Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Threshold Selection in Feature Screening for Error Rate Control

Xu Guo, Haojie Ren, Changliang Zou & Runze Li

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–13:Theory & Methods
https://doi.org/10.1080/01621459.2021.2011735

Taylor & Francis
Taylor & Francis Group

Threshold Selection in Feature Screening for Error Rate Control

Xu Guo*[a], Haojie Ren*[b], Changliang Zou[c], and Runze Li[d]

[a]School of Statistics, Beijing Normal University, Beijing, China; [b]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China; [c]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China; [d]Department of Statistics, The Pennsylvania State University, University Park, PA

ABSTRACT
Hard thresholding rule is commonly adopted in feature screening procedures to screen out unimportant predictors for ultrahigh-dimensional data. However, different thresholds are required to adapt to different contexts of screening problems and an appropriate thresholding magnitude usually varies from the model and error distribution. With an ad-hoc choice, it is unclear whether all of the important predictors are selected or not, and it is very likely that the procedures would include many unimportant features. We introduce a data-adaptive threshold selection procedure with error rate control, which is applicable to most kinds of popular screening methods. The key idea is to apply the sample-splitting strategy to construct a series of statistics with marginal symmetry property and then to utilize the symmetry for obtaining an approximation to the number of false discoveries. We show that the proposed method is able to asymptotically control the false discovery rate and per family error rate under certain conditions and still retains all of the important predictors. Three important examples are presented to illustrate the merits of the new proposed procedures. Numerical experiments indicate that the proposed methodology works well for many existing screening methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2020
Accepted November 2021

KEYWORDS
Empirical distribution; False discovery rate; Feature screening; Per family error rate; Symmetry.

1. Introduction

Ultrahigh-dimensional data analysis are now frequently encountered in diverse fields of scientific research, such as biomedical imaging, functional magnetic resonance imaging, microarrays, and high-frequency financial data, among many others. The ultrahigh dimensionality poses great computational and statistical challenges in statistical inference (Fan, Han, and Liu 2014). To explore the relationship between a response variable Y and p-dimensional predictor vector $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ efficiently, Fan and Lv (2008) introduced a sure independence screening (SIS) procedure to screen out uninfluential predictors as many as possible while retain all influential variables by a computational efficient and stable procedure or algorithm. It is typical that one applies existing regularization methods to further clean up uninfluential predictors based on the model selected by the SIS procedure. As SIS builds on marginal Pearson correlations between the response and the features, various extensions of correlation have been proposed to deal with more general cases. See Fan and Lv (2018) for an updated review on this topic.

A feature screening procedure first defines a nonnegative statistic ω_j measuring the importance of X_j to Y. Generally, this ω_j quantifies the marginal association of X_j and Y in certain sense, and let us assume without loss of generality that the greater ω_j is, the more important X_j is. Suppose that $\widehat{\omega}_j$ is an estimate of ω_j based on the sample $\{X_{ji}, Y_i\}_{i=1}^n$. Thus, $\widehat{\omega}_j$ can be viewed as a marginal utility to rank the importance of X_j at the sample level. We select a set of important predictors with large $\widehat{\omega}_j$:

$$\mathcal{S}(T) = \{j : \widehat{\omega}_j \geq T, \text{ for } 1 \leq j \leq p\}, \tag{1}$$

where T is a threshold. Clearly T controls the model complexity and plays an important role in model selection. To achieve sure screening property, theoretical choice of T has been derived in the literature. In general, the theoretical choice of T depends on unknown quantities related to the joint distribution of $\mathbf{X}$ and Y. Fan and Lv (2008) advocated to retain a fixed number, d, of predictors. A common choice of d is $d = \lfloor n/\log n \rfloor$. This is equivalent to sorting $\widehat{\omega}_j$ from the largest to the smallest and obtaining the order statistics $\widehat{\omega}_{(1)} \leq \widehat{\omega}_{(2)} \leq \cdots \leq \widehat{\omega}_{(p)}$, and then setting $T = \widehat{\omega}_{(p-d+1)}$. This hard thresholding rule has become a standard choice in the literature. However, choosing an appropriate d remains a challenge in practice. On one hand, one can arbitrarily select a conservative one, say a very large d to ensure that all influential features are included with high probability, but an ad-hoc one would in turn include too many noise predictors that should be discarded. On the other hand, a too small d would fail to ensure sure screening property. Zhu et al. (2011) proposed a strategy of choosing the threshold for their proposed sure independent ranking screening procedure by adding auxiliary covariates. Pan, Wang, and Li (2016) considered using the BIC criterion in the linear discriminant analysis setting. Those strategies aim for specific marginal utilities, and

CONTACT Runze Li ril4@psu.edu Department of Statistics, The Pennsylvania State University, University Park, PA; Changliang Zou nk.chlzou@gmail.com School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China.

*The first two authors equally contributed to this work, and are listed in the alphabetic order.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

therefore, it is desirable to develop a general strategy for selecting d under a unified feature screening setting.

There is an obvious tradeoff: a small threshold would include more active features but retain more inactive features as well. Besides the power aspect in terms of the sure screening property, we often would like to know whether the estimated model $\mathcal{S}(T)$ enjoys reproducibility in that the fraction of noise features is controlled (Fan et al. 2020a). However, research on the error rate control in feature screening, such as false discovery rate (FDR), familywise error rate (FWER), or per family error rate (PFER), is very limited. This is partly because the distributions of $\widehat{\omega}_j$'s may vary across j and are often difficult to approximate, let alone all the $\widehat{\omega}_j$'s joint distributions. Among others, Fan, Samworth, and Wu (2009) established an upper bound on the probability of recruiting any inactive variables for the hard thresholding rule under an exchangeable condition. Zhu et al. (2011) suggested a soft thresholding rule by adding artificial auxiliary variables to the data and obtained similar non-asymptotic bound to the one in Fan, Samworth, and Wu (2009). Hao and Zhang (2017) suggested a new concept called oracle p-value and discussed its application for variable screening with FDR control. Recently, Chudik, Kapetanios, and Pesaran (2018) considered FDR control with marginal Pearson correlations. There is clearly lack of a systematic approach to determining the threshold so that certain error rate control can be achieved.

In this work, we propose a simple yet effective selection procedure based on sample-splitting. The method entails reforming the marginal utility statistics into p new statistics with marginal symmetry property, using the empirical distribution of the negative statistics to approximate that of the positive ones. The newly proposed methodology is generic and is applicable for most of existing feature screening statistics. Under a unified framework and some mild conditions, we show that the proposed method is able to control the FDR and PFER asymptotically. We also prove that the procedure can still retain all of the important predictors. The procedures are theoretically illustrated with several commonly used screening approaches, including Fan and Lv (2008)'s SIS procedure, the robust rank correlation screening procedure in Li et al. (2012) and the Kolmogorov filter proposed by Mai and Zou (2012). Numerical experiments indicate that the proposed rule works well in a wide range of settings.

The remainder of our article is structured as follows. In Section 2, we present the basic procedure. Asymptotic justification is provided in Section 3. Numerical studies are conducted in Section 4. Some conclusions and discussions are given in Section 5, and theoretical proofs are delineated in the Appendix. Some technical details and additional numerical results are provided in the supplementary material.

2. Feature Screening with Error Rate Control

2.1. FDR Control

For ease of exposition, we start with the FDR control which is a particularly useful tool to maintain the ability to reliably detect true signals without excessive false positive results when large-scale hypotheses are simultaneously tested (Benjamini and Hochberg 1995). Other error rates will be discussed later on. In what follows, we term the X_j as an *uninformative* predictor if $\omega_j \leq b_n$ for some sequence $b_n \to 0$ as $n \to \infty$. The uninformative set is accordingly $\mathbb{H}_0 = \{j : \omega_j \leq b_n, j = 1, \ldots, p\}$ and its complement $\mathbb{H}_1$ is termed as *informative* set. Denote $p_0 = |\mathbb{H}_0|$, $p_1 = |\mathbb{H}_1|$ and throughout this article we always assume that p_1 is dominated by p, say $p_1 = o(p)$.

Consider a screening procedure defined in (1) which yields a selected set $\mathcal{S}(T)$. A false discovery is made by $\mathcal{S}(T)$ if $j \in \mathcal{S}(T) \bigcap \mathbb{H}_0$. So, the false discovery proportion (FDP) associated with $\mathcal{S}(T)$ is

$$\mathrm{FDP}(\mathcal{S}(T)) = \frac{\#\{j : j \in \mathcal{S}(T) \bigcap \mathbb{H}_0\}}{\#\{j : j \in \mathcal{S}(T)\} \vee 1},$$

and the FDR is accordingly defined as the expectation of $\mathrm{FDP}(\mathcal{S}(T))$. The main goal is to find a data-adaptive threshold T such that

$$\limsup_{n\to\infty} \mathbb{E}\left\{\mathrm{FDP}(\mathcal{S}(T))\right\} \leq \alpha.$$

That is, the asymptotic FDR is controlled at a pre-specified level α.

We first impose some conditions on the marginal utility statistic.

Assumption 1. (i) For $j \in \mathbb{H}_0$ and known number $\gamma > 0$, $n^{\gamma}\widehat{\omega}_j \xrightarrow{d} \mathcal{N}_j$, where $\mathcal{N}_j$ is some nondegenerate variable; (ii) For $j \in \mathbb{H}_1$, $\widehat{\omega}_j \xrightarrow{p} \omega_j > 0$.

Here $\xrightarrow{d}$ and $\xrightarrow{d}$ denote the convergence in distribution and convergence in probability, respectively. It is required that the γ, the convergence rate of $\widehat{\omega}_j$, is known but we do not need any information of $\mathcal{N}_j$. This assumption is quite mild and satisfied by most existing screening procedures. For example, if we take $b_n = o(n^{-\gamma})$, then $\gamma = 1/2$ for the SIS procedure proposed by Fan and Lv (2008), in which ω_j is the absolute value of Pearson correlation coefficient between X_j and Y, while $\gamma = 1$ for the distance correlation SIS (DC-SIS, Li, Zhong, and Zhu 2012).

We next propose a data-driven threshold for controlling FDR of the feature screening procedure based on the marginal utility statistic w_j. We split the whole dataset randomly into two disjoint groups $\mathcal{D}_1 = (\mathbf{X}, Y)_1$ and $\mathcal{D}_2 = (\mathbf{X}, Y)_2$ of *unequal* size $n_1 = n(K-1)/K$ and $n_2 = n/K$, where $K \geq 3$ with assuming that n/K is an integer for simplicity. The marginal utility statistics for the jth variable on $\mathcal{D}_1$ and $\mathcal{D}_2$ are denoted as $\widehat{\omega}_{j1}$ and $\widehat{\omega}_{j2}$, respectively. Define

$$\begin{aligned} W_j &= \mathrm{sgn}\left(n_1^{\gamma}\widehat{\omega}_{j1} - n_2^{\gamma}\widehat{\omega}_{j2}\right)\left(n_1^{\gamma}\widehat{\omega}_{j1} \vee n_2^{\gamma}\widehat{\omega}_{j2}\right) \\ &= \begin{cases} n_1^{\gamma}\widehat{\omega}_{j1}, & \text{if } n_1^{\gamma}\widehat{\omega}_{j1} > n_2^{\gamma}\widehat{\omega}_{j2}, \\ -n_2^{\gamma}\widehat{\omega}_{j2}, & \text{if } n_1^{\gamma}\widehat{\omega}_{j1} \leq n_2^{\gamma}\widehat{\omega}_{j2}. \end{cases} \end{aligned} \tag{2}$$

By Assumption 1-(ii), with probability tending to one, $n_1^{\gamma}\widehat{\omega}_{j1} > n_2^{\gamma}\widehat{\omega}_{j2} > 0$ if $\omega_j > 0, j \in \mathbb{H}_1$, and thus, $W_j > 0$. While for any $j \in \mathbb{H}_0$, W_j is (asymptotically) symmetric around zero due to Assumption 1-(i) and the independence between $\mathcal{D}_1$ and $\mathcal{D}_2$. That is, W_j can be used to discriminate an *uninformative* predictor and an *informative* predictor, and it has *marginal symmetry* property for all the uninformative predictors.

Motivated by the properties of W_j, we propose to choose a threshold $L > 0$ via setting

$$L = \inf\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \alpha\right\}, \tag{3}$$

and select the predictors by $\mathcal{M}(L) = \{j : W_j \geq L\}$, where α is the target FDR level. If the set is empty, we simply set $L = +\infty$. Naturally the $\omega_j, j \in \mathbb{H}_1$ is not too weak. Then $\#\{j : W_j \leq -t\}$ is a good approximation to $\#\{j : W_j \leq -t, j \in \mathbb{H}_0\}$, which further is a good approximation to $\#\{j : W_j \geq t, j \in \mathbb{H}_0\}$ due to the marginal symmetry of W_j for $j \in \mathbb{H}_0$. Theoretically speaking, the key idea is to exploit the following symmetry property (see Lemmas A.1 in the Appendix and Condition 2)

$$\sup_{0 \leq t < M_n} \left| \frac{\sum_{j \in \mathbb{H}_0} \mathbb{I}(W_j \geq t)}{\sum_{j \in \mathbb{H}_0} \mathbb{I}(W_j \leq -t)} - 1 \right| = o_p(1), \tag{4}$$

for some large M_n. It implies that the fraction in (3) is an estimate of the FDP. Since we use the empirical distribution of the negative statistics to approximate that of the positive ones, we refer our procedure to as *REflection via Data Splitting* (REDS).

Let us briefly discuss how to choose K. The way we construct W_j is to ensure that (a) W_j is (asymptotically) symmetric with mean zero for $j \in \mathbb{H}_0$, and (b) W_j is a large positive value for $j \in \mathbb{H}_1$ without imposing other requirements on $\mathcal{N}_j$. Intuitively, with a larger K, the strength of the signal for $j \in \mathbb{H}_1$ would be larger. We then more likely select the important predictors. However, this leads to smaller sample size of n_2, which would further yield a slower convergence rate of $n_2^{\gamma}\widehat{\omega}_{j2}$ for $j \in \mathbb{H}_0$. Consequently, the marginal symmetry of W_j for $j \in \mathbb{H}_0$ may be violated to certain degree and the FDR control of the proposed procedure would be compromised. In practice we recommend $K = 3$, which performs quite well in our numerical study.

Before we pursue further, let us illustrate the idea of the proposed REDS procedure via a toy example. We generate a simulation dataset from a linear model in which $\mathbf{X}$ comes from a multivariate normal distribution with the correlation between X_j and X_k being 0.2 for $j, k \in \mathbb{H}_1$ and being 0 otherwise. In this illustration, we set $n = 100$, $p = 1000$, $p_1 = 40$, and $K = 3$. As in Fan and Lv (2008), we take the absolute value of Pearson correlation as the marginal utility statistics. As a visual representation of the proposed procedure, we depict in Figure 1(a) the scatterplot of the screening statistic pairs for each feature j, $\{(K-1)^{\gamma}\widehat{\omega}_{j1}, \widehat{\omega}_{j2}\}$, with the red triangles and black dots denoting informative predictors and uninformative ones, respectively. A feature j whose point lies below the diagonal line has a positive value of W_j by definition and vice versa. Figure 1(b) depicts the resulting W_j's. Observe that the true signals are primarily above the x-axis, indicating $W_j > 0$, while the uninformative W_j's (black dots) are roughly symmetrically distributed across the horizontal lines. Figure 1(c) shows the corresponding estimate of FDP curve (against t), that is, the fraction in (3), along with the true FDP. The approximation in this case is very good as only very few true informative predictors are present below the horizontal line in Figure 1(b).

2.2. Per Family Error Rate Control

The proposed REDS approach can also be used for providing a rough control of other error rates, such as the *Per Family Error Rate* (PFER), and the *k-FWER*, the probability of making at least k false discoveries. We refer to Romano and Wolf (2007), recently Janson and Su (2016) and the references therein. Here we discuss the PFER only.

The PFER, the expected number of false discoveries, has a clear interpretation, making it a useful criterion in feature screening approaches. Controlling the PFER amounts to find a t so that

$$\mathbb{E}\left\{\sum_{j \in \mathbb{H}_0} \mathbb{I}(W_j \geq t)\right\} = \sum_{j \in \mathbb{H}_0} \Pr(W_j \geq t) \leq k_0,$$

where k_0 is the target PFER level. Again, by the intuition

$$\sum_{j \in \mathbb{H}_0} \Pr(W_j \geq t) \approx \sum_{j \in \mathbb{H}_0} \mathbb{I}(W_j \geq t)$$
$$\approx \sum_{j \in \mathbb{H}_0} \mathbb{I}(W_j \leq -t) \lesssim \sum_{j=1}^{p} \mathbb{I}(W_j \leq -t),$$

we can obtain an approximately correct threshold by

$$L' = \inf\left\{t > 0 : \sum_{j=1}^{p} \mathbb{I}(W_j \leq -t) \leq k_0\right\}. \tag{5}$$

The proposed thresholds L in (3) and L' in (5) are in a similar spirit to that in the knockoff framework introduced by Barber and Candès (2015) and further advanced as "Model-X" knockoffs (Candes et al. 2018), where the context is regression models. The model-X knockoffs framework can control FDR without assuming a specific regression model. However, it requires the joint distribution of predictors to be known. The robustness issue of the model-X knockoffs is theoretically investigated by Fan et al. (2020a) and Barber, Candès, and Samworth (2020). For other recent development of variable selection via knockoffs, see also Barber and Candès (2019) and Fan et al. (2020b). The knockoff procedure operates via constructing "knockoff copies" of each of the p predictors (features) with certain knowledge of the predictors or responses. The signs of test statistics constructed via knockoff would satisfy (or roughly) joint exchangeability and thus, can yield accurate FDR control in finite samples (Candes et al. 2018). However, in the feature screening problems, we usually do not have enough information on $(Y, X_1, \ldots, X_p)^{\top}$, and the exact knockoff copies are generally not available when p is too large compared with n. The knockoff framework is to test conditional hypotheses, while we follow the literature of marginal feature screening (Fan and Lv 2008) and mainly focus on marginal hypotheses. Although conditional and marginal hypotheses are two different concepts, marginal hypotheses can still be informative for the conditional hypotheses in a ultrahigh-dimensional environment, which has been successfully demonstrated by numerous authors in the field of marginal feature screening. Recently, based on marginal linear regression, McKeague and Qian (2015) proposed an adaptive resampling test to detect the presence of significant predictors. See also Wang, McKeague, and Qian (2018) and the references therein for more examples.

The REDS for choosing L and L' provides a unified framework for threshold selection in commonly-used marginal feature screening. No matter $n^{\gamma}\widehat{\omega}_j$ is asymptotically distribution-free or not, provided that $n^{\gamma}\widehat{\omega}_j$ converges to a nondegenerate variable $\mathcal{N}_j$ when $j \in \mathbb{H}_0$ and $\widehat{\omega}_j$ converges to $\omega_j > 0$ when $j \in \mathbb{H}_1$, our approach is basically applicable. Also, notice that

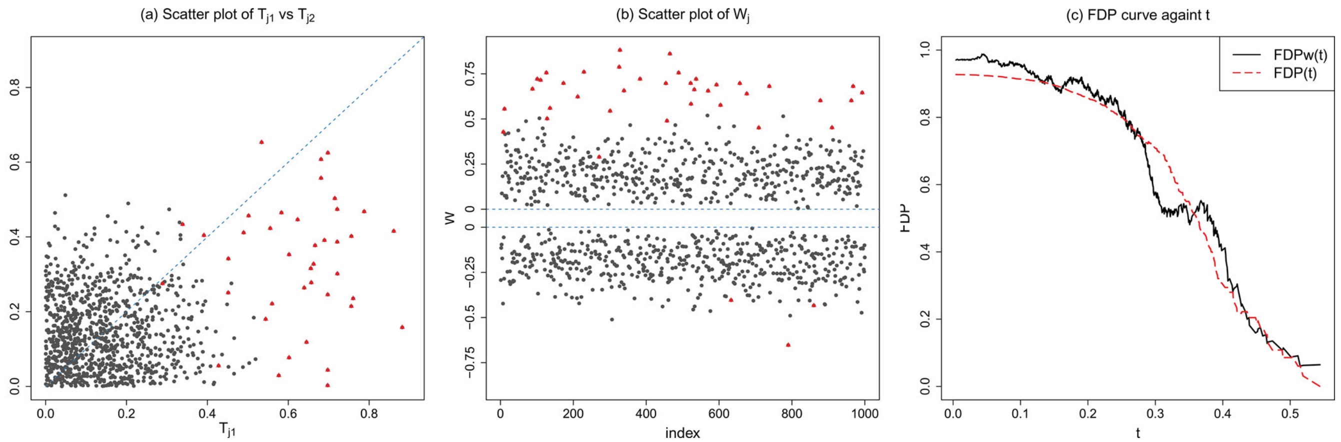

Figure 1. (a) is scatterplot of screening statistic pairs ($T_{j1} = (K-1)^{\gamma}\widehat{\omega}_{j1}, T_{j2} = \widehat{\omega}_{j2}$) with the red triangles and black dots denoting informative predictors and uninformative ones, respectively; (b) is scatterplot of the W_j's; (c) is the corresponding estimate of FDP curve against t (black line) along with the true FDP (red dash-line).

the proposed procedure is computationally efficient–it only uses a one-time split of the data and calculation of the W_j in (2). This is particularly useful in feature screening because one usually expects a rapid scheme which has minimum requirements on computing and storage. By contrast with the knockoff, we here use the sample-splitting idea to achieve a marginal symmetry property. It turns out that the error rate like FDR is able to be controlled in a reasonable range even with marginal symmetry provided that the empirical distribution converges under certain dependence structures (Storey, Taylor, and Siegmund 2004). We note that the sample-splitting idea has been used by many authors in different contexts. For instance, Wasserman and Roeder (2009) firstly divided the data into two independent parts, secondly used one part to narrow down the focus and finally used the remainder to perform inference tasks. Fan, Guo, and Hao (2012a) and Chen, Fan, and Li (2018) adopted this idea to estimate the error variance in high-dimensional regression models.

2.3. Data-Driven Hard Threshold for Screening Methods

The proposed thresholds L in (3) and L' in (5) are based on error rate control. We control the error rates by a joint use of two screening statistics from two splits. In practice, one may prefer to use the original full-sample statistic $\widehat{\omega}_j$. We next demonstrate how to use L and L' to carry out a hard-threshold feature screening. We will focus on L only since we may simply replace L by L' if we want PFER control. For the selected L, feature screening with the hard thresholding $|\mathcal{M}(L)|$ is equivalent to using the threshold T defined by

$$T = \widehat{\omega}_{(p-|\mathcal{M}(L)|+1)},$$

where $\widehat{\omega}_{(1)} \le \cdots \le \widehat{\omega}_{(p)}$ are the order statistics of $\widehat{\omega}_j$'s.

The rationale is as follows. If the FDP is approximately controlled at the level α, there are roughly $|\mathcal{M}(L)|(1-\alpha)$ informative predictors in $\mathcal{M}(L)$. Because $\widehat{\omega}_j$ is expected to be at least as effective as the $\widehat{\omega}_{1j}$, those informative predictors would be retained with an overwhelming probability when we are using $\widehat{\omega}_j$ for screening. That in turn would yield false discoveries no greater than the REDS selection, resulting in a more conservative rule. But due to the use of full-sample information, we can expect that the sure screening property would be better achieved. We will refer this procedure as *hard-thresholding rule with REDS* and the whole algorithm is summarized as follows.

Algorithm 1: Hard-thresholding rule with REDS (HTR)

- *Step 1:* Randomly split the data into two parts and compute $\widehat{\omega}_{kj}$ for $k = 1, 2$ and $j = 1, \ldots, p$;
- *Step 2:* Compute W_j by (2), find the L by (3) and obtain $\mathcal{M}(L) = \{j : W_j > L\}$;
- *Step 3:* Compute $\widehat{\omega}_j$ for $j = 1, \ldots, p$ and select the predictors with the largest $|\mathcal{M}(L)|$ values of $\widehat{\omega}_j$; Output the selected subset as $\mathcal{S}(\widehat{\omega}_{(p-|\mathcal{M}(L)|+1)})$.

In the next section, we will show that with the help of the FDR control of the REDS procedure, the hard-thresholding strategy, HTR, is also able to control the FDR asymptotically. Hence, the REDS has considerable merit in the sense that $\mathcal{M}(L)$ is always informative, especially for complex data and computation-intensive marginal utility statistics where approximate distributions of the $\widehat{\omega}_j$ are not obvious. A preliminary REDS step helps us to provide insight as to the appropriate quantity of the thresholding.

One may be wary of the stability of the one-time split. To reduce randomness occurred in a single sample splitting (Meinshausen, Meier, and Bühlmann 2009), we may employ the “bagging” technique to aggregate results from multiple sample-splitting procedures. Say, we run the first two steps of Algorithm 1 B times with a nominal level α, and obtain a collection of $|\mathcal{M}_k(L)|, k = 1, \ldots, B$. We then aggregate them via some commonly used bagging strategy, such as the voting or averaging. For example, we may use $[B^{-1}\sum_k |\mathcal{M}_k(L)|]$ as our hard threshold. Some simulation results given in the supplementary material show that compared to one-time split, the multiple-splits refinement yields similar FDR and power but slightly smaller variations in the FDP.

3. Theoretical Results

In this section, we study the theoretical properties of the proposed procedures. We start with general setting in Section 3.1 and then work on specific settings corresponding to several commonly-used marginal utility statistics in the literature of feature screening.

3.1. General Settings

In this section, we focus on studying the property of the FDR-based threshold L. We first establish a finite-sample property of W_j.

Proposition 1. Assume that $W_j, 1 \le j \le p$ are well defined. For any $\alpha \in (0,1)$, the REDS selection procedure satisfies

$$\mathrm{FDR} \le \min_{\epsilon \ge 0} \left\{ \alpha(1+5\epsilon) + \Pr\left(\max_{j \in \mathbb{H}_0} \Delta_j > \epsilon \right) \right\},$$

where $\Delta_j = \left| \Pr(W_j > 0 \| W_j|, \mathbf{W}_{-j}) - 1/2 \right|$ and $\mathbf{W}_{-j} = (W_1, \ldots, W_p)^\top \setminus W_j$.

We can interpret Δ_j as measuring the extent to which the symmetry is violated for a specific variable j as well quantifying the effect of the dependence between W_j and $\mathbf{W}_{-j}$ on the FDR. This result concurs with our intuition that controlling the Δ_j's is sufficient to ensure control of the FDR for the REDS method. In the most ideal case where the W_j's are symmetrically distributed and independent, $\Delta_j = 0$ for all $j \in \mathbb{H}_0$, and we obtain the accurate FDR-control result since we can take $\epsilon = 0$. Under asymmetric and dependent scenarios, the Δ_j can still be expected to be small due to the convergence of $n_1^\gamma \widehat{\omega}_{j1}$ and $n_2^\gamma \widehat{\omega}_{j2}$ to a same distribution if n is not too small.

To establish asymptotical property of the proposed procedure, we impose the following technical conditions, which are not the weakest one, but facilitate the technical proofs. Denote $\tilde{F}_{\mathcal{N}_j}(t) = 1 - F_{\mathcal{N}_j}(t)$, where $F_{\mathcal{N}_j}$ is the distribution function of $\mathcal{N}_j$.

Condition 1 (The distribution of marginal utility statistics). (i) Let $H_{nj}(t) = \Pr_{\mathbb{H}_0}(n^\gamma \widehat{\omega}_j > t)$. Then $H_{nj}(t)$ satisfies $H_{nj}(t)/\tilde{F}_{\mathcal{N}_j}(t) \to 1$ uniformly in $j \in \mathbb{H}_0$ and $t \in (0, n^\eta)$ with some η, as $n \to \infty$ and $t \to \infty$; (ii) For j, $\Pr(|\widehat{\omega}_j - \omega_j| \ge \delta_{p,n}) \le d_{p,n}$, where $\delta_{p,n} \to 0$ and $d_{p,n} \to 0$.

Define $\mathcal{C}_\beta \equiv \left\{ j \in \mathbb{H}_1 : \omega_j > 2a\delta_{p,n}/(a-1) \right\}$ and $\beta_{p,n} = |\mathcal{C}_\beta|$, where $a = (K-1)^\gamma$.

Condition 2 (The convergence of empirical distributions). For $\beta_{p,n} \to \infty$, we have

$$\sup_{0 \le t \le G_+^{-1}(\alpha\beta_{p,n}/p)} \left| \{p_0 G_+(t)\}^{-1} \sum_{j \in \mathbb{H}_0} \mathbb{I}\left(W_j \ge t\right) - 1 \right| = o(1),$$

$$\sup_{0 \le t \le G_-^{-1}(\alpha\beta_{p,n}/p)} \left| \{p_0 G_-(t)\}^{-1} \sum_{j \in \mathbb{H}_0} \mathbb{I}\left(W_j \le -t\right) - 1 \right| = o(1),$$

where $G_+(t) = p_0^{-1} \sum_{j \in \mathbb{H}_0} \Pr(W_j \ge t)$ and $G_-(t) = p_0^{-1} \sum_{j \in \mathbb{H}_0} \Pr(W_j \le -t)$.

Remark 1. Condition 1-(i) is stronger than Assumption 1-(i) and is used to establish the marginal symmetry of W_j. It is closely related to the large-deviation theory and can be satisfied by many existing screening procedures. We will discuss this condition in more details for examples in Section 3.2. Condition 1-(ii) is commonly used in the literature of feature screening to derive the sure screening property, for example, in SIS, $\delta_{p,n} = O(n^{-\kappa})$ and $d_{p,n} = O(\exp\{-Cn^{1-2\kappa}/\log n\})$ for $0 < \kappa < 1/2$ and some $C > 0$, while in DC-SIS, $\delta_{p,n} = O(n^{-\kappa})$ and $d_{p,n} = O(\exp\{-Cn^{(1-2\kappa)/3}\})$ for $0 < \kappa < 1/2$ and some $C > 0$ (Fan and Lv 2008; Li, Zhong, and Zhu 2012). Condition 2 sets theoretical minimal requirements for the convergence of the empirical distribution $\{p_0 G_+(t)\}^{-1} \sum_{j \in \mathbb{H}_0} \mathbb{I}\left(W_j \ge t\right)$ to its population one. This is reasonable because we cannot expect that our selection procedure would work well if the empirical distributions are far away from the population ones, since we are using the former to search for thresholds. Similar conditions are popular; see, for instance, Storey, Taylor, and Siegmund (2004). This condition is mainly about the correlation between the uninformative statistics $W_j, j \in \mathbb{H}_0$. When W_k is independent with all the other variables W_j or only dependent with finite number of W_j, then this condition trivially holds. Certainly, W_j's inherit their dependence structure from X_j's which can affect the validity of the convergence. We will explicitly provide some sufficient conditions so that Condition 2 is valid for some specific examples. Denote $T_U = n_1^\gamma \delta_{p,n}$.

Theorem 1. Suppose Conditions 1-2 hold. If $\beta_{p,n} \to \infty$, $p_0 d_{p,n} \to 0$ and $T_U/n^\eta \to 0$ as $(n,p) \to \infty$, then for any $\alpha \in (0,1)$, the FDR of the REDS procedure with threshold L satisfies $\limsup_{(n,p)\to\infty} \mathrm{FDR} \le \alpha$.

Theorem 1 implies that the feature screening procedure with the proposed threshold L can control the FDR level asymptotically for dependent predictors. The condition $\beta_{p,n} \to \infty$ implies that the number of informative predictors with identifiable signal strengths is not too small as $p \to \infty$. This seems to be a necessary condition for FDP control. For example, in the context of multiple testing, Liu and Shao (2014) showed that even with the true p-values, no method is able to control FDP with a high probability if the number of true alternatives is fixed as the number of hypothesis tests goes to infinity. To see this clearer, notice that the key step is to show the validity of (4) in which the convergence of empirical sum such like $\sum_{j \in \mathbb{H}_0} \mathbb{I}(W_j < -t)$ is needed. When t is extremely large, the number of nonzero terms in the summation would be finite and consequently the convergence would fail. The condition that $\beta_{p,n} \to \infty$ helps to rule out the case of having too large t. In fact, we will show that $\Pr\left(\sum_j \mathbb{I}\left(W_j > L\right) \ge \beta_{p,n} \right) \to 1$. This implies that those $\beta_{p,n}$ informative predictors would be identified with probability tending to 1. Thus, there are at least $\beta_{p,n}$ discoveries in $\mathcal{M}(L)$. Its implication is that $\sum_j \mathbb{I}(W_j < -L)$ would tend to infinity by the definition of L. So, $\sum_{j \in \mathbb{H}_0} \mathbb{I}(W_j < -L) \to \infty$ since $\{j : W_j < -L\}$ are more likely to be uninformative predictors.

The condition $\beta_{p,n} \to \infty$ is not too stringent since we only require a few predictors whose marginal utility measures exceeding $2a\delta_{p,n}/(a-1)$ among all the informative predictors. The $p_0 d_{p,n} \to 0$ and $T_U/n^\eta \to 0$ are two technical conditions and can be easily satisfied for most existing screening statistics; this will become clearer in Section 3.2.

Zhu et al. (2011) suggested to independently and randomly generate p auxiliary variables Z from $N(0, 1)$, which is independent of both $\mathbf{X}$ and Y. They proved that when the inactive predictors and the auxiliary variables are exchangeable, the probability of recruiting at least r inactive variables is upper bounded by $(1 - r/2p)^p$. However, the exchangeable condition is not easily satisfied in practice especially when we generate Z without using any information of $\mathbf{X}$, and accordingly such a procedure may not be able to maintain a desired error rate. Our proposed procedure is distinguished from Zhu et al. (2011) in that our procedures do not require exchangeable condition, instead use the sample-splitting scheme to achieve a marginal symmetry property, which is proven particularly useful in the FDR control.

The following theorem establishes the validity of PFER control of the REDS procedure with threshold L'.

Theorem 2. Under conditions of Theorem 1, for any $k_0 > 0$, we have

$$\limsup_{n,p\to\infty}\mathbb{E}\left\{\sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j \ge L')\right\} \le k_0.$$

We next show that Condition 2 holds under some weak dependence structures.

Theorem 3. The Condition 2 is valid under either one of the following two dependence settings, and accordingly the FDR control given in Theorem 1 and the PFER control given in Theorem 2 hold.

(i) Assume that for each W_j, the number of W_k that are dependent with W_j is no more than $r_{p,n}$, where $r_{p,n}/\beta_{p,n} \to 0$ as $(n, p) \to \infty$.

(ii) Assume that

$$\left|\frac{\sum_{j\neq k\in\mathbb{H}_0}\Pr(W_j > t, W_k > t)}{\sum_{j\neq k\in\mathbb{H}_0}\Pr(W_j > t)\Pr(W_k > t)} - 1\right| \le C(\log p)^{-1-\theta}.$$

uniformly in $0 \le t \le G_+^{-1}(\alpha\beta_{p,n}/p)$, where $\theta > 0$ is any small constant.

The first condition (i) imposes the independence between W_j and other $p - r_{p,n}$ W_k's. The $\beta_{p,n}$ is required to diverge faster than $r_{p,n}$. Note that the $r_{p,n}$ is the upper bound on the number of dependent statistics for each statistic W_j, while $\beta_{p,n}$ represents the lower bound of the number of nonzero terms in the summation $\sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j \le -t)$. The requirement on the growth rate of $\beta_{p,n}$ is reflected by the law of large numbers under dependent scenarios. The second one allows W_j being correlated with all the other variables but the average of the correlation coefficients (in terms of $\mathbb{I}(W_j > t)$ and $\mathbb{I}(W_k > t)$) needs to converge to zero at a log-rate. This assumption is similar to the weak dependence structure given in Fan, Han, and Gu (2012b). If the correlation matrix contains many large entries, this condition may not hold; a certain degree of sparseness is needed. Certainly, those two conditions are not the weakest possible. In fact, Condition 2 can also be proved by using the Bernstein-type inequality under some mixing conditions (Merlevède et al. 2009) together with similar arguments given in the proof of Theorem 3.

The next result establishes the FDR control of Algorithm 1.

Theorem 4. Suppose the conditions in Theorem 1 all hold. If $\omega_j > \{(K-1)^\gamma + K^\gamma\}\delta_{p,n}/\{K^\gamma - (K-1)^\gamma\}$ for $j \in \mathbb{H}_1$, then for any $\alpha \in (0, 1)$, the FDR of the HTR procedure satisfies $\limsup_{(n,p)\to\infty}\mathrm{FDR}_{\mathrm{HTR}} \le \alpha$.

It is worth pointing out that in this theorem, we impose a minimum signal condition which is not required for the REDS procedure. For HTR, we use the minimum signal strength condition to ensure that for all $j \in \mathbb{H}_1$, $n^r\widehat{\omega}_j > n_1^r\widehat{\omega}_{1j}$ with probability tending to 1. Accordingly, most of the informative predictors identified by REDS would be retained in $\mathcal{S}(\widehat{\omega}_{(p-|\mathcal{M}(L)|+1)})$. As a contrast, in Theorem 1, the FDR of REDS can be controlled provided that there are certain number of identifiable signals. The HTR procedure needs more stringent conditions to achieve asymptotic FDR control. If there are many weakly informative predictors in $\mathbb{H}_1$, the FDR control is likely to be compromised.

Corollary 1. Suppose the conditions in Theorem 4 all hold and the FDP of HTR satisfies $\limsup_{(n,p)\to\infty}\mathrm{FDP}_{\mathrm{HTR}} \ge \alpha_0$ with $\alpha_0\beta_{p,n} \ge 1$ in probability. If $\omega_j > 2\delta_{p,n}$ for all $j \in \mathbb{H}_1$, then we have $\Pr(\mathcal{S}(\widehat{\omega}_{(p-|\mathcal{M}(L)|+1)}) \supseteq \mathbb{H}_1) \to 1$.

This corollary implies that our HTR procedure is capable of not only controlling the FDR level, but also achieving the sure screening property under suitable conditions.

3.2. Examples

In this section, we demonstrate the theoretical results can be directly applied for commonly-used marginal utility statistics.

Example 3.1. Fan and Lv (2008) proposed sure independent screening (SIS) procedure using Pearson correlation coefficient to measure the dependence between X_j and Y. That is, $\omega_j = |\mathrm{cov}(X_j, Y)/\sqrt{\mathrm{var}(X_j)\mathrm{var}(Y)}|$ and

$$\widehat{\omega}_j = \left|\frac{\sum_{i=1}^n(X_{ji} - \bar{X}_j)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n(X_{ji} - \bar{X}_j)^2\sum_{i=1}^n(Y_i - \bar{Y})^2}}\right|,$$

where $\bar{X}_j$ and $\bar{Y}$ are the sample means of $\{X_{ji}\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively. They showed that under suitable conditions,

$$\Pr(|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}) \le O(\exp\{-Cn^{1-2\kappa}/\log n\}) \quad \text{with } 0 < \kappa < 1/2.$$

Thus, we can take $\delta_{p,n} = cn^{-\kappa}$ and $d_{p,n} = \exp\{-Cn^{1-2\kappa}/\log n\}$. From Lemma S.1 in the supplementary material, we easily verify that Condition 1-(i) holds uniformly for $t \in (0, n^{1/6})$. Hence, Condition 1 holds.

Proposition 2. Suppose that Condition 2 holds and $(X_1, \ldots, X_p, Y)$ are Subgaussian variables with $\max_{j\in\mathbb{H}_0}|X_jY|^3 \le C$ for some constant $C > 0$. The result given in Theorem 4 is valid for the Pearson-correlation based SIS procedure if $p = \exp\{o(n^{\min(1/3,1-2\kappa)})\}$.

The feature dimension p is allowed to grow exponentially fast with the sample size n, which fits for the requirement of ultrahigh-dimensional settings.

Example 3.2. Li et al. (2012) proposed the robust rank correlation screening (RRCS) based on Kendall's τ correlation. Specifically, they take $\omega_j = |\mathbb{E}\{\text{sgn}(X_{j1} - X_{j2})\text{sgn}(Y_1 - Y_2)\}|$, and at the sample level, $\widehat{\omega}_j$ is

$$\widehat{\omega}_j = \left| \frac{2}{n(n-1)} \sum_{1 \le i < l \le n} \text{sgn}(X_{ji} - X_{jl})\text{sgn}(Y_i - Y_l) \right|.$$

The $\widehat{\omega}_j$ is in the form of U-statistic with a bounded kernel function, $\text{sgn}(X_{j1} - X_{j2})\text{sgn}(Y_1 - Y_2)$. Li et al. (2012) established that

$$\Pr(|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}) \le O(\exp\{-Cn^{1-2\kappa}\}) \text{ with } 0 < \kappa < 1/2,$$

and consequently $\delta_{p,n} = cn^{-\kappa}$ and $d_{p,n} = \exp\{-Cn^{1-2\kappa}\}$. Further we can show that Condition 1-(i) holds uniformly for $t \in (0, n^{1/6})$.

Proposition 3. If Condition 2 and $p = \exp\{o(n^{\min(1/3, 1-2\kappa)})\}$ hold, then the result given in Theorem 4 is valid for the RRCS procedure.

Example 3.3. Mai and Zou (2012) developed a Kolmogorov filter (KF) for variable screening in high-dimensional binary classification. For binary response Y taking values ± 1, they considered $\omega_j = \sup_{-\infty<x<\infty} |F_{+j}(x) - F_{-j}(x)|$, where $F_{+j}(x)$ and $F_{-j}(x)$ are the conditional distribution functions of X_j given $Y = 1, -1$, respectively. Correspondingly,

$$\widehat{\omega}_j = \sup_{-\infty<x<\infty} |\widehat{F}_{+j}(x) - \widehat{F}_{-j}(x)|, \quad (6)$$

where $\widehat{F}_{+j}(x)$ and $\widehat{F}_{-j}(x)$ are the empirical conditional distribution functions of X_j given $Y = 1, -1$, respectively.

Mai and Zou (2012) proved that

$$\Pr(|\widehat{\omega}_j - \omega_j| \ge cn^{-\kappa}) \le O(\exp\{-Cn^{1-2\kappa}\}) \text{ with } 0 < \kappa < 1/2.$$

Then we can still take $\delta_{p,n} = cn^{-\kappa}$ and $d_{p,n} = \exp\{-Cn^{1-2\kappa}\}$. In the supplementary material, we show that Condition 1-(i) holds uniformly for $t \in (0, n^{1/2})$. Hence, we have the following result.

Proposition 4. If Condition 2 and $p = \exp\{o(n^{1-2\kappa})\}$ hold, then the result given in Theorem 4 is valid for the Kolmogorov filter.

4. Numerical Study

We illustrate the breadth of applicability of our proposed procedure by studying its performance on simulated data and a real-data example over four different screening methods: SIS (Fan and Lv 2008), RRCS (Li et al. 2012), KF (Mai and Zou 2012), and DC-SIS (Li, Zhong, and Zhu 2012). The W_j in (2) is built when $K = 3$.

4.1. Simulation Study

The performances of the proposed REDS filter and the hard-thresholding rule with REDS ("HTR") are evaluated along with some benchmarks through the comparisons of the FDR, the proportion that all active predictors are selected, that is, $P_a = \Pr(\mathcal{A} \subseteq \mathcal{M}(L))$, the true positive rate (TPR), say the proportion of informative predictors that are correctly identified as such, and the average model size (MS). Here $\mathcal{A}$ is the set of predictors which are truly related to the response. The hard thresholding rule, which selects a set of predictors with a given number d, is considered for comparison. Following the recommendation in Fan and Lv (2008), we choose $d_1 = \lfloor n/\log n \rfloor$ and $d_2 = \lfloor 2n/\log n \rfloor$ as two simple competitors and denote them as "HT-d_1" and "HT-d_2", respectively. All results are obtained with 1000 replications when the nominal level α is 0.2.

Example 1. For the SIS method in Fan and Lv (2008), a linear regression model is set as $Y = \mathbf{X}^\top \boldsymbol{\beta} + \epsilon$. We consider the predictors $\mathbf{X}$ following $N(0, \boldsymbol{\Sigma})$ where the covariance matrix $\boldsymbol{\Sigma} = (\rho^{|i-j|})_{p \times p}$ with $\rho = 0.8$, meanwhile ϵ is from $N(0, 1)$. The $\boldsymbol{\beta}$ is a p-dimensional vector where the first s elements are set as 1 and otherwise 0, that is, $\beta_i = 1$ for $i = 1, \ldots, s$ and $\beta_i = 0$ for $i = s+1, \ldots, p$. We would vary the values of s to assess the performance. Under this model, the informative set is $\mathbb{H}_1 = \{j : w_j > b_n\}$ with $b_n = n^{-2/3}$ and its cardinality is denoted as R. Note that R will be larger than s since there exist some variables correlated with $X_1, \ldots, X_s$ but without relationship with Y in $\mathbb{H}_1$.

We fix the dimension $p = 5000$ and vary $n = 100, 200$ and $s = 15, 25$ to compare the estimated FDR, TPR, P_a and average model size in Table 1. Under all the scenarios, the REDS is able to deliver a quite accurate control and performs better than the BH in terms of FDR control. For example, when $n = 100$ and $s = 25$, the proposed REDS and BH result in the empirical FDRs as 18.5% and 24.1%, respectively, while their corresponding standard errors are 0.5% and 0.3% under 1000 replications. This implies that BH does not control FDR well in this case. This is consistent with our theoretical analysis in Section 3.1. We also observe that the FDR levels of HTR are slightly smaller than REDS's but HTR improves the TPR and P_a over the REDS. Certainly, this is not surprising as HTR uses the full-sample statistics $\widehat{\omega}_j$ for screening without information loss. The HT-d_1 has very small FDR and accordingly it performs not well in terms of P_a under $n = 100$ and $s = 25$ because of the model size d smaller than s. Meanwhile, HT-d_2 is able to yield a large TPR or P_a (as well as a large FDR). Clearly, a larger d should be used but as we argued before how large it depends on unknown quantities, and thus, such ad-hoc choice cannot ensure the sure screening property hold. More simulation results in supplementary material also illustrates that the REDS and HTR can successfully control the FDR at the nominal level.

Example 2. In this example, we consider the RRCS method and the generalized Box-Cox transformation model given by Li et al. (2012): $\lambda^{-1}\{|Y|^\lambda \text{sgn}(Y) - 1\} = \mathbf{X}^\top \boldsymbol{\beta} + \epsilon$ with $\lambda = 0.5$. Here, we consider the predictors $X_1, \ldots, X_p$ following: $\{X_j\}_{j=1}^{[p/3]}$ are from $N(0, \boldsymbol{\Sigma})$, $\{X_j\}_{j=[p/3]+1}^{[2p/3]}$ are iid from the student's $t(3)$ distribution, and $\{X_j\}_{j=[2p/3]+1}^{p}$ are iid from a Poisson distribution with mean one. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{[p/3] \times [p/3]}$ includes two blocks: $\sigma_{ij} = \rho_1$ for $i, j \le R$ and $\sigma_{ij} = \rho_2$ for $i, j > R$ and $\sigma_{ii} = 1$ for $i = 1, \ldots, p$, where $\rho_1 = 0.8$ and $\rho_2 = 0.2$. There are s nonzero elements in $\boldsymbol{\beta}$ where $\beta_1, \cdots, \beta_{s-2}$

Table 1. Comparison results of FDR(%), TPR(%), P_a(%) and the average model size under $p = 5000, n = 100, 200$ and $s = 15, 25$ in Example 1.

		$n = 100$				$n = 200$			
s	Method	FDR	TPR	P_a	MS	FDR	TPR	P_a	MS
$s = 15$	HTR	18.9	63.4	80.9	21.1	17.7	64.9	94.6	22.6
	REDS	22.6	60.4	55.8	21.1	20.6	62.4	82.2	22.6
	BH	25.1	66.7	95.9	23.1	23.0	67.3	100.0	24.3
	HT-d_1	21.4	66.1	95.4	21.0	49.5	69.2	100.0	37.0
	HT-d_2	58.6	69.5	98.2	42.0	73.8	71.7	100.0	74.0
$s = 25$	HTR	13.5	65.2	31.0	27.3	15.2	71.8	80.2	32.0
	REDS	18.5	61.1	9.5	27.3	18.6	69.0	45.5	32.0
	BH	24.1	73.9	52.5	33.9	22.4	75.0	97.4	35.5
	HT-d_1	0.5	61.4	0.0	21.0	26.9	75.2	97.5	37.0
	HT-d_2	39.1	75.3	64.1	42.0	62.2	77.7	99.5	74.0

Table 2. Comparison results of FDR(%), TPR(%), P_a(%) and the average model size under $n = 200$ and $R = 50, 100$ in Example 2.

		$p = 2000$				$p = 5000$			
s	Method	FDR	TPR	P_a	MS	FDR	TPR	P_a	MS
$R = 50$	HTR	15.9	94.9	89.4	60.1	15.6	94.5	88.5	59.7
	REDS	16.8	92.4	67.5	60.1	16.6	91.8	64.3	59.7
	BH	28.6	100.0	99.5	73.9	31.6	99.9	99.4	78.2
	HT-d_1	0.0	75.3	1.4	37.0	0.0	74.9	1.6	37.0
	HT-d_2	29.7	100.0	99.6	74.0	29.7	99.9	99.3	74.0
$R = 100$	HTR	16.3	95.2	91.4	118.1	16.2	94.7	89.9	117.7
	REDS	17.1	92.4	66.3	118.1	17.0	92.3	66.4	117.7
	BH	26.6	100.0	99.8	140.0	29.5	99.9	99.4	146.5
	HT-d_1	0.0	54.1	0.1	37.0	0.0	53.6	0.0	37.0
	HT-d_2	0.0	76.0	3.1	74.0	0.0	75.9	2.7	74.0

Table 3. Comparison results of FDR(%), TPR(%), P_a(%), the average model size, and the probabilities(%) that $X_{[p/3]+1}$ and $X_{[2p/3]+1}$ are selected under $p = 2000$ in Example 3.

(n, R)	Method	FDR	TPR	P_a	$X_{[p/3]+1}$	$X_{[2p/3]+1}$	MS
(200,50)	HTR	20.9	95.6	4.0	15.9	18.2	65.9
	BH	13.2	96.7	2.4	14.1	12.8	58.2
	HT-d_1	0.0	71.2	0.0	0.0	0.0	37.0
	HT-d_2	31.8	97.1	6.6	24.2	24.4	74.0
(200,100)	HTR	20.7	97.6	7.2	25.2	24.8	129.6
	BH	12.9	98.4	3.4	19.9	17.7	115.7
	HT-d_1	0.0	36.3	0.0	0.0	0.0	37.0
	HT-d_2	0.0	72.5	0.0	0.0	0.0	74.0
(400,50)	HTR	20.5	97.9	21.0	37.8	52.1	66.2
	BH	11.3	97.6	12.4	30.4	44.4	57.6
	HT-d_1	22.8	98.0	22.7	41.8	55.6	66.0
	HT-d_2	61.0	99.1	56.0	69.2	81.5	132.0
(400,100)	HTR	21.1	99.2	33.1	52.7	63.0	130.5
	BH	11.4	99.0	22.7	43.6	55.9	114.3
	HT-d_1	0.0	64.7	0.0	0.0	0.0	66.0
	HT-d_2	23.3	99.3	37.6	56.7	67.7	132.0

are chosen randomly from $U(0,1)$, $\beta_{[p/3]+1}, \beta_{[2p/3]+1}$ are set as 2 and otherwise zero. Under this model, we fix the $s = 10$ and notice that the set of predictors which are truly related to the response is $\mathcal{A} = \{j : 1, \ldots, 8, [p/3]+1, [2p/3]+1\}$ and there are R dependent covariates with the first nonzero $s-2$ elements. Accordingly, the informative set is $\mathbb{H}_1 = \{j : 1, \ldots, R, [p/3]+1, [2p/3]+1\}$.

Here, we fix the sample size $n = 200$ and consider the dimension $p = 2000$ or 5000. The results when $R = 50$ and 100 are shown in Table 2. We see that the FDR levels of REDS or HTR are close to the nominal level while maintains a reliable TPR under all the scenarios; it is clearly more effective than the BH and the difference is quite remarkable in terms of FDR control. For example under $p = 5000$ and $R = 50$, the empirical FDR of the proposed REDS and HTR are 16.6% and 15.6%, respectively, with the same standard error 0.3%, but the empirical FDR of the BH is clearly out of control because of its FDR level 31.6% with 0.3% standard error over 1000 simulations. The hard-thresholding method with two different d's shows that more active predictors can be detected only when d is greater than the cardinality of $\mathbb{H}_1$. This again demonstrates the effectiveness of the proposed HTR: it is a data-driven thresholding rule which allows FDR control. Similar results for other values of R are provided in the supplementary material.

Example 3. For the KF method in Mai and Zou (2012), we consider a logistic regression model, in which Y is distributed as Binomial$(1, p(x))$ with $\log\{\frac{p(x)}{1-p(x)}\} = \mathbf{X}^\top \boldsymbol{\beta}$, where $\mathbf{X}$ is generated same as Example 2, while $\beta_1, \ldots, \beta_8$ and $\beta_{[p/3]+1}$ are equal to 1, $\beta_{[2p/3]+1}$ is 2 and otherwise 0. As we have shown that the HTR performs usually better than the REDS in terms of TPR and P_a, in what follows we focus on the HTR only. Note that the Kolmogorov statistic is exactly distribution-free for all the continuous distributions and thus, we obtain the p-values for the BH procedure via simulations. Besides the measurements of FDR, TPR, P_a and the average model size, we also include the probabilities that the active variables $X_{[p/3]+1}$ and $X_{[2p/3]+1}$ are selected, which are from $t(3)$ and Poisson distribution, respectively.

Table 3 presents the comparison results under $p = 2000$ when different sample sizes and R are considered. We observe that the FDRs of the HTR method are close to the nominal level. Meanwhile, the proportion that all active predictors are selected P_a in the HTR procedure largely increases when n is large. It implies that the HTR can achieve the sure screening property when n is large. In contrast, the BH method results in overly conservative FDR levels across all the settings. This can be understood by noticing that some of the covariates in our model are Poisson distributed and thus, the null distribution of the Kolmogorov statistic obtained under a continuous distribution is a misspecified one. Compared to the proposed HTR, although the TPRs of BH are close to those of HTR, the BH method results in lower P_a and lower probabilities that the active covariates $X_{[p/3]+1}$ and $X_{[2p/3]+1}$ could be selected. As we can expect, the hard thresholding methods could result in larger TPR as well as larger FDR only when a larger d is chosen. For example, when we consider the case that $R = 50$ and $n = 200$, the HT-d_2 yields a large TPR as 97.1% with its standard deviation 1.2% but also a large FDR as 31.8% with its standard deviation 0.8% due to the selected model size $d_2 = 74$ much greater than R. Similar analysis also can be found in the previous two examples.

Example 4. In this example, we investigate the performance of the HTR for DC-SIS. The BH method is not readily applicable because approximating the null distribution of the distance correlation needs some permutation procedures which may be too time-consuming. Hence, we only display the results of HTR and the hard-thresholding methods under two following models:

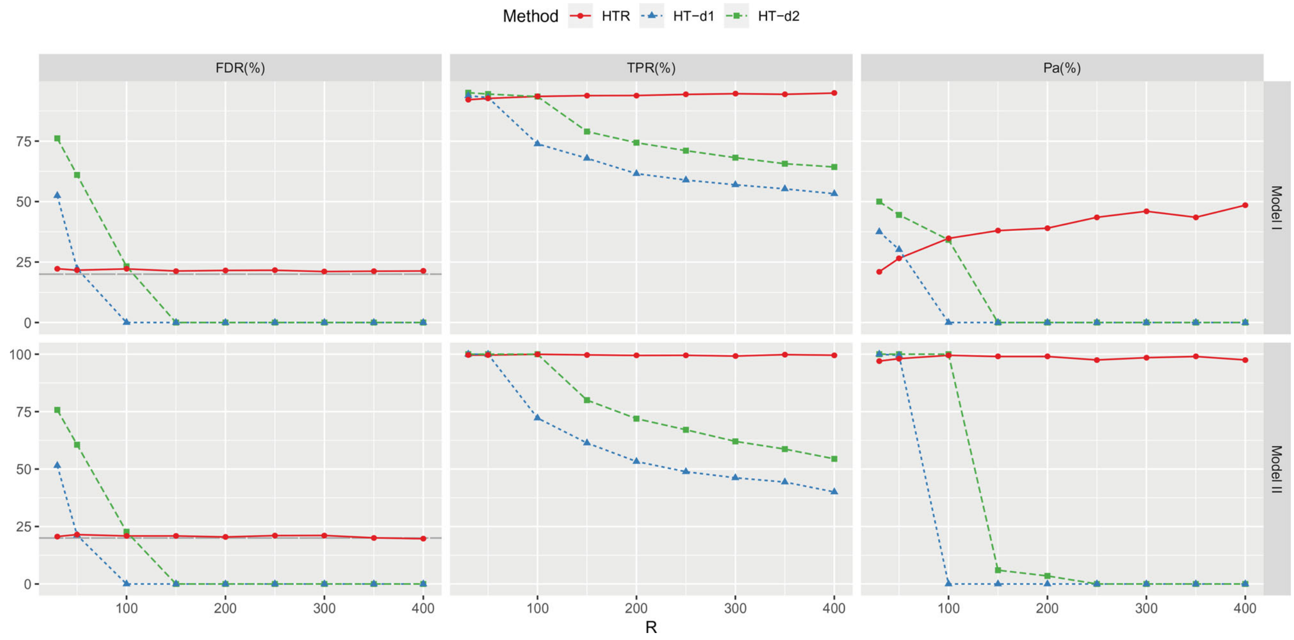

Figure 2. FDR(%), TPR(%), and P_a(%) curves against R under $n = 400$ and $p = 5000$ in Example 4; the gray dashed lines indicate the target FDR level.

- Model I (Nonlinear Regression): $Y = \beta_1 X_1 + \cdots + \beta_8 X_8 + \beta_{[p/3+1]} \sin(X_{[p/3+1]}) + \beta_{[2p/3+1]} X_{[2p/3+1]} + \epsilon$, where $\{X_j\}_{j=[p/3]+1}^{[2p/3]}$ are iid from $U(0,1)$, and the left X_j, ϵ and $\boldsymbol{\beta}$ are same as in Example 2;
- Model II (Poisson Regression): Y is distributed as Poisson $(\mu(x))$ with $\mu(x) = (\mathbf{X}^\top \boldsymbol{\beta})^2$, where $\{X_j\}_{j=[p/3]+1}^{[2p/3]}$ are iid from exp(1), and the left X_j and $\boldsymbol{\beta}$ are chosen same as in Example 2.

Figure 2 depicts the FDR, TPR and P_a curves against R when $p = 5000$ and $n = 400$. It implies that the HTR method performs reasonably well for both models. The FDR varies in an acceptable range of the target level no matter the size of R. Meanwhile, the TPR and the proportion that all active predictors are selected into the model are quite high, which clearly demonstrates the efficiency of our proposed method. As R increases, less active predictors can be detected by the hard-thresholding method with two different d's. This further indicates that the hard-thresholding method is hard to guarantee the sure screening property with an ad-hoc d.

Example 5. Here, we consider the proposed method HTR applied to the PFER control. Followed by the same model settings in Examples 1–4, we fix the dimension $p = 5000$ and the sample size $n = 400$ for four different screening methods SIS, RRCS, KF and DC-SIS, and consider $\mathbf{X}$ generated as Example 2 and the Model (I) for DC-SIS. Table 4 reports some PFER results of the HTR when the target PFER level $k_0 \in [5, 20]$. The validity of the HTR approach in terms of PFER control is clear.

Table 4. Results of PFER over 1000 replications when $k_0 = 5, 10$, and 20 under $p = 5000$ and $n = 400$ for different screening methods.

	$R = 50$			$R = 100$			$R = 200$		
Method \ k_0	5	10	20	5	10	20	5	10	20
SIS	5.7	11.2	21.7	5.7	11.2	21.8	5.6	11.0	21.2
RRCS	4.5	9.4	19.0	4.2	9.3	18.6	4.0	8.6	18.1
KF	5.1	10.4	18.6	4.8	9.8	19.0	4.4	9.4	18.6
DC-SIS	6.2	11.7	22.8	6.2	11.7	22.8	6.1	11.6	22.4

Table 5. Numbers of selected genes for HTR procedure with KF and RRCS method under different target FDR levels.

	Original			Original and Auxiliary		
Method \ α	0.05	0.1	0.2	0.05	0.1	0.2
KF	17	20	53	17(0)	23(1)	56(1)
RRCS	17	25	52	18(0)	27(0)	60(1)

NOTE: Values in parentheses are the numbers of selected auxiliary variables.

4.2. An Application

We apply the proposed HTR procedure with screening methods RRCS and KF to select features for the classification of a diffuse large-B-cell lymphoma (DLBCL-B) dataset. This data are available from *http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi*. There are $p = 661$ genes and $n = 93$ units from two classes: 42 in class 1 and 51 in class 2. For details about the DLBCL-B data, see Hoshida et al. (2007) and Qi et al. (2015).

The scatterplots of statistics W_j for RRCS and KF are presented in Figure 3 when the target FDR level is fixed as $\alpha = 0.1$. We observe that all selected genes (red dots) have large values of W_j, while the unselected ones (black dots) are roughly symmetric across the horizontal lines. Table 5 shows the numbers of selected genes with the HTR procedure under different FDR levels. As we expect, a larger FDR level could result in the selection of more genes. Given the evidence from the simulation results in previous subsections suggesting that the FDR is controlled, it is likely that most of these discovered genes (roughly $1 - \alpha$ percentage) are informative (at least marginally).

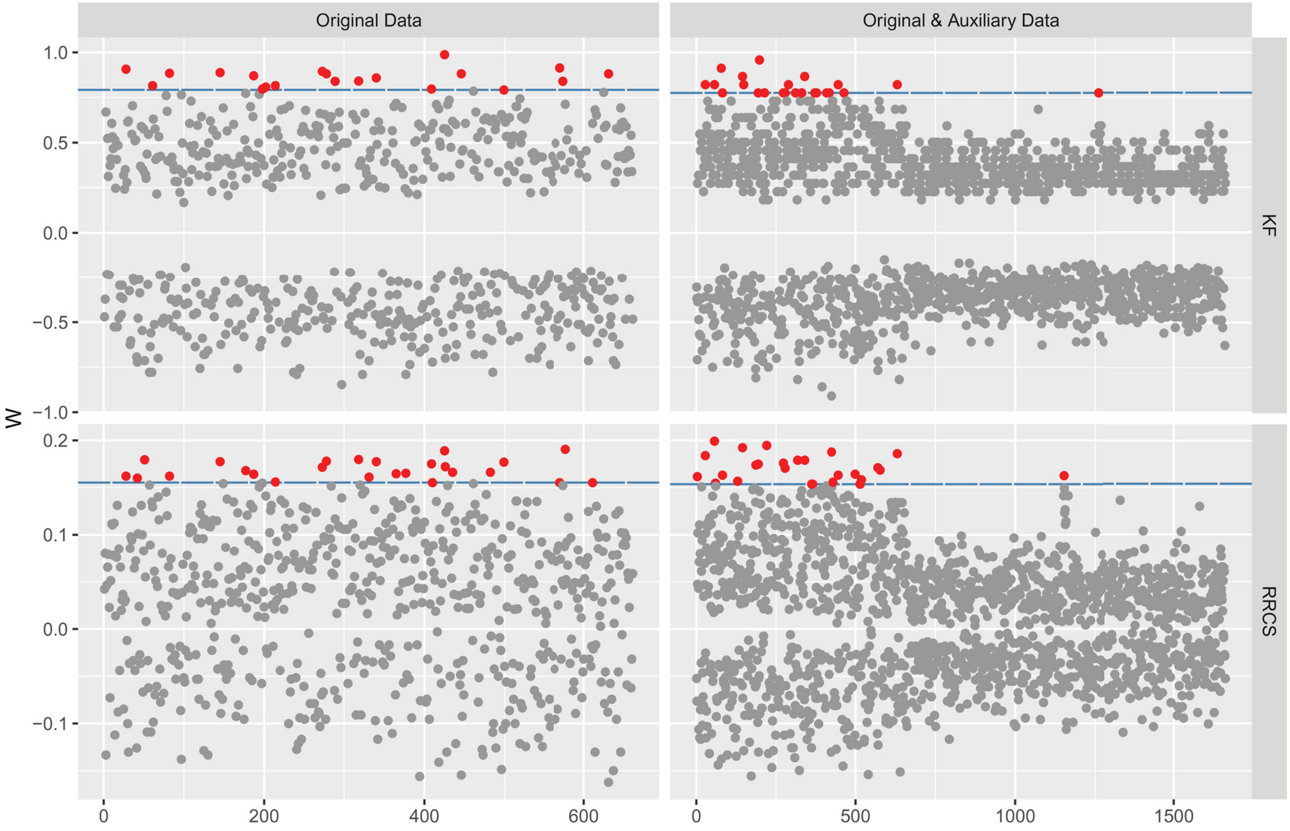

Figure 3. Scatterplots of W_j's for KF and RRCS, respectively, with the red points and blue line denoting selected predictors and the threshold under $\alpha = 0.1$.

To further evaluate the performance, we introduce some simulated variables. Specifically, we independently and randomly generate 1000 auxiliary variables Z from $N(0, 1)$ and select the informative genes from the whole 1661 variables. Since Z is truly inactive by construction, one selected auxiliary variable can be seen as a truly false discovery. Table 5 and Figure 3 show the number of selected variables and the scatterplots of resulting statistics W_j. We rank the marginal utility statistics on the full sample and find that the 19th statistic for KF method and the 29th statistic for RRCS method correspond to an auxiliary variable, respectively. In this case, the traditional hard-thresholding rule with $d_1 = \lfloor n/\log(n) \rfloor = 20$ and $d_2 = \lfloor 2n/\log(n) \rfloor = 41$ may throw away an informative gene due to the selected auxiliary variable. In contrast, our HTR procedure could adaptively decide a large threshold to avoid losing those potentially informative genes.

5. Concluding Remarks

Threshold selection is very important to conduct efficient feature screening procedures. This article proposes a data-driven threshold selection procedure, REDS, to determine the threshold for error rate control under a unified framework. The REDS procedure is easy to implement and computationally efficient. It is shown that our proposed method can asymptotically control the FDR and can still retain all of the important predictors under mild conditions, and thus, it could serve as a useful alternative to the ad-hoc hard thresholding approaches in practice.

In Section 2.1, we construct a test statistic W_j, which satisfies the marginal symmetry for $j \in \mathbb{H}_0$. In fact, an effective W_j should satisfy two main properties: (a) for $j \in \mathbb{H}_0$, W_j is (asymptotically) symmetric with mean zero; (b) for $j \in \mathbb{H}_1$, W_j is a large positive value. Any appropriate forms satisfying these two properties could be taken as suitable choices to implement the REDS method. The W_j given in (2) is the one which can be used without imposing other requirements on $\mathcal{N}_j$. For some specific $\mathcal{N}_j$, more effective and simpler surrogates are available. For instance, if $\mathcal{N}_j$ follows a symmetric distribution, such like the normal distribution as in SIS, we may take $\widetilde{W}_j = n_1^{\gamma}\widehat{\omega}_{j1} \times n_2^{\gamma}\widehat{\omega}_{j2}$. This $\widetilde{W}_j$ is also asymptotically symmetric for $j \in \mathbb{H}_0$, while is a large positive value for $j \in \mathbb{H}_1$ regardless of the signs of ω_j. To make the strength of the signal largest, we can simply take $K = 2$ in this situation. This construction is obviously applicable for the Examples 1–2 discussed in Section 3.2, since we can take ω_j as the original correlation coefficient (without absolute value) and the corresponding $\mathcal{N}_j$'s are normal distributed. Systematic investigation and comparison with different types of W_j certainly warrant future study.

Our unified framework is developed for the marginal screening methods in which the marginal correlations for the important variables must be bounded away from zero. However, sometimes this assumption can be violated, as predictors are often correlated. As a result, unimportant variables that are highly correlated with important predictors will have high priority of being selected. In such situations, the iterative SIS procedure introduced by Fan and Lv (2008) and some other related works such as Wang and Leng (2016) should be considered. It is of interest to further generalize the REDS method to the screening methods which could give effective variable screening without the strong marginal correlation assumption. In addition, it would be also important to identify interactions between predictors. For recent development, see, for instance, Fan et al. (2015) and Kong et al. (2017). It is of great interest to investigate how to adapt the proposed method to interaction screening problems. This would be an interesting topic for future research.

Appendix: Proofs

This appendix contains the proofs of Theorems 1, 3–4 and Corollary 1 in which the technical arguments may be interesting in their own rights. The proofs of Theorem 2 and Propositions 1–4 can be found in the supplementary material, along with a few additional lemmas.

Before we present the proof of Theorem 1, we first state a lemma whose proof is given in the supplementary material. Denote $W_{1j} = n_1^{\gamma}\widehat{\omega}_{j1}$ and $W_{2j} = n_2^{\gamma}\widehat{\omega}_{j2}$, then $W_j = \operatorname{sgn}\left(W_{1j} - W_{2j}\right)\left(W_{1j} \vee W_{2j}\right)$. The next lemma characterizes the closeness between $\Pr(W_j > t)$ and $\Pr(W_j < -t)$, which plays an important role in the proof.

Lemma A.1. Suppose Condition 1-(i) hold. For $t \to \infty$,

$$\left|\frac{\Pr(W_j \geq t)}{\Pr(W_j \leq -t)} - 1\right| = o_p(1),$$

uniformly in $j \in \mathbb{H}_0$ and t.

Proof of Theorem 1. We need to show that Equation (4) holds. Lemmas A.1 and Condition 2 serve this purpose.

By definition, our thresholding rule is equivalent to select the jth feature if $W_j > L$, where

$$L = \inf\left\{t \geq 0 : 1 + \sum_j \mathbb{I}(W_j < -t) \leq \alpha \max\left(\sum_j \mathbb{I}(W_j > t), 1\right)\right\}.$$

We need to establish an asymptotic bound for this L so that the conditions in the theorem hold.

We first show that for any $j \in \mathcal{C}_\beta$, $n_1^{\gamma}\widehat{\omega}_{1j} > n_2^{\gamma}\widehat{\omega}_{2j}$ with probability tending to one. In fact, we have

$$\begin{aligned}
&\Pr\left(n_1^{\gamma}\widehat{\omega}_{1j} < n_2^{\gamma}\widehat{\omega}_{2j} \text{ for some } j \in \mathcal{C}_\beta\right)\\
&\leq \beta_{p,n}\Pr\left(\omega_j < |\widehat{\omega}_{2j} - \omega_j|/(a-1) + a|\widehat{\omega}_{1j} - \omega_j|/(a-1)\right) \to 0
\end{aligned}$$

since $\beta_{p,n}d_{p,n} \to 0$ and $\omega_j > (a+1)\delta_{p,n}/(a-1)$ for all $j \in \mathcal{C}_\beta$. Here $a = (K-1)^{\gamma}$.

Similarly it follows that

$$\begin{aligned}
&\Pr\left(W_j < T_U \text{ for some } j \in \mathcal{C}_\beta\right)\\
&\leq \beta_{p,n}\Pr\left(\omega_j < n_1^{-\gamma}T_U + |\widehat{\omega}_{1j} - \omega_{1j}|\right) \to 0,
\end{aligned}$$

since $\beta_{p,n}d_{p,n} \to 0$ and $\omega_j > n_1^{-\gamma}T_U + \delta_{p,n}$ for all $j \in \mathcal{C}_\beta$.

Thus, we conclude that $\Pr(\sum_j \mathbb{I}(W_j > T_U) \geq \beta_{p,n}) \to 1$. Next note that

$$\begin{aligned}
&\Pr\left(W_j < -T_U \text{ for some } j \in \mathbb{H}_0\right)\\
&\leq p_0\Pr\left(n_1^{\gamma}\widehat{\omega}_{1j} < -T_U, n_1^{\gamma}\widehat{\omega}_{1j} \geq n_2^{\gamma}\widehat{\omega}_{2j}\right)\\
&+ p_0\Pr\left(-n_2^{\gamma}\widehat{\omega}_{2j} < -T_U, n_1^{\gamma}\widehat{\omega}_{1j} < n_2^{\gamma}\widehat{\omega}_{2j}\right)\\
&\leq p_0\Pr\left(|\widehat{\omega}_{1j}| > n_1^{-\gamma}T_U\right) + p_0\Pr\left(|\widehat{\omega}_{2j}| > n_2^{-\gamma}T_U\right) \to 0
\end{aligned}$$

provided that $p_0 d_{p,n} \to 0$. Thus, $\Pr(\sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j < -T_U) = 0) \to 1$.

Let $V_U =: G_-^{-1}(\alpha\beta_{p,n}/p)$. From Condition 2, it follows that

$$\frac{\alpha\beta_{p,n}}{p} = G_-(V_U) = \frac{1}{p_0}\sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j < -V_U)(1 + o(1)).$$

Then we have $V_U \lesssim T_U$ and $\Pr(\sum_j \mathbb{I}(W_j > V_U) \geq \beta_{p,n}) \to 1$. Further we conclude that

$$1 + \sum_j \mathbb{I}(W_j < -V_U) \lesssim \frac{\alpha\beta_{p,n}p_0}{p} \leq \alpha\sum_j \mathbb{I}(W_j > V_U).$$

Thus, we get $L \lesssim V_U$.

On the other hand, notice that

$$\sum_j \mathbb{I}(W_j < -t) \geq \sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j < -t) \approx p_0 G_-(t),$$

$$\alpha\sum_j \mathbb{I}(W_j > t) \leq \alpha\left\{p_1 + \sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j > t)\right\} \approx \alpha\left\{p_1 + p_0 G_+(t)\right\}.$$

Hence, we can conclude that $L \to \infty$ for any fixed $\alpha \in (0,1)$ because $p_1 = o(p)$ as $(p,n) \to \infty$.

Therefore, it follows by Lemma A.1 and Condition 2 that

$$\frac{\sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j > L)}{\sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j < -L)} = 1 + o_p(1). \tag{A.1}$$

Accordingly, we have

$$\begin{aligned}
\text{FDP} = \frac{\sum_{j\in\mathbb{H}_0}\mathbb{I}\left(W_j \geq L\right)}{1 \vee \sum_j \mathbb{I}(W_j \geq L)} &= \frac{1 + \sum_j \mathbb{I}\left(W_j \leq -L\right)}{1 \vee \sum_j \mathbb{I}(W_j \geq L)}\\
&\times \frac{\sum_{j\in\mathbb{H}_0}\mathbb{I}\left(W_j \geq L\right)}{1 + \sum_j \mathbb{I}\left(W_j \leq -L\right)}\\
&\leq \alpha \times R(L).
\end{aligned}$$

Note that $R(L) \leq \sum_{j\in\mathbb{H}_0}\mathbb{I}\left(W_j \geq L\right)/\sum_{j\in\mathbb{H}_0}\mathbb{I}\left(W_j \leq -L\right)$, and thus, $\limsup_{n\to\infty}\text{FDP} \leq \alpha$ in probability by (A.1). Then, for any $\epsilon > 0$,

$$\text{FDR} \leq (1+\epsilon)\alpha R(L) + \Pr\left(\text{FDP} \geq (1+\epsilon)\alpha R(L)\right),$$

from which the second part of this theorem is proved. □

Proof of Theorem 3. We only show the validity of the first formula and the second one hold similarly. We suppress "+" in $G_+(t)$ for notational simplicity. Note that the $G(t)$ is a deceasing and continuous function. Let $a_p = \alpha\beta_{p,n}$, $z_0 < z_1 < \cdots < z_{h_p} \leq 1$ and $t_i = G^{-1}(z_i)$, where $z_0 = a_p/p$, $z_i = a_p/p + b_p\exp(i^{\delta})/p$, $h_p = \{\log((p - a_p)/b_p)\}^{1/\delta}$ with $0 < \delta < 1$ and $b_p/a_p \to 0$. Note that $G(t_i)/G(t_{i+1}) = 1 + o(1)$ uniformly in i. It is therefore, enough to obtain the convergence rate of

$$D_p = \sup_{0\leq i\leq h_p}\left|\frac{\sum_{j\in\mathbb{H}_0}\left\{\mathbb{I}(W_j > t_i) - \Pr(W_j > t_i)\right\}}{p_0 G(t_i)}\right|.$$

Define $\mathcal{M}_j = \{k \in \mathbb{H}_0 :\ W_k \text{ is dependent with } W_j\}$ and further

$$D(t) = \mathbb{E}\left[\left(\sum_{j\in\mathbb{H}_0}\left\{\mathbb{I}(W_j > t) - \Pr(W_j > t)\right\}\right)^2\right].$$

It is noted that under the condition (i)

$$\begin{aligned}
D(t) = \sum_{j\in\mathbb{H}_0}\sum_{k\in\mathcal{M}_j}&\mathbb{E}\left[\left\{\mathbb{I}(W_j > t) - \Pr(W_j > t)\right\}\right.\\
&\left.\left\{\mathbb{I}(W_k > t) - \Pr(W_k > t)\right\}\right] \leq r_{p,n}p_0 G(t).
\end{aligned}$$

Thus, we can have

$$\Pr(D_p \geq \epsilon) \leq \sum_{i=0}^{h_p} \Pr\left(\left|\frac{\sum_{j\in\mathbb{H}_0}[\mathbb{I}(W_j > t_i) - \Pr(W_j > t_i)]}{p_0 G(t_i)}\right| \geq \epsilon\right)$$
$$\leq \frac{1}{\epsilon^2}\sum_{i=0}^{h_p}\frac{1}{p_0^2 G^2(t_i)} D(t_i) \leq \frac{c r_{p,n}}{\epsilon^2}\sum_{i=0}^{h_p}\frac{1}{p_0 G(t_i)}.$$

Moreover, observe that

$$\sum_{i=0}^{h_p}\frac{1}{p_0 G(t_i)} = \frac{p}{p_0}\left(\frac{1}{a_p} + \sum_{i=1}^{h_p}\frac{1}{a_p + b_p e^{i^\delta}}\right)$$
$$\leq c\left(\frac{1}{b_p} + b_p^{-1}\sum_{i=1}^{h_p}\frac{1}{1+e^{i^\delta}}\right) \leq c b_p^{-1}\{1+o(1)\}.$$

Because b_p can be made arbitrarily large as long as $b_p/a_p \to 0$, we have $\Pr(D_p \geq \epsilon) \to 0$ if $r_{p,n}/\beta_{p,n} \to 0$.

Similarly, for the condition (ii), we also have

$$\Pr(D_p \geq \epsilon) \leq \frac{1}{\epsilon^2}\sum_{i=0}^{h_p}\frac{1}{p_0^2 G^2(t_i)} D(t_i) \leq \frac{C(\log p)^{-1-\theta} h_p}{\epsilon^2}.$$

As a result, we obtain $\Pr(D_p \geq \epsilon) \to 0$ by taking $(1+\theta)^{-1} < \delta < 1$.
□

Proof of Theorem 4. Denote $m = |\mathcal{M}(L)|$. The FDP of the hard-thresholding rule with REDS is

$$\mathrm{FDP}_{\mathrm{HTR}} = \frac{\sum_{j\in\mathbb{H}_0}\mathbb{I}\left(j \in \mathcal{S}(\widehat{\omega}_{(p-|\mathcal{M}(L)|+1)})\right)}{\max\{m,1\}}.$$

By the proof of Theorem 1, $\mathrm{FDP}(L) \leq \alpha + o_p(1)$ and $\Pr(\sum_j \mathbb{I}(W_j > L) \geq \beta_{p,n}) \to 1$, we claim that with probability tending to one, $m > 1$ and $\sum_{j\in\mathbb{H}_1}\mathbb{I}(W_j > L) \geq m(1-\alpha)$.

On one hand, recall that $n_1 = n(K-1)/K$. Let $c = (K-1)^\gamma/K^\gamma < 1$. Then

$$\Pr\left(n^\gamma\widehat{\omega}_j < n_1^\gamma\widehat{\omega}_{1j} \text{ for some } j \in \mathbb{H}_1\right)$$
$$\leq p_1\Pr(\widehat{\omega}_j < c\widehat{\omega}_{1j}) = p_1\Pr((1-c)\omega_j < c(\widehat{\omega}_{1j}-\omega_j) - (\widehat{\omega}_j - \omega_j))$$
$$\leq p_1\Pr(\omega_j < c|\widehat{\omega}_{1j}-\omega_j|/(1-c) + |\widehat{\omega}_j-\omega_j|/(1-c))$$
$$\leq p_1\Pr(\omega_j < c|\widehat{\omega}_{1j}-\omega_j|/(1-c) + |\widehat{\omega}_j-\omega_j|/(1-c), |\widehat{\omega}_{1j}-\omega_j|$$
$$> \delta_{p,n}, or, |\widehat{\omega}_{1j}-\omega_j| > \delta_{p,n})$$
$$+ p_1\Pr(\omega_j < (1+c)\delta_{p,n}/(1-c)) \leq p_1 d_{p,n} \to 0.$$

provided that $p_1 d_{p,n} \to 0$ and $\omega_j > (c+1)\delta_{p,n}/(1-c)$. Thus, with probability tending to one, it follows that

$$\sum_{j\in\mathbb{H}_1}\mathbb{I}(n^\gamma\widehat{\omega}_j > L) \geq m(1-\alpha). \tag{A.2}$$

On the other hand, as shown in the proof of Lemma A.1,

$$\Pr(W_j > t) = \tilde{F}_{\mathcal{N}_j}(t)(1+o(1)).$$

uniformly in $t \to \infty$ and j. Hence,

$$\sum_{j\in\mathbb{H}_0}\mathbb{I}(n^\gamma\widehat{\omega}_j > L) \approx \sum_{j\in\mathbb{H}_0}\mathbb{I}(W_j > L) \lesssim m\alpha. \tag{A.3}$$

Combining (A.2) and (A.3), we get if $m > \sum_j \mathbb{I}(n^\gamma\widehat{\omega}_j > L)$,

$$\mathrm{FDP}_{\mathrm{HTR}} \lesssim \frac{m - \sum_{j\in\mathbb{H}_1}\mathbb{I}(n^\gamma\widehat{\omega}_j > L)}{m} \lesssim \alpha,$$

while if $m \leq \sum_j \mathbb{I}(n^\gamma\widehat{\omega}_j > L)$

$$\mathrm{FDP}_{\mathrm{HTR}} \leq \frac{\sum_{j\in\mathbb{H}_0}\mathbb{I}(n^\gamma\widehat{\omega}_j > L)}{m} \lesssim \alpha,$$

from which the assertion holds. □

Proof of Corollary 1. There exists some L^* such that selecting largest $|\mathcal{M}(L)|$ values of $\widehat{\omega}_j$ is equivalent to finding $\{j : n^\gamma\widehat{\omega}_j > L^*\}$. By the condition that $\limsup_{(n,p)\to\infty}\mathrm{FDP}_{\mathrm{HTR}} \geq \alpha_0$ in probability, $\sum_{j\in\mathbb{H}_0}\mathbb{I}(n^\gamma\widehat{\omega}_j > L^*) \geq m\alpha_0$. Because $m > \beta_{p,n}$ with probability tending to one as shown in Theorem 1, we know that $\sum_{j\in\mathbb{H}_0}\mathbb{I}(n^\gamma\widehat{\omega}_j > L^*) \geq \alpha_0\beta_{p,n}$.

Recall that $T_U = n_1^\gamma\delta_{p,n}$ and $c = (K-1)^\gamma/K^\gamma$. Further note that

$$\Pr\left(n^\gamma\widehat{\omega}_j > T_U/c, \text{ for some } j \in \mathbb{H}_0\right)$$
$$\leq p_0\Pr(|\widehat{\omega}_j| > \delta_{p,n}) \to 0.$$

Thus, $\Pr(\sum_{j\in\mathbb{H}_0}\mathbb{I}(n^\gamma\widehat{\omega}_j > T_U/c) = 0) \to 1$.

This then implies that as long as $\alpha_0\beta_{p,n} \geq 1$, we have $L^* \lesssim T_U/c$.
Hence,

$$\Pr(\mathcal{S}(m) \not\supseteq \mathbb{H}_1)$$
$$\leq \Pr\left(n^\gamma\widehat{\omega}_j < L^*, \text{ for some } j \in \mathbb{H}_1\right)$$
$$\leq |\mathbb{H}_1|\Pr(\widehat{\omega}_j - \omega_j < n^{-\gamma}L^* - \omega_j)$$
$$\leq |\mathbb{H}_1|\Pr(\omega_j < n^{-\gamma}L^* + |\widehat{\omega}_j - \omega_j|)$$
$$\leq |\mathbb{H}_1|\Pr(\omega_j < n^{-\gamma}T_U/c + |\widehat{\omega}_j - \omega_j|)$$
$$\leq |\mathbb{H}_1|\Pr\left(\omega_j < n^{-\gamma}T_U/c + |\widehat{\omega}_j - \omega_j|; |\widehat{\omega}_j - \omega_j| \leq \delta_{p,n}\right)$$
$$+ |\mathbb{H}_1|\Pr\left(|\widehat{\omega}_j - \omega_j| > \delta_{p,n}\right)$$
$$\leq |\mathbb{H}_1| d_{p,n} \to 0,$$

from which the results hold. □

Supplementary Materials

This supplementary material contains the proofs of some technical lemmas and corollaries, and additional simulation results.

Acknowledgments

The authors thank the Editor, Associate Editor and anonymous referees for their many helpful comments that have resulted in significant improvements in the article.

Funding

Guo was supported by Beijing NSF No. 1212004 and NNSF of China Grant No. 12071038. Ren was supported by NNSF of China Grant No. 1210010142 and Shanghai Sailing Program. Zou was supported by NNSF of China Grants No. 11925106, 11931001, 11690015, 11971247 and NSF of Tianjin 18JCJQJC46000. Li's research was supported by National Science Foundation grants DMS 1820702, DMS 1953196 and DMS 2015539.

ORCID

Runze Li http://orcid.org/0000-0002-0154-2202

References

Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [3]

——— (2019), "A Knockoff Filter for High-Dimensional Selective Inference," *The Annals of Statistics*, 47, 2504–2537. [3]

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020), "Robust Inference with Knockoffs," *The Annals of Statistics*, 48, 1409–1431. [3]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [2]

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: 'model-X' Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [3]

Chen, Z., Fan, J., and Li, R. (2018), "Error Variance Estimation in Ultrahigh-Dimensional Additive Models," *Journal of the American Statistical Association*, 113, 315–327. [4]

Chudik, A., Kapetanios, G., and Pesaran, M. H. (2018), "A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models," *Econometrica*, 86, 1479–1512. [2]

Fan, J., Guo, S., and Hao, N. (2012a), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 74, 37–65. [4]

Fan, J., Han, F., and Liu, H. (2014), "Challenges of Big Data Analysis," *National Science Review*, 1, 293–314. [1]

Fan, J., Han, X., and Gu, W. (2012b), "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence," *Journal of the American Statistical Association*, 107, 1019–1035. [6]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [1,2,3,5,6,7,10]

——— (2018), "Sure Independence Screening," *Wiley StatsRef: Statistics Reference Online*. New Jersey, NJ: Wiley. [1]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 2013–2038. [2]

Fan, Y., Demirkaya, E., Li, G., and Lv, J. (2020a), "RANK: Large-Scale Inference With Graphical Nonlinear Knockoffs," *Journal of the American Statistical Association*, 115, 362–379. [2,3]

Fan, Y., Kong, Y., Li, D., and Zheng, Z. (2015), "Innovated Interaction Screening for High-Dimensional Nonlinear Classification," *The Annals of Statistics*, 43, 1243–1272. [11]

Fan, Y., Lv, J., Sharifvaghefi, M., and Uematsu, Y. (2020b), "IPAD: Stable Interpretable Forecasting with Knockoffs Inference," *Journal of the American Statistical Association*, 115, 1822–1834. [3]

Hao, N., and Zhang, H. H. (2017), "Oracle P-values and Variable Screening," *Electronic Journal of Statistics*, 11, 3251–3271. [2]

Hoshida, Y., Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2007), "Subclass Mapping: Identifying Common Subtypes in Independent Disease Data Sets," *PloS One*, 2, e1195. [9]

Janson, L. and Su, W. (2016), "Familywise Error Rate Control via Knockoffs," *Electronic Journal of Statistics*, 10, 960–975. [3]

Kong, Y., Li, D., Fan, Y., and Lv, J. (2017), "Interaction Pursuit in High-Dimensional Multi-Response Regression via Distance Correlation," *The Annals of Statistics*, 45, 897–922. [11]

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [2,7]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [2,5,7]

Liu, W., and Shao, Q.-M. (2014), "Phase Transition and Regularized Bootstrap in Large-Scale *t*-Tests with False Discovery Rate Control," *The Annals of Statistics*, 42, 2003–2025. [5]

Mai, Q., and Zou, H. (2012), "The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification," *Biometrika*, 100, 229–234. [2,7,8]

McKeague, I. W., and Qian, M. (2015), "An Adaptive Resampling Test for Detecting the Presence of Significant Predictors," *Journal of the American Statistical Association*, 110, 1422–1433. [3]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "P-values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [4]

Merlevède, F., Peligrad, M., Rio, E. (2009), "Bernstein Inequality and Moderate Deviations Under Strong Mixing Conditions," in *High Dimensional Probability V: The Luminy Vvolume*, eds. C. Houdré, V. Koltchinskii, D. M. Mason., and M. Peligrad, Institute of Mathematical Statistics Institute of Mathematical Statistics Collections, pp. 273–292. [6]

Pan, R., Wang, H., and Li, R. (2016), "Ultrahigh-Dimensional Multiclass Linear Discriminant Analysis by Pairwise Sure Independence Screening," *Journal of the American Statistical Association*, 111, 169–179. [1]

Qi, X., Luo, R., Carroll, R. J., and Zhao, H. (2015), "Sparse Regression by Projection and Sparse Discriminant Analysis," *Journal of Computational and Graphical Statistics*, 24, 416–438. [9]

Romano, J. P. and Wolf, M. (2007), "Control of Generalized Error Rates in Multiple Testing," *The Annals of Statistics*, 35, 1378–1408. [3]

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Series B, 66, 187–205. [4,5]

Wang, H. J., McKeague, I. W., and Qian, M. (2018), "Testing for Marginal Linear Effects in Quantile Regression," *Journal of the Royal Statistical Society*, Series B, 80, 433–452. [3]

Wang, X., and Leng, C. (2016), "High Dimensional Ordinary Least Squares Projection for Screening Variables," *Journal of the Royal Statistical Society*, Series B, 78, 589–611. [10]

Wasserman, L. and Roeder, K. (2009), "High Dimensional Variable Selection," *Annals of statistics*, 37, 2178–2201. [4]

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. [1,2,6]