Electronic Journal of Statistics
Vol. 16 (2022) 498–526
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1951

On sufficient variable screening using log odds ratio filter

Baoying Yang

Department of Statistics, College of Mathematics
Southwest Jiaotong University, Chengdu, China
e-mail: yangbaoying@swjtu.edu.cn

Wenbo Wu*

Department of Management Science and Statistics
The University of Texas at San Antonio, San Antonio, TX
e-mail: wenbo.wu@utsa.edu

Xiangrong Yin†

Department of Statistics, University of Kentucky
319 Multidisciplinary Science Building, Lexington, KY 40536
e-mail: yinxiangrong@uky.edu

Abstract: For ultrahigh-dimensional data, variable screening is an important step to reduce the scale of the problem, hence, to improve the estimation accuracy and efficiency. In this paper, we propose a new dependence measure which is called the log odds ratio statistic to be used under the sufficient variable screening framework. The sufficient variable screening approach ensures the sufficiency of the selected input features in modeling the regression function and is an enhancement of existing marginal screening methods. In addition, we propose an ensemble variable screening approach to combine the proposed fused log odds ratio filter with the fused Kolmogorov filter to achieve supreme performance by taking advantages of both filters. We establish the sure screening properties of the fused log odds ratio filter for both marginal variable screening and sufficient variable screening. Extensive simulations and a real data analysis are provided to demonstrate the usefulness of the proposed log odds ratio filter and the sufficient variable screening procedure.

MSC2020 subject classifications: Primary 62G05, 62J02; secondary 62B99.
Keywords and phrases: Feature screening, dependence measures, sufficient variable screening.

Received September 2021.

Contents

1 Introduction 499

*Corresponding author.
†Professor Xiangrong Yin passed away in August of 2020. He had made significant contributions to the development of this paper.

2 Sufficient variable screening 502
2.1 Framework 502
2.2 Algorithms 503
2.3 Ensemble 504
3 Log odds ratio filter 505
3.1 Motivation 506
3.2 Proposed methodology 506
3.3 The fused log odds ratio filter for sufficient variable screening . . 509
4 Theory 510
4.1 Regularity conditions for marginal screening 510
4.2 Sure screening property for marginal screening 511
4.3 Regularity conditions for sufficient screening 512
4.4 Sure screening property for sufficient screening 514
5 Numerical studies 515
5.1 Simulations 515
5.2 Real data example 522
6 Discussions 524
Acknowledgments 524
Supplementary Material 524
References 524

1. Introduction

Ultrahigh-dimensional data have emerged recently in many areas of modern scientific research, including microarray, genomic, proteomic, brain images and genetic data. There are known challenges such as scalability, noise accumulation, high collinearity, and spurious correlation for analyzing the ultrahigh-dimensional data [10, 7, 12]. To improve the scalability and reduce the noise accumulation, one possible approach is to first reduce the dimensionality of the feature space from a very large scale to a moderate one using a screening procedure and then implement learning algorithms or make inferences based on the much reduced feature space. By doing so, one not only can drastically speed up the computation process, but also can significantly improve the estimation accuracy when the dimensionality of the data is ultrahigh.

Under the assumption that only a small number of variables, which are usually referred as active features, among all observed input features contribute to the response variable, [10] propose the sure independent screening (SIS) method to identify a subset of features that contains the active features. The SIS method is based on the marginal Pearson correlation between an individual feature and the response variable and is designed for the linear regression model under which both response variable and input features follow the Gaussian distribution. Along this direction, many model-based screening procedures have been proposed in recent years under different parametric, semi-parametric, or nonparametric assumptions [see e.g., 12, 8, 18, 2, 11, 20, 24]. Nevertheless, specifying a correct model for ultrahigh-dimensional data remains to be a challenging

task. To tackle this problem, several model-free sure screening procedures have been developed [see e.g., 30, 19, 23, 1, 15, 21, 22, 16, 6, 17] so that the sure screening property can be achieved under much weaker assumptions on the regression function.

While sure screening methods are useful in analyzing ultrahigh-dimensional data, there are some known limitations. First, most screening methods rely on the marginal dependence between input features and the response variable. The marginal screening methods work well only if the noise features are weekly associated with the active features. To deal with strong correlations among the features for model-based screening method, [10] suggest an iterative screening and model fitting procedure which has been demonstrated its usefulness empirically [10, 12, 8], but theoretical justifications for this approach are missing. Another limitation of many screening methods is that they are either proposed based on a specific model or under certain parametric assumptions on the features. Lastly, many screening methods are not invariant to the monotone transformations of the features. That is, the screening results differ with or without making monotone transformations on the features [e.g., 19]. Recent studies have found the Kolmogorov-Smirnov test statistic useful for the variable screening purpose. As proposed by [21], variable screening using the Kolmogorov filter is fully nonparametric and is invariant under monotone transformations. In addition, the fused Kolmogorov filter [22, 16] is shown to be an effective variable screening method when the input features and the response variable are either discrete or continuous.

If we denote the response variable by Y and p-dimensional input features by $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\mathsf{T}} \in \mathbb{R}^p$, in the ultrahigh-dimensional problems where p is very large relative to the sample size, the sure screening methods are to identify a subset of features $\mathbf{X}_{\mathcal{D}}$ such that it contains the true active set of features $\mathbf{X}_{\mathcal{A}}$ (i.e., $\mathcal{A} \subseteq \mathcal{D}$). Hence, sure screening procedures aim to identify majority of the features in $\mathcal{A}^c$ which is the complement of the index set $\mathcal{A}$. In contrast, variable selection procedures more ambitiously try to recover $\mathcal{A}$ exactly [26, 9, 31]. From a different perspective, [28] introduce the concept of sufficient variable selection to deal with the "large p, small n" problem where n is the sample size of the observed data. Let $\mathbf{B}$ be a $p \times q$ matrix with $q \leq p$, where the columns of $\mathbf{B}$ consist of p-dimensional unit vectors $\mathbf{e}_k$ of which the k-th element is 1. The subspace spanned by the columns of $\mathbf{B}$ is called a variable selection space if $Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^{\mathsf{T}}\mathbf{X}$. The intersection of all such variable selection spaces, if exists, is called the central variable selection space and is denoted by $\mathcal{S}^V_{Y|\mathbf{X}}$. It can be shown that the central variable selection space exists under mild conditions [3, 29]. We assume the existence of $\mathcal{S}^V_{Y|\mathbf{X}}$ in this paper. It can be seen that the set of features involved in $\mathcal{S}^V_{Y|\mathbf{X}}$ are equivalent to $\mathbf{X}_{\mathcal{A}}$. Hence, a sufficient variable selection procedure is equivalent to identify $\mathcal{A}$ such that $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$. [27] note that the marginal screening methods identify features in $\mathcal{A}^c$ by evaluating the marginal independence $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c}$ instead of the conditional independence $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}$. Compared with existing marginal screening methods, sufficient screening is particularly useful to improve marginal screening methods in at least two situations:

(i) when the correlations among features are relatively strong; or (ii) when some active features demonstrate weak dependence to the response marginally but strongly associated with the response conditioning on some other correlated features. To achieve sufficient feature screening, [27] propose a variable screening framework based on the conditional independence using distance correlation [25] and Hilbert-Schmidt Independence Criterion [14]. However, [27]'s algorithm requires using a dependence measure to evaluate the association between two random vectors. As a consequence, some classic dependence measures that are defined only to measure the association between two univariate random variables, such as Pearson's correlation or Kolmogorov-Smirnov statistic [21], cannot be used under their framework. In addition, the distance correlation and the Hilbert-Schmidt Independence Criterion are not invariant under monotone transformations.

In this paper, we establish a new sufficient feature screening framework that is suitable for dependence measures defined only for univariate random variables. Therefore, our approach is a generalization of the sufficient screening framework to incorporate any dependence measure without a constraint on the dimension of the random variables (i.e., multivariate *vs* univariate). In addition, we propose an ensemble algorithm to further improve the sufficient screening performance using different dependence measures. The improvement in screening performance brought by our proposed ensemble approach is illustrated by abundant simulation studies. In addition, we propose a new dependence measure, which we call the *log odds ratio statistic*, to assess the statistical association between two random variables. We show that the proposed log odds ratio statistic can be used for variable screening and the log odds ratio filter is fully nonparametric and model-free. It is also invariant under monotone transformation on features. More importantly, it outperforms the fused Kolmogorov filter [22] for the situation when the conditional cumulative distribution function (c.d.f.) $F(y|X_j = x_1)$ and $F(y|X_j = x_2)$ are close to each other for all pairs of (x_1, x_2) and especially when both are close to 0 or 1. By definition, the log odds ratio filter can be applied to the data where the response variable and the input features are either discrete or continuous. Owning their advantages over different situations, the proposed fused log odds ratio filter can be combined with the fused Kolmogorov filter as a complement to each other to achieve better performance under an ensemble approach. We show that the fused log odds ratio filter enjoys sure screening properties for both marginal screening and sufficient variable screening.

The rest of this paper is organized as follows. In Section 2, we introduce a sufficient variable screening framework which can be adopted by any dependence measure defined between two univariate random variables. Based on a new dependence measure, we propose to use the fused log odds ratio filter for sufficient variable screening in Section 3. Sure screening properties of the fused log odds ratio filter are established in Section 4. Section 5 contains simulation studies and a real data application. We conclude with discussions in Section 6. Additional remarks and technical proofs are included in the appendix.

2. Sufficient variable screening

2.1. Framework

For ultrahigh-dimensional data with $p \gg n$, the sparsity assumption assumes that only a small subset of $\mathbf{X}$ are associated with Y. Denote this active set by $\mathcal{A}$, the sparsity assumption is equivalent to assume that $F(Y|\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}^c}) = F(Y|\mathbf{X}_{\mathcal{A}})$, where $\mathcal{A}^c$ is the complement of $\mathcal{A}$. [28] provide an equivalent formulation of the problem as $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}}$ and define the central variable selection subspace $\mathcal{S}^V_{Y|\mathbf{X}}$ which involves features in $\mathbf{X}_{\mathcal{A}}$. Through this definition, [28] discussed the existence and uniqueness of the central variable selection space. The conditional independence $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}}$ indicates that if we can identify $\mathbf{X}_{\mathcal{A}}$, we can eliminate $\mathbf{X}_{\mathcal{A}^c}$ and achieve the goal of variable screening without losing any regression information. Motivated by the conditional independence, a sufficient variable screening method was proposed by [27] using the following lemma.

Lemma 1. *[Proposition 1 of [27]] Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ be any arbitrary random vectors and Y is a random variable. Then, either one of the two conditions:*

$$(i)\ (Y, \boldsymbol{X}_2) \perp\!\!\!\perp \boldsymbol{X}_1, \ \textit{or}$$
$$(ii)\ \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2|Y \ \textit{and}\ Y \perp\!\!\!\perp \boldsymbol{X}_1,$$

implies the condition: (iii) $Y \perp\!\!\!\perp \boldsymbol{X}_1|\boldsymbol{X}_2$.

This lemma sheds light on the sufficient variable screening. Statement (iii) implies that $F(Y|\mathbf{X}_1, \mathbf{X}_2) = F(Y|\mathbf{X}_2)$. Hence, let $\mathbf{X} = (\mathbf{X}_1^\intercal, \mathbf{X}_2^\intercal)^\intercal$, if we can iteratively eliminate $\mathbf{X}_1$ at each step and treat $\mathbf{X}_2$ as a new $\mathbf{X}$, and repeat the process until no additional variable can be eliminated, we can obtain a set of variables that contains $\mathbf{X}_{\mathcal{A}}$. Although statement (iii) is the ultimate goal of sufficient variable screening, it is difficult to measure the conditional independence directly because we do not know $\mathbf{X}_2$ in advance. Lemma 1 enables us to validate statement (iii) through validating statement (i) or statement (ii). To this end, [27] propose two sufficient variable screening approaches based on statements (i) and (ii), which respectively, they call *one-stage* and *two-stage* sufficient variable selection. To test the conditional independence $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|Y$ in statement (ii), [27] adopt a slicing approach by discretizing the values of Y.

In general, the sufficient variable screening framework developed under statements (i) or (ii) of Lemma 1 is only applicable to dependence measures that are defined for measuring the associations between two random vectors. In particular, [27] use distance correlation [DC, 25] and Hilbert-Schmidt Independence Criterion (HSIC) [14] in their study. When a measurement is only defined to measure the dependence between two univariate random variables, e.g. the Kolmogorov statistic [21, 22], we have to modify the framework so that such a measurement can be used. To achieve this goal, we make separate observations from statement (i) and statement (ii) of Lemma 1. Statement (i) implies that $Y \perp\!\!\!\perp \mathbf{X}_1$ and $\mathbf{X}_2 \perp\!\!\!\perp \mathbf{X}_1$. For any arbitrary feature $X_\alpha \in \mathbf{X}_1$ and $X_\beta \in \mathbf{X}_2$, if $Y \not\perp\!\!\!\perp X_\alpha$ or $X_\beta \not\perp\!\!\!\perp X_\alpha$, then $(Y, X_\beta) \not\perp\!\!\!\perp X_\alpha$ and, hence, $Y \not\perp\!\!\!\perp X_\alpha|X_\beta$. On

the other hand, statement (ii) suggests that if $X_\alpha \perp\!\!\!\perp Y$ and $X_\alpha \perp\!\!\!\perp X_\beta | Y$, then $Y \perp\!\!\!\perp X_\alpha | X_\beta$. These relationships involve only univariate random variables X_α, X_β and Y, hence, inspire us to propose the following sufficient variable screening algorithms using the dependence measures defined only for measuring the associations between univariate random variables.

It is noteworthy that Lemma 1 reveals the fundamental differences between the sufficient variable screening methods and the marginal screening methods. While the traditional marginal screening methods [e.g., 10, 8, 22] focus on the marginal independence $Y \perp\!\!\!\perp X_\alpha$ which is the second part of statement (ii) in Lemma 1, the sufficient variable screening methods directly target on the conditional independence in statement (iii). To improve the performance of the marginal screening method, [10] propose to use an iterative procedure by computing the residuals from regressing the response Y over the selected variables. And then the residual is treated as a new response variable to iteratively screen over unselected variables to captured important variables that are missed from the previous step. Since the residuals are obtained based on the previously selected variables, to some extent, it uses the conditional information to avoid missing important variables.

2.2. Algorithms

To make the proposed framework general, let $\mathcal{I}(X, Y)$ denote an arbitrary index that measures the statistical dependence between two univariate random variables X and Y. We first note that using either statement (i) or statement (ii) of Lemma 1 requires to evaluate the marginal independence $Y \perp\!\!\!\perp X_\alpha$ by computing $\mathcal{I}(Y, X_\alpha)$, which conforms to the usual marginal screening methods. To complete the route of using statement (i) of Lemma 1, it additionally requires assessing the marginal independence $X_\beta \perp\!\!\!\perp X_\alpha$ by computing $\mathcal{I}(X_\beta, X_\alpha)$. On the other hand, to follow the path of statement (ii) of Lemma 1, we need to evaluate the conditional independence $X_\alpha \perp\!\!\!\perp X_\beta | Y$ in addition to the marginal independence $Y \perp\!\!\!\perp X_\alpha$. The assessment of this conditional independence is not trivial and we propose to use a slicing approach to overcome the challenge of computing $E[\mathcal{I}(X_\alpha, X_\beta)|Y]$. Define a general partition of the real line,

$$\mathbf{H} = \left\{ [s_{h-1}, s_h) : s_{h-1} < s_h; h = 1, \ldots, H; \bigcup_{h=1}^{H} [s_{h-1}, s_h) \setminus \{s_0\} = \mathbb{R} \right\}, \quad (2.1)$$

where $s_0 = -\infty$ and $s_H = \infty$. We slightly abuse the notation and express all intervals as $[s_{h-1}, s_h)$ by noticing the fact that (s_0, s_1) is open. Note that the definition of the partition $\mathbf{H}$ in (2.1) is arbitrary on the real line and can be used for discretizing any continuous random variable. Let $\mathbf{H}_y$ be a partition of Y with H_y slices $\left[s_{h_y-1}, s_{h_y}\right)$ for $h_y = 1, \ldots, H_y$. We define $\tilde{Y} = h_y$ if and only if $Y \in \left[s_{h_y-1}, s_{h_y}\right)$. If Y is already discrete, we can simply set $\tilde{Y} = Y$. With the

discrete $\tilde{Y}$, we can approximate $E[\mathcal{I}(X_\alpha, X_\beta)|Y]$ as

$$E[\mathcal{I}(X_\alpha, X_\beta)|Y] \approx E[\mathcal{I}(X_\alpha, X_\beta)|\tilde{Y}] = \frac{1}{H_y} \sum_{h_y=1}^{H_y} \left\{ \mathcal{I}(X_\beta, X_\alpha)|\tilde{Y} = h_y \right\}, \quad (2.2)$$

where $\mathcal{I}(X_\beta, X_\alpha)|\tilde{Y} = h_y$ is the index conditioned on $\tilde{Y} = h_y$ (i.e., $\mathcal{I}(X_\beta, X_\alpha)$ is computed within the h_y-th slice of Y). We denote sample estimates of the pivotal quantities by $\widehat{\mathcal{I}}(Y, X_\alpha)$ and $\widehat{\mathcal{I}}(X_\beta, X_\alpha)$ respectively. The algorithm of the sufficient variable screening is as follows.

Sufficient Variable Screening (SVS):

1. Compute $\hat{r}_\alpha = \widehat{\mathcal{I}}(Y, X_\alpha)$, for $\alpha = 1, \ldots, p$. We obtain an estimated index set $\widehat{\mathcal{A}}_1$ as the set of X_α's with the d_1 largest values of $\hat{r}_\alpha$.

2a. Compute $\hat{u}_\beta = \max_{\alpha \in \widehat{\mathcal{A}}_1} \left\{ \widehat{\mathcal{I}}(X_\beta, X_\alpha) \right\}$ for every $\beta \in \widehat{\mathcal{A}}_1^c$. We obtain an estimated index set $\widehat{\mathcal{A}}_2$ as the set of X_β's in $\mathbf{X}_{\widehat{\mathcal{A}}_1^c}$ with the d_2 largest values of $\hat{u}_\beta$.

2b. Or alternatively, use partition (2.1) to slice Y into H_y non-overlapping slices and obtain its discrete surrogate $\tilde{Y}$. Then compute $\widehat{v}_\beta = \max_{\alpha \in \widehat{\mathcal{A}}_1} \left\{ \frac{1}{H_y} \sum_{h_y=1}^{H_y} \left(\widehat{\mathcal{I}}(X_\beta, X_\alpha)|\tilde{Y} = h_y \right) \right\}$ for every $\beta \in \widehat{\mathcal{A}}_1^c$. We obtain an estimated index set $\widehat{\mathcal{A}}_2$ as the set of X_β's in $\mathbf{X}_{\widehat{\mathcal{A}}_1^c}$ with the d_2 largest values of $\widehat{v}_\beta$.

3. The final estimate of the active feature set is $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_1 \cup \widehat{\mathcal{A}}_2$.

In above algorithm, Step 2a and Step 2b follow separate paths led by statement (i) and statement (ii) of Lemma 1 to achieve sufficient variable screening. We call the SVS methods using each approach *SVS–I* and *SVS–II* procedures respectively. While the traditional marginal screening methods focus on only Step 1 by assessing the marginal dependence and estimate the active feature set directly by $\widehat{\mathcal{A}}_1$, the additional step (2a or 2b) ensure the sufficiency of the selected features.

From the practical point of view, for an observed sample data, we need to determine how to partition the response variable Y and choose appropriate values of d_1 and d_2. In our implementation, we set h_y in (2.1) to be the h_y/H_y-th sample quantiles of Y and set $H_y = 2$. [19] suggest to use $d_n = \lceil n/\log n \rceil$ to be the size of $\widehat{\mathcal{A}}$ where n is the sample size and we follow the suggestion of [27] to set $d_1 = 0.95 d_n$ and $d_2 = 0.05 d_n$. Our simulations indicate that these settings generally perform well.

2.3. Ensemble

The proposed SVS framework can be implemented using any index $\mathcal{I}(X, Y)$, such as Pearson correlation, distance correlation [25], Hilbert-Schmidt Independence Criterion [14], and the Kolmogorov statistic [21, 22] among others.

However, different dependence measures enjoy their own advantages for different situations and it is difficult to know in advance which measure is preferred for a specific data. Therefore, we further propose an ensemble SVS approach to combine the use of different dependence measures under the SVS framework. Let $\mathcal{I}_m(X,Y)$ for $m=1,\ldots,M$ be M different dependence measures under consideration. The algorithm of ensemble sufficient variable screening is as follows.

Ensemble Sufficient Variable Screening (ESVS):

1. For each $m=1,\ldots,M$, compute $\hat{r}_\alpha^m=\widehat{\mathcal{I}}_m(Y,X_\alpha)$, for $\alpha=1,\ldots,p$. Denote by $\hat{r}_{(\alpha)}^m$ the order statistics of $\hat{r}_\alpha^m$ such that $\hat{r}_{(1)}^m\le\hat{r}_{(2)}^m\le\cdots\le\hat{r}_{(p)}^m$, let $\varphi(\hat{r}_\alpha^m)$ be the rank of $\hat{r}_\alpha^m$ among all p features such that $\hat{r}_\alpha^m=\hat{r}_{(\varphi(\hat{r}_\alpha^m))}^m$ (i.e., a relatively larger $\hat{r}_\alpha^m$ is corresponding to a higher $\varphi(\hat{r}_\alpha^m)$). Then define a combined rank for each X_α as $\varphi(\hat{r}_\alpha^*)=\max_m\{\varphi(\hat{r}_\alpha^m)\}$. We obtain an estimated index set $\widehat{\mathcal{A}}_1$ as the set of X_α's with the d_1 largest values of $\varphi(\hat{r}_\alpha^*)$.

2a. For each $m=1,\ldots,M$, compute $\hat{u}_\beta^m=\max_{\alpha\in\widehat{\mathcal{A}}_1}\left\{\widehat{\mathcal{I}}_m(X_\beta,X_\alpha)\right\}$ for every $\beta\in\widehat{\mathcal{A}}_1^c$. Let $\varphi(\hat{u}_\beta^m)$ denote the rank of $\hat{u}_\beta^m$ among all features in $\widehat{\mathcal{A}}_1^c$. Then define a combined rank for each X_β as $\varphi(\hat{u}_\beta^*)=\max_m\left\{\varphi(\hat{u}_\beta^m)\right\}$. We obtain an estimated index set $\widehat{\mathcal{A}}_2$ as the set of X_β's with the d_2 largest values of $\varphi(\hat{u}_\beta^*)$.

2b. Use partition (2.1) to slice Y into H_y non-overlapping slices by defining $\tilde{Y}$ and for each $m=1,\ldots,M$, compute

$$\widehat{v}_\beta^m=\max_{\alpha\in\widehat{\mathcal{A}}_1}\left\{\frac{1}{H_y}\sum_{h_y=1}^{H_y}\left(\widehat{\mathcal{I}}_m(X_\beta,X_\alpha)|\tilde{Y}=h_y\right)\right\}$$

for every $\beta\in\widehat{\mathcal{A}}_1^c$. Let $\varphi(\hat{v}_\beta^m)$ denote the rank of $\hat{v}_\beta^m$ among all features in $\widehat{\mathcal{A}}_1^c$. Then define a combined rank for each X_β as $\varphi(\hat{v}_\beta^*)=\max_m\left\{\varphi(\hat{v}_\beta^m)\right\}$. We obtain an estimated index set $\widehat{\mathcal{A}}_2$ as the set of X_β's with the d_2 largest values of $\varphi(\hat{v}_\beta^*)$.

3. The final estimate of the active feature set is $\widehat{\mathcal{A}}=\widehat{\mathcal{A}}_1\cup\widehat{\mathcal{A}}_2$.

We call methods using Step 2a and Step 2b in ESVS algorithm *ESVS–I* and *ESVS–II* procedures respectively.

3. Log odds ratio filter

While many existing dependence measures can be used for the proposed SVS framework, we propose a new measure called the log odds ratio statistic, which is fully nonparametric and invariant under monotone transformation, to be used under the SVS framework.

3.1. Motivation

Before presenting the proposed log odds ratio filter for sufficient variable screening, we provide a brief review of the fused Kolmogorov filter [22]. Based on the fact that a random variable X is independent of Y if and only if the conditional distributions of $X|Y = y$ are identical for all y's, [22] propose using

$$K_j = \sup_{y_1,y_2} \sup_x \left| F_{X_j|Y}(x|Y = y_1) - F_{X_j|Y}(x|Y = y_2) \right| \tag{3.1}$$

to measure the dependence between X_j and Y, where $F_{X_j|Y}$ is the conditional c.d.f. of X_j given Y. To facilitate the empirical estimation, [22] consider a sliced version of (3.1) using the partition defined in (2.1). Given a partition $\mathbf{H}_y$, define a discrete random variable $\tilde{Y} \in \{1, \ldots, H_y\}$ such that $\tilde{Y} = h_y$ if and only if Y is in the h_y-th slice. Then [22] let

$$K_j(\mathbf{H}_y) = \max_{h_1,h_2} \sup_x \left| F_{X_j|\tilde{Y}}(x|\tilde{Y} = h_1) - F_{X_j|\tilde{Y}}(x|\tilde{Y} = h_2) \right|, \tag{3.2}$$

and show that X_j is independent of Y if and only if $K_j(\mathbf{H}_y) = 0$ for all possible choices of $\mathbf{H}_y$. Since $K_j(\mathbf{H}_y)$ depends on a particular choice of partition $\mathbf{H}_y$, motivated by a fusion idea [5], [22] define the fused Kolmogorov filter as

$$K_j^* = \sum_{l=1}^{N} K_j(\mathbf{H}_y^l), \tag{3.3}$$

based on N different partitions $\mathbf{H}_y^l$, for $l = 1, \ldots, N$, where each partition $\mathbf{H}_y^l$ contains H_y^l intervals. [22] showed that the sample estimate $\widehat{K}_j^*$ of (3.3) can be effectively used for marginal variable screening as the fused Kolmogorov filter.

While it is useful in variable screening, the fused Kolmogorov filter can hardly identify informative features when $F_{X_j|Y}(x|Y = y_1)$ and $F_{X_j|Y}(x|Y = y_2)$ are essentially different but very similar, especially when both of them are close to 0 or 1. Consider two scenarios where in the first scenario $F_{X_j|Y}(x|Y = y_1) = 0.01$ and $F_{X_j|Y}(x|Y = y_2) = 0.001$, and in the second scenario $F_{X_j|Y}(x|Y = y_1) = 0.41$ and $F_{X_j|Y}(x|Y = y_2) = 0.401$. Although both differences are 0.009, the difference in the first scenario is much more noteworthy and significant because $F_{X_j|Y}(x|Y = y_1)$ is 10 times larger than $F_{X_j|Y}(x|Y = y_2)$. Therefore, in order to capture the important variables in such a case, we propose to use the difference between $\log\left(\frac{F_{X_j|Y}(x|Y=y_1)}{1-F_{X_j|Y}(x|Y=y_1)}\right)$ and $\log\left(\frac{F_{X_j|Y}(x|Y=y_2)}{1-F_{X_j|Y}(x|Y=y_2)}\right)$ instead of the difference between $F_{X_j|Y}(x|Y = y_1)$ and $F_{X_j|Y}(x|Y = y_2)$ in measuring the statistical dependence of X_j and Y.

3.2. Proposed methodology

We define the log odds ratio statistic as

$$R_{Y|X_j} = \sup_{x_1,x_2} \sup_y \left| \log\left(\frac{F_{Y|X_j}(y|X_j = x_1)}{1 - F_{Y|X_j}(y|X_j = x_1)}\right) - \log\left(\frac{F_{Y|X_j}(y|X_j = x_2)}{1 - F_{Y|X_j}(y|X_j = x_2)}\right) \right|. \tag{3.4}$$

In order to avoid singularity points at $F_{Y|X_j}(y|X_j = x_0) = 0$ or $F_{Y|X_j}(y|X_j = x_0) = 1$ for any x_0, we introduce a threshold constant $\tau > 0$ such that we set $F_{Y|X_j}(y|X_j = x_0)$ to τ if $F_{Y|X_j}(y|X_j = x_0) < \tau$, and set $F_{Y|X_j}(y|X_j = x_0)$ to $1 - \tau$ if $F_{Y|X_j}(y|X_j = x_0) > 1 - \tau$. In general, the magnitude of τ should be small (e.g., 10^{-5}) and it is related to our regularity condition (C2a) which is needed to ensure the sure screening property.

The conditional c.d.f. in (3.4) is based on the conditional distribution of $Y|X$ whereas in the Kolmogorov statistic (3.1) is in the form of $X|Y$. If Y and X are independent, then for any value x_0, $F_{Y|X_j}(y|X_j = x_0)$ are identical, or equivalently, $F_{Y|X_j}(y|X_j = x_0) = F_j(y)$. We choose to use the conditional distribution $Y|X$ because it conforms to the general goal of variable screening, that is, to identify X_j's such that $F_{Y|X_j}(y|X_j)$ is functionally dependent on X_j for some y. Note that, $F_{Y|X_j}(y|X_j = x_1) = \Pr(Y \le y|X_j = x_1)$ and the difference in (3.4) is the difference between two log odds which is equivalent to the log odds ratio. Hence, we call the variable screening procedure based on (3.4) the *log odds ratio filter* (*LogOR filter*; hereafter).

For binary response Y, [12] propose a maximum marginal likelihood screening method under the logistic regression model. Specifically, for each X_j, they consider a model

$$\log\left(\frac{\Pr(Y = 1|X_j)}{1 - \Pr(Y = 1|X_j)}\right) = \beta_0 + \beta_j X_j,$$

and rank variables based on the maximum likelihood estimate $\hat{\beta}_j^{\mathrm{M}}$ of β_j. [12] establish the sure screening property of this approach under certain conditions. Under the logistic regression model, it is well known that the interpretation of β_j is related to the log odds ratio between different values of X_j. However, the LogOR filter (3.4) is a more general approach than the maximum marginal likelihood screening method because the log odds ratio statistic works not only for binary response but also for continuous response and it is completely model-free.

For continuous X_j, we can follow a slicing approach similar to [22] to make the estimation easier. Given a specific partition $\mathbf{H}_{x_j}$ on X_j as defined in (2.1), for each X_j, define a discrete random variable $\tilde{X}_j \in \{1, \ldots, H_{x_j}\}$ such that $\tilde{X}_j = h$ if and only if X_j is in the h-th slice. Accordingly, we define

$$\begin{aligned} R_{Y|X_j}(\mathbf{H}_{x_j}) = \max_{h_1,h_2} \sup_y \Bigg| &\log\left(\frac{F_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_1)}{1 - F_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_1)}\right) \\ &- \log\left(\frac{F_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_2)}{1 - F_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_2)}\right)\Bigg|. \end{aligned} \tag{3.5}$$

The following proposition demonstrates some characteristics of $R_{Y|X_j}(\mathbf{H}_{x_j})$.

Proposition 1. *For $R_{Y|X_j}$ defined in (3.4) and $R_{Y|X_j}(\mathbf{H}_{x_j})$ defined in (3.5), the following statement are true.*

(i) *X_j is independent of Y if and only if $R_{Y|X_j}(\mathbf{H}_{x_j}) = 0$ for all possible choices of $\mathbf{H}_{x_j}$.*

(ii) *Assume that X_j is not independent of Y and for any fixed $x \in \mathbb{R}$, $\Pr(X_j \le x|Y = y)$ is not a constant in y; then $R_{Y|X_j}(\mathbf{H}_{x_j}) \neq 0$ for any choice of $\mathbf{H}_{x_j}$.*

(iii) *Assume that $F_{Y|\tilde{X}_j}(y|X_j = x)$ is continuous and $0 < \tau_1 \le F_{Y|\tilde{X}_j}(y|X_j = x) \le \tau_2 < 1$ for some $0 < \tau_1 \le \tau_2 < 1$. If $\max_{h=1,\ldots,H_{x_j}} \Pr(\tilde{X}_j = h) \to 0$ as $H_{x_j} \to \infty$, then $R_{Y|X_j}(\mathbf{H}_{x_j}) \to R_{Y|X_j}$ as $H_{x_j} \to \infty$. Therefore, if X_j is not independent of Y, $R_{Y|X_j}(\mathbf{H}_{x_j}) > 0$ for sufficiently large H_{x_j}.*

The proof of Proposition 1 is included in the appendix. Proposition 1 indicates that, for continuous X_j, $R_{Y|X_j}(\mathbf{H}_{x_j})$ serves as a good surrogate of $R_{Y|X_j}$ to capture the dependence between X_j and Y.

With an observed random sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\intercal$, an estimate of (3.5) can be obtained as

$$\widehat{R}_{Y|X_j}(\mathbf{H}_{x_j}) = \max_{h_1,h_2} \sup_y \left| \log\left(\frac{\hat{F}_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_1)}{1 - \hat{F}_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_1)} \right) - \log\left(\frac{\hat{F}_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_2)}{1 - \hat{F}_{Y|\tilde{X}_j}(y|\tilde{X}_j = h_2)} \right) \right|, \tag{3.6}$$

where

$$\hat{F}_{Y|\tilde{X}_j}(y|\tilde{X}_j = h) = \frac{1}{n_h} \sum_{i:\tilde{X}_{ij}=h} \mathbb{1}(Y_i \le y), \tag{3.7}$$

and n_h is the number of observations within the h-th slice, and $\tilde{X}_{ij} = h$ if X_{ij} is in the h-th slice. When X_j takes finite discrete values, we can let $\tilde{X}_j = X_j$ without using a partition. When X_j takes infinite discrete values, such as following a Poisson distribution, or X_j is continuous, we can use h/H_{x_j}-th sample quantiles in (2.1) to define the end-points for a partition. To search for the supremum over y, for any given h_1 and h_2, we can search on a set of grid points defined over the support of Y, $\boldsymbol{\Xi}_y = \{y_i : -\infty < y_1 < y_2 < \cdots < y_{k-1} < y_k < \infty\}$, to find the value $y^* \in \boldsymbol{\Xi}_y$ such that (3.6) is maximized. Note that we can also use partition $\mathbf{H}_y$ to define the grid points for $\boldsymbol{\Xi}_y$ and this approach works well in our simulation studies.

To further improve the stability and accuracy for variable screening, we can define a fused LogOR filter as

$$\widehat{R}^*_{Y|X_j} = \sum_{l=1}^{N} \widehat{R}_{Y|X_j}(\mathbf{H}^l_{x_j}), \tag{3.8}$$

which is an ensemble over N different partitions $\mathbf{H}^l_{x_j}$ each containing $H^l_{x_j}$ intervals. Note that the population version of (3.8) is $R^*_{Y|X_j} = \sum_{l=1}^{N} R_{Y|X_j}\left(\mathbf{H}^l_{x_j}\right)$.

To obtain different partitions for the fusion step, we consider multiple uniform slicing schemes under which $\mathbf{H}_{x_j}^l$ has $H_{x_j}^l$ many slices based on the sample quantiles of X_j for $1 \le l \le N$.

3.3. The fused log odds ratio filter for sufficient variable screening

In the marginal variable screening step, the true active marginal feature set is defined as

$$\mathcal{A}_1 = \{j : X_j \not\perp\!\!\!\perp Y\}.$$

Similar to many existing measures, the fused LogOR filter can readily be used for marginal variable screening by selecting input features X_j's that have relatively large $\widehat{R}^*_{Y|X_j}$ values.

In the SVS-I procedure, after obtaining the marginal screening set $\widehat{\mathcal{A}}_1$, it requires to compute $\hat{u}_\beta = \max_{\alpha\in\widehat{\mathcal{A}}_1}\left\{\widehat{\mathcal{I}}(X_\beta, X_\alpha)\right\}$ for every $\beta \in \widehat{\mathcal{A}}_1^c$ to obtain $\widehat{\mathcal{A}}_2$. Hence, to use the fused LogOR filter in this step, we can compute

$$\widehat{U}^*_{X_k|X_j} = \max_{j\in\widehat{\mathcal{A}}_1}\left\{\widehat{R}^*_{X_k|X_j}\right\}. \tag{3.9}$$

On the other hand, in the SVS-II procedure, after obtaining the marginal screening set $\widehat{\mathcal{A}}_1$, it requires to compute

$$\widehat{v}_\beta = \max_{\alpha\in\widehat{\mathcal{A}}_1}\left\{\frac{1}{H_y}\sum_{h_y=1}^{H_y}\widehat{\mathcal{I}}(X_\beta, X_\alpha)|\tilde{Y} = h_y\right\}$$

for every $\beta \in \widehat{\mathcal{A}}_1^c$, where H_y is the number of slices used in the partition (2.1) for Y to define its discrete surrogate $\tilde{Y}$. Therefore, in the sufficient screening step, we use

$$\widehat{V}^*_{X_k|X_j;Y} = \max_{j\in\widehat{\mathcal{A}}_1}\left\{\frac{1}{H_y}\sum_{h_y=1}^{H_y}\left(\widehat{R}^*_{X_k|X_j}|\tilde{Y} = h_y\right)\right\}, \tag{3.10}$$

to incorporate the fused LogOR filter. While there are many choices for the partition $\mathbf{H}_y$ of Y, we use h_y/H_y-th sample quantiles in (2.1) to define the end-points for $\mathbf{H}_y$.

We should note that the fused LogOR filter and the fused Kolmogorov filter have their advantages in different situations and our empirical experiences indicate that it is best to combine them to achieve the supreme screening results. Therefore, we can apply the ensemble algorithm as proposed in Section 2 to combine them in practice. We consider an ensemble of the fused LogOR filter and the fused Kolmogorov filter for they share many common characteristics, such as being fully nonparametric, model-free, and invariant under monotone transformations of response variable and input features. A detailed algorithm on how to combine the fused LogOR filter and the fused Kolmogorov filter is included in the appendix.

4. Theory

In this section, we will show that the fused LogOR filter (3.8) enjoys sure screening properties for both marginal screening and sufficient screening procedures.

4.1. Regularity conditions for marginal screening

Note that the fused LogOR filter is an ensemble over several $\widehat{R}_{Y|X_j}(\mathbf{H}^l_{x_j})$'s which depend on the empirical quantiles of X_j. If we know the distribution of X_j, we can use an oracle uniform partition scheme such that the partition $\mathbf{H}^l_{\mathrm{o}(x_j)}$ contains the intervals defined by the $h/H^l_{x_j}$-th theoretical quantitles of X_j, for $h = 1, \ldots, H^l_{x_j}$. In this situation, since the true distribution of X_j is assumed to be known, we denote the corresponding estimated log odds ratio statistic as $\widehat{R}^{(o)}_{Y|X_j}(\mathbf{H}^l_{\mathrm{o}(x_j)})$ and let

$$\widehat{R}^{(o)}_{Y|X_j} = \sum_{l=1}^{N} \widehat{R}_{Y|X_j}\left(\mathbf{H}^l_{\mathrm{o}(x_j)}\right). \tag{4.1}$$

We call (4.1) as the *oracle* fused logOR filter using the terminology of [22] and its population version is $R^{(o)}_{Y|X_j} = \sum_{l=1}^{N} R_{Y|X_j}\left(\mathbf{H}^l_{\mathrm{o}(x_j)}\right)$.

We define the screening sets obtained using the oracle fused LogOR filter (4.1) and the fused LogOR filter (3.8) as

$$\widehat{\mathcal{A}}_1\left(\widehat{R}^{(o)}_{Y|X_j}\right) = \left\{j : \widehat{R}^{(o)}_{Y|X_j} \text{ is among the } d_n \text{ largest } \widehat{R}^{(o)}_{Y|X_j} \text{ for } j = 1, \ldots, p\right\}, \tag{4.2}$$

and

$$\widehat{\mathcal{A}}_1\left(\widehat{R}^{*}_{Y|X_j}\right) = \left\{j : \widehat{R}^{*}_{Y|X_j} \text{ is among the } d_n \text{ largest } \widehat{R}^{*}_{Y|X_j} \text{ for } j = 1, \ldots, p\right\}, \tag{4.3}$$

respectively, for some predetermined d_n.

We should note that the definitions of the screening feature sets based on (4.2) and (4.3) are equivalent to the commonly used definitions in the literature [10, 8, 19],

$$\widetilde{\mathcal{A}}_1\left(\widehat{R}^{(o)}_{Y|X_j}\right) = \left\{j : \widehat{R}^{(o)}_{Y|X_j} \geq cn^{-\nu}\right\}, \quad \text{and} \quad \widetilde{\mathcal{A}}_1\left(\widehat{R}^{*}_{Y|X_j}\right) = \left\{j : \widehat{R}^{*}_{Y|X_j} \geq cn^{-\nu}\right\},$$

for some predetermined thresholding positive constants c and ν. [19] discussed the connection and equivalence of these definitions.

In order to establish the sure screening properties for $\widehat{\mathcal{A}}_1\left(\widehat{R}^{(o)}_{Y|X_j}\right)$ and $\widehat{\mathcal{A}}_1\left(\widehat{R}^{*}_{Y|X_j}\right)$, we need to assume the following regularity conditions.

(C1) There exists a set $\mathcal{D}$ such that $\mathcal{A}_1 \subset \mathcal{D}$ and

$$\Delta_{\mathcal{D}} = \min_l \left(\min_{j \in \mathcal{D}} R_{Y|X_j} \left(\mathbf{H}^l_{\mathrm{o}(x_j)} \right) - \max_{j \notin \mathcal{D}} R_{Y|X_j} \left(\mathbf{H}^l_{\mathrm{o}(x_j)} \right) \right) > 0.$$

(C2) Let $H^{\min}_{x_j} = \min_l \{H^l_{x_j}\}$. For any b_1, b_2 such that $P\left(X_j \in [b_1, b_2)\right) \leq 2/H^{\min}_{x_j}$, we have

$$0 < \tau_1 \leq F_{Y|X_j}\left(y | X_j \in [b_1, b_2)\right) \leq \tau_2 < 1, \tag{C2a}$$

for some constants $0 < \tau_1 \leq \tau_2 < 1$; and

$$\left| \log \left(\frac{F_{Y|X_j}(y|X_j = x_1)}{1 - F_{Y|X_j}(y|X_j = x_1)} \right) - \log \left(\frac{F_{Y|X_j}(y|X_j = x_2)}{1 - F_{Y|X_j}(y|X_j = x_2)} \right) \right| \leq \frac{\Delta_{\mathcal{D}}}{8}, \tag{C2b}$$

for all y, j and $x_1, x_2 \in [b_1, b_2)$.

Condition (C1) is the key condition which is commonly used in the literature for establishing the sure screening property of marginal screening methods. It assures that the important variables in the active set $\mathcal{A}_1$ are also marginally more important than the noise variables. In the appendix, we provide a discussion on how Condition (C1) can be satisfied in general.

Condition (C2a) requires the conditional c.d.f. to be bounded away from 0 and 1 such that the log odds ratio statistic (3.4) is well defined. This condition can be easily met by setting a small threshold value as described in Section 3. Condition (C2b) requires the sample quantiles of X_j's to be close enough to the population quantiles. Note that no other distributional and moment assumptions are needed to establish the marginal sure screening property of the fused LogOR filter.

4.2. Sure screening property for marginal screening

Theorem 1. *Assume conditions (C1) and (C2). If* $H^l_{x_j} \leq \lceil \log n \rceil$ *for all* l *and* $d_n \geq |\mathcal{D}|$, *then, for the screening sets* $\widehat{\mathcal{A}}_1 \left(\widehat{R}^{(o)}_{Y|X_j} \right)$ *and* $\widehat{\mathcal{A}}_1 \left(\widehat{R}^{*}_{Y|X_j} \right)$ *defined in (4.2) and (4.3), we have*

$$\Pr\left(\mathcal{A}_1 \subset \widehat{\mathcal{A}}_1 \left(\widehat{R}^{(o)}_{Y|X_j} \right) \right) \geq 1 - \eta \quad \textit{and} \quad \Pr\left(\mathcal{A}_1 \subset \widehat{\mathcal{A}}_1 \left(\widehat{R}^{*}_{Y|X_j} \right) \right) \geq 1 - \eta, \tag{4.4}$$

where $\eta = CNp\left(\log^2 n\right) \exp\left(-\frac{n\tau_*^2 \Delta^2_{\mathcal{D}}}{\log n} \right) + CN\left(\log^2 n\right) \exp\left(-\frac{n}{\log^2 n} \right)$ *for some generic positive constant* C *and* $\tau_* = \min(\tau_1, 1 - \tau_2)$.

The proof of Theorem 1 is provided in the appendix. According to (4.4), the fused LogOR filter has the same order as the oracle fused LogOR filter indicating that the proposed slicing scheme using sample quantitles is as good

as using the oracle information about the theoretical quantiles. In addition, Theorem 1 provides guidance on choosing the partition $\mathbf{H}^l_{x_j}$ because it requires $H^l_{x_j} \leq \lceil \log n \rceil$. Therefore, combining this requirement and suggestions in the literature [5, 22], we set $H^l_{x_j} = 3, \ldots, \lceil \log n \rceil$ to ensure a sufficient sample size within each slice.

From the results, we can see that both the oracle fused LogOR filter and the fused LogOR filter possess the sure screening property with a probability tending to 1 if

$$\Delta_{\mathcal{D}} \gg \sqrt{\frac{\log n \log(pN \log n)}{n}}.$$

For $N \leq \log n$, if there exists a constant $o < \kappa < 1$ such that $\Delta_{\mathcal{D}} \gg n^{-\kappa}$, above condition on $\Delta_{\mathcal{D}}$ is equivalent to $\log p \ll n^{\xi}$ for any $\xi \in (0, 1-2\kappa)$. This requirement on the order of p is same as many existing methods require [e.g., 10]. However, the fused LogOR filter is fully nonparametric and, hence, is more flexible.

It is worth noting that η is independent of d_n. Therefore, as long as we choose a sufficiently large d_n such that $d_n \geq |\mathcal{D}|$, the sure screening property is guaranteed. In our simulation, we use the conventional value $d_n = \lceil n/\log n \rceil$ suggested by [19].

4.3. Regularity conditions for sufficient screening

In the sufficient screening step of the SVS-I procedure, the major difference is to replace $\widehat{R}^*_{Y|X_j}$ by $\widehat{R}^*_{X_k|X_j}$. To establish the sure screening property of using $\widehat{U}^*_{X_k|X_j}$, we need to make assumptions similar to conditions (C1) and (C2) based on the conditional c.d.f. $F(X_k|X_j)$ instead of $F(Y|X_j)$. Then we could obtain similar results as Theorem 1. Given its similarity, we omit the discussions for this case, but focus on the SVS-II procedure which is a more complicated case because $\widehat{V}^*_{X_k|X_j;Y}$ in (3.10) involves partitions of both X_j and Y.

The second step in SVS-II procedure is based on selecting additional features that are marginally independent of the response. We can define the oracle filter for (3.10) as

$$\widehat{V}^{(o)}_{X_k|X_j;Y} = \max_{j \in \widehat{\mathcal{A}}_1} \left\{ \frac{1}{H_y} \sum_{h_y=1}^{H_y} \widehat{R}^{(o)}_{X_k|X_j} | \tilde{Y} = h_y \right\}, \tag{4.5}$$

of which the corresponding population parameter is

$$V^{(o)}_{X_k|X_j;Y} = \max_{j \in \mathcal{A}_1} \left\{ E\left[R^{(o)}_{X_k|X_j} | Y \right] \right\}.$$

Note that the population index set for the sufficient variable screening step is

$$\mathcal{A}_2 = \{ k : X_k \not\perp\!\!\!\perp X_j | Y \text{ for all } j \text{ such that } X_j \not\perp\!\!\!\perp Y \}.$$

We define the screening sets using $\widehat{V}^{(o)}_{X_k|X_j;Y}$ and $\widehat{V}^{*}_{X_k|X_j;Y}$ as

$$\widehat{\mathcal{A}}_2\Big(\widehat{V}^{(o)}_{X_k|X_j;Y}\Big) = \Big\{k : \widehat{V}^{(o)}_{X_k|X_j;Y} \text{ is among the } d'_n \text{ largest } \widehat{V}^{(o)}_{X_k|X_j;Y} \text{ for } j \in \widehat{\mathcal{A}}_1\Big\}, \tag{4.6}$$

and

$$\widehat{\mathcal{A}}_2\left(\widehat{V}^{*}_{X_k|X_j;Y}\right) = \left\{k : \widehat{V}^{*}_{X_k|X_j;Y} \text{ is among the } d'_n \text{ largest } \widehat{V}^{*}_{X_k|X_j;Y} \text{ for } j \in \widehat{\mathcal{A}}_1\right\}, \tag{4.7}$$

for some pre-determined d'_n.

In order to establish the sure screening properties of $\widehat{\mathcal{A}}_2\left(\widehat{V}^{(o)}_{X_k|X_j;Y}\right)$ and $\widehat{\mathcal{A}}_2\left(\widehat{V}^{*}_{X_k|X_j;Y}\right)$ for $\mathcal{A}_2$, we need to assume the following regularity conditions.

(C1*) There exists a set $\mathcal{D}_2$ such that $\mathcal{A}_2 \subset \mathcal{D}_2$ and

$$\Delta_{\mathcal{D}_2} = \min_{k\in\mathcal{D}_2} V^{(o)}_{X_k|X_j;Y} - \max_{k\notin\mathcal{D}_2} V^{(o)}_{X_k|X_j;Y} > 0.$$

(C2*) The observations are i.i.d. and conditions (C1) and (C2) hold within each slice of Y. Denote the logOR statistic $R_{X_k|X_j}\left(\mathbf{H}^l_{\mathrm{o}(x_j)}\right)$ within the slice h_y by $R_{X_k|X_j,h_y}\left(\mathbf{H}^l_{\mathrm{o}(x_j)}\right)$. Then, for any h_y, there exists a set $\mathcal{D}$ such that $\mathcal{A}_1 \subset \mathcal{D}$ and

$$0 \le \Delta^{h_y}_{\mathcal{D}} = \min_l\left(\min_{j\in\mathcal{D}} R_{X_k|X_j,h_y}\left(\mathbf{H}^l_{\mathrm{o}(x_j)}\right) - \max_{j\notin\mathcal{D}} R_{X_k|X_j,h_y}\left(\mathbf{H}^l_{\mathrm{o}(x_j)}\right)\right) \le \frac{\Delta_{\mathcal{D}_2}}{N}, \tag{C2*a}$$

Let $\Delta_{\mathcal{D}^*} = \max_{h_y}\Delta^{h_y}_{\mathcal{D}}$.

Moreover, within each slice of Y (i.e., given $\tilde{Y} = h_y$), consider a pair of random variables (X_k, X_j). Denote the conditional c.d.f. of $X_k|X_j$ within the slice h_y by $F_{X_k|X_j,h_y}(\cdot)$. Let $H^{\min}_{x_j} = \min_l\{H^l_{x_j}\}$ where $H^l_{x_j}$ is the number of slices considered in the partition $\mathbf{H}^l_{x_j}$. Then for any b_1, b_2 such that $\Pr\left(X_j \in [b_1, b_2)\right) \le 2/H^{\min}_{x_j}$, we have

$$0 < \tau_1 \le F_{X_k|X_j,h_y}\left(x_0|X_j \in [b_1, b_2)\right) \le \tau_2 < 1, \tag{C2*b}$$

for some constants $0 < \tau_1 \le \tau_2 < 1$ and denote $\min(\tau_1, 1-\tau_2)$ by τ_*; and

$$\left|\log\left(\frac{F_{X_k|X_j,h_y}(x_0|X_j = x_1)}{1 - F_{X_k|X_j,h_y}(x_0|X_j = x_1)}\right) - \log\left(\frac{F_{X_k|X_j,h_y}(x_0|X_j = x_2)}{1 - F_{X_k|X_j,h_y}(x_0|X_j = x_2)}\right)\right| \le \frac{\Delta^{h_y}_{\mathcal{D}}}{8}, \tag{C2*c}$$

for all x_0, j and $x_1, x_2 \in [b_1, b_2)$.

(C3*) For any interval $[k_1, k_2)$, we have

$$\inf_{y\in[k_1,k_2)} \left(R^{(o)}_{X_k|X_j}|Y=y\right) \leq \left(R^{(o)}_{X_k|X_j}|Y\in[k_1,k_2)\right) \leq \sup_{y\in[k_1,k_2)} \left(R^{(o)}_{X_k|X_j}|Y=y\right). \tag{C3*a}$$

Furthermore, for any $\epsilon > 0$, if $1/H_y - \epsilon \leq \Pr\{Y \in [k_1, k_2)\} \leq 1/H_y + \epsilon$, then for any $y_1, y_2 \in [k_1, k_2)$,

$$\left|\left(R^{(o)}_{X_k|X_j}|Y=y_1\right) - \left(R^{(o)}_{X_k|X_j}|Y=y_2\right)\right| \leq \epsilon/2. \tag{C3*b}$$

Assumption (C1*) ensures that the proposed screening statistic $V^{(o)}_{X_k|X_j;Y}$ has a clear separation between the features in $\mathcal{D}_2$ and features $\mathcal{D}_2^c$. Assumption (C2*) ensures the rank consistency of $\widehat{R}^{(o)}_{X_k|X_j,h_y}$ and $\widehat{R}^{*}_{X_k|X_j,h_y}$ in each slice by assuming the same conditions of Theorem 1 within each slice. Hence, it is generally true if conditions of Theorem 1 hold. Finally, condition (C3*) assumes that when we slice Y for conditioning using partition $\mathbf{H}_y$, for any given values y_1 and y_2 within a particular slice, the conditional dependencies $R^{(o)}_{X_k|X_j}|Y = y_1$ and $R^{(o)}_{X_k|X_j}|Y = y_2$ are close to each other (i.e. their difference is bounded). As a consequence, it ensures that the expected conditional association $E\left[R^{(o)}_{X_k|X_j}\middle|\tilde{Y}\right]$ based on the discretized $\tilde{Y}$ well approximates the expected conditional association $E\left[R^{(o)}_{X_k|X_j}|Y\right]$ based on the original continuous Y. This condition is commonly used in sufficient dimension reduction literature [13, 4] when the slicing is used.

4.4. Sure screening property for sufficient screening

Theorem 2. *Assume conditions (C1*)–(C3*). If $H^l_{x_j} \leq \lceil \log n \rceil$ for all l and $d'_n \geq |\mathcal{A}_2|$, then, for the screening sets $\widehat{\mathcal{A}}_2\left(\widehat{V}^{(o)}_{X_k|X_j;Y}\right)$ and $\widehat{\mathcal{A}}_2\left(\widehat{V}^{*}_{X_k|X_j;Y}\right)$ defined in (4.6) and (4.7), we have*

$$\Pr\left(\mathcal{A}_2 \subset \widehat{\mathcal{A}}_2\left(\widehat{V}^{(o)}_{X_k|X_j;Y}\right)\right) \geq 1-\varsigma \quad \textit{and} \quad \Pr\left(\mathcal{A}_2 \subset \widehat{\mathcal{A}}_2\left(\widehat{V}^{*}_{X_k|X_j;Y}\right)\right) \geq 1-\varsigma, \tag{4.8}$$

where

$$\varsigma = CNp^2(\log^2 n)\exp\left(-Cn\frac{\tau_*^2\Delta^2_{\mathcal{D}^*}}{\log n}\right) + CNp^2\exp\left(-C\frac{n}{\log^2 n}\right) + Cp^2\exp\left(-Cn\Delta^2_{\mathcal{D}^*}\right)$$

for some generic positive constant C.

The results of Theorem 2 indicate that the SVS step using the fused LogOR filter also enjoys the sure screening property if

$$\Delta_{\mathcal{D}^*} \gg \sqrt{\frac{\log n \log(p^2 N \log n)}{n}}.$$

5. Numerical studies

In this section, we evaluate the performance of the proposed fused LogOR filter under sufficient variable screening framework through simulations and a real data example.

5.1. Simulations

In [22], the fused Kolmogorov filter has been compared with several other existing screening methods in the literature, including marginal correlation screening [10], nonparametric independence screening [8], distance correlation screening [19], rank correlation screening [18], empirical likelihood screening [2], and the quantile adaptive screening [15]. The fused Kolmogorov filter demonstrated superior performance over these methods due to its unique characteristics such as being fully nonparametric, model-free and invariant to monotone transformation. Our proposed fused LogOR filter shares the same advantages as the fused Kolmogorov filter and is expected to be more sensitive to the tail-distribution, i.e. $F(Y|X_j)$ is closer to 0 or 1. Hence, we focus on comparing the fused LogOR filter to the fused Kolmogorov filter [22] in our simulations. Our empirical results indicate that the fused Kolmogorov filter and the fused LogOR filter is advantageous in different situations and one is not consistently better than the other. Hence, we consider the ensemble filter which combines the two filters together following the ensemble procedure described in Section 2 to take the advantages of the both filters. We denote these methods as "K", "LogOR", "Ens" respectively. Not only we demonstrate the effectiveness of the fused LogOR filter for marginal variable screening, we also demonstrate that the sufficient variable screening procedures described are useful to improve the marginal variable screening results. In addition, we will show that the ensemble of the fused Kolmogorov filter and the fused LogOR filter significantly improve the performance of variable screening for the cases where neither approach is unable to identify all active features.

For each of our simulated models, we repeat each experiment 200 times and report the following criteria to evaluate the variable screening results.

- $\mathcal{P}_i$: the proportion that an individual active predictor is selected out of the total number of replicates.
- $\mathcal{P}_a$: the proportion that all active predictors are selected out of the total number of replicates.

Note that the results are better when $\mathcal{P}_i$ and $\mathcal{P}_a$ are closer to 1. For each simulated model, we consider various settings of $n = 200, 400$ and $p = 500, 2000, 5000$.

TABLE 1
Simulation results for models (M1a)–(M1c)

Model	p	n	Method	$\mathcal{P}_i$ X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$\mathcal{P}_a$ **All**
(M1a)	500	200	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.980	0.965
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
	2000	200	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.965	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.950	0.910
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
	5000	200	K	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990
			LogOR	0.895	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.940	0.850
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.980
(M1b)	500	200	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.980	0.965
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
	2000	200	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.965	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.950	0.910
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
	5000	200	K	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990
			LogOR	0.895	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.940	0.850
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.980
(M1c)	500	200	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.980	0.965
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
	2000	200	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.965	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.950	0.910
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
	5000	200	K	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990
			LogOR	0.895	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.940	0.850
		400	K	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			LogOR	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.980

Example 1. Consider the semiparametric model $T_y(Y) = \mathbf{T}(\mathbf{X})^\intercal \boldsymbol{\beta} + \varepsilon$, where $\mathbf{T} = (T_1, \ldots, T_p)$ and $T_y, T_1, \ldots, T_p$ are strict monotone univariate transformations. Let $\boldsymbol{\beta} = 0.8 \times (\mathbf{1}_{10}, \mathbf{0}_{p-10})^\intercal$, $\mathbf{T}(\mathbf{X}) \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ follows an autoregressive structure with elements $\sigma_{ij} = 0.7^{|i-j|}$, and ε be a standard normal random variable. We consider the following setting of $T_y(Y)$ and $\mathbf{T}(\mathbf{X})$:

$$T_y(Y) = Y,\ T_j(X_j) = X_j; \tag{M1a}$$

$$T_y(Y) = Y,\ T_j(X_j) = \frac{1}{2}\log(X_j); \tag{M1b}$$

$$T_y(Y) = \log(Y),\ T_j(X_j) = X_j. \tag{M1c}$$

The models in Example 1 are considered by [22] and they are designed under strict monotone univariate transformations. From the results presented in Table 1, we observe that both the fused Kolmogorov filter and the fused LogOR filter perform well for all models and there is no obvious difference between the two. Therefore, the fused LogOR filter shares the same advantages of the fused Kolmogorov filter in the marginal screening.

Since the proposed fused logOR filter is more robust to detect the active features that are associated with the response variable at the tail of the conditional distribution, In the next example, we consider models with different conditional cumulative $F_{Y|X_j}(y|X_j)$.

Example 2. Let X_1 be a Bernoulli(0.5) random variable and $X_2, \ldots, X_p$ be i.i.d. standard normal random variables. Then $Y = \varepsilon_1$ if $X_1 = 1$ and $Y = \varepsilon_2$ if $X_1 = 0$. We consider the following settings for ε_1 and ε_2:

$$\varepsilon_1 \sim N(0,1),\ \text{and } \varepsilon_2 \sim t(1); \tag{M2a}$$

$$\varepsilon_1 \sim N(0,1),\ \text{and } \varepsilon_2 \sim 0.5N(0,1) + 0.5N(-1,3). \tag{M2b}$$

In this example, since X_1 is the only active predictor, we report only the $\mathcal{P}_i$ in Table 2 from which we observe that the proposed fused LogOR filter significantly outperforms the fused Kolmogorov filter.

TABLE 2
Simulation results for models (M2a) and (M2b)

Model	n	Method	$p = 500$	$p = 2000$	$p = 5000$
(M2a)	200	K	0.070	0.010	0.005
		LogOR	0.830	0.605	0.350
	400	K	0.120	0.020	0.005
		LogOR	1.000	0.985	0.980
(M2b)	200	K	0.190	0.085	0.035
		LogOR	0.820	0.645	0.475
	400	K	0.555	0.335	0.215
		LogOR	0.995	0.985	0.930

In the following example, we consider a slightly more complicated case than Example 2 by introducing additional active predictors. We will also show that in

TABLE 3
Simulation results for Example 3.

		$p = 500$			$p = 2000$			$p = 5000$		
n	Method	X_1	X_2	**All**	X_1	X_2	**All**	X_1	X_2	**All**
200	K	0.020	1.000	0.020	0.000	0.945	0.000	0.000	0.976	0.000
	LogOR	0.695	0.250	0.170	0.445	0.135	0.050	0.164	0.152	0.035
	Ens	0.600	1.000	0.600	0.365	0.920	0.325	0.117	0.929	0.105
	ESVS-I	0.590	1.000	0.590	0.345	0.915	0.300	0.117	0.917	0.094
	ESVS-II	0.590	1.000	0.590	0.345	0.915	0.300	0.117	0.917	0.094
400	K	0.070	1.000	0.070	0.005	1.000	0.005	0.012	1.000	0.012
	LogOR	1.000	0.210	0.210	0.960	0.080	0.070	0.951	0.084	0.072
	EM	1.000	1.000	1.000	0.945	1.000	0.945	0.879	1.000	0.879
	ESVS-I	0.995	1.000	0.995	0.945	1.000	0.945	0.843	1.000	0.843
	ESVS-II	0.995	1.000	0.995	0.945	1.000	0.945	0.843	1.000	0.843

this example, the ensemble of the two filters can improve the variable screening results.

Example 3. Let X_1 be a Bernoulli(0.5) random variable, and $X_2, \ldots, X_p$ are i.i.d. standard normal random variables. Then $Y = 0.5X_2 + \varepsilon_1$ if $X_1 = 1$ and $Y = 0.5X_2 + \varepsilon_2$ if $X_1 = 0$ where $\varepsilon_1 \sim N(0,1)$ and $\varepsilon_2 \sim t(1)$.

From Table 3, we can see that neither of the fused Kolmogorov filter nor the fused LogOR filter is able to identify both active predictors as marginal screening procedures. Hence, we consider the ensemble of the two as well as the ensemble SVS procedures. As a result, the ensemble filter brings the substantial improvement, especially when the sample size is large enough. In this case, the SVS procedures perform similarly to marginal screening procedure using the ensemble filter.

Example 4. Let X_1 be a Bernoulli(0.5) random variable, and $(X_2, \ldots, X_p)^\intercal$ follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$ for $i, j = 2, \ldots, p$ with $\sigma_{ii} = 1$; $\sigma_{i5} = \sigma_{5i} = \rho^\kappa$ for $i \neq 5$; and $\sigma_{ij} = \rho$ for $i \neq j, i \neq 5, j \neq 5$. Consider a general model:

$$Y = \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 T(X_5) + \begin{cases} \epsilon_1, & \text{if } X_1 = 1, \\ \epsilon_2, & \text{if } X_1 = 0; \end{cases}$$

where $\varepsilon_1 \sim N(0,1)$ and $\varepsilon_2 \sim t(1)$. We consider the following settings:

$$\kappa = 0.5,\ T(X_5) = X_5,\ \beta_2 = \beta_3 = \beta_4 = 1,\ \beta_5 = -3\rho^\kappa; \tag{M4a}$$

$$\kappa = 0.7,\ T(X_5) = \exp(X_5),\ \beta_2 = \beta_3 = \beta_4 = 0.5,\ \beta_5 = -1.5e^{0.5}\rho^\kappa; \tag{M4b}$$

$$\kappa = 0.7,\ T(X_5) = X_5^3,\ \beta_2 = \beta_3 = \beta_4 = 0.5,\ \beta_5 = -0.3\rho^\kappa. \tag{M4c}$$

In all models of Example 4, the predictors $X_2, \ldots, X_p$, except for X_5 are equally correlated with coefficient ρ, while X_5 has correlation ρ^κ with all other $p - 2$ predictors. In these models, X_1–X_4 are active predictors and X_5 is also an active variable that is marginally independent of the response. The results of Example 4 are gathered in Tables 4–6 which demonstrate that the ESVS-I and ESVS-II procedures significantly improve the marginal screening procedures.

TABLE 4
Simulation results for model (M4a).

			$\rho = 0.5$						$\rho = 0.8$					
p	n	Method	X_1	X_2	X_3	X_4	X_5	**All**	X_1	X_2	X_3	X_4	X_5	**All**
500	200	K	0.010	0.885	0.875	0.895	0.025	0.000	0.020	0.495	0.455	0.430	0.055	0.000
		LogOR	0.440	0.215	0.250	0.245	0.025	0.000	0.685	0.090	0.115	0.105	0.050	0.000
		Ens	0.330	0.860	0.830	0.825	0.020	0.005	0.590	0.410	0.330	0.395	0.050	0.000
		ESVS-I	0.325	0.840	0.830	0.815	1.000	0.225	0.590	0.385	0.330	0.385	0.975	0.045
		ESVS-II	0.325	0.840	0.830	0.815	0.995	0.225	0.590	0.390	0.335	0.385	0.990	0.050
	400	K	0.010	1.000	0.995	0.995	0.090	0.005	0.065	0.685	0.755	0.780	0.065	0.005
		LogOR	0.940	0.190	0.240	0.255	0.105	0.005	0.995	0.120	0.160	0.150	0.080	0.000
		Ens	0.915	0.985	0.985	0.995	0.065	0.065	0.980	0.640	0.700	0.700	0.030	0.025
		ESVS-I	0.910	0.985	0.985	0.995	1.000	0.875	0.980	0.630	0.690	0.705	1.000	0.470
		ESVS-II	0.910	0.990	0.985	0.995	1.000	0.880	0.980	0.640	0.695	0.700	1.000	0.470
2000	200	K	0.005	0.730	0.735	0.715	0.010	0.000	0.005	0.220	0.250	0.220	0.005	0.000
		LogOR	0.245	0.090	0.125	0.100	0.005	0.000	0.480	0.035	0.020	0.030	0.000	0.000
		Ens	0.125	0.670	0.660	0.650	0.010	0.000	0.390	0.155	0.210	0.190	0.000	0.000
		ESVS-I	0.120	0.655	0.660	0.650	0.995	0.025	0.370	0.145	0.200	0.185	0.925	0.000
		ESVS-II	0.120	0.655	0.660	0.650	0.995	0.025	0.370	0.150	0.200	0.185	0.905	0.000
	400	K	0.010	0.965	0.980	0.975	0.005	0.000	0.015	0.575	0.570	0.575	0.010	0.000
		LogOR	0.745	0.050	0.070	0.105	0.030	0.000	0.915	0.020	0.030	0.050	0.005	0.000
		Ens	0.630	0.960	0.980	0.965	0.005	0.000	0.865	0.505	0.490	0.515	0.005	0.000
		ESVS-I	0.615	0.955	0.980	0.960	1.000	0.575	0.860	0.495	0.480	0.510	1.000	0.225
		ESVS-II	0.615	0.955	0.980	0.960	1.000	0.575	0.860	0.500	0.480	0.510	1.000	0.225
5000	200	K	0.000	0.609	0.682	0.573	0.000	0.000	0.011	0.151	0.162	0.197	0.000	0.000
		LogOR	0.061	0.061	0.146	0.097	0.012	0.000	0.291	0.000	0.011	0.034	0.000	0.000
		Ens	0.036	0.597	0.561	0.524	0.012	0.000	0.174	0.069	0.127	0.093	0.000	0.000
		ESVS-I	0.036	0.597	0.548	0.524	1.000	0.012	0.162	0.069	0.127	0.093	0.791	0.000
		ESVS-II	0.036	0.597	0.548	0.524	1.000	0.012	0.162	0.069	0.127	0.093	0.791	0.000
	400	K	0.000	0.975	0.987	0.914	0.000	0.000	0.000	0.402	0.451	0.475	0.000	0.000
		LogOR	0.634	0.121	0.048	0.048	0.012	0.000	0.891	0.024	0.061	0.000	0.000	0.000
		Ens	0.536	0.939	0.951	0.902	0.012	0.000	0.853	0.292	0.353	0.426	0.000	0.000
		ESVS-I	0.524	0.939	0.951	0.902	1.000	0.439	0.853	0.292	0.353	0.414	1.000	0.085
		ESVS-II	0.524	0.939	0.951	0.902	1.000	0.439	0.853	0.292	0.353	0.414	1.000	0.085

TABLE 5
Simulation results for model (M4b).

			$\rho = 0.5$						$\rho = 0.8$					
p	n	Method	X_1	X_2	X_3	X_4	X_5	**All**	X_1	X_2	X_3	X_4	X_5	**All**
500	200	K	0.000	0.805	0.785	0.825	0.180	0.000	0.000	0.305	0.335	0.360	0.185	0.000
		LogOR	0.360	0.185	0.190	0.140	0.110	0.000	0.325	0.145	0.135	0.110	0.070	0.000
		Ens	0.275	0.720	0.710	0.705	0.175	0.005	0.225	0.280	0.295	0.245	0.140	0.000
		ESVS-I	0.270	0.720	0.695	0.695	0.490	0.025	0.225	0.260	0.275	0.240	0.415	0.000
		ESVS-II	0.270	0.720	0.700	0.695	0.570	0.025	0.225	0.255	0.280	0.235	0.410	0.000
	400	K	0.000	0.985	0.985	1.000	0.360	0.000	0.000	0.600	0.655	0.665	0.385	0.000
		LogOR	0.905	0.145	0.195	0.185	0.150	0.000	0.850	0.105	0.125	0.145	0.145	0.000
		Ens	0.850	0.960	0.965	0.990	0.275	0.220	0.760	0.460	0.545	0.550	0.270	0.010
		ESVS-I	0.850	0.960	0.965	0.990	0.990	0.770	0.760	0.440	0.555	0.530	0.925	0.090
		ESVS-II	0.850	0.960	0.965	0.990	0.990	0.770	0.760	0.445	0.550	0.530	0.950	0.095
2000	200	K	0.000	0.525	0.545	0.555	0.060	0.000	0.000	0.135	0.140	0.135	0.065	0.000
		LogOR	0.185	0.060	0.055	0.060	0.020	0.000	0.130	0.025	0.025	0.030	0.015	0.000
		Ens	0.135	0.420	0.430	0.475	0.050	0.000	0.100	0.115	0.090	0.095	0.030	0.000
		ESVS-I	0.135	0.405	0.430	0.475	0.215	0.005	0.095	0.115	0.085	0.090	0.140	0.000
		ESVS-II	0.135	0.405	0.430	0.475	0.275	0.000	0.095	0.115	0.085	0.090	0.130	0.000
	400	K	0.000	0.925	0.890	0.920	0.115	0.000	0.000	0.380	0.365	0.335	0.125	0.000
		LogOR	0.705	0.045	0.045	0.085	0.095	0.000	0.550	0.025	0.035	0.055	0.020	0.000
		Ens	0.595	0.870	0.840	0.870	0.095	0.020	0.475	0.245	0.250	0.225	0.085	0.000
		ESVS-I	0.585	0.870	0.835	0.865	0.895	0.330	0.470	0.225	0.245	0.220	0.685	0.000
		ESVS-II	0.585	0.870	0.835	0.865	0.925	0.340	0.470	0.230	0.245	0.225	0.730	0.000
5000	200	K	0.000	0.421	0.373	0.361	0.000	0.000	0.000	0.071	0.112	0.112	0.014	0.000
		LogOR	0.048	0.000	0.048	0.024	0.012	0.000	0.028	0.000	0.028	0.028	0.014	0.000
		Ens	0.024	0.313	0.289	0.277	0.000	0.000	0.014	0.042	0.084	0.098	0.000	0.000
		ESVS-I	0.024	0.313	0.289	0.277	0.096	0.000	0.014	0.042	0.084	0.098	0.098	0.000
		ESVS-II	0.024	0.313	0.289	0.277	0.192	0.000	0.014	0.042	0.084	0.098	0.071	0.000
	400	K	0.000	0.879	0.783	0.771	0.072	0.000	0.000	0.214	0.202	0.178	0.035	0.000
		LogOR	0.578	0.024	0.000	0.024	0.024	0.000	0.523	0.011	0.011	0.035	0.000	0.000
		Ens	0.518	0.783	0.711	0.674	0.072	0.000	0.441	0.154	0.131	0.107	0.011	0.000
		ESVS-I	0.518	0.771	0.711	0.674	0.746	0.144	0.428	0.142	0.119	0.107	0.476	0.000
		ESVS-II	0.518	0.771	0.711	0.674	0.843	0.132	0.428	0.142	0.119	0.107	0.452	0.000

TABLE 6
Simulation results for model (M4c).

			$\rho = 0.5$						$\rho = 0.8$					
p	n	Method	X_1	X_2	X_3	X_4	X_5	**All**	X_1	X_2	X_3	X_4	X_5	**All**
500	200	K	0.000	0.865	0.830	0.855	0.145	0.000	0.000	0.545	0.505	0.510	0.125	0.000
		LogOR	0.350	0.195	0.210	0.195	0.130	0.000	0.265	0.145	0.145	0.130	0.070	0.000
		Ens	0.225	0.765	0.750	0.770	0.120	0.015	0.200	0.460	0.400	0.400	0.115	0.000
		ESVS-I	0.205	0.760	0.735	0.765	0.520	0.045	0.195	0.455	0.395	0.395	0.390	0.000
		ESVS-II	0.205	0.760	0.735	0.765	0.465	0.035	0.195	0.455	0.390	0.390	0.360	0.000
	400	K	0.000	0.990	0.990	1.000	0.350	0.000	0.000	0.790	0.870	0.805	0.345	0.000
		LogOR	0.920	0.185	0.225	0.210	0.150	0.005	0.865	0.185	0.195	0.150	0.135	0.000
		Ens	0.865	0.980	0.975	0.990	0.220	0.185	0.800	0.645	0.770	0.770	0.200	0.070
		ESVS-I	0.850	0.975	0.975	0.985	0.995	0.805	0.800	0.655	0.765	0.765	0.920	0.310
		ESVS-II	0.850	0.975	0.975	0.985	0.975	0.780	0.800	0.645	0.760	0.760	0.910	0.285
2000	200	K	0.000	0.655	0.595	0.680	0.025	0.000	0.000	0.265	0.260	0.295	0.015	0.000
		LogOR	0.140	0.100	0.090	0.095	0.015	0.000	0.160	0.070	0.045	0.035	0.015	0.000
		Ens	0.080	0.520	0.475	0.590	0.025	0.000	0.100	0.200	0.220	0.190	0.020	0.000
		ESVS-I	0.075	0.520	0.460	0.550	0.225	0.000	0.100	0.205	0.220	0.180	0.175	0.000
		ESVS-II	0.075	0.520	0.460	0.550	0.215	0.000	0.100	0.195	0.220	0.180	0.090	0.000
	400	K	0.000	0.960	0.955	0.960	0.090	0.000	0.000	0.560	0.580	0.565	0.090	0.000
		LogOR	0.705	0.070	0.060	0.080	0.055	0.000	0.725	0.035	0.005	0.050	0.050	0.000
		Ens	0.600	0.940	0.910	0.930	0.040	0.010	0.580	0.470	0.440	0.430	0.060	0.000
		ESVS-I	0.590	0.930	0.910	0.930	0.865	0.410	0.575	0.470	0.435	0.420	0.680	0.015
		ESVS-II	0.590	0.930	0.910	0.930	0.855	0.425	0.575	0.470	0.430	0.420	0.750	0.030
5000	200	K	0.000	0.540	0.482	0.471	0.011	0.000	0.000	0.232	0.151	0.139	0.000	0.000
		LogOR	0.034	0.057	0.034	0.022	0.000	0.000	0.046	0.046	0.023	0.011	0.000	0.000
		Ens	0.011	0.459	0.391	0.367	0.000	0.000	0.023	0.162	0.081	0.093	0.000	0.000
		ESVS-I	0.011	0.459	0.367	0.344	0.080	0.000	0.023	0.162	0.081	0.093	0.093	0.000
		ESVS-II	0.011	0.459	0.367	0.344	0.126	0.000	0.023	0.162	0.081	0.093	0.034	0.000
	400	K	0.000	0.881	0.892	0.940	0.047	0.000	0.646	0.012	0.024	0.012	0.012	0.000
		LogOR	0.535	0.071	0.059	0.083	0.011	0.000	0.000	0.414	0.378	0.317	0.024	0.000
		Ens	0.416	0.869	0.809	0.904	0.011	0.011	0.573	0.353	0.268	0.256	0.000	0.000
		ESVS-I	0.416	0.857	0.809	0.892	0.809	0.178	0.573	0.353	0.268	0.231	0.573	0.000
		ESVS-II	0.416	0.857	0.809	0.892	0.678	0.166	0.573	0.353	0.268	0.231	0.487	0.000

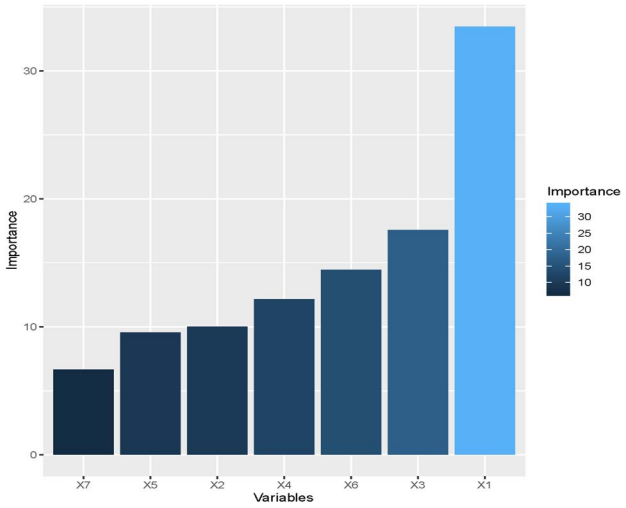

FIG 1. *Variable importance of the NO2 dataset*

5.2. Real data example

In this section, we use the NO2 dataset to illustrate the methods. The NO2 dataset was a subsample of $n = 500$ observations from a data set collected by the Norwegian Public Roads Administration that is originated in a study where air pollution at a road is related to traffic volume and meteorological variables. This dataset is available at http://lib.stat.cmu.edu/datasets/NO2.dat. The response variable consist of hourly values of the logarithm of the concentration of NO2 (particles), measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The seven predictor variables are the logarithm of the number of cars per hour, temperature 2 meter above ground (°C), wind speed (meters/second), the temperature difference between 25 and 2 meters above ground (°C), wind direction (degrees between 0 and 360), hour of day and day number from October 1. 2001. Figure 1 seems indicating that the seventh predictor is the least important.

Following the approach in [22], in addition to the 7 predictors in the original dataset, we added 493, 1993, 4993 independent noise variables following the standard norm distribution, such that $p = 500$, $p = 2000$, $p = 5000$, respectively. We apply the fused Kolmogorov filter, the fused LogOR filter, and ensemble of the two for marginal variable screening, and the ESVS-I, and ESVS-II for sufficient variable screening. We random split the dataset as the training set ($n_1 = 400$) and the validation set ($n_2 = 100$) with 100 replications. In each replication, we examine whether the screening methods can distinguish the useful predictors from the noise variables. For the fused filters, we consider the combination of $G_i = 3, \ldots, 6$, as in the simulation studies. For each screening method we select the top 7 predictors as the active predictors, treating other 493, 1993, or 4993

as noise predictors. We report the variable screening results in Table 7. From the results, it seems that all marginal screening methods have difficulty identifying X_7 as an active predictor but the proposed SVS procedures significantly improve the chance of selecting it. We also observe that the fused LogOR filter is more competent to select X_2 and X_5 than the fused Kolmogorov filter.

TABLE 7
Variable screening results of the NO2 dataset

p	Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	**All**
500	K	1.000	0.100	1.000	1.000	0.240	1.000	0.100	0.010
	LogOR	1.000	0.860	1.000	0.810	0.710	0.990	0.010	0.010
	Ens	1.000	0.860	1.000	0.810	0.710	1.000	0.010	0.010
	ESVS-I	1.000	0.950	1.000	1.000	0.670	1.000	0.500	0.300
	ESVS-II	1.000	0.950	1.000	1.000	0.700	1.000	0.470	0.300
2000	K	1.000	0.030	1.000	0.980	0.070	1.000	0.020	0.000
	LogOR	1.000	0.730	1.000	0.730	0.600	0.920	0.000	0.000
	Ens	1.000	0.730	1.000	0.730	0.580	1.000	0.000	0.000
	ESVS-I	1.000	0.880	1.000	0.960	0.540	1.000	0.350	0.160
	ESVS-II	1.000	0.890	1.000	1.000	0.570	1.000	0.270	0.160
5000	K	1.000	0.000	1.000	0.940	0.020	1.000	0.010	0.000
	LogOR	1.000	0.650	1.000	0.560	0.440	0.880	0.000	0.000
	Ens	1.000	0.640	1.000	0.560	0.420	1.000	0.000	0.000
	ESVS-I	1.000	0.790	1.000	0.970	0.390	1.000	0.240	0.090
	ESVS-II	1.000	0.800	1.000	1.000	0.420	1.000	0.170	0.090

We further examine how variable screening step helps predicting the response variable by fitting a generalized additive model (GAM) and a random forest (RF) model using the selected top 7 predictors. We report the mean and standard deviation of prediction mean squared error (PMSE) on the validation data based on 100 replications.

TABLE 8
Average PMSE (and its standard deviation) using the selected top 7 predictors over 100 replications

Method		$p = 500$ GAM	$p = 500$ RF	$p = 2000$ GAM	$p = 2000$ RF	$p = 5000$ GAM	$p = 5000$ RF
K	average	0.276	0.050	0.287	0.051	0.290	0.052
	sd	(0.043)	(0.020)	(0.041)	(0.020)	(0.043)	(0.021)
LogOR	average	0.265	0.047	0.282	0.051	0.294	0.054
	sd	(0.043)	(0.020)	(0.052)	(0.022)	(0.054)	(0.024)
Ens	average	0.266	0.047	0.282	0.051	0.293	0.054
	sd	(0.043)	(0.020)	(0.052)	(0.022)	(0.054)	(0.023)
ESVS-I	average	0.257	0.045	0.265	0.047	0.269	0.049
	sd	(0.040)	(0.019)	(0.043)	(0.020)	(0.042)	(0.021)
ESVS-II	average	0.257	0.045	0.263	0.047	0.267	0.050
	sd	(0.040)	(0.019)	(0.042)	(0.020)	(0.042)	(0.021)

As a reference, which can be treated as an oracle approach, we compute the average and standard deviation of the PMSE using original 7 variables over the validation data with 100 observations randomly selected from the 500 data points. It turned out to be that the average PMSE using original 7 variables is 0.2467 for GAM and 0.0418 for RF with 0.035 and 0.017 as standard deviations

respectively. From Table 8, we can observe that as a marginal screening method, the fused LogOR filter and the ensemble fitler outperform the fused Kolmogorov filter by itself. Both SVS procedures further improve the prediction accuracy.

6. Discussions

In this paper, we propose a general sufficient variable screening framework that works for any dependence measure that is defined to measure the statistical association between two univariate random variables. Two separate sufficient variable screening procedures are proposed to overcome the limitations of the marginal screening methods when the active variables are marginally independent of the response variable. In addition, an ensemble approach is proposed to combine advantages of different dependence measures to further boost the screening performance. In addition, a new dependence measure, the log odds ratio statistic, is proposed for variable screening which enjoys the sure screening properties for both marginal and sufficient variable screening. The fused logOR filter overcomes the challenge for the fused Kolmogorov filter when the conditional c.d.f is close to 0 or 1. It has been demonstrated empirically that the ensemble of the fused logOR filter and the fused Kolmogorov filter delivers superior screening results in most cases. While under the current ensemble framework, all candidate screening methods are treated equally in the ensemble step, obtaining an optimal weighting over all candidate screening methods is an interesting direction for future research.

Acknowledgments

The authors thank the editor, associate editor, and referees for their thoughtful and insightful comments and suggestions.

Supplementary Material

Supplementary Material of "On sufficient variable screening using log odds ratio filter"
(doi: 10.1214/22-EJS1951SUPP; .pdf). In this supplementary file, we provide a discussion on how Condition (C1) of Theorem 1 can be satisfied in general and technical proofs for the propositions and theorems. We also provide a detailed algorithm on how we combine the proposed log odds ratio filter and the Kolmogorov filter to achieve supreme screening results.

References

[1] Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. (2013). Ultrahigh dimensional feature screening via rkhs embeddings. *Journal of Machine Learning Research*, 31:126–134. 16th International Conference on Artificial Intelligence and Statistics, AISTATS 2013; Conference date: 29-04-2013 Through 01-05-2013.

[2] Chang, J., Tang, C. Y., and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.*, 41(4):2123–2148.

[3] Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.

[4] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100(470):410–428.

[5] Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association*, 109(506):815–827.

[6] 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.

[7] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, 36(6):2605–2637.

[8] 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. PMID: 22279246.

[9] 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.

[10] 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.

[11] Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284. PMID: 25309009.

[12] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.*, 38(6):3567–3604.

[13] Gannoun, A. and Saracco, J. (2003). An asymptotic theory for sir_α method. *Statistica Sinica*, 13(2):297–310.

[14] Gretton A., Bousquet O., S. A. S. B. (2005). Measuring statistical dependence with hilbert-schmidt norms. *Algorithmic Learning Theory. ALT 2005. Lecture Notes in Computer Science*, 3734(6):2769–2794.

[15] He, X., Wang, L., and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.*, 41(1):342–369.

[16] Kim, A. K. H. and Shin, S. J. (2017). The cumulative kolmogorov filter for model-free screening in ultrahigh dimensional data. *Statistics & Probability Letters*, 126:238 – 243.

[17] Kong, J., Wang, S., and Wahba, G. (2015). Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine*, 34(10):1708–1720.

[18] Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *Ann. Statist.*, 40(3):1846–1877.

[19] Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via dis-

tance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139. PMID: 25249709.

[20] Liu, J., Li, R., and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109(505):266–274. PMID: 24678135.

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

[22] Mai, Q. and Zou, H. (2015). The fused kolmogorov filter: A nonparametric model-free screening method. *Ann. Statist.*, 43(4):1471–1497.

[23] Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434.

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

[25] Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794.

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

[27] Yang, B., Yin, X., and Zhang, N. (2019). Sufficient variable selection using independence measures for continuous response. *Journal of Multivariate Analysis*, 173:480 – 493.

[28] Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 77:879–892.

[29] Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8):1733 – 1757.

[30] 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.

[31] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.