Journal of the American Statistical Association

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

RaSE: A Variable Screening Framework via Random Subspace Ensembles

Ye Tian & Yang Feng

To cite this article: Ye Tian & Yang Feng (2021): RaSE: A Variable Screening Framework via Random Subspace Ensembles, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1938084

To link to this article: https://doi.org/10.1080/01621459.2021.1938084

RaSE: A Variable Screening Framework via Random Subspace Ensembles

Ye Tian

Department of Statistics, Columbia University

and

Yang Feng

Department of Biostatistics, School of Global Public Health, New York University

Corresponding author Yang Feng yang.feng@nyu.edu

Abstract

Variable screening methods have been shown to be effective in dimension reduction under the ultra-high dimensional setting. Most existing screening methods are designed to rank the predictors according to their individual contributions to the response. As a result, variables that are marginally independent but jointly dependent with the response could be missed. In this work, we propose a new framework for variable screening, *Random Subspace Ensemble (RaSE)*, which works by evaluating the quality of random subspaces that may cover multiple predictors. This new screening framework can be naturally combined with any subspace evaluation criterion, which leads to an array of screening methods. The framework is capable to identify signals with no marginal effect or with high-order interaction effects. It is shown to enjoy the sure screening property and rank consistency. We also develop an iterative version of RaSE screening with theoretical support. Extensive simulation studies and real-data analysis show the effectiveness of the new screening framework.

Keywords: Variable screening; Random subspace method; Ensemble learning; Sure screening property; Rank consistency; High dimensional data; Variable selection

1 Introduction

With the rapid advancement of computing power and technology, high-dimensional data become ubiquitous in many disciplines such as genomics, image analysis, and tomography. With high-dimensional data, the number of variables p could be much larger than the sample size n. What makes statistical inference possible is the sparsity assumption, which assumes only a few variables have contributions to the response. Under this sparsity assumption, there has been a rich literature on the topic of variable selection, including LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and MCP (Zhang, 2010). Despite the success of these methods in many applications, for the ultra-high dimensional scenario where the dimension p grows exponentially with n, they may not work well due to the "curse of dimensionality" in terms of simultaneous challenges to computational expediency, statistical accuracy, and algorithmic stability (Fan et al., 2009).

To conquer these difficulties, Fan and Lv (2008) proposed a novel procedure called *sure independence screening* (SIS) with solid theoretical support. In the past decade, the power of feature screening has been well recognized and a myriad of screening methods have been proposed. The existing screening methods can be broadly classified into two categories, model-based methods and model-free ones. Model-based screening methods rely on specific models, such as SIS (Fan and Lv, 2008) and its extensions to generalized linear models (Fan et al., 2009), Cox model (Fan et al., 2010; Zhao and Li, 2012), non-parametric independence screening method based on additive models (Fan et al., 2011; Cheng et al., 2014) and screening via high-dimensional ordinary least-square projection (HOLP) (Wang and Leng, 2016). Recently, model-free approaches become more popular because of less stringent requirements. Examples of such approaches include the sure independent ranking and screening (SIRS) (Zhu et al., 2011), the screening method based on distance correlation (DC-SIS) and its iterative version (Li et al., 2012; Zhong and Zhu, 2015), screening procedure via martingale difference correlation (MDC-SIS) (Shao and Zhang, 2014), screening via Kolmogorov filter (Mai and Zou, 2013, 2015), the screening approach for discriminant analysis (MV-SIS) (Cui et al., 2015), interaction pursuit via Pearson correlation (IP) and the distance correlation (IPDC) (Fan et al., 2016; Kong et al., 2017), the screening method based

on ball correlation (Pan et al., 2018), the nonparametric screening under conditional strictly convex loss (Han, 2019), and the screening method via covariate information number (CIS) (Nandy et al., 2020).

For variables that are marginally independent but jointly dependent with the response, many existing screening methods could miss them. This issue has been recognized in the literature (Fan and Lv, 2008; Fan et al., 2009; Zhu et al., 2011; Zhong and Zhu, 2015) and iterative screening procedures were developed, which were shown to be effective empirically. However, to the best of our knowledge, there is not much theoretical development for the iterative screening methods. In addition, some iterative screening methods (e.g. iterative SIS) are coupled with a variable selection method like LASSO or SCAD, making its performance dependent on the specific choice of the regularization method. Besides, some other iterative screening methods (e.g. iterative SIRS and iterative DC-SIS) recruit variables step by step through residuals until a pre-specified number of variables are picked. Thus, their success hinges on a key tuning parameter, that is, how many variables to recruit in each step, making these procedures potentially less robust.

These issues mentioned above motivate us to propose a new screening framework which goes beyond marginal utilities. In the new framework, we investigate multiple features at the same time, via the random subspace method (Ho, 1998). Tian and Feng (2021) proposed a new *Random Subspace Ensemble* classification method based on a similar idea, *RaSE*, according to a specific aggregation framework first introduced in Cannings and Samworth (2017). They advocated applying RaSE on sparse classification problems. The main idea of RaSE can be simply described as follows. First, B_1B_2 random subspaces are generated from a specific distribution on subspaces, which are evenly divided into B_1 groups. Next, the best subspace within each group is picked according to some criterion and a base learner is trained in that subspace. Hence we obtain B_1 base learners, each of which corresponds to a subspace. Finally, these B_1 base learners are aggregated on average and the ensemble will be used in prediction. The vanilla RaSE algorithm is reviewed in Algorithm 3 in Appendix A.1. It's important to note that there is a by-product of RaSE, which is the selected proportion of each variable within B_1 selected subspaces. In

this work, we will use this selected proportion to do variable screening, and call this the *RaSE screening* framework.

We highlight the merits of RaSE screening framework as follows. First, by looking at different feature subspaces, variables marginally independent but jointly dependent with the response can be identified. Second, instead of proposing only a single screening approach, the flexible framework of RaSE allows us to use any criterion function for comparing subspaces, leading to an array of screening methods. One possible way to construct such a criterion function is to choose a base learner and a specific measure for comparing the subspaces. For example, if we know linear methods are suitable for the data, then we can apply RaSE by picking subspaces achieving lower BIC under linear models. If *k*-nearest neighbor (*k*NN) is believed to perform better, we can apply RaSE by choosing subspaces with the smallest cross-validation error on *k*NN. Third, under general conditions, we show the sure screening property and rank consistency for RaSE screening framework. Finally, we develop a novel iterative RaSE screening framework with sure screening property established without the need to use a variable selection step or specifying the number of variables to recruit in each step.

The rest of this paper is organized as follows. Section 2 introduces the vanilla RaSE screening framework and its iterative version in detail, and discusses the relationship between RaSE and marginal screening methods. In Section 3, we present the theoretical properties for vanilla RaSE and iterative RaSE screening, including sure screening property and rank consistency. In Section 4, extensive simulation studies and real-data analysis are conducted to demonstrate the power of our new screening framework. We summarize our contributions and point out some promising future avenues in Section 5. The supplementary materials include all the technical proofs as well as additional details.

2 RaSE: A General Variable Screening Framework

In what follows, we consider predictors $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ and response y. For regression problems, y takes value from the real line $\mathbb{R}$, while for classification problems, y takes value from an integer set $\{1, \ldots, K\}$, where $K > 1$ is a known

integer. Denote the training data as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Denote by $S_{\text{Full}} = \{1, \ldots, p\}$ the full feature set. The signal set $S^* \subseteq S_{\text{Full}}$ is defined as the set S with minimal cardinality satisfying $y \mid \boldsymbol{x}_S \perp\!\!\!\perp \boldsymbol{x}_{S_{\text{Full}} \setminus S}$. Denote $p^* = |S^*|$. $[a]$ is used to represent the largest integer no larger than a.

To introduce the RaSE framework, we denote the $B_1 B_2$ random subspaces as $\{S_{b_1 b_2}, b_1 = 1 \ldots, B_1, b_2 = 1 \ldots, B_2\}$, the b_1-th group of subspaces as $\{S_{b_1 b_2}\}_{b_2=1}^{B_2}$, and the selected B_1 subspaces as $\{S_{b_1 *}\}_{b_1=1}^{B_1}$. The objective function corresponding to the specific criterion to choose subspaces is written as $\text{Cr}_n : \mathcal{S} \to \mathbb{R}$, where $\mathcal{S}$ is the collection of all subspaces. Assume a smaller value of Cr_n leads to a better subspace. Although the original RaSE (Tian and Feng, 2021) was introduced to solve classification problems, we now consider the general prediction framework, including both classification and regression.

2.1 Vanilla RaSE screening framework

Following the idea of Tian and Feng (2021), we use the proportion of each feature among the selected B_1 subspaces as the importance measure. Therefore, a natural screening procedure is to rank variables based on this proportion vector, then pick the variables with the largest proportions. The RaSE screening framework is summarized in Algorithm 1.

Algorithm 1: Vanilla RaSE screening

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, subspace distribution $\mathcal{D}$, criterion function Cr_n, integers B_1 and B_2, number of variables N to select

Output: the selected proportion of each feature $\boldsymbol{\eta}$, the selected subset $\hat{S}$

1 Independently generate random subspaces $S_{b_1 b_2} \sim \mathcal{D}, 1 \le b_1 \le B_1, 1 \le b_2 \le B_2$

2 **for** $b_1 \leftarrow 1$ *to* B_1 **do**

3 Select the optimal subspace $S_{b_1*} = S_{b_1 b_2^*}$, where $b_2^* = \arg\min_{1 \le b_2 \le B_2} \mathrm{Cr}_n(S_{b_1 b_2})$

4 end

5 Output the selected proportion of each feature $\boldsymbol{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)^T$ where

$$\hat{\eta}_j = B_1^{-1} \sum_{b_1=1}^{B_1} 1(j \in S_{b_1*}), j = 1, \ldots, p$$

6 Output $\hat{S} = \{1 \le j \le p : \hat{\eta}_j \text{ is among the } N \text{ largest of all}\}$

In the algorithm, the subspace distribution $\mathcal{D}$ is chosen as a *hierarchical uniform distribution* over the subspaces by default. Specifically, with D as the upper bound of the subspace size, we first generate the subspace size d from the uniform distribution over $\{1, \ldots, D\}$. Then, the subspace S_{11} follows the uniform distribution over all size-d subspaces $\{S \subseteq S_{\text{Full}} : |S| = d\}$. In practice, the subspace distribution can be adjusted if we have prior information about the data structure.

Algorithm 1 is not the end of the story because it ranks all the variables but does not determine how many variables to keep. To facilitate the theoretical analysis, we define the final feature subset to be selected as

$$\hat{S}_\alpha = \{1 \le j \le p : \hat{\eta}_j \text{ is among the } [\alpha D / c_{2n}] \text{ largest of all}\}, \quad (1)$$

where c_{2n} is a constant (to be specified in the next section) depending on n, B_2, D, and the criterion Cr which is a population counterpart of Cr_n. Here, α can be any constant larger than 1, which will appear in the upper bound introduced in the sure screening theorem of Section 3.

2.2 Iterative RaSE screening

As we mentioned in the introduction, the existing iterative screening methods have various tuning components such as the number of variables to recruit in each step

and/or a specific variable selection method. We propose the iterative RaSE screening in Algorithm 2 to tackle these issues.

The main idea of iterative RaSE screening is to update the subspace distribution based on the selected proportion in the preceding steps and not to conduct variable screening until the final step. To understand the details in the algorithm, we introduce a new subspace distribution.

Algorithm 2: Iterative RaSE screening (RaSE_T)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, initial subspace distribution $\mathcal{D}^{[0]}$, criterion function Cr_n, integers B_1 and B_2, the number of iterations T, positive constant C_0, number of variables N to select

Output: the selected proportion of each feature $\boldsymbol{\eta}^{[T]}$, the selected subset $\hat{S}$

1 for $t \leftarrow 0$ ***to*** T **do**

2 Independently generate random subspaces $S_{b_1 b_2}^{[t]} \sim \mathcal{D}^{[t]}, 1 \le b_1 \le B_1, 1 \le b_2 \le B_2$

3 for $b_1 \leftarrow 1$ ***to*** B_1 **do**

4 Select the optimal subspace $S_{b_1 *}^{[t]} = S_{b_1 b_2^*}^{[t]}$, where $b_2^* = \arg\min_{1 \le b_2 \le B_2} \mathrm{Cr}_n(S_{b_1 b_2}^{[t]})$

5 end

6 Update $\boldsymbol{\eta}^{[t]}$ where $\hat{\eta}_j^{[t]} = B_1^{-1} \sum_{b_1=1}^{B_1} 1(j \in S_{b_1 *}^{[t]}), j = 1, \ldots, p$

7 Update $\mathcal{D}^{[t+1]} \leftarrow$ hierarchical restrictive multinomial distribution $\mathcal{R}(\mathcal{U}_0, p, \boldsymbol{\eta}^{[t]})$,

where $\tilde{\eta}_j^{[t]} \propto [\hat{\eta}_j^{[t]} 1(\hat{\eta}_j^{[t]} > C_0 / \log p) + \frac{C_0}{p} 1(\hat{\eta}_j^{[t]} \le C_0 / \log p)]$ and $\sum_{j=1}^p \tilde{\eta}_j^{[t]} = 1$

8 end

9 Output the selected proportion of each feature $\boldsymbol{\eta}^{[T]}$

10 Output $\hat{S} = \{1 \le j \le p : \hat{\eta}_j^{[T]} \text{ is among the } N \text{ largest of all}\}$

Note that each subspace S can be equivalently represented as $\boldsymbol{J} = (J_1, \ldots, J_p)^T$, where $J_j = 1(j \in S), j = 1, \ldots, p$. A subspace following the *hierarchical restrictive multinomial distribution* $\mathcal{R}(\mathcal{U}, p, \boldsymbol{\eta})$, where $\sum_{j=1}^{p} \tilde{\eta}_j = 1$ and $\tilde{\eta}_j \ge 0$, is equivalent to the procedure:

1. Draw d from distribution $\mathcal{U}$ on $\{1, \ldots, D\}$;
2. Draw $\boldsymbol{J} = (J_1, \ldots, J_p)^T$ from a restrictive multinomial distribution with parameter $(p, d, \boldsymbol{\eta})$, where the restriction is $J_j \in \{0,1\}$.

For example, the hierarchical uniform distribution belongs to this family where $\mathcal{U}$ is the uniform distribution $\mathcal{U}_0$ on $\{1, \ldots, D\}$ and $\tilde{\eta}_j = \frac{1}{p}$ for all $j = 1, \ldots, p$.

With the hierarchical restrictive multinomial distribution in hand, we can depict the iterative algorithm more precisely. At iteration t, the algorithm updates the subspace distribution of next round $\mathcal{D}^{[t+1]}$ by the hierarchical restrictive multinomial distribution $\mathcal{R}(\mathcal{U}_0, p, \boldsymbol{\eta}^{[t]})$, where $\tilde{\eta}_j^{[t]} \propto [\hat{\eta}_j^{[t]} 1(\hat{\eta}_j^{[t]} > C_0 / \log p) + \frac{C_0}{p} 1(\hat{\eta}_j^{[t]} \le C_0 / \log p)]$ and $\hat{\eta}_j^{[t]}$ is the proportion of variable j in the B_1 selected subspaces $\{S_{b_1^*}^{[t]}\}_{b_1=1}^{B_1}$.

2.3 Connections with marginal screening and interaction detection

Before closing this section and moving into theoretical analysis, we want to point out the connection of RaSE screening approach with the classical marginal screening methods as well as the important problem of interaction detection.

First of all, it is easy to observe that when D = 1 in Algorithm 1, with proper measure, RaSE screening method reduces to the marginal screening approaches. In this sense, RaSE screening method can be seen as an extension of classical marginal screening frameworks by evaluating subspaces instead of individual predictors. In addition, when there are signals with no marginal contribution, one intuitive idea is to screen all possible interaction terms, which demand extremely high computational costs. For example, screening all the order-d interactions leads to a computational cost of $O(p^d)$. Instead of screening all possible interactions, RaSE randomly chooses some feature subspaces and explores their contributions to the response via a specific criterion. The carefully designed mechanism of generating random subspaces along with the iterative step greatly alleviate the requirement on computation.

Second, there has been a great interest in studying screening methods for interaction detection (Hao and Zhang, 2014; Fan et al., 2016; Kong et al., 2017). The proposed RaSE screening framework works in a different fashion, by evaluating the contribution of variables through the joint contributions in different subspaces. A simulation example (Example 4) where we have 4-way interactions among predictors will be studied to show the effectiveness of RaSE.

3 Theoretical Analysis

In this section, we investigate the theoretical properties of RaSE screening method to help readers understand how it works and why it can succeed in practice. We are not claiming that the assumptions we make are the weakest and conclusions we obtain are the strongest.

Before moving forward, we first define some notations. For two numbers a and b, we denote $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. For two numerical sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we denote $a_n = o(b_n)$ or $a_n \ll b_n$ if $\lim_{n\to\infty} |a_n / b_n| = 0$. Denote $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $\limsup_{n\to\infty} |a_n / b_n| < \infty$. When $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold at the same time, we write it as $a_n \asymp b_n$. Denote Euclidean norm for a length-p vector $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ as $\| \boldsymbol{x} \|_2 = \sqrt{\sum_{j=1}^{p} x_j^2}$. $\mathbf{1}_p$ represents a length-p vector with all entries 1. For a $p \times p'$ matrix

$A=(a_{ij})_{p\times p'}$, define the 1-norm $\|A\|_1=\sup_j\sum_{i=1}^{p}|a_{ij}|$, the operator norm $\|A\|_2=\sup_{\boldsymbol{x}:\|\boldsymbol{x}\|=1}\|A\boldsymbol{x}\|_2$, the infinity norm $\|A\|_\infty=\sup_i\sum_{j=1}^{p'}|a_{ij}|$ and the maximum norm $\|A\|_{\max}=\sup_{i,j}|a_{ij}|$. We also denote the minimal and maximal eigenvalues of a square matrix A as $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. Besides, we use different probability notations $\mathbf{P},\mathbb{P},\mathrm{P}$ to represent probabilities w.r.t. randomness from subspaces, randomness from training samples, and all randomness, respectively. And we use the same fonts $\mathbf{E},\mathbb{E},\mathrm{E}$ to represent the corresponding expectations. In addition, throughout this section, we assume $p^*=|S^*|$ is fixed.

3.1 Sure screening property

First, note that the success of RaSE relies on the large selected proportions of all signals. According to Algorithm 1, the selected proportion of signal *j* depends on the comparison of two different types of subspaces, namely "covering signal *j*" or "not covering signal *j*". To understand when RaSE can succeed, we also need to compare subspaces "covering a subset $\bar{S}_j\ni j$" or "not covering $\bar{S}_j$", which is essential when signal *j* has no marginal effect. Next, we analyze the joint distribution of these two types of subspaces given the number of B_2 subspaces covering some $\bar{S}_j\ni j$, in the following useful lemma.

Lemma 1. Let $\{S_{1b_2}\}_{b_2=1}^{B_2}\overset{i.i.d.}{\sim}\mathcal{R}(\mathcal{U}_0,p,p^{-1}\mathbf{1}_p)$. For any set $\bar{S}_j\ni j$ with cardinality $|\bar{S}_j|\le D$, let $p_j=\mathbf{P}(S_{11}\supseteq\bar{S}_j)=D^{-1}\sum_{d=|\bar{S}_j|}^{D}\frac{\binom{p-|\bar{S}_j|}{d-|\bar{S}_j|}}{\binom{p}{d}}$. Given $N_j:=\#\{b_2:S_{1b_2}\supseteq\bar{S}_j\}=k$, dividing $\{S_{1b_2}\}_{b_2=1}^{B_2}$ into $\{S_{1b_2}^{(j)}\}_{b_2=1}^{k}$ and $\{S_{1b_2}^{(-j)}\}_{b_2=1}^{B_2-k}$, where $S_{1b_2}^{(j)}\supseteq\bar{S}_j$ and $S_{1b_2}^{(-j)}\not\supseteq\bar{S}_j$,

(i) $\{S_{1b_2}^{(j)}\}_{b_2=1}^{k}$ independently follow the distribution

$$\mathbf{P}(S^{(j)}=S)=\left[D\cdot p_j\binom{p}{|S|}\right]^{-1}\cdot 1(S\supseteq\bar{S}_j);\qquad(2)$$

(ii) $\{S_{1b_2}^{(-j)}\}_{b_2=1}^{B_2-k}$ independently follow the distribution

$$\mathbf{P}(S^{(-j)} = S) = \left[D(1-p_j)\binom{p}{|S|} \right]^{-1} \cdot 1(S \not\supseteq \bar{S}_j); \qquad (3)$$

(iii) $\{S_{1b_2}^{(j)}\}_{b_2=1}^{k} \perp\!\!\!\perp \{S_{1b_2}^{(-j)}\}_{b_2=1}^{B_2-k}$.

The proof of Lemma 1 can be found in Appendix B. It shows us that given $N_j := \#\{b_2 : S_{1b_2} \supseteq \bar{S}_j\} = k$, $\{S_{1b_2}^{(j)}\}_{b_2=1}^{k}$ and $\{S_{1b_2}^{(-j)}\}_{b_2=1}^{B_2-k}$ are independent. And each $S_{1b_2}^{(j)}, S_{1b_2}^{(-j)}$ follows a "weighted" hierarchical uniform distribution by adjusting the sampling weight based on the cardinality of subspace.

Now, we introduce a concentration of Cr_n on its population version Cr for a collection of subsets. In particular, for any D, there exists a sequence $\{\epsilon_n := \epsilon(n, D)\}_{n=1}^{\infty}$ and positive constant $c_{1n} \to 0$ such that

$$\mathbb{P}\left(\sup_{S:|S|\le D} |\mathrm{Cr}_n(S) - \mathrm{Cr}(S)| > \epsilon_n \right) \le c_{1n} \quad (4)$$

holds for any n. Such a sequence $\{\epsilon_n\}_{n=1}^{\infty}$ always exists, though we would like it to be small to have a uniform concentration as described in the following assumption, which is important to establish the sure screening property of RaSE.

Assumption 1. For any $j = 1, \ldots, p$, there exists a subset $\bar{S}_j \ni j$, and we denote $\delta_j(S) := \delta_j(n, D, S) = \mathbf{P}_{S^{(j)}}(\mathrm{Cr}(S) - \mathrm{Cr}(S^{(j)}) < 2\epsilon_n \mid S)$, where $S^{(j)}$ follows the distribution in (2) w.r.t. $\bar{S}_j$. It holds that

$$D \ge \sup_{j \in S^*} |\bar{S}_j|, B_2 \inf_{j \in S^*} p_j \gtrsim 1, \limsup_{n, D, B_2 \to \infty} \left\{ B_2 \sup_{j \in S^*} \mathbf{E}_{S^{(-j)}} \left[\delta_j (S^{(-j)})^{\frac{1}{2} B_2 p_j} \right] \right\} < \infty,$$

where $S^{(-j)}$ follows the distribution in (3) and $p_j = \mathbf{P}(S_{11} \supseteq \bar{S}_j) = D^{-1} \sum_{d=|\bar{S}_j|}^{D} \frac{\binom{p-|\bar{S}_j|}{d-|\bar{S}_j|}}{\binom{p}{d}}$.

Remark 1. In Assumption 1, δ_j measures the strength of signal j via comparing the two types of feature subspaces introduced in Lemma 1. From the assumption, we need a large B_2 when δ_j is small.

Theorem 1 (Sure screening property). Define

$$c_{2n} := c_2(n, B_2, D) := (1-c_{1n})\left(1-\sup_{j\in S^*}\mathbf{E}_{S^{(-j)}}\left[\delta_j (S^{(-j)})^{\frac{1}{2}B_2 p_j}\right]\right)^{B_2}\left(1-\exp\left\{-\frac{3}{28}B_2 \inf_{j\in S^*} p_j\right\}\right).$$

For any $\alpha > 1$, let $\hat{S}_\alpha = \{1 \le j \le p : \hat{\eta}_j \text{ is among the } [\alpha D / c_{2n}] \text{ largest of all}\}$. Under Assumption 1, when $B_1 \gg \log p^*$ and $n \to \infty$, we have

(i) $\mathrm{P}(S^* \subseteq \hat{S}_\alpha) \ge 1 - p^* \exp\left\{-2B_1 c_{2n}^2 \left(1-\frac{1}{\alpha}\right)^2\right\} \to 1$;

(ii) The selected model size $|\hat{S}_\alpha| \lesssim D$.

Next, we would like to analyze the restriction on B_2 imposed by Assumption 1, which depends on δ_j. We first introduce a useful notion called *detection complexity*.

Definition 1 (Detection complexity). We say feature $j \in S^*$ is detectable in complexity d, if there exists a subset $\bar{S}_j \ni j$ with cardinality d and another subset $S_j^0 \subseteq S_{\text{Full}} \setminus \{j\}$ with cardinality p_j^0, such that

$$\inf_{S\in \mathcal{S}, S'\in \mathcal{S}'} \left[\mathrm{Cr}(S) - \mathrm{Cr}(S')\right] > 2\epsilon_n,$$

where $\mathcal{S} = \mathcal{S}(j, D) = \{S : |S| \le D, |S \cap (S^* \cup S_j^0)| < d\}$, $\mathcal{S}' = \mathcal{S}'(j, D) = \{S : |S| \le D, S \supseteq \bar{S}_j\}$, and ϵ_n satisfies (4). We define the detection complexity of j, which is denoted by d_j, as the minimal integer d to make $j \in S^*$ detectable in complexity d.

Remark 2. The detection complexity d_j actually indicates the difficulty to identify signal j. When $\mathrm{d}_j = 1$, $\bar{S}_j$ is actually equal to $\{j\}$ and $\mathcal{S} = \{S : |S| \le D, S \cap (S^* \cup S_j^0) = \varnothing\}$. It implies that the given criterion function performs better at subsets covering j than at subsets not intersecting with $S^* \cup S_j^0$. S_j^0 is introduced to avoid cases that some noises might have strong marginal effects. This condition is similar to marginal conditions in literature, for examples, see Fan and Lv (2008); Fan et al. (2011); Zhu et al. (2011); Li et al. (2012); Shao and Zhang (2014); Cui et al. (2015); Pan et al. (2018); Nandy et al. (2020). The difference is that here we state it via subspaces instead of single features used in existing works. And when $\mathrm{d}_j \ge 2$, the

definition of detection complexity allows us to consider the joint contribution of multiple features. See Examples 1 and 6 in our numerical studies as examples.

Now we introduce an assumption under which the restriction on B_2 can be explicitly calculated.

Assumption 2. All signals in S^* are detectable in complexity $\mathtt{d}$, where $\mathtt{d} = \max_{j \in S^*} \mathtt{d}_j$.

Intuitively speaking, this assumption requires all signals to be detectable under the same level, which equals the largest detection complexity of signals. In some sense, it is necessary for the sure screening property. Signal j with large detection complexity is associated with a larger set $\bar{S}_j$, requiring a larger B_2 to sample subsets that cover $\bar{S}_j$ with sufficiently high probability.

Proposition 1. Under Assumption 2, when $B_2 \asymp \left(\frac{p}{D}\right)^{\mathtt{d}}$, Assumption 1 hold.

In the ideal case, in Assumption 1, we can set $\bar{S}_j = \{j\}$ for all $j \in S^*$, implying $\mathtt{d} = 1$, which leads to the weakest restriction on B_2, that is, $B_2 \asymp \frac{p}{D}$. If a signal j does not have marginal contribution to the response, we have $\mathtt{d}_j \geq 2$, requiring a larger order of B_2 to satisfy Assumption 1. This motivates the iterative RaSE screening (Algorithm 2) which usually has a less stringent restriction on B_2, making the framework more applicable to high-dimensional settings.

Next, we study the sure screening property for iterative RaSE, and discuss how the restriction on B_2 can be relaxed. For simplicity, we only study the one-step iteration, i.e. the case when T = 1. It's not very hard to generalize the conditions and conclusions to the general case when T > 1. To better state the results, we first generalize Lemma 1 to understand the distribution of two aforementioned types of subspaces after one iteration.

Lemma 2. For any set $\bar{S}_j \ni j$ with cardinality $|\bar{S}_j| \leq D$, let $\{S_{1b_2}\}_{b_2=1}^{B_2} \overset{i.i.d.}{\sim}$ some distribution $\mathcal{F}$ such that $\mathbf{P}_{S_{11} \sim \mathcal{F}}(S_{11} \supseteq \bar{S}_j) \in (0,1)$. Given $N_j := \#\{b_2 : S_{1b_2} \supseteq \bar{S}_j\} = k$, dividing $\{S_{1b_2}\}_{b_2=1}^{B_2}$ into $\{S_{1b_2}^{(j)}\}_{b_2=1}^{k}$ and $\{S_{1b_2}^{(-j)}\}_{b_2=1}^{B_2-k}$, where $S_{1b_2}^{(j)} \supseteq \bar{S}_j$ and $S_{1b_2}^{(-j)} \not\supseteq \bar{S}_j$,

(i) $\{S_{1b_2}^{(j)}\}_{b_2=1}^{k}$ independently follow the distribution

$$\mathbf{P}(S^{(j)}=S)=\mathbf{P}_{S_{11}\sim\mathcal{F}}(S_{11}=S)\cdot\frac{1(S\supseteq\bar{S}_j)}{\mathbf{P}_{S_{11}\sim\mathcal{F}}(S_{11}\supseteq\bar{S}_j)}; \qquad (5)$$

(ii) $\{S_{1b_2}^{(-j)}\}_{b_2=1}^{B_2-k}$ independently follow the distribution

$$\mathbf{P}(S^{(-j)}=S)=\mathbf{P}_{S_{11}\sim\mathcal{F}}(S_{11}=S)\cdot\frac{1(S\not\supseteq\bar{S}_j)}{\mathbf{P}_{S_{11}\sim\mathcal{F}}(S_{11}\not\supseteq\bar{S}_j)}; \qquad (6)$$

(iii) $\{S_{1b_2}^{(j)}\}_{b_2=1}^{k}\perp\!\!\!\perp\{S_{1b_2}^{(-j)}\}_{b_2=1}^{B_2-k}$.

We omit the proof of Lemma 2 as it is very similar to Lemma 1. Next, we introduce the following technical assumption analogous to Assumption 1.

Assumption 3. Suppose the signal set S^* can be decomposed as $S^*=S_{[0]}^*\cup S_{[1]}^*$, where $S_{[0]}^*$ and $S_{[1]}^*$ satisfy the following conditions:

(i) (The first-step detection) For any $j\in S_{[0]}^*$, denote $\delta_j^{[0]}(S)=\mathbf{P}_{S^{(j)}}(\mathrm{Cr}(S)-\mathrm{Cr}(S^{(j)})<2\epsilon_n\mid S)$, $p^{[0]}=\mathbf{P}(j\in S_{11})=\frac{D+1}{2p}$, where $S^{(j)}$ follows the distribution in (2) w.r.t. $\bar{S}_j=\{j\}$. Then,

$$B_2\gtrsim\frac{p}{D},\limsup_{n,D,B_2\to\infty}\left\{B_2\sup_{j\in S_{[0]}^*}\mathbf{E}_{S^{(-j)}}\left[\delta_j^{[0]}(S^{(-j)})^{\frac{1}{2}B_2p^{[0]}}\right]\right\}<\infty,$$

where $S^{(-j)}$ follows the distribution in (3) w.r.t. $\bar{S}_j$.

(ii) (The second-step detection) Denote $\delta_j^{[1]}(S)=\mathbf{P}_{S^{(j)}}(\mathrm{Cr}(S)-\mathrm{Cr}(S^{(j)})<2\epsilon_n\mid S)$, $p^{[1]}=\frac{(D+1)C_0}{2(D+C_0)p}$, where $S^{(j)}$ follows the distribution in (5) w.r.t. $\bar{S}_j=\{j\}$, and C_0 is a constant from Algorithm 2. $\Upsilon=\{\mathcal{R}(\mathcal{U}_0,p,\boldsymbol{\eta})\}$ is a family of hierarchical restrictive multinomial distributions satisfying

$$\inf_{j\in S_{[0]}^*}\tilde{\eta}_j\geq\frac{c_2^*}{(D+C_0)},$$

for a constant $c_2^* > 0$. Then,

$$B_2 \gtrsim p, \limsup_{n,D,B_2\to\infty}\left\{B_2 \sup_{\mathcal{F}\in\Upsilon}\sup_{j\in S^*}\mathbf{E}_{S^{(-j)}}\left[\delta_j^{[1]}(S^{(-j)})^{\frac{1}{2}B_2 p^{[1]}}\right]\right\} < \infty,$$

where $S^{(-j)}$ follows the distribution in (6) w.r.t. $\bar{S}_j = \{j\}$ and $\{S_{b1b2}\}_{b_1,b_2} \overset{i.i.d.}{\sim} \mathcal{F} \in \Upsilon$.

Remark 3. Condition (i) is a relaxed version of Assumption 1, which replaces S^* by a subset $S^*_{[0]}$. This can be seen as a first-step detection condition for RaSE screening method to capture $S^*_{[0]}$. The remaining signals in $S^*_{[1]}$ that might be missed in the first step will be captured in the second step. The family of distributions Υ is introduced to incorporate the randomness in the first step of RaSE screening. This type of stepwise detection condition is very common in the literature (Jiang and Liu, 2014; Li and Liu, 2019; Zhou et al., 2020; Tian and Feng, 2021).

Theorem 2 (Sure screening property for one-step iterative RaSE screening). Define

$$c_{2n}^{[l]} := c_2^{[l]}(n, B_2, D) := (1-c_{1n})\left(1-\sup_{j\in S^*}\mathbf{E}_{S^{(-j)}}\left[\delta_j^{[l]}(S^{(-j)})^{\frac{1}{2}B_2 p^{[l]}}\right]\right)^{B_2}\left(1-\exp\left\{-\frac{3}{28}B_2(p^{[l]})^2\right\}\right),$$

where $l = 0, 1$. For $\hat{S}_\alpha^{[1]} = \{1 \le j \le p : \hat{\eta}_j^{[1]} \text{ is among the } [\alpha D / c_{2n}^{[1]}] \text{ largest of all}\}$, where $\alpha > 1$, under Assumption 3, if $c_{2n}^{[0]} > c_2^*$ and $B_1 \gg \log p^*$, we have

(i) $\mathrm{P}(S^* \subseteq \hat{S}_\alpha^{[1]}) \ge 1 - p^* \exp\left\{-2B_1(c_{2n}^{[0]} - c_2^*)^2\right\} - p^* \exp\left\{-2B_1(c_{2n}^{[1]})^2\left(1-\frac{1}{\alpha}\right)^2\right\} \to 1$, as $n \to \infty$;

(ii) $|\hat{S}_\alpha^{[1]}| \lesssim D$.

The lower bound in (i) comes from the two steps of Algorithm 2, which is very intuitive. The general iterative RaSE screening algorithm with any $T \ge 1$ can be studied similarly by imposing analogous conditions, which we leave as future work.

The restriction on B_2 can be discussed in a similar fashion as the vanilla RaSE screening for some specific scenarios. For instance, similar to Definition 1, we can define the detection complexity of the second step based on the distribution of

subsets from the first step. If a similar assumption like Assumption 2 (see Assumption 5 in Appendix A.2 for the precise statement) holds, then we can expect that there exist $B_2 \asymp p/D$ in the first step and $B_2 \lesssim D^{|S^*_{[0]}|}(\log p)^{|S^*_{[1]}|-1} p$ in the second step to make Assumption 3 hold (see Proposition 3 in Appendix A.2 for a precise description), which relaxes the requirement shown in Proposition 1 ($B_2 \asymp (p/D)^{|S^*_{[0]}|+1}$) to a great extent. In Section 4, an array of simulations and real data analyses will show the effectiveness of iterative RaSE screening.

3.2 Rank consistency

Next, we study another important property of the RaSE screening, namely the rank consistency. First, we impose the following assumption.

Assumption 4. Suppose the following conditions hold:

(i) Denote $\tilde{\delta}_j(S) = \mathbf{P}_{S^{(-j)}}(\mathrm{Cr}(S) - 2\epsilon_n < \mathrm{Cr}(S^{(-j)}) \mid S)$ and $\delta_j(S) := \delta_j(n, D, S) = \mathbf{P}_{S^{(j)}}(\mathrm{Cr}(S) - \mathrm{Cr}(S^{(j)}) < 2\epsilon_n \mid S)$, where $S^{(-j)}$ follows the distribution in (3) with respect to $\bar{S}_j = \{j\}$ while $S^{(j)}$ follows the distribution in (2) with respect to some subset $\bar{S}_j \ni j$. We have

$$\gamma(n, D, B_2) := (1 - c_{1n})\left(1 - 2\exp\left\{-\frac{3}{28} B_2 \inf_{j \in S^*} p_j\right\}\right)$$
$$\cdot \left[\left(1 - \sup_{j \in S^*} \mathbf{E}_{S^{(-j)}}[\delta_j(S^{(-j)})^{\frac{1}{2}B_2 p_j}]\right)^{B_2} + \left(1 - \sup_{j \notin S^*} \mathbf{E}_{S^{(j)}}\left[\tilde{\delta}_j(S^{(j)})^{B_2 - \frac{3}{2}B_2 p^{[0]}}\right]\right)^{\frac{3}{2}B_2 p^{[0]}} - 1\right]$$
$$- c_{1n} > 0,$$

where $S^{(j)}$ and $S^{(-j)}$ follow the distributions in (2) and (3), respectively.

(ii) $B_1 \gg \gamma(n, D, B_2)^{-2} \vee \log p$.

Remark 4. Condition (i) is introduced to make sure the signals are separable from the noises. Here, $\tilde{\delta}_j$ is a parallel definition to δ_j, measuring the noise level via comparing the two types of feature subspaces introduced in Lemma 1. A related condition can be found in Assumption (C3) of Cui et al. (2015).

Theorem 3 (Rank consistency). Under Assumption 4,

$$\mathrm{P}\left(\inf_{j\in S^*}\hat{\eta}_j > \sup_{j\notin S^*}\hat{\eta}_j\right) \geq 1 - p\exp\left\{-\frac{1}{2}B_1\gamma^2(n,D,B_2)\right\} \to 1,$$

as $n, B_1, B_2 \to \infty$.

In addition, under Assumption 2 with $\mathtt{d}=1$, when B_2 is restricted to some level, we have Assumption 4 holds by default.

Proposition 2. Under Assumption 2 with $\mathtt{d}=1$, there exist constants $C_2 > C_1 > 0$, such that, when $B_2 \in (C_1 p/D, C_2 p/D)$, Assumption 4 holds.

4 Numerical Studies

In this section, we will investigate the performance of RaSE screening methods via extensive simulations and real data experiments. Each setting is replicated 200 times. In simulations, we evaluate different screening approaches by calculating the 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size (MMS) to include all signals. The smaller the quantile is, the better the screening approach is. For real data, since S^* is unknown, we compare different methods by investigating the performance of the corresponding post-screening procedure. That is, after screening, we keep the same number of variables for each screening method, then the same model is fitted based on those selected variables and their prediction performance on an independent test data is reported.

We compare RaSE screening methods with SIS (Fan and Lv, 2008), ISIS (Fan and Lv, 2008; Fan et al., 2009), SIRS (Zhu et al., 2011), DC-SIS (Li et al., 2012), MDC-SIS (Shao and Zhang, 2014), MV-SIS (Cui et al., 2015), HOLP (Wang and Leng, 2016), IPDC (Kong et al., 2017), and CIS (Nandy et al., 2020).

All the experiments are conducted in R. We implement RaSE screening methods in `RaSEn` package. R package `SIS` is used to implement SIS. Corresponding to one-step iterative RaSE, we report the results of ISIS with two screening steps and one selection step (Saldana and Feng, 2018).[1] R package `screening` (https://github.com/wwrechard/screening) is used to implement HOLP. We conduct

SIRS, DC-SIS and MV-SIS through R package `VariableSelection`. IPDC is implemented by calling the function `dcor` in R package `energy`. We implement MDC-SIS through function `mdd` in R package `EDMeasure` to calculate the martingale difference divergence. CIS is implemented via R codes shared in Nandy et al. (2020).

We combine RaSE framework with various criteria to choose subspaces, including minimizing BIC (RaSE-BIC) and eBIC (RaSE-eBIC) in linear model or logistic regression model, minimizing the leave-one-out MSE/error in *k*-nearest neighbor (*k*NN) (RaSE-*k*NN), and minimizing the 5-fold cross-validation MSE/error in support vector machine (SVM) with RBF kernel (RaSE-SVM). We add a subscript 1 to RaSE to denote the one-step iterative RaSE (e.g. RaSE_1-BIC). In practice, we can choose the criterion based on the model we prefer in the post-screening procedure. For example, if we would like to use linear model in post-screening, we could set minimizing BIC of linear model as the criterion. If we want to fit a non-linear model in post-screening, minimizing cross-validation error in *k*NN or SVM with RBF kernel can be good choices. Some exploratory analysis can help us choose a proper post-screening method.

For all RaSE methods, we fix $B_1 = 200$ and $B_2 = 20 \times [p/D]$, motivated by Proposition 1. In addition, following Weng et al. (2019), we fix $D = [\sqrt{n}]$, which is motivated from the fact that many estimators are $\sqrt{n}$-consistent. And we verify the effectiveness of this choice in Example 1. For Example 1, we also investigate the impact of B_1, B_2 and D on the median MMS. For RaSE-*k*NN and RaSE_1-*k*NN, k is set to be 5. For RaSE-eBIC and RaSE_1-eBIC, we set the penalty parameter $\gamma = 0.5$ (Chen and Chen, 2008, 2012).

All the codes used in numerical experiments can be found on GitHub (https://github.com/ytstat/RaSE-screening-codes).

4.1 Simulations

Example 1 (Example II in Fan and Lv (2008)). We generate data from the following model:

$$y = 5x_1 + 5x_5 + 5x_3 - \frac{15}{\sqrt{2}} x_4 + \epsilon,$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)^T \sim N(\boldsymbol{0}, \Sigma), \Sigma = (\sigma_{ij})_{p\times p}, \sigma_{ij} = 0.5^{1(i\neq j)}, \epsilon \sim N(0,1)$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. The signal set $S^* = \{1,2,3,4\}$. n = 100 and p = 1000.

In this example, there is no correlation between y and x_4, further leading to the independence due to normality, therefore methods based on the marginal effect will fail to capture x_4. However, after projecting y on the space which is perpendicular with any signals from x_1, x_2 and x_3, the correlation appears between the projected y and x_4, which motivates the ISIS. Besides, the proposed RaSE methods are also expected to succeed since it works with feature subsets instead of a single variable.

We present the results in the left panel of Table 1. From the results, it can be seen that all the marginal screening methods do not perform well in the sense that they need a large model to cover all 4 signals. ISIS performs much better because it can detect the signals with a smaller model than SIS with one step iteration. For RaSE screening methods with no iteration, as analyzed in Proposition 1, we have $\mathrm{d} = 2$ since x_4 has no marginal contribution to y, leading to a theoretical requirement $B_2 \asymp (p/D)^2$, where $(p/D)^2 = 10^4$. Despite the current small B_2 setting, RaSE-BIC and RaSE-eBIC still perform better than SIS and other marginal screening methods. After one iteration, RaSE$_1$-BIC and RaSE$_1$-eBIC improve a lot compared to their vanilla counterparts, with RaSE$_1$-eBIC achieving the best performance.

Note that iterations can usually improve the performance of vanilla RaSE at small quantiles, but possibly lead to worse performance at large quantiles. See RaSE-kNN and RaSE$_1$-kNN for examples. This phenomenon happens because iterative RaSE is very aggressive and the success of the second step is based on the accurate capture of some signals in the first step. If the first step fails to identify enough signals but captures many noises, these noises will be selected more frequently in the second step.

To further study the impact of (B_1, B_2), we run this example for 200 times under different (B_1, B_2) settings, where we range B_1 from 100 to 1000 with increment 100 and B_2 from 1000 to 97000 with increment 6000. The median of MMS with RaSE-

BIC and RaSE$_1$-BIC is summarized in Figure 1. It shows that in general, larger (B_1, B_2) leads to better performance. The performance is stable in terms of B_1 when B_2 is large. On the other hand, the performance improves continuously as B_2 grows. In particular, for RaSE-BIC, when $B_2 \geq 10^4$, it can capture S^* very well, which agrees with Proposition 1. These results indicate that we can further improve the performance of RaSE screening if we have sufficient computational resources. RaSE$_1$-BIC can always achieve a great performance with a small B_2, showing its effectiveness in relaxing the restriction on B_2.

We also run this example 200 times to plot the median of MMS for RaSE-BIC and RaSE$_1$-BIC under different (D, B_2) while fixing $B_1 = 200$ in Figure 3 in Appendix A.3, where D ranges from 2 to 40 with increment 2 and B_2 from 200 to 5000 with increment 300. The subfigure (a) shows that for RaSE-BIC, for a given B_2, the impact of D is not monotonic. RaSE-BIC has a good and stable performance when D is around $\sqrt{n} = 10$, which verifies the effectiveness of our choice for D. The subfigure (b) shows that the performance of RaSE$_1$-BIC is very robust with respect to D, as long as D and B_2 are not very small.

To compare the computational time of different methods, we list the average running time in 200 replications of Example 1 in Table 2. All codes were run on NYU Greene clusters (2x Intel Xeon Platinum 8268 24C 205W 2.9GHz Processor) with 40 cores and 50 GB memory.[2] It can be seen that RaSE methods have heavier computational burdens than other screening methods since their success leverages generating a large number of subspaces. This can be alleviated with parallel computing and more powerful machines.

Example 2 (Latent clusters). We generate data from the following linear model:

$$y = 0.5(\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 + \tilde{x}_4 + \tilde{x}_5 + \epsilon),$$

where $\boldsymbol{x} = (\tilde{x}_1, \ldots, \tilde{x}_p)^T \sim N(\boldsymbol{0}, \Sigma)$, $\epsilon \sim t_2$, $\Sigma = (\sigma_{ij})_{p\times p} = (0.5^{|i-j|})_{p\times p}$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. Generate $z \sim \text{Unif}(\{-3, 3\}) \perp\!\!\!\perp \boldsymbol{x}$ and $\boldsymbol{x} = \boldsymbol{x} + z\mathbf{1}_p$. The signal set $S^* = \{1, 2, 3, 4, 5\}$. n = 200 and p = 2000.

Figure 2 shows the scatterplots of y vs. x_1 (left panel) and y vs. x_{10} (right panel). We expect the methods based on Pearson correlation to deteriorate due to the partial cancellation of signals by the averaging of two clusters. For such kind of data, *k*NN could be a favorable approach. The performances of various methods are presented in the right panel of Table 1. SIS and MDC-SIS perform well at 5% and 25% quantiles. RaSE-*k*NN and RaSE-SVM perform quite well with their performances further improved by their respective one-step iterative versions.

Example 3 (Example 1.c in Li et al. (2012)). We generate data from the following model:

$$y = 2\beta_1 x_1 x_2 + 3\beta_2 1(x_{12} < 0) x_{22} + \epsilon,$$

where

$\beta_j = (-1)^U (4\log n/\sqrt{n} + |Z|), j = 1,2, U \sim \text{Bernoulli}(0.4), Z \sim N(0,1), \epsilon \sim N(0,1), \boldsymbol{x} \sim N(\boldsymbol{0}, \Sigma)$ where $\Sigma = (\sigma_{ij})_{p\times p} = (0.8^{|i-j|})_{p\times p}, U \perp\!\!\!\perp Z, \epsilon \perp\!\!\!\perp \boldsymbol{x}$, and $(U,Z) \perp\!\!\!\perp (\epsilon, \boldsymbol{x})$. Note that we regenerate (U, Z) for each replication, so the results might differ from those in Li et al. (2012). The signal set $S^* = \{1,2,12,22\}$. n = 200 and p = 2000.

The left panel of Table 3 exhibits the results of different screening methods. Due to the interaction term and indicator function, approaches based on linear models like SIS, ISIS, HOLP, and RaSE with BIC and eBIC do not perform very well. CIS and RaSE$_1$-*k*NN achieve a very good performance at 5%, 25% and 50% quantiles. RaSE-*k*NN performs well at 5% and 25% quantiles but worse at others. RaSE-SVM performs well at the first two quantiles. The iteration step improves the performances of RaSE-*k*NN and RaSE-SVM significantly, and RaSE$_1$-SVM outperforms all the other methods except at 95% quantile.

Example 4 (Interactions). We generate data from the following model:

$$y = 3\sqrt{|x_1|} + 2\sqrt{|x_1|}x_2^2 + 4\sin(x_1)\sin(x_2)\sin^2(x_3) + 12\sin(x_1)|x_2|\sin(x_3)x_4^2 + 0.5\epsilon,$$

where $x_1, \ldots, x_p \overset{i.i.d.}{\sim} N(0,1), \epsilon \sim N(0,1)$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. The signal set $S^* = \{1,2,3,4\}$. n = 300 and p = 1000.

This example evaluates the capability of different screening methods in terms of selecting high-order interactions. The results are summarized in the right panel of Table 3. It can be observed that RaSE-*k*NN, RaSE$_1$-*k*NN, RaSE-SVM, RaSE$_1$-SVM, IPDC, and CIS achieve an acceptable performance, particularly for the lower quantiles. IPDC and CIS perform better at 75% and 95% quantiles than all RaSE methods but worse at the other three quantiles than RaSE$_1$-*k*NN. The remaining methods do not perform well on any of the 5 quantiles. It shows that RaSE framework equipped with minimizing cross-validation MSE on *k*NN or kernel SVM is promising to capture high-order interactions.

Example 5 (Gaussian mixture, Example 1 in Cannings and Samworth (2017)). We generate data from the following model:

$$y \sim \text{Bernoulli}(0.5),\ \boldsymbol{x} \mid y = r \sim \frac{1}{2}N(\boldsymbol{\mu}_r, \Sigma) + \frac{1}{2}N(-\boldsymbol{\mu}_r, \Sigma), r = 0,1,$$

where $\boldsymbol{\mu}_0 = (2, -2, 0, \ldots, 0)^T$, $\boldsymbol{\mu}_1 = (2, 2, 0, \ldots, 0)^T$, Σ is an identity matrix. The signal set $S^* = \{1, 2\}$. n = 200 and p = 2000.

From the scatterplots in Figure 4 in Appendix A.3, the marginal screening methods are expected to fail because all signals are marginally independent with y. The only way to capture the signals is to measure the joint contribution of (x_1, x_2). We summarize the results in the left panel of Table 4.

The table shows that the marginal methods fail as we expected. RaSE with BIC and eBIC fail as well because the data points from the two classes are not linearly separable (Figure 4). SIRS, RaSE$_1$-*k*NN and RaSE$_1$-SVM achieve the best performance with very accurate feature ranking.

Example 6 (Multinomial logistic regression, Case 2 in Section 4.5 of Fan et al. (2009)). We first generate $\tilde{x}_1, \ldots, \tilde{x}_4 \overset{i.i.d.}{\sim} \text{Unif}([-\sqrt{3}, \sqrt{3}])$ and $\tilde{x}_5, \ldots, \tilde{x}_p \overset{i.i.d.}{\sim} N(0,1)$, then let $x_1 = \tilde{x}_1 - \sqrt{2}\tilde{x}_5$, $x_2 = \tilde{x}_2 + \sqrt{2}\tilde{x}_5$, $x_3 = \tilde{x}_3 - \sqrt{2}\tilde{x}_5$, $x_4 = \tilde{x}_4 + \sqrt{2}\tilde{x}_5$ and $x_j = \tilde{x}_j$ for $j = 5, \ldots, p$. The response is generated from

$$\text{P}(y = r \mid \boldsymbol{x}) \propto \exp\{f_r(\boldsymbol{x})\}, r = 1, \ldots, 4,$$

where $f_1(\boldsymbol{x}) = -a\tilde{x}_1 + a\tilde{x}_4$, $f_2(\boldsymbol{x}) = a\tilde{x}_1 - a\tilde{x}_2$, $f_3(\boldsymbol{x}) = a\tilde{x}_2 - a\tilde{x}_3$ and $f_4(\boldsymbol{x}) = a\tilde{x}_3 - a\tilde{x}_4$ with $a = 5/\sqrt{3}$. The signal set $S^* = \{1,2,3,4,5\}$. n = 200 and p = 2000.

In this example, x_5 is marginally independent of y, therefore the marginal methods are expected to fail to capture x_5. Results are summarized in the right panel of Table 4.

We observe that ISIS, RaSE$_1$-BIC, and, RaSE$_1$-eBIC lead to better performances. Without iteration, RaSE-BIC still performs competitively compared to other non-iterative approaches. Similar to Example 1, the iteration usually improves the performance of vanilla RaSE at small quantiles, but leads to worse performance at large quantiles possibly due to the aggressiveness of iterative RaSE.

To justify the effectiveness of RaSE methods in dealing with more complicated predictors, we add two additional examples in Appendix A.3.2. In Example 7, we consider realistic predictors, with the same conditional model $y|\boldsymbol{x}$ as in Example 1. In Example 8, we use a mix of continuous and discrete variables, with the same conditional model as in Example 2. While we have similar findings as in Examples 1 and 2, the performance of most approaches become slightly worse, showing the challenges for analyzing real data.

4.2 Real data experiments

In this section, we investigate the performance of RaSE screening methods on two real data sets. Each data set is randomly divided into training data and test data. As suggested by Fan and Lv (2008), we select variables via different screening methods on training data, then the LASSO, kNN and SVM are fitted based on the selected variables on training data, and finally we evaluate different screening methods based on their corresponding post-screening performance on test data. As benchmarks, we also fit LASSO, kNN and SVM models on the training data without screening. Following Fan and Lv (2008), we choose the top $[n/\log n]$ variables for all screening methods, i.e. let $N = [n/\log n]$ in Algorithms 1 and 2. Note that we could also choose $[\alpha D]$ variables for any $\alpha > 1$, which is motivated by (1). Another possibility is to use data-driven strategies. For instance, we could sample out a separate validation data set and use the post-screening validation MSE/classification error to determine N.[3]

We randomly divide the whole data set into 90% training data and 10% test data in each of 200 replications, and apply various screening methods on training data. Each time, both training and test data are standardized by using the center and scale of training data.

4.2.1 Colon cancer data set

This data set was collected by Alon et al. (1999) and consists of 2000 genes measured on 62 patients, of which 40 are diagnosed with colon cancer (class 1) and 22 are healthy (class 0). The information on each gene is represented as a continuous variable. The prediction results are summarized in the left panel of Table 5.

The table shows that SIS, ISIS, MDC-SIS, CIS, RaSE-BIC, RaSE$_1$-BIC, RaSE-eBIC, and RaSE$_1$-eBIC improve the performance of vanilla LASSO. In addition, RaSE-BIC with LASSO achieves the best performance among all post-screening procedures based on LASSO. Besides, RaSE-*k*NN with *k*NN and RaSE$_1$-*k*NN with *k*NN lead to better results than those of vanilla *k*NN. RaSE-SVM and RaSE$_1$-SVM also improve the performance of vanilla SVM, demonstrating the effectiveness of RaSE to improve various vanilla methods.

For results of RaSE methods, we also gather the top 10 selected features in 200 replications and calculate the percentages of selection of these top features out of 200 replications. The 10 features with the highest percentages (selection rates) are plotted in Figure 5 in Appendix A.3. We notice that the first few features have high or moderately high selection rates (100% or $> 50\%$, respectively), implying that they are frequently selected in different replications. These results demonstrate that RaSE-based variable screening methods are reasonably stable.

4.2.2 Rat eye expression data set

This data set was used by Scheetz et al. (2006); Fan et al. (2011); Wang and Leng (2016); Zhong and Zhu (2015); Nandy et al. (2020) among others. It contains the gene expression values corresponding to 18976 probes from the eyes of 120 twelve-week-old male F2 rats. Among the 18976 probes, TRIM32 is the response,

which is responsible to cause Bardet-Biedl syndrome. We follow Wang and Leng (2016) to focus on the top 5000 genes with the highest sample variance. Therefore, the final sample size is 120 and there are 5000 predictors. The right panel of Table 5 shows the test average mean squared error (MSE) coupled with the standard deviation for each post-screening procedure.

The results show that SIS, ISIS, RaSE-BIC and RaSE-eBIC with LASSO achieve comparable performance, which are better than that of the vanilla LASSO. RaSE-*k*NN with *k*NN and RaSE$_1$-*k*NN with *k*NN enhance the vanilla *k*NN method as well. RaSE-SVM with SVM and RaSE$_1$-SVM with SVM only slightly improve the vanilla SVM for this data set. Note that MV-SIS is not directly applicable to this data set. It is possible to discretize all the variables to make MV-SIS work. See Section 4.2 in Cui et al. (2015) for details.

Similar to the colon data set, we also demonstrate the stability of RaSE methods in Figure 6.

5 Discussion

In this article, we propose a very general screening framework named RaSE screening, based on the random subspace ensemble method. We can equip it with any criterion function for comparing subspaces. By comparing subspaces instead of single predictors, RaSE screening can capture signals without marginal effects on response. Besides, an iterative version of the RaSE screening framework is introduced to enhance the performance of vanilla RaSE and relax the restriction on B_2. In the theoretical analysis, we establish sure screening property for both vanilla and iterative RaSE frameworks under some general conditions. The rank consistency is also proved for the vanilla RaSE. We investigate the relationship between the signal strength and the appropriate choice of B_2, which shows that in some sense the weaker the signal is, a larger B_2 is necessary for RaSE to succeed. In the numerical studies, the effectiveness of RaSE and its iterative version is verified through multiple simulation examples and real data analyses.

The success of RaSE leverages on proper choices of Cr (the criterion), B_1 (the number of subspace groups), B_2 (the number of subspace candidates in each

group), and D (the maximum subspace size). While we have studied their impacts on the performance of RaSE, there exists potential improvement for choosing these " tuning" parameters. For example, the subspace distribution at each iteration step could be further generalized, e.g., we can choose D from the empirical distribution of the sizes of the selected B_1 subspaces.

There are many other interesting problems worth further studying. The first question is that whether there is an adaptive way to automatically select the number of iterations (T). A possible solution is cross-validation and to stop the iteration process when the performance of RaSE on validation data stops improving further. Another interesting topic is to use different B_2 values in different iteration steps, which might further speed up the computation.

Acknowledgements

We are grateful to the editor, the AE, and anonymous reviewers for their insightful comments which have greatly improved the scope and quality of the manuscript. Feng's research is partially supported by NSF CAREER Grant DMS-2013789. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750.

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574.

Cheng, M.-Y., Honda, T., Li, J., and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics*, 42(5):1819–1849.

Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. In *Borrowing strength: Theory powering applications–a Festschrift for Lawrence D. Brown*, pages 70–86. Institute of Mathematical Statistics.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038.

Fan, Y., Kong, Y., Li, D., and Lv, J. (2016). Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*.

Han, X. (2019). Nonparametric screening under conditional strictly convex loss for ultrahigh dimensional sparse data. *Annals of Statistics*, 47(4):1995–2022.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844.

Jiang, B. and Liu, J. S. (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, 42(5):1751–1786.

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(2):897–922.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Li, Y. and Liu, J. S. (2019). Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association*, 114(525):271–286.

Mai, Q. and Zou, H. (2013). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234.

Mai, Q. and Zou, H. (2015). The fused kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497.

Nandy, D., Chiaromonte, F., and Li, R. (2020). Covariate information number for feature screening in ultrahigh-dimensional supervised problems. *Journal of the American Statistical Association*, pages 1–40. DOI: https://doi.org/10.1080/01621459.2020.1864380.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2018). A generic sure independence screening procedure. *Journal of the American Statistical Association*.

Saldana, D. F. and Feng, Y. (2018). Sis: an r package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software*, 83(2):1–25.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., and Casavant, T. L. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.

Tian, Y. and Feng, Y. (2021). Rase: Random subspace ensemble classification. *Journal of Machine Learning Research*, 22(45):1–93.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3(78):589–611.

Weng, H., Feng, Y., and Qiao, X. (2019). Regularization after retention in ultrahigh dimensional linear regression models. *Statistica Sinica*, 29(1):387–407.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis*, 105(1):397–411.

Zhong, W. and Zhu, L. (2015). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation*, 85(11):2331–2345.

Zhou, T., Zhu, L., Xu, C., and Li, R. (2020). Model-free forward screening via cumulative divergence. *Journal of the American Statistical Association*, 115(531):1393–1405.

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(496):1464–1475.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

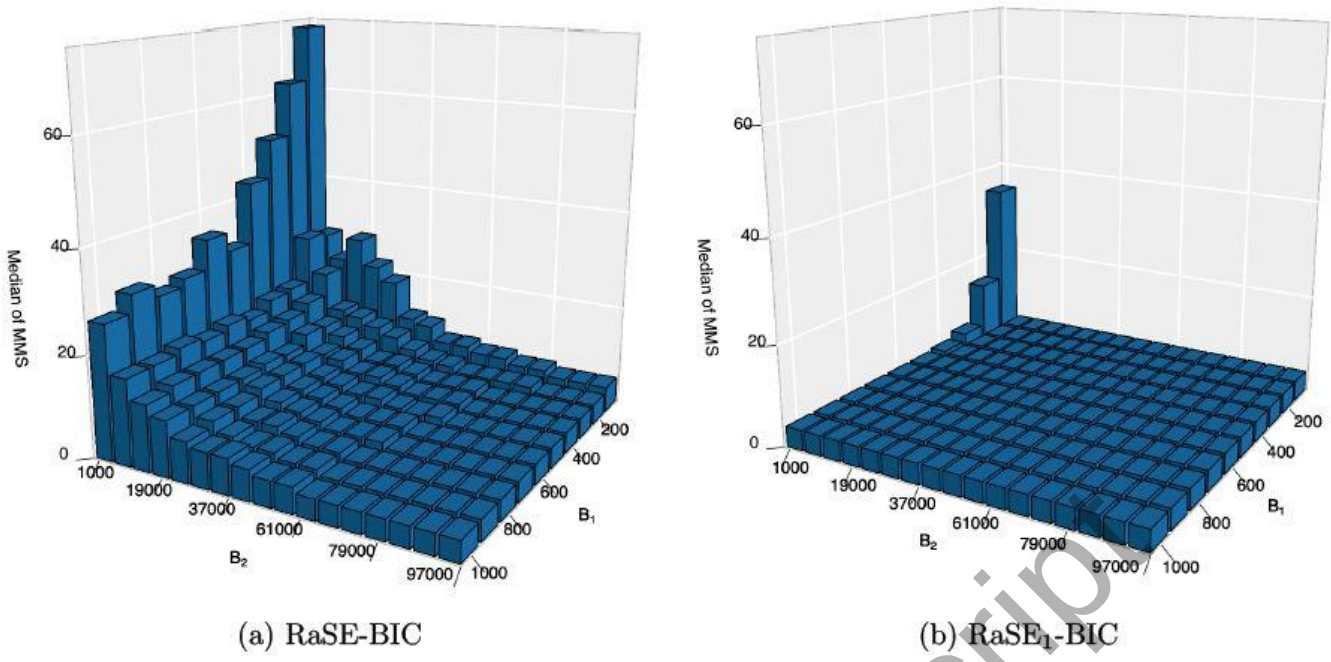

Fig. 1 Median MMS to capture S^* ($|S^*|=4$) as (B_1, B_2) varies for RaSE-BIC (a) and RaSE$_1$-BIC (b) in Example 1.

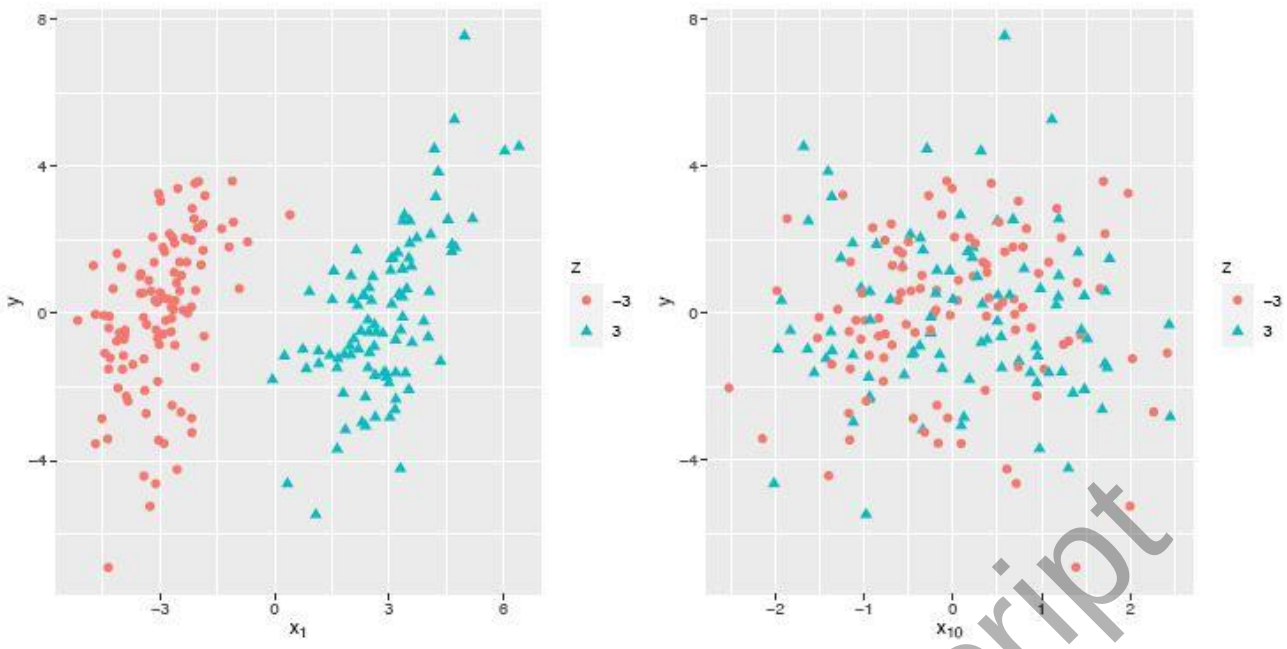

Fig. 2 Scatterplots of y vs. x_1 and y vs. x_{10} for Example 2 (n = 200).

Table 1 Quantiles of MMS in Examples 1 and 2.

Method/MMS	Example 1					Example 2				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	227	317	397	647	922	6	28	105	592	1855
ISIS	14	15	15	15	25	172	861	1415	1825	1963
SIRS	87	370	594	762	949	6	1158	1492	1774	1964
DC-SIS	96	358	610	776	942	6	1083	1460	1752	1976
HOLP	912	949	969	986	999	45	196	576	1252	1906
IPDC	224	442	700	869	980	59	210	386	678	1517
MDC-SIS	146	287	512	734	937	6	20	93	999	1908
CIS	203	434	601	780	940	2000	2000	2000	2000	2000
RaSE-BIC	5	12	37	126	650	6	358	1514	1821	1956
RaSE_1-BIC	4	4	4	16	55	13	834	1507	1797	1969
RaSE-eBIC	6	21	42	489	852	8	26	1323	1789	1935
RaSE_1-eBIC	4	4	4	4	14	907	1485	1739	1878	1971
RaSE-*k*NN	22	88	233	312	883	5	5	6	76	1190
RaSE_1-*k*NN	6	80	422	694	921	5	5	5	13	1846
RaSE-SVM	13	59	150	336	842	5	5	5	6	68
RaSE_1-SVM	4	4	82	126	542	5	5	5	5	11

Table 2 Average (over 200 replications) computational time in seconds for various methods in Example 1. For simplicity, for RaSE methods, we use criteria to differentiate them and the subscript "1" denotes the one-step iterative version of the corresponding RaSE-based methods.

Other methods	SIS	ISIS	SIRS	DC-SIS	HOLP	IPDC	MDC-SIS	CIS
Time (s)	0.01	0.67	0.28	1.30	0.02	0.49	0.28	1.18
RaSE methods	BIC	BIC_1	eBIC	$eBIC_1$	*k*NN	kNN_1	SVM	SVM_1
Time (s)	1.99	4.03	2.01	3.94	6.74	13.66	150.41	305.77

Table 3 Quantiles of MMS in Examples 3 and 4.

Method/MMS	Example 3					Example 4				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	184	810	1370	1732	1957	264	570	709	885	984
ISIS	362	1008	1482	1775	1945	293	626	810	911	978
SIRS	54	741	1294	1634	1920	487	737	867	935	992
DC-SIS	25	456	1222	1638	1923	44	304	603	814	949
HOLP	326	954	1475	1774	1975	316	586	767	886	974
IPDC	128	429	920	1397	1899	7	19	68	158	528
MDC-SIS	52	165	504	1331	1872	189	482	736	889	979
CIS	4	5	8	55	548	5	33	136	352	789
RaSE-BIC	637	1242	1619	1842	1959	355	693	825	914	986
RaSE_1-BIC	714	1196	1550	1839	1974	424	661	824	918	981
RaSE-eBIC	484	1137	1496	1794	1951	302	553	784	913	987
RaSE_1-eBIC	725	1330	1617	1806	1948	480	686	860	930	986
RaSE-*k*NN	5	33	168	1321	1855	5	15	68	290	889
RaSE_1-*k*NN	4	5	8	125	1528	4	8	51	446	910
RaSE-SVM	4	18	504	1282	1848	4	15	132	468	938
RaSE_1-SVM	4	4	5	14	1141	4	30	232	645	898

Table 4 Quantiles of MMS in Examples 5 and 6.

Method/MMS	Example 5					Example 6				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	515	1090	1414	1746	1947	170	471	910	1436	1932
ISIS	445	1001	1470	1784	1967	7	7	7	8	8
SIRS	2	2	2	2	2	821	1242	1551	1813	1966
DC-SIS	451	960	1385	1706	1913	765	1155	1526	1775	1947
MV-SIS	379	957	1366	1692	1895	199	706	1258	1660	1909
HOLP	495	1065	1381	1712	1936	—	—	—	—	—
IPDC	495	1010	1344	1673	1908	879	1425	1722	1884	1988
MDC-SIS	462	1038	1332	1708	1948	163	498	1064	1628	1917
CIS	2000	2000	2000	2000	2000	229	736	1195	1652	1941
RaSE-BIC	506	1081	1487	1804	1946	8	14	20	26	1525
RaSE_1-BIC	464	968	1360	1692	1927	5	5	5	6	14
RaSE-eBIC	425	1045	1424	1705	1965	26	346	894	1406	1919
RaSE_1-eBIC	480	988	1370	1727	1938	5	7	10	14	1184
RaSE-*k*NN	2	3	5	6	8	38	202	294	1470	1925
RaSE_1-*k*NN	2	2	2	2	2	27	376	967	1486	1828
RaSE-SVM	2	4	6	8	26	11	39	118	343	1743
RaSE_1-SVM	2	2	2	2	2	5	5	118	1133	1792

Table 5 Average test classification error rate with standard deviations (in parentheses) for colon cancer data set and average test mean square errors (MSEs) with standard deviations (in parentheses) for rat eye expression data set. We boldface the values corresponding to the best performances and italicize the values corresponding to the subsequent two best performances.

Screening	Post-screening	Cancer	Eye
—	LASSO	0.1792(0.1427)	0.0103(0.0091)
SIS		0.1633(0.1407)	*0.0091(0.0068)*
ISIS		0.1767(0.1444)	*0.0091(0.0068)*
SIRS		0.2800(0.1734)	0.0132(0.0123)
DC-SIS		0.3000(0.1998)	0.0124(0.0118)
MV-SIS		0.2958(0.1826)	—
HOLP		0.1825(0.1491)	0.0228(0.0269)
IPDC		0.1917(0.1464)	0.0129(0.0132)
MDC-SIS		0.1600(0.1406)	0.0103(0.0071)
CIS		0.1550(0.1332)	0.0194(0.0231)
RaSE-BIC		**0.1192(0.1277)**	**0.0090(0.0066)**
RaSE$_1$-BIC		*0.1417(0.1324)*	0.0123(0.0104)
RaSE-eBIC		0.3083(0.2118)	0.0092(0.0069)
RaSE$_1$-eBIC		0.1458(0.1397)	0.0122(0.0098)
—	*k*NN	0.2258(0.1653)	0.0166(0.0206)
RaSE-*k*NN		0.1533(0.1340)	0.0131(0.0158)
RaSE$_1$-*k*NN		0.1867(0.1500)	0.0133(0.0161)
—	SVM	0.2025(0.1503)	0.0160(0.0243)
RaSE-SVM		*0.1375(0.1277)*	0.0158(0.0231)
RaSE$_1$-SVM		0.1858(0.1477)	0.0158(0.0232)

Notes

[1] For more details, please refer to the toy example in Appendix A.3.1.

[2] For SIS, ISIS, SIRS, DC-SIS and HOLP, since the package implementing them does not provide the option to use multi-cores, we ran them with a single core only.

[3] For RaSE methods, sometimes there might be less than $[n/\log n]$ variables which have positive selected proportion. In this case, we randomly choose from the remaining variables with 0 selected proportions to have the desired number of selected variables.