary variables for thresholding was first proposed by Luo, Stefanski, and Boos (2006) to tune the entry significance level in forward selection, and then extended by Wu, Boos, and Stefanski (2007) to control the false selection rate of forward regression in the linear model. We adopt this idea in our setup as follows. We independently and randomly generate d auxiliary variables $\mathbf{z} \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}_d)$ such that $\mathbf{z}$ is independent of both $\mathbf{x}$ and Y. The normality is not critical here, as we shall see later. Regard the $(p + d)$-dimensional vector $(\mathbf{x}^{\mathrm{T}}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$ as the predictors and Y as the response. We calculate ω_k for $k = 1, \ldots, p + d$. Since $\mathbf{z}$ is truly inactive by construction, we have $\min_{k\in\mathcal{A}}\omega_k > \max_{\ell=1,\ldots,d}\omega_{p+\ell}$ by Theorem 1, and given a random sample $\{(\mathbf{x}_i, \mathbf{z}_i, Y_i), i = 1, \ldots, n\}$, it holds in probability that $\min_{k\in\mathcal{A}}\widehat{\omega}_k > \max_{\ell=1,\ldots,d}\widehat{\omega}_{p+\ell}$ by Theorem 2. Define $C_d = \max_{\ell=1,\ldots,d}\widehat{\omega}_{p+\ell}$, which can be viewed as a benchmark that separates the active predictors from the inactive ones. This leads to the selection,

$$\widehat{\mathcal{A}}_1 = \{k : \widehat{\omega}_k > C_d\}. \tag{2.9}$$

We call (2.9) the soft thresholding selection.

The next theorem gives an upper bound on the probability of recruiting any inactive variables by the above soft thresholding selection. It can be viewed as an analogue of theorem 1 in the article of Fan, Samworth, and Wu (2009), while the exchangeability condition imposed in this theorem is similar in spirit to their condition (A1). This result shows how the soft thresholding rule performs.

Theorem 3. Let $r \in \mathbb{N}$, the set of natural numbers. We assume the exchangeability condition, that is, the inactive predictors $\{X_j, j \in \mathcal{I}\}$ and the auxiliary variables $\{Z_j, j = 1, \ldots, d\}$ are exchangeable in the sense that both the inactive and auxiliary variables are equally likely to be recruited by the soft thresholding procedure. Then

$$P(|\widehat{\mathcal{A}}_1 \cap \mathcal{I}| \ge r) \le \left(1 - \frac{r}{p + d}\right)^d,$$

where $|\cdot|$ denotes the cardinality of a set.

An issue of practical interest in soft thresholding is the choice of number of auxiliary variables d. Intuitively, a small d value may introduce much variability, whereas a large d value requires heavier computation. Empirically, we choose $d = p$, and our numerical experience has suggested that this choice works

quite well. Choosing an optimal d, however, is out of the scope of this article and is a potential direction for future research.

In addition to soft thresholding, we also consider a hard thresholding rule proposed by Fan and Lv (2008), which retains a fixed number of predictors with the largest N values of ω_k's; that is,

$$\widehat{\mathcal{A}}_2 = \{k : \widehat{\omega}_k > \widehat{\omega}_{(N)}\}, \tag{2.10}$$

where N is usually chosen to be $[n/\log n]$ and $\widehat{\omega}_{(N)}$ denotes the Nth largest value among all $\widehat{\omega}_k$'s.

In practice, the data determine whether the soft or hard thresholding comes into play. To better understand the two thresholding rules, we conducted a simulation study. The results are not reported here but in an earlier version of this article available at the authors' websites. We make the following observations from our simulation study. When the signal in the data is sparse (a small p_1), the hard thresholding rule often dominates the soft selection rule. On the other hand, when there are many active predictors (a large p_1), the soft thresholding becomes more dominant. While the hard thresholding is fully determined by the sample size, soft thresholding takes into account the effect of signals in the data, which is helpful when p_1 is relatively large. Consequently, we propose to combine the soft and hard thresholding, and construct the final active predictor index set as

$$\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_1 \cup \widehat{\mathcal{A}}_2, \tag{2.11}$$

where the union of the two sets is taken.

2.5 Iterative Feature Screening

An inherent issue with any feature screening procedure based on a marginal utility measure is that the method may miss those predictors which are marginally unrelated but jointly related to the response. To overcome this problem, we develop an iterative version of our proposed screening method. It is similar in spirit to the family of iterative SIS methods. However, unlike iterative SIS which breaks the correlation structure among predictors through the correlation between the residuals of the response and the remaining predictors, our method computes the correlation between the original response Y and the residual of the remaining $\mathbf{x}$. This is because the residual of Y is not available in a model-free context. However, we can compute the residual of $\mathbf{x}$, where the residual is defined as the projection of the remaining of $\mathbf{x}$ onto the orthogonal complement space of the predictors selected in the previous steps. More specifically, our iterative procedure is given as follows.

Step 1. We first apply our proposed screening procedure for $\mathbf{y}$ and $\mathbf{X}$, where $\mathbf{X}$ denotes the $n \times p$ data matrix that stacks n sample observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Suppose $p_{(1)}$ predictors are selected, where $p_{(1)} < N = [n/\log n]$. We denote the set of indices of the selected predictors by $\widehat{\mathcal{A}}_{(1)}$, and the associated $n \times p_{(1)}$ data matrix by $\mathbf{X}_{\widehat{\mathcal{A}}_{(1)}}$.

Step 2. Let $\widehat{\mathcal{I}}_{(1)}$ denote the complement of $\widehat{\mathcal{A}}_{(1)}$, and $\mathbf{X}_{\widehat{\mathcal{I}}_{(1)}}$ denote the remaining $n \times (p - p_{(1)})$ data matrix. Next, we define the predictor residual matrix

$$\mathbf{X}_r = \{\mathbf{I}_n - \mathbf{X}_{\widehat{\mathcal{A}}_1}(\mathbf{X}_{\widehat{\mathcal{A}}_1}^{\mathrm{T}}\mathbf{X}_{\widehat{\mathcal{A}}_1})^{-1}\mathbf{X}_{\widehat{\mathcal{A}}_1}^{\mathrm{T}}\}\mathbf{X}_{\widehat{\mathcal{I}}_{(1)}}.$$

Apply again our proposed screening procedure for $\mathbf{y}$ and $\mathbf{X}_r$. Suppose $p_{(2)}$ predictors are selected, and the resulting index set is denoted by $\widehat{\mathcal{A}}_{(2)}$. Update the total selected predictor set by $\widehat{\mathcal{A}}_{(1)} \cup \widehat{\mathcal{A}}_{(2)}$.

Step 3. Repeat Step 2 $M - 1$ times until the total selected number of predictors $p_{(1)} + \cdots + p_{(M)}$ exceeds the prespecified number $N = [n/\log n]$. The final selected predictor set is $\widehat{\mathcal{A}}_{(1)} \cup \cdots \cup \widehat{\mathcal{A}}_{(M)}$.

For the iterative procedure, we fix the number of total selected predictors $N = [n/\log n]$. In our simulations, we consider an $M = 2$ iterative procedure and choose $p_{(1)} = [N/2]$, which works well for our example. Some guidelines on selecting these parameters in an iterative feature screening procedure can be found in the work of Fan, Samworth, and Wu (2009).

3. NUMERICAL STUDIES

3.1 General Setup

In this section we assess the finite sample performance of the proposed method and compare it with existing competitors via Monte Carlo simulations. For brevity, we refer to our approach as *sure independent ranking and screening* (SIRS). Throughout, we set the sample size $n = 200$ and the total number of predictors $p = 2000$. We repeat each scenario 1000 times. For the soft thresholding, we set the number of auxiliary variables $d = p$. We generate the predictors $\mathbf{x}$ from a normal distribution with mean zero. Unless otherwise specified, we consider two covariance structures of $\mathbf{x}$: $\mathbf{\Sigma}_1 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.8^{|i-j|}$; and $\mathbf{\Sigma}_2 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ii} = 1$, $\sigma_{ij} = 0.4$ if both $i, j \in \mathcal{A}$ or $i, j \in \mathcal{I}$, and $\sigma_{ij} = 0.1$ otherwise.

To evaluate the performance of the proposed method, we employ mainly two criteria. The first criterion measures accuracy of ranking the predictors (with no thresholding). For that purpose, we record the minimum number of predictors in a ranking that is required to ensure the inclusion of all the truly active predictors. We denote this number by $\mathcal{R}$. The second criterion focuses on accuracy of feature screening when applying the proposed thresholding rule to the ranked predictors. Unlike feature selection, where it is important to simultaneously achieve both a high true positive and a low false positive, feature screening is more concerned with retaining all the truly active predictors. This is because screening usually serves as a preliminary massive reduction step, and is often followed by a conventional feature selection for further refinement. For that reason, we record the proportion that all the truly active predictors are correctly identified after thresholding in 1000 repetitions, and denote this proportion by $\mathcal{S}$. A ranking and screening procedure is deemed competent if it yields an $\mathcal{R}$ value that is close to the true number of active predictors p_1, and an $\mathcal{S}$ value that is close to 1.

3.2 Linear Models

A large number of well-known variable screening and selection approaches, such as linear SIS (Fan and Lv 2008), Lasso (Tibshirani 1996), stepwise regression, and forward regression (Wang 2009) have been proposed in the literature. We thus begin with a class of linear models. Our simulations reveal the following two key observations. First, when the model is indeed linear homoscedastic with a normal error, SIRS has a comparable performance to the model-based methods which correctly specify the model. Second, when the true model deviates from the imposed model assumptions (e.g., the variance is

heteroscedastic or the error distribution is heavily tailed), our method clearly outperforms the model-based methods.

Example 1. In the first example, we consider a classical linear model with varying squared multiple correlation coefficient R^2, variance structure, and error distribution:

$$Y = c\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \sigma\varepsilon, \tag{3.1}$$

where $\boldsymbol{\beta} = (1, 0.8, 0.6, 0.4, 0.2, 0, \ldots, 0)^{\mathrm{T}}$ takes grid values. We consider two predictor covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ as specified in Section 3.1. We also examine two variance structures: $\sigma = \sigma_1$, a constant, and $\sigma = \sigma_2 = \exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})$, with $\boldsymbol{\gamma} = (0, \ldots, 0, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$ and ones appear in the 20th, 21st, and 22nd positions. Thus, σ_1 leads to a constant variance model, and we choose $\sigma_1^2 = 6.83$ for $\boldsymbol{\Sigma}_1$, and $\sigma_1^2 = 4.92$ for $\boldsymbol{\Sigma}_2$, which equals $\mathrm{var}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ at the population level for the corresponding $\mathbf{x}$. σ_2 leads to a nonconstant variance model. We consider two error ε distributions, a standard normal N(0, 1), and a t-distribution with one degree of freedom that has a heavy tail. We vary the constant c in front of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ to control the signal-to-noise ratio. For the constant variance model σ_1, we choose $c = 0.5$, 1, and 2, with the corresponding $R^2 = 20\%, 50\%$, and 80%, respectively. For the nonconstant variance model σ_2, R^2 are all very small ($<0.01\%$).

We first evaluate our proposed utility measure in terms of accuracy in ranking the predictors. We also compare our method (SIRS) with another ranking procedure, linear SIS of Fan and Lv (2008). Table 1 reports the median of the $\mathcal{R}$ values. For $\sigma = \sigma_1$, the number of truly actives $p_1 = 5$ and for $\sigma = \sigma_2$, $p_1 = 8$. It is seen that, when the model is linear, homoscedastic (σ_1), and the error follows a standard normal distribution N(0, 1), linear SIS performs the best, with the $\mathcal{R}$ measure being very close to p_1. However, the method breaks down for the heteroscedastic variance (σ_2) or the heavy-tailed error distribution (t_1). By contrast, our proposed procedure is comparable to linear SIS for the homoscedastic normal error, but is consistently superior with either the heteroscedastic variance or the heavy-tailed error distribution. Notably, our screening measure uses only the ranks of the observed response values, which partly explains why our method performs well for a heavy-tailed error (t_1). In addition, we observe that our method performs well across a wide range of signal-to-noise ratios (σ_1 with varying c), and the results for $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are similar.

Next we evaluate our feature screening method with the proposed thresholding rule (2.11). We also compare with some commonly used and linear-model-based feature selection approaches, including linear SIS, Lasso, stepwise regression, and forward regression. For stepwise regression, we use 0.05 as the inclusion probability and 0.10 as the exclusion probability. For Lasso and forward regression, we find that the BIC criterion proposed in the literature does not yield a satisfactory performance in our setup. Therefore, for those two methods, as well as linear SIS, we choose the same number of predictors as our proposed screening using the thresholding rule (2.11). The proportion $\mathcal{S}$ is reported in Table 2, which indicates that the SIRS performs competently across different scenarios, with the proportion $\mathcal{S}$ close to 1. As expected, SIRS outperforms other methods for error being t-distribution with one degree of freedom (i.e., the Cauchy distribution) since other methods require finite error variance. It is also expected that all the selection methods except for SIRS cannot identify the active predictors in the variance of random error. Thus, when the error is heteroscedastic, the proportions shown in Table 2 for all methods except SIRS are almost zero. To make favorable comparison toward the model-based methods when the error is heteroscedastic, we further summarize the proportion that all active predictors (X_1–X_5) contained in the regression function are correctly identified out of 1000 replications in Table 3, from which it can be seen that SIRS performs very well, while all other methods perform unsatisfactorily. This is because the random error in this case contains some very extreme values (outliers), and the SIRS is robust to the outliers because it only uses the ranks of the observed response values.

Example 2. In this example, we continue to employ the linear model (3.1). In addition, we set $\sigma = 1$, $c = 1$, and $\boldsymbol{\beta} = (1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$, so that there are $p_1 = 3$ truly active predictors. What differs in this example is that we consider a more challenging covariance structure for the normally distributed $\mathbf{x}$ where $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}_3 = (\sigma_{ij})_{p\times p}$ with entries $\sigma_{ii} = 1, i = 1, \ldots, p$, and $\sigma_{ij} = 0.4, i \neq j$. We note that condition (C1) is not satisfied in this setup. In addition, we generate the error ε from a t distribution with 1, 2, 3, and 30 degrees of freedom. We remark that t_1 is the Cauchy distribution, t_1 and t_2 have infinite variance, t_3 has finite variance, and t_{30} is almost indistinguishable from a standard normal distribution. As such we

Table 1. The ranking criterion $\mathcal{R}$ for Example 1—minimum number of predictors required to ensure the inclusion of all the truly active predictors. The numbers reported are the median of $\mathcal{R}$ out of 1000 replications

			$\boldsymbol{\Sigma}_1$			$\boldsymbol{\Sigma}_2$		
ε	σ	Method	$c = 0.5$	$c = 1$	$c = 2$	$c = 0.5$	$c = 1$	$c = 2$
N(0, 1)	σ_1	SIRS	5	5	5	7	5	5
		SIS	5	5	5	6	5	5
	σ_2	SIRS	9	11	18	8	9	8
		SIS	1739	1735	1646	1571	1447	1210
t_1-dist	σ_1	SIRS	5	5	5	5	5	5
		SIS	1358	566	31	1608	1257	337
	σ_2	SIRS	10	9	12	10	9	9
		SIS	1735	1732	1757	1687	1678	1666

Table 2. The selection criterion $\mathcal{S}$ for Example 1—proportion that all the truly active predictors (X_1–X_5 for $\sigma = \sigma_1$ and X_1–X_5, X_{20}–X_{22} for $\sigma = \sigma_2$) are correctly identified out of 1000 replications. Reported are our proposal (SIRS), linear SIS, Lasso, stepwise regression (Step), and forward regression (FR)

			$\boldsymbol{\Sigma}_1$			$\boldsymbol{\Sigma}_2$		
ε	σ	Method	$c = 0.5$	$c = 1$	$c = 2$	$c = 0.5$	$c = 1$	$c = 2$
N(0, 1)	σ_1	SIRS	0.953	1.000	1.000	0.778	0.998	1.000
		SIS	0.965	1.000	1.000	0.832	0.999	1.000
		Lasso	0.032	0.230	0.618	0.197	0.576	0.926
		Step	0.001	0.007	0.066	0.002	0.034	0.306
		FR	0.015	0.111	0.382	0.000	0.015	0.307
	σ_2	SIRS	0.993	0.989	0.814	0.918	0.900	0.891
		SIS	0.000	0.000	0.001	0.010	0.033	0.058
		Lasso	0.000	0.000	0.000	0.000	0.000	0.001
		Step	0.000	0.000	0.000	0.000	0.000	0.000
		FR	0.000	0.000	0.000	0.000	0.000	0.000
t_1-dist	σ_1	SIRS	0.996	1.000	1.000	0.883	0.992	1.000
		SIS	0.052	0.231	0.513	0.014	0.118	0.357
		Lasso	0.002	0.004	0.036	0.002	0.025	0.080
		Step	0.000	0.000	0.001	0.000	0.000	0.001
		FR	0.000	0.007	0.016	0.000	0.000	0.003
	σ_2	SIRS	0.932	0.990	0.974	0.844	0.895	0.887
		SIS	0.000	0.000	0.000	0.004	0.005	0.006
		Lasso	0.000	0.000	0.000	0.000	0.000	0.000
		Step	0.000	0.000	0.000	0.000	0.000	0.000
		FR	0.000	0.000	0.000	0.000	0.000	0.000

have a model that gradually approaches a normal distribution when the degrees of freedom increase.

Table 4 reports the ranking criterion $\mathcal{R}$ and Table 5 reports the selection criterion $\mathcal{S}$. Again we observe a qualitative pattern similar to Example 1. That is, when the error is close to normal (t_{30}), the model-based SIS, Lasso, stepwise regression, and forward regression perform very well, and our model-free procedure yields a comparable outcome. When the error deviates from a normal distribution (t with decreasing degrees of freedom), however, the performance of all the model-based alternatives quickly deteriorates, while our method continues to perform well.

As shown above, when the model is correctly specified (e.g., Example 1 with $c = 0.5$ and a normal error), or sufficiently close to the true model (as seen in the trend of Example 2 as the error degree of freedom increases), the model-based solution is more competent than our model-free solution. This is not surprising because the former is equipped with additional model information. In practice, which solution to resort to depends on the amount of knowledge and confidence an analyst has about the model. Our approach can be used in conjunction with, rather than as an alterative to, many model-based feature screening and selection solutions.

Table 3. The selection criterion $\mathcal{S}^*$ for Example 1 with heteroscedastic error—proportion that all active predictors (X_1–X_5) contained in the regression function are correctly identified out of 1000 replications. Reported are linear SIS, Lasso, stepwise regression (Step), and forward regression (FR)

		$\boldsymbol{\Sigma}_1$			$\boldsymbol{\Sigma}_2$		
ε	Method	$c = 0.5$	$c = 1$	$c = 2$	$c = 0.5$	$c = 1$	$c = 2$
N(0, 1)	SIRS	0.993	0.999	1.000	0.931	0.970	0.994
	SIS	0.000	0.004	0.012	0.013	0.042	0.091
	Lasso	0.000	0.000	0.000	0.000	0.000	0.008
	Step	0.000	0.000	0.000	0.000	0.000	0.000
	FR	0.000	0.000	0.000	0.000	0.000	0.000
t_1-dist	SIRS	0.932	0.990	1.000	0.848	0.944	0.980
	SIS	0.000	0.000	0.000	0.004	0.005	0.007
	Lasso	0.000	0.000	0.000	0.000	0.000	0.000
	Step	0.000	0.000	0.000	0.000	0.000	0.000
	FR	0.000	0.000	0.000	0.000	0.000	0.000

Table 4. The ranking criterion $\mathcal{R}$ for Example 2. The quintuplet in each parenthesis consists of the minimum, the first quartile, median, third quartile, and maximum value of $\mathcal{R}$ out of 1000 data replications

ε	SIRS					SIS				
t_1-dist	(3	4	9	28	1368)	(4	623	1126	1593	1999)
t_2-dist	(3	3	3	5	680)	(3	3	7	36	1935)
t_3-dist	(3	3	3	3	210)	(3	3	3	4	650)
t_{30}-dist	(3	3	3	3	30)	(3	3	3	3	7)

3.3 Nonlinear Models and Discrete Response

Our next goal is to demonstrate that the proposed model-free approach offers a useful and robust procedure in the sense that it works for a large variety of different models when there is little knowledge about the underlying true model. Toward that end, we consider two sets of examples that cover a wide range of commonly used parametric and semiparametric models. The first set involves a continuous response, including the transformation model, the multiple-index model, and the heteroscedastic model.

Example 3. The response is continuous. The error ε follows a standard normal distribution. $\boldsymbol{\beta} = (2 - U_1, \ldots, 2 - U_{p_1}, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_1 = (2 - U_1, \ldots, 2 - U_{p_1/2}, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0, \ldots, 0, 2 + U_{p_1/2+1}, \ldots, 2 + U_{p_1}, 0, \ldots, 0)^{\mathrm{T}}$, and U_k's follow a uniform distribution on $[0, 1]$. We vary the number of active predictors p_1 to reflect different sparsity levels. The predictor $\mathbf{x}$ follows a mean-zero normal distribution with two covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ as given in Section 3.1.

3.a. A transformation model: $Y = \exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2 + \varepsilon\}$.

3.b. A multiple-index model: $Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + \exp\{\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}\} + \varepsilon$.

3.c. A heteroscedastic model: $Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + \exp\{(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}) + \varepsilon\}$.

Table 6 reports the ranking criterion $\mathcal{R}$ and Table 7 reports the selection proportion criterion $\mathcal{S}$ after applying the thresholding rule (2.11) to the ranked predictors. For a wide range of models under investigation, $\mathcal{R}$ is often equal or close to the actual number of truly active predictors p_1, whereas $\mathcal{S}$ is equal or close to 1, indicating a very high accuracy in both ranking and selection. In addition, our method clearly outperforms the alternative approaches which assume the linear homoscedastic model while the true models are not linear homoscedastic in this example.

We have also examined a set of models with a discrete response, including the logistic model, the probit model, the Poisson log-linear model, and the proportional hazards model (with a binary censoring indicator). Due to the space limitation, we only reported those results in an earlier version of this article. Again, our extensive simulations show that the SIRS performs very well for the variety of discrete response models we have examined.

Table 5. The selection criterion $\mathcal{S}$ for Example 2. The caption is the same as Table 2

ε	SIRS	SIS	Lasso	Step	FR
t_1-dist	0.961	0.076	0.027	0.002	0.004
t_2-dist	0.997	0.913	0.849	0.640	0.647
t_3-dist	0.998	0.995	0.995	0.982	0.987
t_{30}-dist	1.000	1.000	1.000	1.000	1.000

3.4 Iterative Screening

We next briefly examine the proposed iterative version of our marginal screening approach. The example is based upon a configuration in the article by Fan and Lv (2008).

Example 4. We employ the linear model (3.1), with $\boldsymbol{\beta} = (5, 5, 5, -15\rho^{1/2}, 0, \ldots, 0)^{\mathrm{T}}$, $c = 1$, $\sigma = 1$, and ε follows a standard normal distribution. We draw $\mathbf{x}$ from a mean-zero normal population with the covariance $\boldsymbol{\Sigma}_4 = (\sigma_{ij})_{p\times p}$ with entries $\sigma_{ii} = 1$, for $i = 1, \ldots, p$, $\sigma_{i4} = \sigma_{4i} = \rho^{1/2}$ for $i \neq 4$, and $\sigma_{ij} = \rho$, for $i \neq j$, $i \neq 4$, and $j \neq 4$. That is, all predictors except for X_4 are equally correlated with correlation coefficient ρ, while X_4 has correlation $\rho^{1/2}$ with all other $p - 1$ predictors. By design X_4 is independent of Y, so that our method cannot pick it up except by chance, whereas X_4 is indeed an active predictor when $\rho \neq 0$. We also vary the value of ρ to be 0, 0.1, 0.5, and 0.9, with a larger ρ yielding a higher collinearity.

We compare both the non-iterative and the iterative versions of our screening method. For the iterative procedure, we choose $M = 2$ iterations with $p_{(1)} = [N/2]$ and $N = [n/\log(n)]$. This simple choice performs very well in this example. Table 8 reports the proportion criterion $\mathcal{S}$, where the iterative procedure dramatically improves over its non-iterative counterpart.

3.5 A Real Data Analysis

As an illustration, we apply the proposed screening method to the analysis of microarray diffuse large-B-cell lymphoma (DLBCL) data of Rosenwald et al. (2002). Given that DLBCL is the most common type of lymphoma in adults and has only about 35 to 40 percent survival rate after the standard chemotherapy, there has been continuous interest to understand the genetic factors that influence the survival outcome. The outcome in the study was the survival time of $n = 240$ DLBCL patients after chemotherapy. Measurements of $p = 7399$ genes obtained from cDNA microarrays for each individual patient were the predictors. Given such a large number of predictors and small sample size, feature screening seems a necessary initial step as a prelude to any other sophisticated statistical modeling that does not cope well with such high dimensionality.

All predictors are standardized to have mean zero and variance 1. We form the bivariate response consisting of the observed survival time and the censoring indicator. We use a data split of Li and Luan (2005) and Lu and Li (2008), which divides the data into a training set with $n_1 = 160$ patients and a testing set with 80 patients. We apply the proposed screening method to the training data. Among 200 trials of the thresholding rule (2.11), 196 times the hard thresholding rule dominates. Therefore, we choose $[n_1/\log(n_1)] = 31$ genes in our final set. This result seems to agree with the analysis of this

Table 6. The ranking criterion $\mathcal{R}$ for Example 3. The caption is the same as Table 4

p_1	Model	Method	$\boldsymbol{\Sigma}_1$					$\boldsymbol{\Sigma}_2$				
4	3.a.	SIRS	(4	4	4	4	5)	(4	4	4	4	4)
		SIS	(4	4	4	6	690)	(4	4	4	12	1808)
	3.b.	SIRS	(4	4	4	4	5)	(4	4	4	4	4)
		SIS	(4	4	6	12	1962)	(4	4	6	60	1996)
	3.c.	SIRS	(4	4	4	4	5)	(4	4	4	4	4)
		SIS	(4	5	7	23	1739)	(4	4	25	207	1998)
8	3.a.	SIRS	(8	8	8	8	10)	(8	8	8	8	8)
		SIS	(8	25	78	214	1784)	(8	48	177	518	2000)
	3.b.	SIRS	(8	8	8	8	11)	(8	8	8	8	8)
		SIS	(8	147	458	1061	1997)	(8	99	349	825	1981)
	3.c.	SIRS	(8	8	8	8	10)	(8	8	8	8	8)
		SIS	(9	171	496	1097	1999)	(8	113	398	896	1988)
16	3.a.	SIRS	(16	16	16	16	22)	(16	16	16	16	16)
		SIS	(29	463	845	1358	2000)	(18	456	881	1310	2000)
	3.b.	SIRS	(16	16	17	18	34)	(16	16	16	16	16)
		SIS	(35	1207	1676	1881	2000)	(25	559	1019	1517	1999)
	3.c.	SIRS	(16	16	17	18	34)	(16	16	16	16	16)
		SIS	(70	1286	1705	1890	2000)	(20	560	1047	1500	2000)

same dataset in the literature: only a small number of genes are relevant, and according to our simulations, the hard thresholding is more dominant in this scenario. Based on those selected genes, we fit a Cox proportional hazards model. We evaluate the prediction performance of this model following the approach of Li and Luan (2005) and Lu and Li (2008). That is, we apply the screening approach and fit a Cox model for the training data. We then compute the risk scores for the testing data and divide it to a low-risk group and a high-risk group, where the cutoff value is determined by the median of the estimated scores from the training set. Figure 1(a) shows the Kaplan–Meier estimate of survival curves for the two risk groups of patients in the testing data. The two curves are well separated, with the log-rank test yielding a p-value equal to 0.0025, indicating a good prediction of the fitted model.

Both Li and Luan (2005) and Lu and Li (2008) used a univariate Cox model to screen the predictors. Applying their screening approach, while retaining as many as 31 genes, yields a subset of genes among which 12 overlap with the ones identified by our method. As a simple comparison, we also fit a Cox model based on the genes selected by their marginal screening method, and evaluate its prediction performance. Figure 1(b) is constructed in the same fashion as Figure 1(a) except that the genes are selected by the univariate Cox model. The figure shows that the two curves are less well separated, with the p-value of the log-rank test equal to 0.1489, suggesting an inferior predictive performance compared to our method.

We remark that, without any information about the appropriate model form for this dataset, our model-free screening result seems more reliable compared to a model-based procedure. We also note that choosing the Cox model after screen-

Table 7. The selection criterion $\mathcal{S}$ for Example 3. The caption is the same as Table 2

Model	Method	$\boldsymbol{\Sigma}_1$			$\boldsymbol{\Sigma}_2$		
		$p_1 = 4$	$p_1 = 8$	$p_1 = 16$	$p_1 = 4$	$p_1 = 8$	$p_1 = 16$
3.a.	SIRS	1.000	1.000	1.000	1.000	1.000	1.000
	SIS	0.963	0.330	0.002	0.878	0.310	0.034
	Lasso	0.118	0.000	0.000	0.475	0.003	0.000
	Step	0.008	0.000	0.000	0.014	0.000	0.000
	FR	0.035	0.000	0.000	0.004	0.000	0.000
3.b.	SIRS	1.000	1.000	1.000	1.000	1.000	1.000
	SIS	0.868	0.084	0.001	0.741	0.191	0.025
	Lasso	0.082	0.000	0.000	0.247	0.002	0.000
	Step	0.004	0.000	0.000	0.043	0.000	0.000
	FR	0.058	0.000	0.000	0.031	0.000	0.000
3.c.	SIRS	1.000	1.000	1.000	1.000	1.000	1.000
	SIS	0.810	0.065	0.000	0.603	0.169	0.024
	Lasso	0.041	0.000	0.000	0.151	0.000	0.000
	Step	0.003	0.000	0.000	0.011	0.000	0.000
	FR	0.028	0.000	0.000	0.006	0.000	0.000

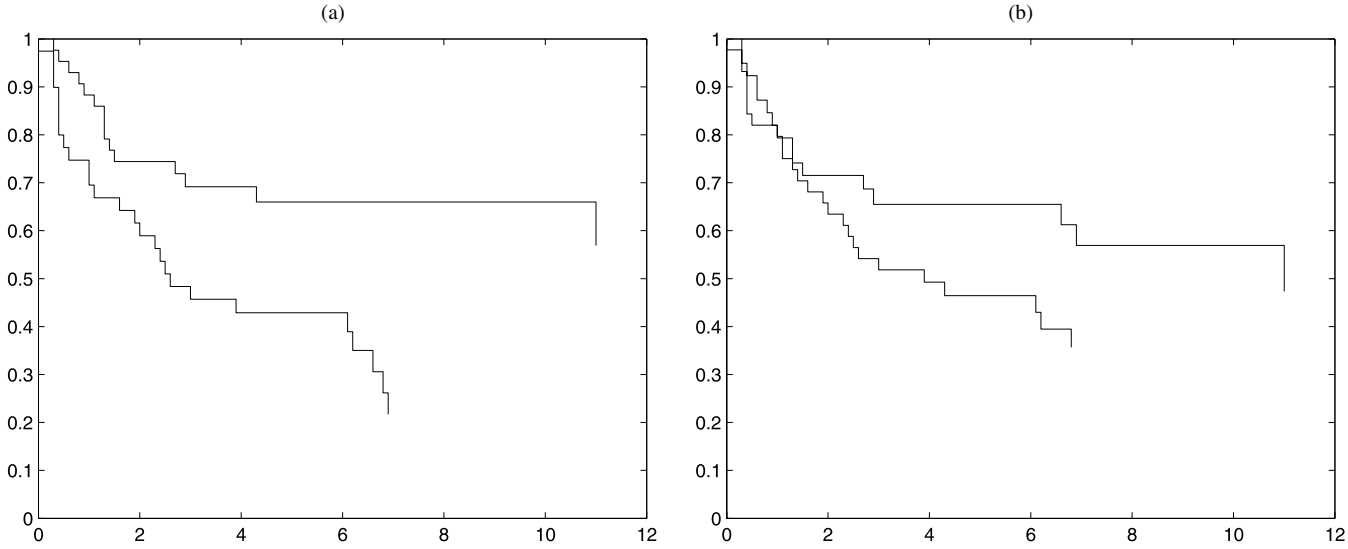

Figure 1. The Kaplan–Meier estimate of survival curves for the two risk groups in the testing data. Panel (a) is based on the proposed feature screening, and (b) is based on the univariate Cox model screening.

ing only serves as a simple illustration in this example. More refined model building and selection could be employed after feature screening, while the model-free nature of our screening method grants full flexibility in subsequent modeling.

APPENDIX: PROOFS

Proof of Theorem 1

Without loss of generality, we assume that the basis matrix $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ satisfies $\boldsymbol{\beta}^{\mathrm{T}} \operatorname{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta} = \mathbf{I}_K$, where $\mathbf{I}_K$ is a $K \times K$ identity matrix. In this case, the linearity condition (2.7) is simplified as $E(X_k \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}) = \operatorname{cov}(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$. For ease of presentation, we denote the matrix $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ by $\mathbf{v}^2$ for a vector $\mathbf{v}$.

Consider the left side of (2.8). Because $\mathbf{x}$ is independent of Y given $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$ and $\widetilde{Y}$ is an independent copy of Y, it follows that $\mathbf{x}$ is independent of Y and $\widetilde{Y}$ given $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$. This, together with the simplified linearity condition and the law of iterated expectations, yields that

$$\begin{aligned} &E\{X_k \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\} \\ &\quad = E\big[E\{X_k \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}\} \mid \widetilde{Y}\big] \\ &\quad = E\big[E(X_k \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}) E\{\mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}\} \mid \widetilde{Y}\big] \\ &\quad = \operatorname{cov}(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}) E\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}. \end{aligned} \tag{A.1}$$

Then one can obtain that

$$\begin{aligned} &\max_{k \in \mathcal{I}} E\big\{E^2(X_k \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y})\big\} \\ &\quad = \max_{k \in \mathcal{I}} \big(\operatorname{cov}(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta} E\big[E^2\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}\big] \\ &\qquad \times \boldsymbol{\beta}^{\mathrm{T}} \operatorname{cov}(\mathbf{x}_{\mathcal{A}}, X_k)\big) \\ &\quad \le \lambda_{\max}\big(E\big[E^2\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}\big]\big) \\ &\qquad \times \max_{k \in \mathcal{I}} \{\operatorname{cov}(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}} \operatorname{cov}(\mathbf{x}_{\mathcal{A}}, X_k^{\mathrm{T}})\}, \end{aligned} \tag{A.2}$$

where the first equality follows from (C2). Then it is straightforward to verify that

$$\begin{aligned} &\lambda_{\max}\big(E\big[E^2\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}\big]\big) \\ &\quad \le \sum_{j=1}^{K} E\big[E^2\{\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}\big] \\ &\quad \le \sum_{j=1}^{K} \lambda_{\max}\big(\operatorname{cov}^{-1/2}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}) E\big[E^2\{\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}\big] \\ &\qquad \times \operatorname{cov}^{-1/2}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\big) \\ &\quad \le K\lambda_{\max}\{\operatorname{cov}^{-1}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\} \\ &\qquad \times \lambda_{\max}\big(E\big[E^2\{\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}\big]\big) \\ &\quad = K\lambda_{\max}\big(E\big[E^2\{\mathbf{x}_{\mathcal{A}} \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\}\big]\big) \\ &\qquad / \lambda_{\min}\{\operatorname{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\}. \end{aligned} \tag{A.3}$$

Here the second inequality follows because $\boldsymbol{\beta}^{\mathrm{T}} \operatorname{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta} = \mathbf{I}_K$, and the third inequality holds due to the fact that $\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{B}\mathbf{C}) \le \lambda_{\max}(\mathbf{B})\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{C})$ for any matrix $\mathbf{B} \ge 0$. After some algebra, we have

$$\begin{aligned} &\max_{k \in \mathcal{I}} \{\operatorname{cov}(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}} \operatorname{cov}(\mathbf{x}_{\mathcal{A}}, X_k)\} \\ &\quad = \sum_{j=1}^{K} \max_{k \in \mathcal{I}} \{\operatorname{cov}(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}, X_k) \operatorname{cov}(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}_j\} \\ &\quad \le \sum_{j=1}^{K} \{\boldsymbol{\beta}_j^{\mathrm{T}} \operatorname{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{I}}^{\mathrm{T}}) \operatorname{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}_j\} \\ &\quad \le K\lambda_{\max}\{\operatorname{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{I}}^{\mathrm{T}}) \operatorname{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\} \\ &\qquad / \lambda_{\min}\{\operatorname{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\}. \end{aligned} \tag{A.4}$$

Table 8. The selection criterion $\mathcal{S}$ for Example 4—proportion that all the truly active predictors are correctly identified out of 1000 replications. ISIRS denotes the iterative version of the proposed SIRS method

Method	$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
ISIRS	0.925	1.000	1.000	0.940
SIRS	1.000	0.005	0.000	0.000

Then Condition (C1), together with (A.2), (A.3), and (A.4), entails (2.8).

Proof of Corollary 1

It follows from the definition in (2.4) that $\omega_k = 0$ is equivalent to $E\{X_k \mathbf{1}(Y < y)\} = 0$ for any $y \in \Psi_y$. Because Y relates to $\mathbf{x}$ only through linear combinations $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$, it follows that there exists some $y \in \Psi_y$ such that $E\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}\mathbf{1}(Y < y)\} \neq \mathbf{0}$. Consequently, (A.1) implies that $E\{X_k \mathbf{1}(Y < y)\} = 0$ if and only if $\operatorname{cov}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}, X_k) = \mathbf{0}$, which completes the proof of Corollary 1.

Proof of Theorem 2

To enhance readability, we divide the proof into two main steps.

Step 1. We first show that, under condition (C3),

$$P\Big(\sup_{k=1,\ldots,p} |\widehat{\omega}_k - \omega_k| > \varepsilon\Big) \le 2p\exp\{n\log(1-\varepsilon s_\varepsilon/2)/3\}. \tag{A.5}$$

Note that $\hat{\omega}_k$ can be expressed as follows:

$$\begin{aligned}\widehat{\omega}_k &= \frac{2}{n(n-1)(n-2)}\sum_{j<i<l}^{n}\big\{X_{jk}X_{ik}\mathbf{1}(Y_j < Y_l)\mathbf{1}(Y_i < Y_l)\\ &\quad + X_{lk}X_{ik}\mathbf{1}(Y_l < Y_j)\mathbf{1}(Y_i < Y_j)\\ &\quad + X_{jk}X_{lk}\mathbf{1}(Y_j < Y_i)\mathbf{1}(Y_l < Y_i)\big\}\\ &\stackrel{\text{def}}{=} \frac{6}{n(n-1)(n-2)}\sum_{j<i<l}^{n} h(X_{jk}, Y_j; X_{ik}, Y_i; X_{lk}, Y_l).\end{aligned}$$

Thus, $\widehat{\omega}_k$ is a standard U-statistic. With Markov's inequality, we can obtain that, for any $0 < t < s_0 k^*$, where $k^* = [n/3]$,

$$P(\widehat{\omega}_k - \omega_k \ge \varepsilon) \le \exp\{-t\varepsilon\}\exp\{-t\omega_k\}E[\exp\{t\widehat{\omega}_k\}].$$

Through 5.1.6 in the book by Serfling (1980), the U-statistic $\widehat{\omega}_k$ can be represented as an average of averages of independent and identically distributed random variables; that is, $\widehat{\omega}_k = (n!)^{-1}\sum_{n!} w(X_{1k}, Y_1; \ldots, X_{nk}, Y_n)$, where each $w(X_{1k}, Y_1; \ldots, X_{nk}, Y_n)$ is an average of $k^* = [n/3]$ independent and identically distributed random variables, and $\sum_{n!}$ denotes summation over $n!$ permutations $i_1, \ldots, i_n$ of $(1, \ldots, n)$. We denote that $\psi_h(s) = E[\exp\{sh(X_{jk}, Y_j; X_{ik}, Y_i; X_{lk}, Y_l)\}]$ for $0 < s < s_0$. Since the exponential function is convex, it follows by Jensen's inequality that

$$\begin{aligned}E[\exp\{t\widehat{\omega}_k\}] &= E\bigg[\exp\bigg\{t(n!)^{-1}\sum_{n!} w(X_{1k}, Y_1; \ldots, X_{nk}, Y_n)\bigg\}\bigg]\\ &\le (n!)^{-1}\sum_{n!} E\big[\exp\{tw(X_{1k}, Y_1; \ldots, X_{nk}, Y_n)\}\big]\\ &= \psi_h^{k^*}(t/k^*).\end{aligned}$$

Combining the above two results, we obtain that

$$\begin{aligned}P(\widehat{\omega}_k - \omega_k \ge \varepsilon) &\le \exp\{-t\varepsilon\}[\exp\{-t\omega_k/k^*\}\psi_h(t/k^*)]^{k^*}\\ &= [\exp\{-s\varepsilon\}\exp\{-s\omega_k\}\psi_h(s)]^{k^*},\end{aligned} \tag{A.6}$$

where $s = t/k^*$. Note that $E\{h(X_{jk}, Y_j; X_{ik}, Y_i; X_{lk}, Y_l)\} = \omega_k$, and with Taylor expansion, $\exp\{sY\} = 1 + sY + s^2 Z/2$ for any generic random variable Y, where $0 < Z < Y^2\exp\{s_1 Y\}$, and s_1 is a constant between 0 and s. It follows that

$$\begin{aligned}\exp\{-s\omega_k\}\psi_h(s) \le 1 + s^2\big[E\{h^4(X_{jk}, Y_j; X_{ik}, Y_i; X_{lk}, Y_l)\}\\ \times E\exp\{2s_1(h-\omega_k)\}\big]^{1/2}/2.\end{aligned}$$

By invoking Condition (C3), it follows that there exists a constant C (independent of n and p) such that $\max_{1\le k\le p}\exp\{-s\omega_k\}\psi_h(s) \le 1 + Cs^2$; that is,

$$\max_{1\le k\le p}\exp\{-s\omega_k\}\psi_h(s) = 1 + O(s^2).$$

Recall that $0 < s = t/k^* < s_0$. For a sufficiently small s, which can be achieved by selecting a sufficiently small t, we have that $\exp(-s\varepsilon) = 1 - \varepsilon s + O(s^2)$ and therefore,

$$\max_{1\le k\le p}[\exp(-s\varepsilon)\exp(-s\omega_k)\psi_h(s)] \le 1 - \varepsilon s/2. \tag{A.7}$$

Combining the results (A.6) and (A.7), we show that, for any $\varepsilon > 0$, there exists a sufficiently small s_ε such that $\max_{1\le k\le p}\{P(\widehat{\omega}_k - \omega_k \ge \varepsilon)\} \le (1 - \varepsilon s_\varepsilon/2)^{n/3}$. Here we use the notation s_ε to emphasize s depending on ε. Similarly, we can prove that $\max_{1\le k\le p}\{P(\widehat{\omega}_k - \omega_k \le -\varepsilon)\} \le (1 - \varepsilon s_\varepsilon/2)^{n/3}$. Therefore,

$$P\Big(\sup_{k=1,\ldots,p} |\widehat{\omega}_k - \omega_k| > \varepsilon\Big) \le 2p\exp\{n\log(1-\varepsilon s_\varepsilon/2)/3\}. \tag{A.8}$$

This completes the proof of Step 1.

Step 2. We next show that

$$P\Big(\max_{k\in\mathcal{I}}\widehat{\omega}_k < \min_{k\in\mathcal{A}}\widehat{\omega}_k\Big) \ge 1 - 4p\exp\big\{n\log\big(1-\delta s_{\delta/2}/4\big)/3\big\}. \tag{A.9}$$

Recall the assumption that $\delta = \min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k > 0$. Thus,

$$\begin{aligned}&P\Big(\min_{k\in\mathcal{A}}\widehat{\omega}_k \le \max_{k\in\mathcal{I}}\widehat{\omega}_k\Big)\\ &= P\Big(\min_{k\in\mathcal{A}}\widehat{\omega}_k - \min_{k\in\mathcal{A}}\omega_k + \delta \le \max_{k\in\mathcal{I}}\widehat{\omega}_k - \max_{k\in\mathcal{I}}\omega_k\Big)\\ &\le P\Big(\sup_{k\in\mathcal{A}}|\widehat{\omega}_k - \omega_k| \ge \delta/2\Big)\\ &\quad + P\Big(\sup_{k\in\mathcal{I}}|\widehat{\omega}_k - \omega_k| \ge \delta/2\Big).\end{aligned} \tag{A.10}$$

By using (A.8) with $\varepsilon = \delta/2$, (A.9) holds.

Proof of Theorem 3

Denote $p^* = p - |\mathcal{A}|$. For a fixed $r \in \mathbb{N}$, the event that $|\widehat{\mathcal{A}}_1 \cap \mathcal{I}| \ge r$ means there are at least r elements in $\{\widehat{\omega}_k : k \in \mathcal{I}\}$ greater than all values of $\{\widehat{\omega}_k : k = p+1, \ldots, p+d\}$. Because the auxiliary variables $\mathbf{z}$ and the inactive predictors $\mathbf{x}_{\mathcal{I}}$ are equally likely to be recruited given Y, it follows that

$$\begin{aligned}P(|\widehat{\mathcal{A}}_1 \cap \mathcal{I}| \ge r) &= \frac{p^*!}{(p^*-r)!}(p^* - r + d)!\big/(p^* + d)!\\ &\le \bigg(1 - \frac{r}{p^* + d}\bigg)^d.\end{aligned}$$

The result of Theorem 3 follows.

[Received September 2010. Revised May 2011.]

REFERENCES

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384. [1464]

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. [1464]

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489. [1465]

Choi, N. H., Shedden, K., Sun, Y., and Zhu, J. (2009), "Penalized Regression Methods for Ranking Multiple Genes by Their Strength of Unique Association With a Quantitative Trait," technical report, University of Michigan. [1465]

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Ser. B*, 34, 187–220. [1465]

Donoho, D. L. (2000), "High-Dimensional Data: The Curse and Blessings of Dimensionality," in *American Mathematical Society Conference Mathematical Challenges of 21st Century*, UCLA, Los Angeles. [1464]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Property," *Journal of the American Statistical Association*, 96, 1348–1360. [1464]

——— (2006), "Statistical Challenges With High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the International Congress of Mathematicians*, Vol. III, eds. M. Sanz-Sole, J. Soria, J. L. Varona, and J. Verdera, Zurich: European Mathematical Society, pp. 595–622. [1464]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911. [1464-1469,1471]

——— (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148. [1464]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1464]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 1829–1853. [1464,1467,1468]

Fang, K. T., Kotz, S., and Ng, K. W. (1990), *Symmetric Multivariate and Related Distributions*, London: Chapman & Hall. [1467]

Hall, P., and Li, K. C. (1993), "On Almost Linearity of Low Dimensional Projection From High Dimensional Data," *The Annals of Statistics*, 21, 867–889. [1467]

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21, 157–178. [1465]

Härdle, W., Liang, H., and Gao, J. T. (2000), *Partially Linear Models*, Germany: Springer Physica-Verlag, New York. [1465]

Li, H., and Luan, Y. (2005), "Boosting Proportional Hazards Models Using Smoothing Spline, With Application to High-Dimensional Microarray Data," *Bioinformatics*, 21, 2403–2409. [1471,1472]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussion), *Journal of the American Statistical Association*, 86, 316–342. [1467]

Lu, W., and Li, L. (2008), "Boosting Methods for Nonlinear Transformation Models With Censored Survival Data," *Biostatistics*, 9, 658–667. [1471,1472]

Luo, X., Stefanski, L. A., and Boos, D. D. (2006), "Tuning Variable Selection Procedure by Adding Noise," *Technometrics*, 48, 165–175. [1465,1467]

Pettitt, A. N. (1982), "Inference for the Linear Model Using a Likelihood Based on Ranks," *Journal of Royal Statistical Society, Ser. B*, 44, 234–243. [1465]

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Hermelink, H. K., Smeland, E. B., and Staudt, L. M. (2002), "The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma," *The New England Journal of Medicine*, 346, 1937–1947. [1471]

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley. [1474]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [1464,1468]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [1468]

Wu, Y., Boos, D. D., and Stefanski L. A. (2007), "Controlling Variable Selection by the Addition of Pseudo Variables," *Journal of the American Statistical Association*, 102, 235–243. [1465,1467]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, 68, 49–67. [1464]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1464]