Electronic Journal of Statistics
Vol. 17 (2023) 2139–2179
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2150

Sufficient variable screening with high-dimensional controls

Chenlu Ke

Department of Statistical Sciences and Operations Research, Virginia Commonwealth University
e-mail: kec2@vcu.edu

Abstract: Variable screening for ultrahigh-dimensional data has attracted extensive attention in the past decade. In many applications, researchers learn from previous studies about certain important predictors or control variables related to the response of interest. Such knowledge should be taken into account in the screening procedure. The development of variable screening conditional on prior information, however, has been less fruitful, compared to the vast literature for generic unconditional screening. In this paper, we propose a model-free variable screening paradigm that allows for high-dimensional controls and applies to either continuous or categorical responses. The contribution of each individual predictor is quantified marginally and conditionally in the presence of the control variables as well as the other candidates by reproducing-kernel-based R^2 and partial R^2 statistics. As a result, the proposed method enjoys the sure screening property and the rank consistency property in the notion of sufficiency, with which its superiority over existing methods is well-established. The advantages of the proposed method are demonstrated by simulation studies encompassing a variety of regression and classification models, and an application to high-throughput gene expression data.

MSC2020 subject classifications: Primary 62B10, 62B86; secondary 62G05, 62G08, 62H30.
Keywords and phrases: Conditional independence, rank consistency, reproducing kernel Hilbert space, sure screening.

Received August 2022.

1. Introduction

Ultrahigh-dimensional data have become increasingly prevalent nowadays in diverse fields such as biomedical sciences, finance and social sciences. Since a vital task of contemporary statistical analysis is to extract core information by identifying low-dimensional sparse presentations of the predictive variables, variable selection becomes an indispensable part of the analysis pipeline. However, impeded by the complications embedded in ultrahigh-dimensional data, it is often beyond the hope in practice to recover all the truly important predictors with no error. Traditional variable selection and regularization methods are no longer applicable due to statistical and computational issues associated with increasing data volume. Recent years have seen rising attention to variable screening as a less ambitious yet efficient way to remove most irrelevant information before more sophisticated modeling can be pursued. Variable screening was first

introduced by Fan and Lv (2008) for linear models to fast filter out redundant variables through marginal independence learning based on the Pearson correlation. The screening mechanism asymptotically almost surely identifies all important predictors, and thus is called sure independence screening (SIS). Since conjecturing about the dependence structure is presumably challenging in high-dimensional spaces, more flexible approaches have emerged to avoid model specifications. For instance, Li, Zhong and Zhu (2012) and Balasubramanian, Sriperumbudur and Lebanon (2013) improved SIS using distance correlation (DCOR; Székely, Rizzo and Bakirov, 2007) and a more general class of dependence measures, namely Hilbert-Schmidt independence criterion (HSIC; Gretton et al., 2005), respectively. Model-free SIS procedures focusing on categorical responses include Kolmogorov filter (Mai and Zou, 2013, 2015) and MV-SIS (Cui, Li and Zhong, 2015), among others.

Conditional variable screening is a pertinent addition to the toolbox of analyzing ultrahigh-dimensional data when prior information is available or when potential confoundings exist. In many applications, researchers know from previous investigations that certain variables are responsible for the outcomes or should be controlled for in related studies. This knowledge should be taken into account so that these variables can assist in the selection of the other important predictors while being shielded from screening. Variable screening then relies on the learning of conditional dependence between potential predictors and the response variable given a control set already contained in the model. Compared to the rich literature in unconditional variable screening, the development in conditional variable screening has been less yielding. Measuring conditional dependence is in general a hard problem. It was recently revealed that a valid test for conditional independence does not have power against any alternative for continuous random vectors, unless the test is carefully chosen by some domain knowledge (Shah and Peters, 2020). Extending unconditional screening methods to complement conditional screening is therefore nontrivial and most of their adaptions remain elusive. Limited work has expanded to generalized linear model (Barut, Fan and Verhasselt, 2016) and varying coefficient linear model (Fan, Ma and Dai, 2014; Liu, Li and Wu, 2014; Yang, Yang and Li, 2020). Model-free approaches have also been developed based on conditional DCOR (CDCOR; Wang et al., 2015; Wen et al., 2018) and Blum–Kiefer–Rosenblatt correlation (Zhou, Liu and Zhu, 2020), which are mainly designed for continuous responses. While it is common to have multiple or even high-dimensional control variables in the analysis of ultrahigh-dimensional data, the conditional set is often restricted to very low dimension in existing screening methods because common nonparametric approaches for estimating moments of conditional distributions, such as kernel smoothing and k-nearest neighbors, suffer from the curse of dimensionality.

Interrelations, redundancy and noise also add to the difficulty of variable screening. A common but not negligible issue of the aforementioned unconditional or conditional screening methods is that important predictors making little or no marginal contribution (or conditional contribution given the control variables) to the response variable cannot be detected, while spurious variables that are highly correlated with some important predictors may be falsely se-

lected (Fan and Lv, 2008). Intuitively, this issue can be solved if each predictor is evaluated after adjusting for the joint effect of the other candidates and the control variables, which unfortunately also becomes an arduous conditional independence problem as the conditional vector is ultrahigh-dimensional. For unconditional screening, Yang, Yin and Zhang (2019) proposed some approachable sufficient conditions for identifying irrelevant predictors by incorporating the joint information of all variables. However, a marginally silent predictor may not survive those conditions if it does not have a stronger correlation with the rest variables compared to the other candidates. False discoveries of spurious variables are exacerbated in the meanwhile. Other available solutions are mostly iterative algorithms (Fan and Lv, 2008; Balasubramanian, Sriperumbudur and Lebanon, 2013; Liu, Li and Wu, 2014, etc.) which at each step select variables that explain the "residuals" from the previous iteration. Nevertheless, defining "residuals" without model specification is not obvious and no theoretical support has been provided. A gap still exists in addressing the predicament, especially for model-free conditional screening.

In this paper, we develop a model-free sufficient variable screening paradigm that allows prior information to be integrated. The impetus is the general lack of flexible and reliable conditional screening tools for ultrahigh-dimensional data, creating an impediment to promising applications in practice. Two jointly sufficient conditions for identifying null variables are introduced to assess the contribution of each individual predictor marginally and conditionally in the presence of the others as well as the control set. The assessments are carried out via reproducing-kernel-based R^2 and partial R^2 measures, which inherit the interpretability of the classical R^2 statistics but assume no underlying model structure. Deviation bounds for the empirical measures can be found uniformly in spite of the dimension of the control variables. Consequently, the proposed screening procedure satisfies the sure screening property and the rank consistency property in a "large p small n" setting. In a nutshell, our proposal has the following **merits** compared to existing methods. Firstly, the framework is developed with the notion of sufficiency – meaning that there is no loss of information in reducing the dimension. All the truly important predictors will be selected with high probability, including those that are individually independent but jointly dependent of the response variable when conditioning on the control variables. Secondly, arbitrarily many control variables can be adjusted for in the screening procedure that is implemented with the proposed model-free utility measures. Thirdly, the unified framework applies to either continuous or categorical response variables with finite or diverge numbers of categories. Lastly, the method is very general and can work with a broad class of independence measures. When there is no available prior information, the procedure automatically performs unconditional sufficient variable screening, which also surpasses existing unconditional screening approaches in detecting important predictors that are marginally independent of the response.

The rest of this paper is organized as follows. In Section 2, we develop the framework of conditional sufficient variable screening along with the kernel-based utility measures, and study related properties. Numerical studies compar-

ing our proposal with existing alternatives are provided in Section 3, followed by a short discussion in Section 4 to close the paper. All technical proofs are deferred to the appendix, along with a discussion addressing the computational aspects of the proposed procedure.

2. Conditional sufficient variable screening

2.1. Methodology

Let $Y \in \mathbb{R}$ be the response variable, $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be the vector of predictors that are subject to screening, and $\mathbf{W} \in \mathbb{R}^{p_0}$ be the vector of control variables such as known important predictors, confounders or exposure variables. Ideally, we would like to identify the smallest index set $\mathcal{A} \subseteq \{1, \ldots, p\}$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | \{\mathbf{X}_{\mathcal{A}}, \mathbf{W}\},$$

where $\mathbf{X}_{\mathcal{A}} := \{X_j : j \in \mathcal{A}\}$. Conditioning on $\mathbf{W}$, the selection of $\mathbf{X}_{\mathcal{A}}$ is "minimal sufficient" since no information about the regression is impaired in reducing the predictors to the utmost. The primary goal of conditional variable screening is to achieve sufficiency; that is, to find a reduced index set that covers $\mathcal{A}$ with relatively small cardinality (realistically assumed to be smaller than the sample size). This aim is less ambitious than exact variable selection that recovers $\mathcal{A}$ precisely, but in exchange for faster dimension reduction of massive data. After the majority of irrelevant variables are eliminated, more accurate variable selection can be further conducted. The following proposition lays the cornerstone for conditional sufficient variable screening.

Proposition 1. *Let $\mathbf{X}_{-j}$ denote the vector of all predictors excluding X_j ($j = 1, \ldots, p$), then*

1. *$j \notin \mathcal{A}$ if and only if condition*

$$(a)\ X_j \perp\!\!\!\perp Y | (\mathbf{X}_{-j}, \mathbf{W})$$

   holds, $j = 1, \ldots, p$; and
2. *the following pair of conditions (b1) and (b2) implies condition (a):*

$$(b1)\ X_j \perp\!\!\!\perp Y;\ (b2)\ (\mathbf{X}_{-j}, \mathbf{W}) \perp\!\!\!\perp Y | X_j.$$

The first statement in Proposition 1 streamlines a one-by-one screening procedure where each individual predictor should be assessed in the presence of the control variables as well as the other predictors. However, the conditional independence in (a) is rather difficult to verify, especially given the ultrahigh dimension of the conditioning vector. To circumvent the hurdle, one can check the conditions in the second statement instead, where $(b1)$ is simply a marginal independence condition and $(b2)$ only involves a single conditional variable. Since the pair of conditions $(b1)$ and $(b2)$ is stronger than condition (a), it is ensured that the excluded predictors are truly unimportant and subsequently, the remaining set will cover $\mathbf{X}_{\mathcal{A}}$.

Nonetheless, the biggest challenge in ultrahigh dimensional screening is that variables tend to be correlated by chances (Fan and Lv, 2010), so an unimportant predictor can be spuriously associated with the response due to its correlation with the important ones. Therefore, instead of eliminating unimportant predictors through independence tests, variable screening procedures select predictors that contribute most to the response, or equivalently, those "violate" condition (a) the most. Since violation of condition (a) implies violation of condition $(b1)$ or condition $(b2)$, X_j is retained if its marginal contribution to Y is large or if the conditional contribution of $(\mathbf{X}_{-j}, \mathbf{W})$ to Y given X_j is large. Examining only condition $(b1)$ is not sufficient as there might exist important predictors that are marginally independent but collectively dependent of the response. This occurs when an important predictor X_{j^*} is correlated with other significant ones in the model, but its effect on the response is offset by the joint effect of the others, so any marginal change in that predictor does not lead to a difference in the response. Since the joint effect of $(\mathbf{X}, \mathbf{W})$ can be decomposed into the marginal effect of an arbitrary predictor X_j and the conditional effect of $(\mathbf{X}_{-j}, \mathbf{W})$ given X_j already contained in the model, the conditional effect maxes out if X_j has absolutely no marginal contribution. As a result, the conditional effect of $(\mathbf{X}_{-j^*}, \mathbf{W})$ given X_{j^*} is in fact tantamount to the full effect of $(\mathbf{X}_{\mathcal{A}}, \mathbf{W})$. For a spurious predictor X_j, the conditional effect of $(\mathbf{X}_{-j}, \mathbf{W})$ given X_j is undermined due to the indirect marginal effect of X_j. For an unimportant predictor $X_{j'}$ that is completely independent of the response, the conditional effect of $(\mathbf{X}_{-j'}, \mathbf{W})$ given $X_{j'}$ is also maximized, but in this case $X_{j'}$ must be independent of $(\mathbf{X}_{\mathcal{A}}, \mathbf{W})$ and hence can be easily distinguished from marginally irrelevant but jointly important predictors. Assessing condition $(b2)$ is thus indispensable for further capturing important predictors whose marginal contribution to the response is very weak or nonexistent.

The mainstream of existing conditional screening methods evaluates an insufficient condition

$$(a')\ X_j \perp\!\!\!\perp Y|\mathbf{W}$$

without considering the effect of $\mathbf{X}_{-j}$. As a consequence, active predictors that are individually independent but collectively dependent of the response variable given the control set are falsely ruled out. In other words, only part of the active index set, namely $\mathcal{A}_1 := \{j \in \mathcal{A} : X_j \not\perp\!\!\!\perp Y|\mathbf{W}\} \subseteq \mathcal{A}$, can be identified. Besides, $\mathbf{W}$ as the conditional vector in (a') is often restricted to be low-dimensional. Note that by leaving out $\mathbf{X}_{-j}$ in Proposition 1, conditions $(b1)$ and

$$(b2')\ \mathbf{W} \perp\!\!\!\perp Y|X_j$$

jointly imply condition (a'), which suggests that one can utilize conditions $(b1)$ and $(b2')$ to recover $\mathcal{A}_1$ given high-dimensional controls. That is, X_j is retained if the marginal contribution of X_j is large or if the conditional contribution of $\mathbf{W}$ given X_j is large, either of which can lead to a strong joint effect of $(X_j, \mathbf{W})$. In the meanwhile, the collective effect of $(X_j, \mathbf{W})$ can also be decomposed into the constant effect of $\mathbf{W}$ and the conditional effect of X_j given $\mathbf{W}$, so a strong joint effect implies a strong conditional effect of X_j given $\mathbf{W}$.

In the following subsections, we propose utility measures to evaluate conditions ($b1$) and ($b2$), based on which a conditional sufficient variable screening procedure that allows $\mathbf{W}$ to be high-dimensional is then developed.

2.2. Kernel-based ANOVA statistics

Analysis of variance (ANOVA) plays an important role in statistical inference for linear models, and related statistics can be used to measure the marginal or the conditional contribution of individual predictors. The famous SIS (Fan and Lv, 2008) ranks predictors by the R^2 statistic of univariate linear regression, and its extension, conditional SIS (CSIS; Barut, Fan and Verhasselt, 2016), is based on a measure closely related to partial R^2 to adjust for the effect of the known important predictors. Elicited by these work, we introduce a model-free ANOVA decomposition through the theory of reproducing kernel Hilbert space (RKHS), which leads to generalized R^2 and partial R^2 statistics that can be employed to assess conditions ($b1$) and ($b2$) in Proposition 1.

Let us start with a brief review of RKHS. Let $\mathbb{H}_K$ denote a RKHS of real-valued functions defined on $\mathcal{X}$ with respect to the reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. That is, for any $\mathbf{x} \in \mathcal{X}$ and $f \in \mathbb{H}_K$, $K(\cdot, \mathbf{x}) \in \mathbb{H}_K$ and $\langle f, K(\cdot, \mathbf{x})\rangle_{\mathbb{H}_K} = f(\mathbf{x})$. A reproducing kernel is positive definite and every positive definite kernel has an associated RKHS according to Moore-Aronszajn theorem. The map $\phi_K(\mathbf{x}) = K(\cdot, \mathbf{x}) : \mathcal{X} \to \mathbb{H}_K$ is called the canonical feature map of K. Let $\mathcal{M}(\mathcal{X})$ be the set of all Borel probability measures on $\mathcal{X}$. For any $\mathbf{P} \in \mathcal{M}(\mathcal{X})$, the kernel embedding of $\mathbf{P}$ into $\mathbb{H}_K$ is defined by the Borel integral $\mu_K(\mathbf{P}) = \int \phi_K(\mathbf{x}) d\mathbf{P}(\mathbf{x}) : \mathcal{M}(\mathcal{X}) \to \mathbb{H}_K$. For any $f \in \mathbb{H}_K$, we have $\langle f, \mu_K(\mathbf{P})\rangle_{\mathbb{H}_K} = \int f(\mathbf{x}) d\mathbf{P}(\mathbf{x})$. The kernel embedding is well defined if $\int K^{\frac{1}{2}}(\mathbf{x}, \mathbf{x}) d\mathbf{P}(\mathbf{x}) < \infty$ by Riesz representation theorem.

Let $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{Z}$ be three random vectors on $\mathbb{R}^{p_1}$, $\mathbb{R}^{p_2}$ and $\mathbb{R}^{p_3}$, respectively. Denote the probability distribution of $\mathbf{U}$, $\mathbf{U}|\mathbf{V}$ and $\mathbf{U}|(\mathbf{V}, \mathbf{Z})$ by $\mathbf{P}_{\mathbf{U}}$, $\mathbf{P}_{\mathbf{U}|\mathbf{V}}$ and $\mathbf{P}_{\mathbf{U}|(\mathbf{V},\mathbf{Z})}$. For a selected positive definite kernel $K : \mathbb{R}^{p_1} \times \mathbb{R}^{p_1} \to \mathbb{R}$ and the associated RKHS $\mathbb{H}_K$, the total variation in $\phi_K(\mathbf{U})$ at the population level can be measured by the total sum of squares (SSTO) in RKHS norm as:

$$SSTO_K(\mathbf{U}) := E_{\mathbf{U}} \left\| \phi_K(\mathbf{U}) - \mu_K(\mathbf{P}_{\mathbf{U}}) \right\|^2_{\mathbb{H}_K}.$$

It can be shown through RKHS operations that

$$SSTO_K(\mathbf{U}) = E_{\mathbf{U}} K(\mathbf{U}, \mathbf{U}) - E_{\mathbf{U},\mathbf{U}'} K(\mathbf{U}, \mathbf{U}'),$$

where $\mathbf{U}'$ is an independent and identically distributed copy of $\mathbf{U}$. We have the following kernel decomposition of the total variation in $\phi_K(\mathbf{U})$ into the regression sum of squares (SSR) and the error sum of squares (SSE) for the regression of $\mathbf{U}$ on $\mathbf{V}$:

$$SSTO_K(\mathbf{U}) = SSR_K(\mathbf{U}|\mathbf{V}) + SSE_K(\mathbf{U}|\mathbf{V}),$$

where

$$SSR_K(\mathbf{U}|\mathbf{V}) := E_{\mathbf{V}} \left\| \mu_K(\mathbf{P}_{\mathbf{U}|\mathbf{V}}) - \mu_K(\mathbf{P}_{\mathbf{U}}) \right\|^2_{\mathbb{H}_K},$$
$$SSE_K(\mathbf{U}|\mathbf{V}) := E_{(\mathbf{U},\mathbf{V})} \left\| \phi_K(\mathbf{U}) - \mu_K(\mathbf{P}_{\mathbf{U}|\mathbf{V}}) \right\|^2_{\mathbb{H}_K}.$$

Alternative formulas for easier computation of the two components are

$$SSR_K(\mathbf{U}|\mathbf{V}) = E_{\mathbf{V}} E_{\mathbf{U}|\mathbf{V},\mathbf{U}'|\mathbf{V}} K(\mathbf{U},\mathbf{U}') - E_{\mathbf{U},\mathbf{U}'} K(\mathbf{U},\mathbf{U}'),$$
$$SSE_K(\mathbf{U}|\mathbf{V}) = E_{\mathbf{U}} K(\mathbf{U},\mathbf{U}) - E_{\mathbf{V}} E_{\mathbf{U}|\mathbf{V},\mathbf{U}'|\mathbf{V}} K(\mathbf{U},\mathbf{U}'),$$

where $(\mathbf{U}', \mathbf{V}')$ is an independent and identically distributed copy of $(\mathbf{U}, \mathbf{V})$ and $E_{\mathbf{U}|\mathbf{v},\mathbf{U}'|\mathbf{v}}(\cdot)$ denotes conditional expectation $E(\cdot|\mathbf{V} = \mathbf{v}, \mathbf{V}' = \mathbf{v})$. In fact, the kernel regression sum of squares $SSR_K(\mathbf{U}|\mathbf{V})$ is identical to the expected conditional characteristic function-based independence criterion (ECCFIC; Ke and Yin, 2020) between $\mathbf{U}$ and $\mathbf{V}$, and $SSTO_K(\mathbf{U})$ is the ECCFIC between $\mathbf{U}$ and itself. Then the proportion of total variation in $\phi_K(\mathbf{U})$ explained by $\mathbf{V}$ is

$$R^2_K(\mathbf{U}|\mathbf{V}) := \frac{SSR_K(\mathbf{U}|\mathbf{V})}{SSTO_K(\mathbf{U})}.$$

By Theorem 5 in Ke and Yin (2020), $0 \le R^2_K(\mathbf{U}|\mathbf{V}) \le 1$, where $R^2_K(\mathbf{U}|\mathbf{V}) = 0$ if and only if $\mathbf{U} \perp\!\!\!\perp \mathbf{V}$ and $R^2_K(\mathbf{U}|\mathbf{V}) = 1$ if and only if $\mathbf{U}$ is a measurable function of $\mathbf{V}$, assuming K is characteristic (Fukumizu et al., 2009). Examples of characteristic kernels include Gaussian, Laplacian, inverse multiquadratics, among others.

In a similar vein, the marginal effect associated with $\mathbf{V}$ given $\mathbf{Z}$ already contained in the model to explain $\mathbf{U}$ can be measured by the extra sum of squares

$$\begin{aligned} SSR_K(\mathbf{U}|\mathbf{V};\mathbf{Z}) &:= E_{\mathbf{V}} \left\| \mu_K(\mathbf{P}_{\mathbf{U}|(\mathbf{V},\mathbf{Z})}) - \mu_K(\mathbf{P}_{\mathbf{U}|\mathbf{Z}}) \right\|^2_{\mathbb{H}_K} \\ &= E_{(\mathbf{V},\mathbf{Z})} E_{\mathbf{U}|(\mathbf{V},\mathbf{Z}),\mathbf{U}'|(\mathbf{V},\mathbf{Z})} K(\mathbf{U},\mathbf{U}') - E_{\mathbf{Z}} E_{\mathbf{U}|\mathbf{Z},\mathbf{U}'|\mathbf{Z}} K(\mathbf{U},\mathbf{U}'). \end{aligned}$$

The following equalities can be easily observed:

$$\begin{aligned} SSR_K(\mathbf{U}|\mathbf{V};\mathbf{Z}) &= SSR_K(\mathbf{U}|(\mathbf{V},\mathbf{Z})) - SSR_K(\mathbf{U}|\mathbf{Z}) \\ &= SSR_K(\mathbf{U}|\mathbf{Z};\mathbf{V}) + SSR_K(\mathbf{U}|\mathbf{V}) - SSR_K(\mathbf{U}|\mathbf{Z}), \end{aligned}$$

which align well with the properties of typical extra sum of squares in linear model. Furthermore, the partial R^2 that measures the proportion of the remaining variation in $\phi_K(\mathbf{U})$ after regressing on $\mathbf{Z}$ that is explained by adding $\mathbf{V}$ to the model is given by

$$R^2(\mathbf{U}|\mathbf{V};\mathbf{Z}) := \frac{SSR_K(\mathbf{U}|\mathbf{V};\mathbf{Z})}{SSE_K(\mathbf{U}|\mathbf{Z})}.$$

It can be shown analogously that $0 \le R^2(\mathbf{U}|\mathbf{V};\mathbf{Z}) \le 1$ with $R^2(\mathbf{U}|\mathbf{V};\mathbf{Z}) = 0$ if and only if $\mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{Z}$. The two R^2-type statistics in RKHS are nonlinear

generalizations of the classical R^2 and partial R^2 since they require no linearity or distribution assumptions.

In the next, we develop sample estimators for the above ANOVA statistics given observed data $\{\mathbf{U}_i, \mathbf{V}_i, \mathbf{Z}_i\}_{i=1}^n$. Firstly, $SSTO_K(\mathbf{U})$ can be simply estimated by

$$SSTO_{K,n}(\mathbf{U}) := \frac{1}{n}\sum_{t=1}^{n} K(\mathbf{U}_t, \mathbf{U}_t) - \frac{1}{n^2}\sum_{t_1,t_2=1}^{n} K(\mathbf{U}_{t_1}, \mathbf{U}_{t_2}).$$

If $\mathbf{V}$ is continuous, the Nadaraya-Watson estimator of $SSR_K(\mathbf{U}|\mathbf{V})$, relying on a product smoothing kernel $G : \mathbb{R}^{p_2} \to \mathbb{R}$ and a tuning bandwidth $h = h(n) \in \mathbb{R}$, is given by

$$SSR_{K,G,n}(\mathbf{U}|\mathbf{V}) := \frac{1}{n^3}\sum_{t_1,t_2,t_3=1}^{n} \frac{G_{t_1t_2}G_{t_1t_3}K_{t_2t_3}}{\frac{1}{n^2}\sum_{s_1,s_2=1}^{n} G_{t_1s_1}G_{t_1s_2}} - \frac{1}{n^2}\sum_{t_1,t_2=1}^{n} K_{t_1t_2},$$

where $G_{t_1t_2} = G_h(\mathbf{V}_{t_1} - \mathbf{V}_{t_2})$, $G_h(\cdot) = h^{-q}G(\cdot/h)$, and $K_{t_1t_2} = K(\mathbf{U}_{t_1}, \mathbf{U}_{t_2})$. Then the extra sum of squares can be estimated by

$$SSR_{K,\widetilde{G},G,n}(\mathbf{U}|\mathbf{V};\mathbf{Z}) := SSR_{K,\widetilde{G},n}(\mathbf{U}|(\mathbf{V},\mathbf{Z})) - SSR_{K,G,n}(\mathbf{U}|\mathbf{Z}),$$

using one of its properties, where $\widetilde{G} : \mathbb{R}^{p_2+p_3} \to \mathbb{R}$ is a product smoothing kernel applied on $\mathbf{V}$ and $\mathbf{Z}$ jointly with a tuning bandwidth $\tilde{h} = \tilde{h}(n) \in \mathbb{R}$. If $\mathbf{V}$ is categorical with L levels $\{\mathbf{v}^{(l)}\}_{l=1}^L$ and within each level we have n_l observations $\{\mathbf{U}_i^{(l)}, \mathbf{v}^{(l)}, \mathbf{Z}_i^{(l)}\}_{i=1}^{n_l}$, then a natural estimator of $SSR_K(\mathbf{U}|\mathbf{V})$ is given by

$$SSR_{K,n}(\mathbf{U}|\mathbf{V}) := \frac{1}{n}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{t_1,t_2=1}^{n_l} K(\mathbf{U}_{t_1}^{(l)}, \mathbf{U}_{t_2}^{(l)}) - \frac{1}{n^2}\sum_{t_1,t_2=1}^{n} K(\mathbf{U}_{t_1}, \mathbf{U}_{t_2}).$$

And the extra sum of squares can be estimated by

$$SSR_{K,G,n}(\mathbf{U}|\mathbf{V};\mathbf{Z}) := SSR_{K,G,n}(\mathbf{U}|\mathbf{Z};\mathbf{V}) + SSR_{K,n}(\mathbf{U}|\mathbf{V}) - SSR_{K,G,n}(\mathbf{U}|\mathbf{Z}),$$

where $SSR_{K,G,n}(\mathbf{U}|\mathbf{Z};\mathbf{V}) := \sum_{l=1}^{L}\frac{n_l}{n}SSR_{K,G,n}(\mathbf{U}^{(l)}|\mathbf{Z}^{(l)})$ is the weighted sum of within-level regression sums of squares. Sample R^2 and sample partial R^2 can be calculated accordingly, based on the type of $\mathbf{V}$. With a slight abuse of notation in exchange for simplicity, hereafter we will indiscriminately use $R_n^2(\mathbf{U}|\mathbf{V})$ and $R_n^2(\mathbf{U}|\mathbf{V};\mathbf{Z})$ to refer to their respective estimators for either type of $\mathbf{V}$.

The proposed kernel-based ANOVA statistics are closely related to the well-known class of Hilbert-Schmidt independence criterion (HSIC; Gretton et al., 2005) under the RKHS framework (Ke and Yin, 2020). To assess the correlation between two random vectors $\mathbf{U}$ and $\mathbf{V}$, HSIC measures the discrepancy between the joint distribution $\mathbf{P}_{(\mathbf{U},\mathbf{V})}$ and the product of the marginal distributions $\mathbf{P}_\mathbf{U}\mathbf{P}_\mathbf{V}$. Thus, the two random vectors are treated symmetrically. Distance correlation (DCOR; Székely, Rizzo and Bakirov, 2007) is a special case of HSIC

correlation when a distance kernel is used (Sejdinovic et al., 2013). From the perspective of regression, we consider the discrepancy between the conditional distribution $\mathbf{P}_{\mathbf{U}|\mathbf{V}}$ and the unconditional distribution $\mathbf{P}_{\mathbf{U}}$ instead. The resulting R^2 statistic attains zero if and only if the response is independent of the predictors, and reaches one if and only if the predictors fully explain the response. In contrast, although HSIC equal zero indicates independence and vice versa, it is not clear under what circumstances HSIC approaches its upper bound or how the random vectors are related when the upper bound is attained. Therefore, compared with HSIC, the kernel-based R^2 statistic better quantifies the contribution of the predictors to the response because it characterizes both independence and functional dependence in a supervised way. The kernel-based partial R^2 statistic as a supervised analogue of conditional DCOR (CDCOR; Wang et al., 2015) also boasts the same advantage. Besides, DCOR and CDCOR only apply to continuous random vectors, but the kernel-based ANOVA statistics can allow either continuous or categorical vectors. More benefits of using the proposed statistics as filters in variable screening are to demonstrate in the next subsection.

2.3. Utility measures and deviation bounds

Returning to the context of conditional sufficient variable screening, the following kernel-based R^2 and partial R^2 statistics are employed to assess conditions $(b1)$ and $(b2)$ in Proposition 1, respectively:

1) $w_j^M := R_K^2(X_j|Y) = \dfrac{SSE_K(X_j|Y)}{SSTO_K(X_j)}$ and

2) $w_j^C := R_{\widetilde{K}}^2((\mathbf{X}_{-j}, \mathbf{W})|Y; X_j) = \dfrac{SSR_{\widetilde{K}}(\mathbf{X}_{-j}, \mathbf{W})|Y; X_j)}{SSE_{\widetilde{K}}(\mathbf{X}_{-j}, \mathbf{W})|X_j)}$,

for selected reproducing kernels $K : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ and $\widetilde{K} : \mathbb{R}^{p+p_0-1}\times\mathbb{R}^{p+p_0-1} \to \mathbb{R}$. The sample utility measures are denoted by $\widehat{w}_j^M := R_n^2(X_j|Y)$ and $\widehat{w}_j^C := R_n^2((\mathbf{X}_{-j}, \mathbf{W})|Y; X_j)$ for observed data $\{\mathbf{X}_i, Y_i, \mathbf{W}_i\}_{i=1}^n$. The marginal utility measure is the kernel R^2 for the inverse regression of $X_j|Y$, whereas the conditional utility measure is the kernel partial R^2 associated with Y given X_j already contained in the inverse model to explain $(\mathbf{X}_{-j}, \mathbf{W})$. Note that the relation between X_j and the other predictors is adjusted for in the conditional utility measure, which helps distinguish true signal from pure noise that is fully independent of Y and $(\mathbf{X}_{\mathcal{A}}, \mathbf{W})$ because the denominator $SSE_{\widetilde{K}}(\mathbf{X}_{-j}, \mathbf{W})|X_j)$ becomes larger if X_j is less correlated with the other predictors. The idea of inverse regression has been implanted in many dimension reduction methods for high-dimensional data (Li, 1991; Cook and Weisberg, 1991; Li and Wang, 2007, etc.) to tackle the difficulty of handling the conditional distribution of $Y|\mathbf{X}$ in a standard forward way. Here with inverse regression, the response variable is allowed to be either continuous or categorical, and estimating conditional expectation via Nadaraya-Watson is only applied to a single conditional variable (continuous Y or X_j) or bivariate conditional vector (X_j, Y) so that the

curse of dimensionality is avoided. As a payoff, dimension-free deviation bounds can be found for the sample utility measures as demonstrated in Theorem 1 below. Note that Ke and Yin (2020) studied finite sample properties such as $\sqrt{n}$-consistency for kernel regression sum of squares with fix dimension, so their results do not adapt to the ultrahigh-dimensional setting considered in this paper, where the dimension can diverge with the sample size. Let $f_{\mathbf{U}}(\mathbf{u})$ denote the density function of $\mathbf{U}$, and $f_{\mathbf{U}|\mathbf{V}}(\mathbf{u}|\mathbf{v})$ denote the conditional density function of $\mathbf{U}$ given $\mathbf{V}=\mathbf{v}$. The following regularity conditions are imposed in Theorem 1 to facilitate the technical proof although they are certainly not the weakest ones.

(C1) The characteristic kernels K and $\widetilde{K}$ are bounded.

(C2) The smoothing kernels $G:\mathbb{R}\to\mathbb{R}$ and $\widetilde{G}:\mathbb{R}^2\to\mathbb{R}$ are products of univariate kernel $g:\mathbb{R}\to\mathbb{R}$ such that $\int_{\mathbb{R}} u^i g(u)du=\delta_{i0}$ $(i=0,1)$ and $g(u)=O((1+|u|^4)^{-1})$, where δ_{ij} is Kronecker's delta.

(C3) $h,\tilde{h}\to 0$ and $nh^2, n\tilde{h}^4\to\infty$ as $n\to\infty$.

(C4) X_j is continuous and $f_{X_j}(x_j)$ is bounded away from zero, for $j=1,\ldots,p$. In addition, the first partial derivatives of $f_{X_j}(x_j)$, $f_{\mathbf{X},\mathbf{W}}(\mathbf{x},\mathbf{w})$ and $f_{\mathbf{X}_{-j},\mathbf{W}|X_j}(\mathbf{x}_{-j},\mathbf{w}|x_j)$ with respect to x_j are uniformly bounded by some constants that do not depend on x_j, for each j.

(C5) Y is categorical with $L=O(n^{\kappa})$ levels $\{y^{(1)},y^{(2)},\ldots,y^{(L)}\}$ for some $\kappa\geq 0$. Let $P_l:=P(Y=y^{(l)})$ for $l=1,\ldots,L$, then there exists $c_0>0$ such that $\min_{1\leq l\leq L}P_l\geq 2c_0/L$.

(C6) Y is continuous and $f_Y(y)$ as well as $f_{X_j,Y}(x_j,y)$ are bounded away from zero, for $j=1,\ldots,p$. In addition, the first partial derivatives of $f_Y(y)$, $f_{X_j,Y}(x_j,y)$ and $f_{X_j|Y}(x_j|y)$ with respect to y are uniformly bounded by some constants that do not rely on y, and the first partial derivatives of $f_{\mathbf{X},\mathbf{W},Y}(\mathbf{x},\mathbf{w},y)$ and $f_{\mathbf{X}_{-j},\mathbf{W}|X_j,Y}(\mathbf{x}_{-j},\mathbf{w}|x_j,y)$ with respect to (x_j,y) are uniformly bounded by some constants that do not rely on y and x_j, for each j.

Condition (C1) is also adopted in Balasubramanian, Sriperumbudur and Lebanon (2013) for reproducing kernels. Condition (C2) implies that the smoothing kernel function is bounded from above and satisfies some moment conditions, which holds for many well-known kernel functions. Condition (C3) requires the bandwidth to be chosen appropriately according to n. Conditions (C4) and (C6) impose smoothness conditions on the density functions, which can be relaxed by assuming local Lipschitz properties (Li, Zhu and Zhu, 2011; Ke and Yin, 2020). Conditions (C2)–(C4) and (C6) are commonly assumed in applications of Nadaraya-Watson estimators (Chen, Cook and Zou, 2015; Yin and Yuan, 2020). Condition (C5) allows a diverging number of levels for a categorical response but the proportion of each level should not be too small (see also Cui, Li and Zhong, 2015).

Theorem 1. *If Y is categorical, then under conditions (C1)–(C5),*

$$P\left(|\widehat{w}_j^M - w_j^M|>\epsilon\right)\leq 2L\exp\left\{-\frac{a_1 n\epsilon^2}{L^3}\right\}$$

$$\text{and } P\left(|\widehat{w}_j^C - w_j^C| > \epsilon\right) \le 2Ln \exp\left\{-\frac{a_2 n\epsilon^2}{L^3}\right\}$$

for any $\epsilon > 0$*, where* $j = 1, \ldots, p$*, and* $a_1, a_2 > 0$ *are some constants depending on* c_0*. If* Y *is continuous, then under conditions (C1)–(C4) and (C6),*

$$P\left(|\widehat{w}_j^M - w_j^M| > \epsilon\right) \le 2n \exp\left\{-a_3 n\epsilon^2\right\}$$
$$\text{and } P\left(|\widehat{w}_j^C - w_j^C| > \epsilon\right) \le 2n \exp\left\{-a_4 n\epsilon^2\right\}$$

for any $\epsilon > 0$*, where* $j = 1, \ldots, p$*, and* $a_3, a_4 > 0$ *are some constants.*

If we let the deviation of the sample marginal measure to vanish as n increases by setting $\epsilon = c_1 n^{-\gamma_1}$ for some positive constants c_1 and γ_1, then with a categorical response we have

$$P\left(\max_{1\le j\le p} |\widehat{w}_j^M - w_j^M| > c_1 n^{-\gamma_1}\right) \le O\left(p \exp\left\{-b_1 n^{1-2\gamma_1-3\kappa} + \kappa \log n\right\}\right),$$

where $b_1 > 0$ is a constant depending on c_0 and c_1. Similarly, taking the deviation of the sample conditional measure to be $\epsilon = c_2 n^{-\gamma_2}$ for some positive constants c_2 and γ_2,

$$P\left(\max_{1\le j\le p} |\widehat{w}_j^C - w_j^C| > c_2 n^{-\gamma_2}\right) \le O\left(p \exp\left\{-b_2 n^{1-2\gamma_2-3\kappa} + (1+\kappa) \log n\right\}\right),$$

where $b_2 > 0$ is a constant depending on c_0 and c_2. Assuming $1-2\gamma-3\kappa \in (0, 1]$ where $\gamma := \max\{\gamma_1, \gamma_2\}$, the probability that the sample utility measures deviate from the true contribution of the predictors decays exponentially with n, as long as $\log p = o(n^{1-2\gamma-3\kappa})$. If $\kappa = 0$, meaning the response variable has a fixed number of categories, or if the response variable is continuous, the order can be relaxed to $\log p = o(n^{1-2\gamma})$. The consistency is attained regardless of the dimension of $\mathbf{W}$, with no string attached to the underlying model structure. This salient property, together with the flexibility in the response variable and the generalized ANOVA interpretation, makes the proposed utility measures very appealing for ultrahigh-dimensional data.

2.4. The screening procedure

By Proposition 1, predictors making discernibly marginal or conditional contribution should be retained. Therefore, the active index set can be estimated by

$$\widehat{\mathcal{A}} := \left\{1 \le j \le p : \widehat{w}_j^M \ge c_1 n^{-\gamma_1} \text{ or } \widehat{w}_j^C \ge c_2 n^{-\gamma_2}\right\},$$

where c_1, c_2, γ_1 and γ_2 are constants relying on the strength of the true signal, which will be defined soon in condition (C7). We henceforth refer to the above screening procedure as kernel-based conditional sufficient variable screening, or KCSVS for short. In the next, we show that the proposed procedure embraces the sure screening property as well as the rank consistency property.

Let $\mathcal{A}_M := \{j \in \mathcal{A} : X_j \not\perp\!\!\!\perp \mathbf{Y}\}$, $\mathcal{A}_C := \mathcal{A} \setminus \mathcal{A}_M$. The following conditions ensure that important predictors are detectable and are distinguishable from noise.

(C7) There exist $c_1, c_2 > 0$ and $\gamma_1, \gamma_2 \in [0, 1/2)$, such that

$$\min_{j\in\mathcal{A}_M} w_j^M \geq 2c_1 n^{-\gamma_1} \text{ and } \min_{j\in\mathcal{A}_C} w_j^C \geq 2c_2 n^{-\gamma_2}.$$

(C8) There exist $c_3, c_4 > 0$ and $\gamma_3, \gamma_4 \in [0, 1/2)$, such that

$$\min_{j\in\mathcal{A}_M} w_j^M - \max_{j\notin\mathcal{A}_M} w_j^M \geq 2c_3 n^{-\gamma_3} \text{ and } \min_{j\in\mathcal{A}_C} w_j^C - \max_{j\notin\mathcal{A}_C} w_j^C \geq 2c_4 n^{-\gamma_4}.$$

Condition (C7) is typically assumed in the literature of variable screening (Fan and Lv, 2008; Li, Zhong and Zhu, 2012; Yang, Yin and Zhang, 2019, etc.) requiring that the true signal cannot diminish too fast as n diverges. Condition (C8) further restricts the decay rate of the discrepancy between the true signal and the noise (Zhu et al., 2011; Cui, Li and Zhong, 2015; Liu et al., 2022, etc.). The two properties of KCSVS are exhibited in Theorems 2 and 3, respectively.

Theorem 2 (Sure Screening)**.** *Let $s := |\mathcal{A}|$ and $\gamma := \max\{\gamma_1, \gamma_2\}$. If Y is categorical, then under conditions (C1)–(C5) and (C7), we have*

$$P\left(\mathcal{A} \subset \widehat{\mathcal{A}}\right) \geq 1 - O\left(s \exp\left\{-bn^{1-2\gamma-3\kappa} + (1+\kappa)\log n\right\}\right)$$

for $\kappa \in [0, \frac{1}{3} - \frac{2\gamma}{3})$, where b is a positive constant depending on c_0, c_1 and c_2. If Y is continuous, then under conditions (C1)–(C4), (C6) and (C7), we have

$$P\left(\mathcal{A} \subset \widehat{\mathcal{A}}\right) \geq 1 - O\left(s \exp\left\{-\tilde{b}n^{1-2\gamma} + \log n\right\}\right),$$

where $\tilde{b}$ is a positive constant depending on c_1 and c_2.

As implied by Theorems 2, KCSVS selects all important predictors asymptotically almost surely, including those that are conditionally independent with the response given the control variables.

Theorem 3 (Rank Consistency)**.** *Let $\tilde{\gamma} := \max\{\gamma_3, \gamma_4\}$. If Y is categorical and $\log p = o(n^{1-2\tilde{\gamma}-3\kappa})$ for $\kappa \in [0, \frac{1}{3} - \frac{2\tilde{\gamma}}{3})$, then under conditions (C1)–(C5) and (C8),*

$$\liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_M} \widehat{w}_j^M - \max_{j\notin\mathcal{A}_M} \widehat{w}_j^M\right\} > 0 \text{ and } \liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_C} \widehat{w}_j^C - \max_{j\notin\mathcal{A}_C} \widehat{w}_j^C\right\} > 0$$

almost surely. If Y is continuous and $\log p = o(n^{1-2\tilde{\gamma}})$, then under conditions (C1)–(C4), (C6) and (C8),

$$\liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_M} \widehat{w}_j^M - \max_{j\notin\mathcal{A}_M} \widehat{w}_j^M\right\} > 0 \text{ and } \liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_C} \widehat{w}_j^C - \max_{j\notin\mathcal{A}_C} \widehat{w}_j^C\right\} > 0$$

almost surely.

Theorem 3 guarantees that important predictors can be separated from nulls by some thresholds as long as there is truly a gap between the two sets of variables in terms of the signal strength.

There is no established way of determining such threshold values in a finite sample setting. As it is commonly assumed that the cardinality of the truly important set is small, one may specify a model size $d < n$ and select $\widehat{\mathcal{A}}^* := \widehat{\mathcal{A}}^*_M \cup \widehat{\mathcal{A}}^*_C$, where

$$\widehat{\mathcal{A}}^*_M(d_1) := \{1 \le j \le p : \widehat{w}^M_j \text{ is among the first } d_1 \text{ largest of all}\},$$
$$\widehat{\mathcal{A}}^*_C(d_2) := \{j \notin \widehat{\mathcal{A}}^*_M(d_1) : \widehat{w}^C_j \text{ is among the first } d_2 \text{ largest of all}\},$$

for $d_1 + d_2 = d$. Typical choices of d are $[n/\log(n)]$, $2[n/\log(n)]$, $3[n/\log(n)]$, and $n-1$ (Fan and Lv, 2008; Li, Zhong and Zhu, 2012). And we can simply set $d_1 = d_2 = [d/2]$, in which case the marginal and conditional utility measures are weighted equally in the selection of $\widehat{\mathcal{A}}^*$. Let $\{r^M_j\}^p_{j=1}$ and $\{r^C_j\}^p_{j=1}$ be the two rankings of variables by $\{\widehat{w}^M_j\}^p_{j=1}$ and $\{\widehat{w}^C_j\}^p_{j=1}$, respectively. A joint ranking $\{r_j\}^p_{j=1}$ can be acquired by ascending $(r^M_j \wedge r^C_j, r^M_j \vee r^C_j)$. Then selecting the top d variables is identical to the trivial choice of $\widehat{\mathcal{A}}^*$ with $d_1 = d_2$. The sure screening property ensures that the probability of selecting all the active predictors is close to one when d is sufficiently large. If we only select $\widehat{\mathcal{A}}^* := \widehat{\mathcal{A}}^*_M(d)$, kernel-based univariate screening (KUS) is performed, which can be regarded as a model-free generalization of linear SIS (Fan and Lv, 2008).

3. Numerical studies

In this section, we test the performance of the proposed procedure through simulation studies and real data analysis.

3.1. Simulation studies

The following methods are included for comparisons with KCSVS in the simulation studies: NIS (Fan, Ma and Dai, 2014), CCSIS (Liu, Li and Wu, 2014), CSIS (Barut, Fan and Verhasselt, 2016), CDCSIS (Wen et al., 2018), BKRSIS (Zhou, Liu and Zhu, 2020) and KUS as a benchmark. Among aforementioned, only KCSVS and CSIS can handle a categorical response, or a high-dimensional control set, but CSIS still requires $p_0 < n - 1$. For each predictor CSIS fits a generalized linear model with the predictor and the control variables, and selects predictors with the largest absolute coefficients (for a normal or poisson response) or deviances (for a multiclass response). NIS and CCSIS are developed for varying coefficient model where the coefficient varies with an exposure variable. KCSVS and KUS are conducted with the Gaussian kernel being the reproducing kernel as well as the smoothing kernel for density estimation. The bandwidths of the two Gaussian kernels are set to heuristic median pairwise distance (Gretton et al., 2008) and $h = (n(q+2)/4)^{-1/(q+4)}\sigma^2$ (Silverman,

1986), respectively, where q is the dimension of the random vector conditional on which an expectation is estimated and σ^2 can be estimated by its average marginal sample variance. Discussions on how to choose the reproducing kernel and associated parameters can be found in Fukumizu et al. (2009) and Gretton et al. (2012). All variables are standardized prior to screening. We report the following statistics based on 200 replicates:

- the τ^{th} quantiles of the minimum model size (MMS) that includes all active predictors, denoted as M_τ, for $\tau = 5\%,\ 25\%,\ 50\%,\ 75\%,\ 95\%$;
- the proportion of selecting a certain active predictor X_j, denoted as P_j^s, and the proportion of including all active predictors, denoted as P_a, given a model size d.

MMS is defined as $\min\{M_1 + M_2\}$ such that $\mathcal{A} \subseteq \{1 \le j \le p : \widehat{w}_j^M$ is among the first M_1 largest of all$\} \cup \{1 \le j \le p : \widehat{w}_j^C$ is among the first M_2 largest of all$\}$. By default, we set $n = 300$, $p = 10000$, and $d = 2[n/\log n] = 105$. The size of the control set p_0 varies from 1 to 2000.

Example 1. Let $\mathbf{X} \in \mathbb{R}^{p+p_0+1}$ be distributed as $N_{p+p_0+1}(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 1$, $\sigma_{i5} = \sigma_{5i} = 0$ for $i \neq 5$, and $\sigma_{ij} = 0.5$ otherwise. That is, Σ is compound symmetry except that X_5 is independent with other variables. Given $\mathbf{X}$, we simulate a variety of generalized linear/nonlinear models as follows.

Model 1: (Linear) $Y = 0.8(X_1+X_2+X_3-1.5X_4+X_5)+\epsilon$, where $\epsilon \sim N(0,1)$.
Model 2: (Heterogeneity) $Y = \frac{2}{3}(X_1-0.5X_4)+\sin(\frac{\pi}{2}(X_2-0.5X_4))+\frac{1}{3}(X_3-0.5X_4+1)^2+|X_5|\epsilon$, where $\epsilon \sim N(0,1)$.
Model 3: (Probit) $Y = \operatorname{argmax}_{l=1}^4 Y^{(l)}$, where $Y^{(l)} = (-1)^{l+1}(X_l-0.5X_4)+\epsilon_l$ for $l = 1, 2, 3$, $Y^{(4)} = X_5 + \epsilon_4$, and $\boldsymbol{\epsilon} = (\epsilon_1 \cdots \epsilon_4) \sim N_4(\mathbf{0}, I)$.
Model 4: (Poisson) $Y \sim \text{Poisson}(\lambda)$, where $\lambda = e^{5g(\mathbf{X})-2}/(1+e^{5g(\mathbf{X})-3})$ and $g(\mathbf{X}) = X_1 + X_2 + X_3 - 1.5X_4 + X_5$.

All models are designed such that $X_4 \perp\!\!\!\perp Y | X_5$. Conditioning on the following covariate subsets of size p_0 for each model, the screening procedures are performed on the remaining first p predictors:

(a) $W = X_5$, $p_0 = 1$;
(b) $\mathbf{W} = (X_5, X_{p+2}, X_{p+3}, \cdots, X_{p+p_0})$, $p_0 = 2000$;
(c) $W = X_1 + X_5$, $p_0 = 1$;
(d) $\mathbf{W} = (X_1 + X_5, X_2)$, $p_0 = 2$;
(e) $\mathbf{W} = (X_1 + X_5, X_2, X_{p+4}, X_{p+5}, \cdots, X_{p+p_0+1})$, $p_0 = 2000$.

Theses control sets are deliberately chosen to examine the screening procedures when (a): there exists an important variable (X_4) that is conditionally independent of the response given the control set; (c): only a compound of some important variables is available as the control variable; (d): multiple control variables are considered; (b) and (e): the high-dimensional control set contains a large amount of noise in addition to known important variables.

The results are summarized in Table 1. Note that not all methods are included for comparisons in every model as some are not applicable to certain

TABLE 1

Quantiles of MMS M_τ's and average selection proportions P_j^s's and P_a's for models in Example 1 based on 200 replicates. A cell for P_j^s displays a dash if the corresponding variable X_j is assigned to the control set and thus protected from screening.

Model	p_0	s	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_2^s	P_3^s	P_4^s	P_a
1(a)	1	4	NIS	9988.9	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	0.000	0.000
			CCSIS	9998.0	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	0.000	0.000
			CSIS	10000.0	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	0.000	0.000
			CDCSIS	9995.9	10000.0	10000.0	10000.0	10000.0	1.000	1.000	0.995	0.000	0.000
			BKRSIS	9998.0	10000.0	10000.0	10000.0	10000.0	1.000	0.995	0.995	0.000	0.000
			KUS	9692.4	9980.0	9999.0	10000.0	10000.0	0.995	1.000	0.980	0.000	0.000
			KCSVS	4.0	4.0	5.0	12.0	68.0	0.980	0.990	0.965	1.000	0.935
1(b)	2000	4	KCSVS	4.0	4.0	5.0	12.0	68.0	0.980	0.990	0.965	1.000	0.935
1(c)	1	3	NIS	3.0	3.0	3.0	12.5	715.3	–	1.000	1.000	0.870	0.870
			CCSIS	3.0	3.0	7.0	185.0	5091.1	–	1.000	1.000	0.720	0.720
			CSIS	3.0	3.0	3.0	3.0	775.7	–	1.000	1.000	0.915	0.915
			CDCSIS	3.0	7.0	77.0	1171.8	8049.9	–	1.000	1.000	0.530	0.530
			BKRSIS	3.0	3.0	4.0	26.0	2660.6	–	1.000	1.000	0.800	0.800
			KUS	9692.4	9980.0	9999.0	10000.0	10000.0	–	1.000	0.980	0.000	0.000
			KCSVS	3.0	3.0	3.0	6.0	50.1	–	0.990	0.965	1.000	0.955
1(d)	2	2	CSIS	2.0	2.0	2.0	38.8	905.0	–	–	0.810	1.000	0.810
			KUS	9692.4	9980.0	9999.0	10000.0	10000.0	–	–	0.980	0.000	0.000
			KCSVS	2.0	2.0	2.0	2.0	44.1	–	–	0.965	1.000	0.965
1(e)	2000	2	KCSVS	2.0	2.0	2.0	2.0	44.1	–	–	0.965	1.000	0.965
2(a)	1	4	NIS	9582.0	9989.0	10000.0	10000.0	10000.0	1.000	0.985	0.995	0.000	0.000
			CCSIS	9994.0	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	0.000	0.000
			CSIS	9994.6	10000.0	10000.0	10000.0	10000.0	1.000	0.995	1.000	0.000	0.000
			CDCSIS	9993.0	10000.0	10000.0	10000.0	10000.0	0.995	1.000	1.000	0.000	0.000
			BKRSIS	9987.9	10000.0	10000.0	10000.0	10000.0	0.990	1.000	0.995	0.000	0.000
			KUS	9629.8	9977.0	9999.0	10000.0	10000.0	0.995	1.000	1.000	0.000	0.000
			KCSVS	4.0	4.0	4.0	6.0	24.2	0.980	1.000	0.995	1.000	0.980
2(b)	2000	4	KCSVS	4.0	4.0	4.0	6.0	24.2	0.985	1.000	0.995	1.000	0.980
2(c)	1	3	NIS	73.0	4114.8	7980.5	9796.0	9999.0	–	0.985	1.000	0.060	0.060
			CCSIS	814.5	6272.5	9566.5	9982.2	10000.0	–	1.000	1.000	0.020	0.020
			CSIS	97.8	4171.8	8755.5	9965.5	10000.0	–	1.000	1.000	0.055	0.055
			CDCSIS	1204.4	6845.5	9570.5	9978.2	10000.0	–	1.000	1.000	0.010	0.010
			BKRSIS	169.8	3193.0	7969.0	9885.5	10000.0	–	1.000	1.000	0.045	0.045
			KUS	9629.8	9977.0	9999.0	10000.0	10000.0	–	1.000	1.000	0.000	0.000
			KCSVS	3.0	3.0	3.0	3.0	7.0	–	1.000	0.995	1.000	0.995
2(d)	2	2	CSIS	2.0	2.0	2.0	2.0	4.1	–	–	0.990	0.995	0.985
			KUS	9629.7	9977.0	9999.0	10000.0	10000.0	–	–	1.000	0.000	0.000
			KCSVS	2.0	2.0	2.0	2.0	6.0	–	–	0.995	1.000	0.995
2(e)	2000	2	KCSVS	2.0	2.0	2.0	2.0	6.0	–	–	0.995	1.000	0.995
3(a)	1	4	CSIS	9766.5	9997.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	0.000	0.000
			KUS	8070.8	9799.5	9985.0	10000.0	10000.0	0.995	1.000	0.995	0.000	0.000
			KCSVS	4.0	4.0	5.0	8.0	40.2	0.980	1.000	0.990	0.990	0.960
3(b)	2000	4	KCSVS	4.0	4.0	5.0	8.0	40.2	0.980	1.000	0.990	0.990	0.960
3(c)	1	3	CSIS	522.9	5000.5	8605.5	9814.2	9997.0	–	1.000	1.000	0.015	0.015
			KUS	8070.8	9799.5	9985.0	10000.0	10000.0	–	1.000	0.995	0.000	0.000
			KCSVS	3.0	3.0	3.0	5.0	27.3	–	1.000	0.990	0.990	0.980
3(d)	2	2	CSIS	2.0	2.0	2.0	2.0	8.0	–	–	1.000	0.995	0.995
			KUS	8069.9	9799.5	9985.0	10000.0	10000.0	–	–	0.995	0.000	0.000
			KCSVS	2.0	2.0	2.0	4.0	26.3	–	–	0.990	0.990	0.980
3(e)	2000	2	KCSVS	2.0	2.0	2.0	4.0	26.3	–	–	0.990	0.990	0.980
4(a)	1	4	NIS	9943.8	9999.0	10000.0	10000.0	10000.0	0.995	1.000	0.995	0.000	0.000
			CCSIS	9978.9	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	0.000	0.000
			CSIS	9997.0	10000.0	10000.0	10000.0	10000.0	0.990	1.000	0.990	0.000	0.000
			CDCSIS	9990.0	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	0.000	0.000
			BKRSIS	9993.0	10000.0	10000.0	10000.0	10000.0	0.985	0.995	1.000	0.000	0.000
			KUS	9884.0	9996.8	10000.0	10000.0	10000.0	0.985	0.995	0.985	0.000	0.000
			KCSVS	4.0	4.0	5.5	11.2	75.1	0.965	0.980	0.975	1.000	0.920
4(b)	2000	4	KCSVS	4.0	4.0	6.0	13.2	68.3	0.961	0.978	0.983	1.000	0.922
4(c)	1	3	NIS	3.0	4.0	15.0	126.0	2821.9	–	1.000	0.990	0.745	0.735
			CCSIS	3.0	4.0	63.5	1056.8	6479.1	–	1.000	1.000	0.515	0.515
			CSIS	3.0	3.0	7.0	121.0	1844.0	–	0.995	1.000	0.745	0.740
			CDCSIS	3.0	16.2	453.5	3352.0	8515.0	–	1.000	1.000	0.410	0.410
			BKRSIS	3.0	5.0	53.0	1085.8	6009.7	–	0.975	0.975	0.570	0.545
			KUS	9884.0	9996.8	10000.0	10000.0	10000.0	–	0.995	0.985	0.000	0.000
			KCSVS	3.0	3.0	3.0	6.0	41.0	–	0.980	0.975	1.000	0.955
4(d)	2	2	CSIS	2.0	2.0	10.5	279.0	4609.8	–	–	0.675	1.000	0.675
			KUS	9884.0	9996.8	10000.0	10000.0	10000.0	–	–	0.985	0.000	0.000
			KCSVS	2.0	2.0	2.0	3.0	15.0	–	–	0.975	1.000	0.975
4(e)	2000	2	KCSVS	2.0	2.0	2.0	3.0	15.0	–	–	0.975	1.000	0.975

designs. Since KUS does not leverage any prior information, its performance does not vary from (a) to (b), or from (d) to (e), so duplicated results are omitted. Conditioning on (a) $W = X_5$, only KCSVS successfully detects X_4 for the variety of models studied in this example, and therefore it significantly outperforms the other methods with high selection probabilities and minimum model sizes close to the true size $s = 4$. Even when the control set is dominated by irrelevant variables besides X_5 in (b), KCSVS remains equally powerful. All the conditional screening methods are able to utilize the intermediate variable (c) $W = X_1 + X_5$ to improve screening for Model 1 and Model 4, but for Model 2 and Model 3, where the response variables do not directly rely on W, only KCSVS recovers the truly active set effectively. With more accurate information added to (d) $\mathbf{W} = (X_1 + X_5, X_2)$, the performance of CSIS catches up with KCSVS. In the presence of redundant information in (e), KCSVS still achieves sure screening and offers a strong advantage over CSIS as CSIS is no longer applicable when the dimension of $\mathbf{W}$ exceeds the sample size. It is also clear that KCSVS improves KUS as the conditional utility measure weighs in.

Example 2. In this example, we study another generalized linear model assuming first-order autoregression covariance structure for $\mathbf{X}$, as well as a multi-class classification model with independent non-normal predictors.

Model 5: (Linear) Let $\mathbf{X} \in \mathbb{R}^{p+p_0}$ be distributed as $N_{p+p_0}(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$. $Y = 5X_1 + 5X_2 + 2(X_3 + X_4)^2 + 5X_5\mathbf{1}_{\{X_5>0\}} + \exp\{X_6 + 2\sin(\pi X_7/2)\} + 5|X_8|\ln(1 + |X_8|) + (X_9 - 1)^3 + 5\sin(1/X_{10}) + \epsilon$, where $\epsilon \sim N(0, 1)$.

Model 6: (Multiclass) The response Y is generated with $P(Y = l) = 1/5$, $l = 1, \ldots, 5$. Given $Y = l$, X_{2l-1} and X_{2l} follow a normal mixture $0.5N(1, 0.2^2) + 0.5N(-1, 0.2^2)$ independently, and other variables follow the Cauchy distribution independently.

Screening is conditional on the following control vectors:

(a) $\mathbf{W} = (X_1, X_3, X_5, X_7, X_9)$, $p_0 = 5$;
(b) $\mathbf{W} = (X_1, X_3, X_5, \cdots, X_{2p_0-1})$, $p_0 = 2000$;
(c) $\mathbf{W} = (X_2, X_4, X_6, X_8, X_{10})$, $p_0 = 5$;
(d) $\mathbf{W} = (X_2, X_4, X_6, \cdots, X_{2p_0})$, $p_0 = 2000$.

For each model, all the important predictors are correlated with the response variable given any of the control vectors. In other words, the goal of this example is to assess KCSVS when $\mathcal{A} = \mathcal{A}_1$. Model 6 generates a 5-category response with a balanced design, which resembles Model 7 in Mai and Zou (2015).

As demonstrated in Table 2, both KUS and KCSVS work reasonably well in all scenarios and lead CSIS by large margins. CSIS can barely identify the nonlinear active predictors in Model 5 and Model 6. In contrast, KCSVS as a model-free screening procedure shows superior adaptability to different types and structures of data. Moreover, we again observe that KCSVS is flexible about the dimension of the control vector and is robust to the noise therein.

TABLE 2
Quantiles of MMS M_τ's and average selection proportions P_j^s's and P_a's for models in Example 2 based on 200 replicates.

Model	p_0	s	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_2^s	P_4^s	P_6^s	P_8^s	P_{10}^s	P_a
5(a)	5	5	CSIS	1045.3	3342.8	6042.5	8289.5	9692.4	0.985	0.080	0.990	0.075	0.615	0.000
			KUS	5.0	5.0	5.0	8.0	68.4	1.000	0.995	1.000	0.980	1.000	0.975
			KCSVS	5.0	5.0	5.0	8.0	68.4	1.000	0.980	1.000	0.955	0.995	0.935
5(b)	2000	5	KCSVS	5.0	5.0	5.0	7.0	71.3	1.000	0.980	1.000	0.955	0.995	0.935
6(a)	5	5	CSIS	5841.7	8078.2	8942.5	9623.0	9961.3	0.020	0.010	0.010	0.020	0.010	0.000
			KUS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
			KCSVS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
6(b)	2000	5	KCSVS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
Model	p_0	s	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_3^s	P_5^s	P_7^s	P_9^s	P_a
5(c)	5	5	CSIS	403.5	2084.2	5291.5	7893.8	9263.4	0.990	0.065	0.600	0.315	1.000	0.010
			KUS	5.0	5.0	5.0	5.0	11.1	1.000	1.000	0.995	0.995	1.000	0.990
			KCSVS	5.0	5.0	5.0	5.0	11.1	1.000	0.995	0.995	0.995	1.000	0.985
5(d)	2000	5	KCSVS	5.0	5.0	5.0	5.0	13.0	1.000	0.995	0.995	0.995	1.000	0.985
6(c)	5	5	CSIS	5873.4	7646.0	8844.0	9531.5	9928.0	0.015	0.025	0.010	0.025	0.015	0.000
			KUS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
			KCSVS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
6(d)	2000	5	KCSVS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000

Example 3. This example is to evaluate KCSVS for varying coefficient model. Let $(\mathbf{U}^*, \mathbf{X})$ be generated from $N_{p+p_0}(\mathbf{0}, \Sigma)$ for some Σ specified later, where $\mathbf{U}^*\mathbb{R}^{p_0}, \mathbf{X} \in \mathbb{R}^p$. Let $\mathbf{U} = \Phi(\mathbf{U}^*)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N_{p_0}(\mathbf{0}, I)$. Define $\beta_1(\mathbf{U}) = 2\cos(\pi U_1/2) + 2$, $\beta_2(\mathbf{U}) = 2U_2 + 2$, $\beta_3(\mathbf{U}) = (2 - U_3)^2$, $\beta_4(\mathbf{U}) = -2\sin^2(2\pi U_4) - 2$ and $\beta_5(\mathbf{U}) = -0.5\sum_{j=1}^4 \beta_j(\mathbf{U})$.

Model 7: (Linear) Let $(U_1^*, U_2^*, U_3^*, U_4^*)$ be distributed as $N_4(\mathbf{0}, \Sigma_1)$ and $(U_5^*, \ldots, U_{p_0}^*, \mathbf{X})$ be distributed as $N_{p+p_0-4}(\mathbf{0}, \Sigma_1)$ independently, where $\Sigma_1 = (\sigma_{ij})$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5$ for $i \neq j$. Consider $Y = \sum_{j=1}^5 \beta_j(\mathbf{U})X_j + \epsilon$, where $\epsilon \sim N(0, 1)$.

Model 8: (Linear) The same as Model 7 except that $(\mathbf{U}^*, \mathbf{X}) \sim N_{p+p_0}(\mathbf{0}, \Sigma_1)$.

Model 9: (Probit) Let $(\mathbf{U}^*, \mathbf{X})$ be generated as in Model 7 and let $Y = \operatorname{argmax}_{l=1}^4 Y^{(l)}$, where $Y^{(l)} = (-1)^{l+1}\beta_l(\mathbf{U})(X_l - 0.5X_5) + \epsilon_l$ for $l = 1, 2, 3, 4$, and $\boldsymbol{\epsilon} = (\epsilon_1 \cdots \epsilon_4) \sim N_4(\mathbf{0}, I)$.

Model 10: (Interaction) Let $(\mathbf{U}^*, \mathbf{X}) \sim N_{p+p_0}(\mathbf{0}, \Sigma_2)$, where $\Sigma_2 = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$. Consider $Y = 3\beta_1 X_1 X_2 + \beta_2(X_{12} + 1)^2 + 3\beta_3 \sin(\pi X_{22}/2) + \beta_4 \exp(|X_{33}|) + \epsilon$, where $\epsilon \sim N(0, 1)$.

Model 7 and Model 9 are designed such that $X_5 \perp\!\!\!\perp Y | (U_1, \ldots, U_4)$. In Model 8, $\mathbf{U}$ and $\mathbf{X}$ are correlated, while in the other models, the important exposure variables $(U_1, \ldots, U_4)$ are independent or almost uncorrelated with the predictors of primary interest and other nuisance exposure variables. Screening procedures are conducted given the following control vectors:

(a) $W = U_1$, $p_0 = 1$;
(b) $\mathbf{W} = (U_1\ U_2\ U_3\ U_4)$, $p_0 = 4$;
(c) $\mathbf{W} = \mathbf{U}$, $p_0 = 2000$.

These model are adapted from examples in Liu, Li and Wu (2014) and

Yang, Yang and Li (2020), where only one exposure variable is involved. We consider more generally numerous exposure variables in each of the regression models and different covariance structures of variables. Only KCSVS and CSIS can handle the multiple exposure variables in (b), and KCSVS is further investigated with the very high-dimensional control vector in (c). Also note that for Model 9, methods other than KCSVS and CSIS are not applicable to the categorical response.

The results are summarized in Table 3. For Model 7, all methods except for KCSVS fail to discover the hidden variable X_5. The same observation holds in Model 8 when conditioning on U_1 and in Model 9. Moreover, the competitors barely identify X_4 in Models 7 and 8 because X_4 is weakly correlated with the response due to the model design. Given all the important exposure variables, the performance of KCSVS further improves and surpasses CSIS. In addition, KCSVS demonstrates competitive ability to disentangle predictors from interaction or nonlinear terms in Model 10. This simulation example indicates that KCSVS is an effective screening procedure for varying coefficient model containing multiple or even high-dimensional exposure variables.

Computation efficiency is an important issue for screening procedures. Although KCSVS relies on two utility measures and calculating the conditional measures involves ultrahigh-dimensional vectors, its computation complexity can be optimized to $O(n^2p)$, which in theory is comparable to single-measure screening procedures that adopt RKHS-based indexes such as DCOR, HSIC and CDCOR (see Appendix B for more details). In our simulation studies, KCSVS is in fact shown to be more efficient than CDCSIS, which is implemented using the *cdcsis* package in R. Moreover, KCSVS as the only applicable method for very high-dimensional controls remains equally efficient for a wide range of p_0. A summary of the computational time for different methods is presented in Appendix B.

In summary, there are several important takeaways from the above simulation studies.

- Only KCSVS among all methods can capture hidden important predictors that are conditionally independent with the response variable given the control set. When the model contains no such predictors, KCSVS still successes in distinguishing marginally important predictors from irrelevant ones.
- KCSVS can handle control vectors of very high dimension, and utilize important control variables (or even unimportant ones that are correlated with the response) to enhance screening. It also demonstrates better resilience to inaccurate prior information than the other methods.
- KCSVS effectively and efficiently adapts to a variety of models including linear/nonlinear regression, classification, and varying coefficient models.

TABLE 3
Quantiles of MMS M_τ's and average selection proportions P_j^s's and P_a's for models in Example 3 based on 200 replicates.

Model	p_0	s	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_2^s	P_3^s	P_4^s	P_5^s	P_a
7(a)	1	5	NIS	9819.0	10000.0	10000.0	10000.0	10000.0	1.000	1.000	0.990	0.000	0.000	0.000
			CCSIS	9836.9	9997.8	10000.0	10000.0	10000.0	1.000	1.000	0.995	0.000	0.000	0.000
			CSIS	9701.7	9990.0	10000.0	10000.0	10000.0	1.000	1.000	0.995	0.000	0.000	0.000
			CDCSIS	9991.9	10000.0	10000.0	10000.0	10000.0	1.000	1.000	0.995	0.000	0.000	0.000
			BKRSIS	9930.8	10000.0	10000.0	10000.0	10000.0	1.000	1.000	0.995	0.000	0.000	0.000
			KUS	9012.8	9958.2	9996.0	10000.0	10000.0	1.000	1.000	0.970	0.000	0.000	0.000
			KCSVS	5.0	5.0	5.0	8.2	101.3	1.000	1.000	0.945	0.990	0.980	0.925
7(b)	4	5	CSIS	9988.9	10000.0	10000.0	10000.0	10000.0	1.000	1.000	0.995	0.000	0.000	0.000
			KCSVS	5.0	5.0	5.0	8.2	101.3	1.000	1.000	0.945	0.990	0.980	0.925
7(c)	2000	5	KCSVS	5.0	5.0	5.0	8.0	93.2	1.000	1.000	0.945	0.990	0.985	0.925
8(a)	1	5	NIS	676.7	4032.0	6870.0	9429.2	9995.0	1.000	1.000	0.995	0.105	0.030	0.010
			CCSIS	3752.6	8099.8	9680.0	9963.0	10000.0	1.000	1.000	1.000	0.030	0.000	0.000
			CSIS	446.8	3747.0	7059.5	9374.2	9973.1	1.000	1.000	1.000	0.130	0.050	0.025
			CDCSIS	598.2	4229.0	8031.0	9737.0	9998.0	1.000	1.000	1.000	0.095	0.035	0.020
			BKRSIS	1639.1	6072.8	8938.0	9858.0	9997.0	1.000	1.000	1.000	0.050	0.010	0.005
			KUS	9743.4	9978.0	9999.0	10000.0	10000.0	1.000	1.000	0.990	0.000	0.000	0.000
			KCSVS	5.0	5.0	5.0	6.0	15.2	1.000	1.000	0.985	1.000	0.995	0.980
8(b)	4	5	CSIS	5.0	5.0	9.5	70.2	1068.5	1.000	1.000	1.000	0.895	0.850	0.780
			KCSVS	5.0	5.0	5.0	6.0	15.2	1.000	1.000	0.985	1.000	0.995	0.980
8(c)	2000	5	KCSVS	5.0	5.0	5.0	6.0	15.2	1.000	1.000	0.985	1.000	0.995	0.980
9(a)	1	5	CSIS	9896.4	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	1.000	0.000	0.000
			KUS	9245.4	9976.8	9998.5	10000.0	10000.0	1.000	0.985	0.985	0.985	0.000	0.000
			KCSVS	5.0	5.0	5.0	8.0	72.8	1.000	0.980	0.980	0.975	1.000	0.935
9(b)	4	5	CSIS	9930.9	10000.0	10000.0	10000.0	10000.0	1.000	1.000	1.000	1.000	0.000	0.000
			KCSVS	5.0	5.0	5.0	8.0	72.8	1.000	0.980	0.980	0.975	1.000	0.935
9(c)	2000	5	KCSVS	5.0	5.0	5.0	8.0	72.8	1.000	0.980	0.980	0.975	1.000	0.935
Model	p_0	s	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_2^s	P_{12}^s	P_{22}^s	P_{33}^s	P_a
10(a)	1	5	NIS	26.0	64.8	162.5	459.0	1225.9	0.595	0.640	1.000	1.000	0.865	0.380
			CCSIS	5.0	8.8	16.0	42.0	271.6	0.940	0.950	1.000	1.000	0.980	0.880
			CSIS	357.7	1741.2	4023.0	5996.5	9104.6	0.165	0.210	0.990	0.445	0.215	0.015
			CDCSIS	1237.2	4141.5	6939.0	8715.8	9779.5	0.065	0.090	1.000	0.825	0.085	0.000
			BKRSIS	493.7	2140.2	3784.5	6319.2	9164.5	0.145	0.210	1.000	0.795	0.155	0.015
			KUS	5.0	5.0	5.0	6.0	15.0	1.000	0.995	1.000	1.000	1.000	0.995
			KCSVS	5.0	5.0	5.0	6.0	15.0	0.995	0.995	1.000	1.000	1.000	0.990
10(b)	4	5	CSIS	1123.2	4476.5	6677.0	8727.2	9703.8	0.070	0.080	1.000	0.820	0.085	0.000
			KCSVS	5.0	5.0	5.0	6.0	15.0	0.995	0.995	1.000	1.000	1.000	0.990
10(c)	2000	5	KCSVS	5.0	5.0	5.0	6.0	15.0	0.995	0.995	1.000	1.000	1.000	0.990

3.2. Diffuse large-B-cell lymphoma data

Diffuse large-B-cell lymphoma (DLBCL) is the most common type of blood cancer and demonstrates genetic and biological heterogeneous. DLBCL molecular subtypes include germinal center B-cell-like (GCB), activated B-cell-like (ABC) and Type-III (unclassified), among which ABC is a more acute subtype associated with far worse survival prognosis than GCB. Classifying DLBCL is therefore a critical step towards developing personalized therapies for DLBCL patients. DLBCL subtypes are typically determined by hierarchical clustering based on the similarity of DLBCL gene expression to activated peripheral blood B cells or normal germinal center B-cells (Alizadeh et al., 2000). It would be interesting to identify a batch of core genes that characterizes such an classification (closely related to survival). For example, the B-cell lymphoma 6 (BCL6)

gene is known to be frequently translocated and hypermutated in DLBCL and associated with GCB subtype as a marker of germinal center differentiation (Dalla-Favera et al., 1999; Dunleavy and Wilson, 2014). Aside from genetic information, some clinical measurements can be easily acquired to assist in the classification. International Prognostic Index (IPI) is a widely-used scoring system for the prognosis of DLBCL after chemotherapy on the basis of five clinical characteristics. Rosenwald et al. (2002) and Li (2006) suggest utilizing both IPI and genetic information to stratify patients for therapeutic trials and enhance survival prediction.

Rosenwald et al. (2002) applied a hierarchical clustering algorithm to group 240 lymphoma samples collected from DLBCL patients into the three subtypes based on 100 cDNA expressions (including BCL6). Data is available at `https://llmpp.nih.gov/DLBCL/`. Our goal is to discover influential genes for DLBCL classification through KCSVS, in which IPI and the expression of BCL6 are treated as prior information. In total, 219 cases with positive survival time and the IPI risk recorded are included for our analysis. We add 1,901 independently and normally distributed noise variables to make up a total of 2,000 variables subject to the screening conditional on IPI and BCL6. The dataset is split into a training cohort of 147 samples and a validation cohort of 72 samples following Rosenwald et al. (2002). We first conduct KCSVS to locate $d = 2[72/\log(72)] = 34$ genes, followed by model-free dimension reduction on the selected genes and the two control variables via sliced inverse regression (SIR; Li, 1991), and then perform linear discriminant analysis (LDA) using the first two SIR directions. The performance of the fitted model is evaluated using the testing samples. Kaplan-Meier curves for the overall survival of the predicted subtypes are presented in Fig. 1. In particular, the survival curves for predicted GCB and ABC are compared using the log-rank test. For both training and testing cohorts, the ABC subgroup identified by the KCSVS+SIR+LDA model is indeed associated with worse prognosis than the GCB subgroup (p-values < 0.001).

To further examine the classification accuracy of the KCSVS+SIR+LDA model, we repeat the above modeling procedure 100 times, but in each trial we randomly split the data at the ratio of 147:72 for training versus testing. The model is benchmarked against CSIS followed by penalized LDA (PenLDA). We display the distribution of the misclassification rate in Fig. 2, which suggests that KCSVS+SIR+LDA makes better predictions of the subtypes than CSIS+PenLDA. On average KCSVS captures 18 "true" genes (i.e., genes that were used to determine the subtypes in the original dataset) out of the 34 selected genes. In summary, KCSVS detects influential genes that lead to not only better discrimination of the DLBCL subtypes but also biologically sensible findings. Among the three genes that are selected in all replicates by KCSVS, CD10 is a known marker for GCB (Dunleavy and Wilson, 2014) while CCND2 is often up-regulated in ABC (Blenk et al., 2007). In fact, since gene expression profiling analysis is not practical in clinical laboratory, immunohistochemistry algorithms based on tissue microarray have been proposed to predict DLBCL subtypes, and the most commonly used algorithm in routine practice was made up of three markers including CD10 and BCL6 (Hans et al., 2004).

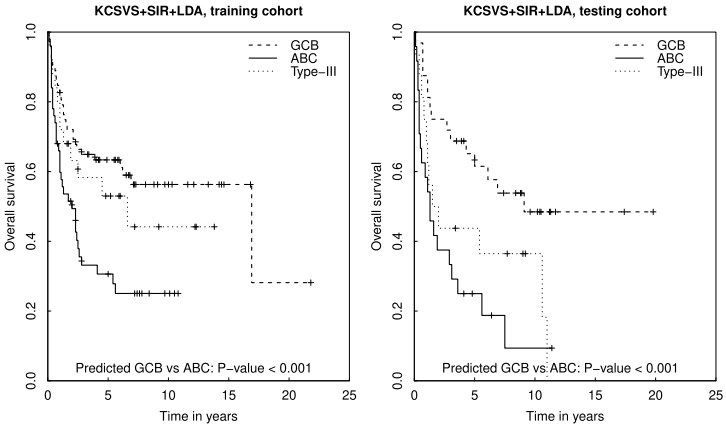

FIG 1. *Kaplan–Meier curves of the overall survival for DLBCL patients in the training cohort (left panel) and the testing cohort (right panel) assigned to the three subtypes by the KCSVS+SIR+LDA model. The survival curves for the predicted GCB and ABC subgroups are compared using the log-rank test.*

4. Discussion

In this paper, we develop a model-free conditional sufficient variable screening framework that allows for high-dimensional control variables and applies to either continuous or categorical responses. The framework is built upon the RKHS theory to avoid model specification while preserving nice interpretation from the perspective of regression. We proposed kernel-based R^2 and partial R^2 as nonlinear generalizations of the typical R^2 statistics to quantify the marginal contribution of a variable as well as its conditional contribution given the control set and the other variables. Dimension-free deviation bounds are found for the kernel R^2 statistics, which is the key to achieve conditional sufficient variable screening for ultrahigh-dimensional data. The sure screening property and the rank consistency property of the proposed procedure are established. The major advantage of KCSVS over existing competitors is that KCSVS can discover hidden important predictors that are individually independent but jointly dependent with the response variable conditioning on arbitrarily many control variables. As demonstrated numerically, KCSVS is a powerful screening tool that copes well with assorted regression or classification models, and varying quality of prior information. We conjecture that control variables that are carefully chosen by domain experts would compound the eminence of KCSVS for particular applications in practice.

There are two byproducts of the proposed procedure. In the absence of $\mathbf{W}$, the conditional utility measure becomes $w_j^C = R^2_{\widetilde{K}}(\mathbf{X}_{-j}|Y; X_j)$ and the proposed procedure performs kernel-based unconditional sufficient variable screening (KSVS). Yang, Yin and Zhang (2019) and Yuan et al. (2022) also developed sufficient variable screening procedures that evaluate each predictor

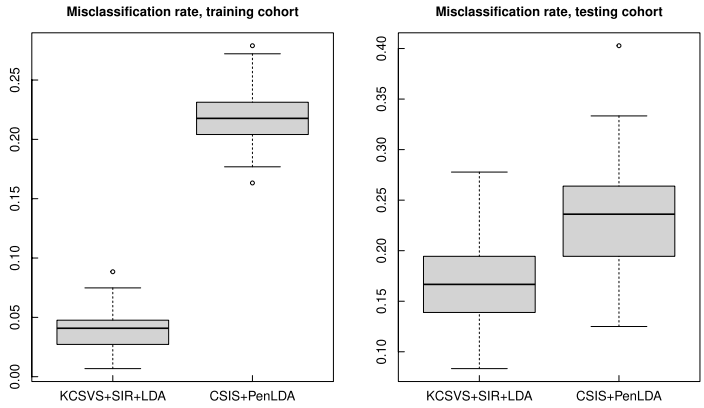

FIG 2. *Boxplots of the training (left panel) and the testing (right panel) classification errors for the two competing models applied to the DLBCL data over 100 random partitions.*

marginally and conditionally. To identify important predictors that make no marginal contribution, they consider $X_j \perp\!\!\!\perp (\mathbf{X}_{-j}, Y)$ for continuous responses and $X_j \perp\!\!\!\perp \mathbf{X}_{-j}|Y$ for categorical responses. However, since the relation between X_j and $\mathbf{X}_{-j}$ dominates the two conditions due to the ultrahigh dimensionality of $\mathbf{X}_{-j}$, an important predictor may not survive if it does not have a strong correlation with the rest variables. In contrast, KSVS relies on condition $(b2)$ $\mathbf{X}_{-j} \perp\!\!\!\perp Y|X_j$, which is known to be most violated by a marginally silent predictor, and the correlation between X_j and $\mathbf{X}_{-j}$ is adjusted for in kernel partial R^2. As a result, KSVS is more powerful in terms of discovering important predictors that are marginally independent of the response. Theoretical properties and simulation studies for KSVS are provided in Appendix D. The proposed procedure can also be modified easily to accommodate cases when it is believed that $\mathcal{A} = \mathcal{A}_1$, as mentioned in Section 2.1. The conditional utility measure is replaced by $w_j^{C^*} = R^2_{\widetilde{K}}(\mathbf{W}|Y; X_j)$ in the procedure to recover $\mathcal{A}_1$ more precisely. We refer to this adapted procedure as kernel-based univariate conditional screening (KUCS) as candidate variables enter the reproducing kernel regression model that already contains $\mathbf{W}$ one at a time. Although KUSC shares the same target on $\mathcal{A}_1$ with most existing conditional screening methods, it gets rid of the dimension limitation of $\mathbf{W}$. It can be shown that KUCS achieves both sure screening and rank consistency with regard to $\mathcal{A}_1$, which is deferred to Appendix C along with some numerical results due to limited space.

Our procedure is developed using kernel R^2 statistics, but other appropriate utility measures can also be implanted to the framework. For example, the marginal utility measure can be replaced by $R^2_K(Y|X_j)$ (the kernel R^2 for regression $Y|X_j$), DCOR, HSIC, martingale difference divergence (Shao and Zhang, 2014), or projection correlation (Liu et al., 2022), to name a few, and the conditional utility measure can be substituted with CDCOR. Admittedly, this paper

poses some open (though less critical) questions besides what it solves. Recall that since the screening procedure aims to achieve sufficiency as opposed to minimal sufficiency, we simply treat the marginal and the conditional utility measures as equally important and select a generous number of variables. On the one hand, adjusting the relative weight of $\widehat{\mathcal{A}}_M^*$ and $\widehat{\mathcal{A}}_C^*$ may improve the performance of KCSVS, which would require a data-driven tuning process. On the other hand, one may consider incorporating a strategy to control false discoveries. False discovery rate (FDR) control often involves creating a set of synthetic predictors that are designed to mimic the statistical properties of the original predictors, but are not directly related to the response. By comparing the effects of the original predictors with those of the synthetic predictors, we can distinguish truly important predictors from those that are unimportant or merely spurious correlated. FDR control through knockoff features (Barber and Candès, 2015, 2019) is one such technique developed for high-dimensional variable selection and has been applied to ultrahigh-dimensional unconditional screening by Liu et al. (2022). However, the construction of conditional knockoff features given high-dimensional controls remains a challenging task. These aforementioned topics will be pursued in future research to round out the proposed framework.

Appendix A: Main proofs

A.1. Proof of Theorem 1

The proof is composed of four parts and each part justifies one of the inequalities. The following lemma is vital to the proof of Theorem 1.

Lemma 1 (Deviation bound for U-statistics, Hoeffding, 1963)**.** *Let* $g(\mathbf{U}_1, \ldots, \mathbf{U}_r)$ *be a kernel of a U-statistic* U_n*, i.e.,* $U_n := \frac{1}{(n)_r} \sum_{i_r^n} g(\mathbf{U}_{i_1}, \ldots, \mathbf{U}_{i_r})$*, where* $n > r$*,* $(n)_r := \frac{n!}{(n-r)!}$ *and* $\sum_{i_r^n}$ *is taken over all r-tuples* $\{i_1, \ldots, i_r\}$ *drawn without replacement from* $\{1, \ldots, n\}$*. If* $b_1 \leq g(\mathbf{U}_1, \ldots, \mathbf{U}_r) \leq b_2$*, then for any* $\epsilon > 0$*, the following bound holds:*

$$P\{|U_n - EU_n| \geq \epsilon\} \leq 2\exp\{-2w\epsilon^2/(b_2 - b_1)^2\},$$

where $w := [n/r]$*, the largest integer contained in* n/r*.*

This lemma finds a uniform bound for any U-statistic of arbitrary dimensional data, as long as the associated kernel is bounded. We repeatedly use this result throughout the proofs of the four inequalities.

(I) If Y is categorical, then under conditions (C1) and (C5), for any $\epsilon \in (0, 1)$,

$$P\{|\widehat{w}_j^M - w_j^M| \geq \epsilon\} \leq 2L\exp\left\{-\frac{a_1 n}{L^3}\epsilon^2\right\},$$

where $j = 1, \ldots, p$, and $a_1 > 0$ is a constant depending on c_0.

Proof. We aim to show the uniform consistency of the denominator and the numerator of $\widehat{w}_j^M$ under regularity conditions respectively. Because the denominator of $\widehat{w}_j^M$ has a similar form as the numerator, we deal with its numerator only below. Let

$$\begin{aligned}\widehat{\mathcal{H}} :=& SSR_{K,n}(X_j|Y)\\ =& \sum_{l=1}^{L}\frac{n_l}{n}\frac{1}{n_l^2}\sum_{i_1,i_2=1}^{n_l}K(X_{i_1,j}^{(l)},X_{i_2,j}^{(l)})-\frac{1}{n^2}\sum_{i_1,i_2=1}^{n}K(X_{i_1,j},X_{i_2,j})\\ :=& \sum_{l=1}^{L}\hat{P}_l V_{n_l}^{(l)}-V_n^{(0)},\end{aligned}$$

where $V_{n_l}^{(l)}$ $(l=0,\ldots,L)$ are V-statistics. Let $U_{n_l}^{(l)}$ $(l=0,\ldots,L)$ be corresponding U-statistics with $E_l := EU_{n_l}^l$ $(l=0,\ldots,L)$. Under condition (C1), without loss of generality, we assume that the kernel K is bounded above by 1. Hence, $0\le E_l\le 1$ for $l=0,\ldots,L$. Denote $\mathcal{H} := SSR_K(X_j|Y)=\sum_{l=1}^{L}P_lE_l-E_0$. For any $\epsilon\in(0,1)$,

$$\begin{aligned}&P\left\{|\widehat{\mathcal{H}}-\mathcal{H}|\ge\epsilon\right\}\\ =&P\left\{\left|\sum_{l=1}^{L}\hat{P}_l\left(V_{n_l}^{(l)}-E_l\right)+\sum_{l=1}^{L}\left(\hat{P}_l-P_l\right)E_l-\left(V_n^{(0)}-E_0\right)\right|\ge\epsilon\right\}\\ \le&P\left\{\sum_{l=1}^{L}\hat{P}_l\left|V_{n_l}^{(l)}-E_l\right|\ge\frac{\epsilon}{3}\right\}+P\left\{\sum_{l=1}^{L}\left|\hat{P}_l-P_l\right|E_l\ge\frac{\epsilon}{3}\right\}\\ &+P\left\{\left|V_n^{(0)}-E_0\right|\ge\frac{\epsilon}{3}\right\}\\ :=&T_1+T_2+T_3.\end{aligned}$$

Let us consider T_1 first.

$$\begin{aligned}T_1\le&P\left\{L\max_l\hat{P}_l\left|V_{n_l}^{(l)}-E_l\right|\ge\frac{\epsilon}{3}\right\}\\ \le&P\left\{\max_l\left|V_{n_l}^{(l)}-E_l\right|\ge\frac{\epsilon}{3L},\ \min_l\hat{P}_l\ge\frac{c_0}{L}\right\}+P\left\{\min_l\hat{P}_l<\frac{c_0}{L}\right\}\\ \le&P\left\{\max_l\left|V_{n_l}^{(l)}-E_l\right|\ge\frac{\epsilon}{3L},\ \min_l n_l\ge\frac{c_0 n}{L}\right\}\\ &+P\left\{\max_l\left|\hat{P}_l-P_l\right|\ge\frac{c_0}{L}\right\}\\ \le&\sum_{l=1}^{L}P\left\{\left|V_{n_l}^{(l)}-E_l\right|\ge\frac{\epsilon}{3L},n_l\ge\frac{c_0 n}{L}\right\}+\sum_{l=1}^{L}P\left\{\left|\hat{P}_l-P_l\right|\ge\frac{c_0}{L}\right\}\\ :=&\sum_{l=1}^{L}T_{11}^{(l)}+\sum_{l=1}^{L}T_{12}^{(l)},\end{aligned}$$

where the third inequality holds because $\max_l \left|\hat{P}_l - P_l\right| \geq P_l - \hat{P}_l \geq 2c_0/L - c_0/L = c_0/L$ by condition (C5).

$$\begin{aligned}
T_{11}^{(l)} &= P\left\{\left|\frac{n_l-1}{n_l}U_{n_l}^{(l)} + \frac{1}{n_l^2}\sum_{i=1}^{n_l} K(X_{i,j}^{(l)}, X_{i,j}^{(l)}) - E_l\right| \geq \frac{\epsilon}{3L}, n_l \geq \frac{c_0 n}{L}\right\} \\
&\leq P\left\{\frac{n_l-1}{n_l}\left|U_{n_l}^{(l)} - E_l\right| + \left|\frac{1}{n_l^2}\sum_{i=1}^{n_l} K(X_{i,j}^{(l)}, X_{i,j}^{(l)}) - \frac{1}{n_l}E_l\right| \geq \frac{\epsilon}{3L}, \right. \\
&\qquad \left. n_l \geq \frac{c_0 n}{L}\right\} \\
&\leq P\left\{\left|U_n^{(l)} - E_l\right| \geq \frac{\epsilon}{3L} - \frac{2}{n_l}, n_l \geq \frac{c_0 n}{L}\right\} \\
&\leq P\left\{\left|U_n^{(l)} - E_l\right| \geq \frac{\epsilon}{6L}, n_l \geq \frac{c_0 n}{L}\right\}, \text{ for } n \text{ large enough} \\
&\leq 2\exp\left\{-\frac{c_0 n\epsilon^2}{36L^3}\right\},
\end{aligned}$$

where the last inequality follows from Lemma 1. Also,

$$T_{12}^{(l)} \leq 2\exp\left\{-\frac{2c_0^2 n}{L^2}\right\}.$$

Similarly, we can show $T_2 \leq P\left\{\max_l \left|\hat{P}_l - P_l\right| \geq \frac{\epsilon}{3L}\right\} \leq 2L\exp\left\{-\frac{2n\epsilon^2}{9L^2}\right\}$ and $T_3 \leq 2\exp\left\{-\frac{n\epsilon^2}{36}\right\}$. Combining T_1, T_2 and T_3, we have

$$P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \geq \epsilon\right\} \leq 2L\exp\left\{-\frac{a_1 n\epsilon^2}{L^3}\right\},$$

where a_1 is a positive constant depending on c_0. □

(II) If Y is continuous, then under conditions (C1)–(C3) and (C6), for any $\epsilon > 0$,

$$P\{|\widehat{w}_j^M - w_j^M| \geq \epsilon\} \leq 2n\exp\left\{-a_3 n\epsilon^2\right\},$$

where $j = 1, \ldots, p$, and $a_3 > 0$ is a constant.

Proof. For a given $j \in \{1, \ldots, p\}$, let $\gamma(y) := E(d_{1234}|Y_1 = y, Y_2 = y)$, where $d_{1234} = K(X_{1j}, X_{2j}) - K(X_{3j}, X_{4j})$, then $\mathcal{H} := \mathcal{H}_K^2(X_j|Y) = E\gamma(Y)$. The kernel regression estimator

$$\begin{aligned}
\widehat{\mathcal{H}} &:= SSR_{K,G,n}(X_j|Y) \\
&= \frac{1}{n^5}\sum_{t_1,\ldots,t_5=1}^{n} \frac{G_{t_1t_2}G_{t_1t_3}d_{t_2t_3t_4t_5}}{\hat{f}_Y^2(y_{t1})} \\
&= \frac{1}{n}\sum_{t_1=1}^{n} \frac{f_Y^2(y_{t_1})}{\hat{f}_Y^2(y_{t_1})}\widehat{\gamma}(y_{t_1}),
\end{aligned}$$

where $f_Y(\cdot)$ is the density function of Y, $\hat{f}_Y(y_{t_1}) := \frac{1}{n}\sum_{s=1}^n G_{t_1 s}$ and

$$\widehat{\gamma}(y_{t_1}) := \frac{1}{n^4}\sum_{t_2,t_3,t_4,t_5=1}^{n} \frac{G_{t_1t_2}G_{t_1t_3}d_{t_2t_3t_4t_5}}{f_Y^2(y_{t_1})}.$$

Without loss of generality, we assume that $f_Y(y)$ is bounded below by some $c > 0$ by condition (C6). We first show some intermediate results.

(R1) $P\{|\hat{f}_Y(y_t) - f_Y(y_t)| \geq \epsilon\} \leq 2\exp\{-n\epsilon^2/2\}$.

Note that $\hat{f}_Y(y_t) = \frac{1}{nh}G(0) + \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{s\neq t} G_{ts}\right)$ and $\frac{1}{nh}G(0) = o(1)$ by conditions (C2) and (C3). Denote $U_{n-1} := \frac{1}{n-1}\sum_{s\neq t} G_{ts}$. Then

$$\begin{aligned} EU_{n-1} &= \int h^{-1}G(\frac{y_t - y}{h})f_Y(y)dy \\ &= \int G(u)f_Y(y_t + hu)du = f_Y(y_t) + O(h^2) \end{aligned}$$

by Taylor expansion and conditions (C2) and (C6), and $O(h^2) = o(1)$ by conditions (C2) and (C3). Hence,

$$\begin{aligned} &P\left\{\left|\hat{f}_Y(y_t) - f_Y(y_t)\right| \geq \epsilon\right\} \\ =& P\left\{\left|\frac{n-1}{n}U_{n-1} - EU_{n-1} + o(1)\right| \geq \epsilon\right\} \\ =& P\left\{\left|\frac{n-1}{n}(U_{n-1} - EU_{n-1}) + o(1)\right| \geq \epsilon\right\} \\ \leq& P\left\{|U_{n-1} - EU_{n-1}| \geq \frac{\epsilon}{2}\right\} \text{ for } n \text{ sufficiently large} \\ \leq& 2\exp\left\{-\frac{n\epsilon^2}{2}\right\} \text{ by Lemma 1.} \end{aligned}$$

(R2) $P\{|\frac{1}{n}\sum_{t_1=1}^n \widehat{\gamma}(y_{t_1}) - \mathcal{H}| \geq \epsilon\} \leq 2\exp\{-n\epsilon^2/8\}$.

Denote the corresponding U-statistic of $\frac{1}{n}\sum_{t_1=1}^n \widehat{\gamma}(y_{t_1})$ as $\widetilde{\mathcal{H}}$, that is,

$$\widetilde{\mathcal{H}} := C_n^5 \sum_{t_1<\cdots<t_5} \frac{1}{5!}\sum_{\pi} g_{i_1i_2i_3i_4i_5},$$

where $g_{12345} := G_{12}G_{13}d_{2345}/f_Y^2(y_1)$ and $\sum_\pi$ represents summation over the 5! permutations of $(i_1,\ldots,i_5)$ of $(t_1,\ldots,t_5)$. Under conditions (C2) and (C3), $\frac{1}{n}\sum_{t_1=1}^n \widehat{\gamma}(y_{t_1}) = \widetilde{\mathcal{H}} + o(1)$. We will show in the next that $E\widetilde{\mathcal{H}} = \mathcal{H} + o(1)$ in two parts. Firstly,

$$\begin{aligned} \Gamma_1 := \int& h^{-2}G(\frac{y_1 - y_2}{h})G(\frac{y_1 - y_3}{h})K(x_2, x_3) \\ & f_Y^{-1}(y_1)f_{X_jY}(x_2, y_2)f_{X_jY}(x_3, y_3)dx_2dx_3dy_1dy_2dy_3 \end{aligned}$$

$$
\begin{aligned}
&= \int K(x_2, x_3) f_{X_j|Y}(x_2|y_1 + hu) f_{X_j|Y}(x_3|y_1 + hv) dx_2 dx_3 \\
&\quad G(u)G(v) f_Y(y_1 + hu) f_Y(y_1 + hv) du dv f_Y^{-1}(y_1) dy_1 \\
&= \int K(x_2, x_3) f_{X_j|Y}(x_2|y_1) f_{X_j|Y}(x_3|y_1) dx_2 dx_3 f_Y(y_1) dy_1 \\
&\quad + O_p(h^2)
\end{aligned}
$$

by Taylor expansion and conditions (C2) and (C6). Similarly, we can show

$$
\begin{aligned}
\Gamma_2 := &\int K(x_4, x_5) f_{X_j}(x_4) f_{X_j}(x_5) dx_4 dx_5 \\
&h^{-2} G(\frac{y_1 - y_2}{h}) G(\frac{y_1 - y_3}{h}) f_Y^{-1}(y_1) f_Y(y_2) f_Y(y_3) dy_1 dy_2 dy_3 \\
= &\int K(x_4, x_5) f_{X_j}(x_4) f_{X_j}(x_5) dx_4 dx_5 f_Y(y_1) dy_1 + O_p(h^2).
\end{aligned}
$$

Therefore, $E\widetilde{\mathcal{H}} = \Gamma_1 + \Gamma_2 = \mathcal{H} + o(1)$. Then

$$
\begin{aligned}
&P\left\{\left|\frac{1}{n}\sum_{t_1=1}^{n} \widehat{\gamma}(y_{t_1}) - \mathcal{H}\right| \geq \epsilon\right\} \\
=&P\left\{\left|\widetilde{\mathcal{H}} - E\widetilde{\mathcal{H}} + o(1)\right| \geq \epsilon\right\} \\
\leq&P\left\{\left|\widetilde{\mathcal{H}} - E\widetilde{\mathcal{H}}\right| \geq \frac{\epsilon}{2}\right\} \text{ for } n \text{ sufficiently large} \\
\leq&2\exp\left\{-\frac{n\epsilon^2}{8}\right\} \text{ by Lemma 1.}
\end{aligned}
$$

Now, for arbitrary $\epsilon \in (0, 1)$,

$$
\begin{aligned}
&P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \geq \epsilon\right\} \\
\leq&P\left\{\left|\frac{1}{n}\sum_{t_1=1}^{n} \frac{f_Y^2(y_{t_1})}{\hat{f}_Y^2(y_{t_1})}\widehat{\gamma}(y_{t_1}) - \mathcal{H}\right| \geq \epsilon\right\} \\
\leq&P\left\{\left|\frac{1}{n}\sum_{t_1=1}^{n} \widehat{\gamma}(y_{t_1}) - \mathcal{H} + \frac{1}{n}\sum_{t_1=1}^{n}\left(\frac{f_Y^2(y_{t_1})}{\hat{f}_Y^2(y_{t_1})} - 1\right)\widehat{\gamma}(y_{t_1})\right| \geq \epsilon\right\} \\
\leq&P\left\{\left|\frac{1}{n}\sum_{t_1=1}^{n} \widehat{\gamma}(y_{t_1}) - \mathcal{H}\right| \geq \frac{\epsilon}{2}\right\} \\
&+ P\left\{\left|\frac{1}{n}\sum_{t_1=1}^{n}\left(\frac{f_Y^2(\mathbf{y}_{t_1})}{\hat{f}_Y^2(y_{t_1})} - 1\right)\widehat{\gamma}(y_{t_1})\right| \geq \frac{\epsilon}{2}\right\} \\
:=&T_1 + T_2.
\end{aligned}
$$

By (R2), $T_1 \le 2\exp\{-n\epsilon^2/640\}$. Moreover,

$$\begin{aligned}
T_2 \le & P\left\{\max_{t_1}\left|\left(\frac{f_Y^2(y_{t_1})}{\hat{f}_Y^2(y_{t_1})}-1\right)\widehat{\gamma}(y_{t_1})\right| \ge \frac{\epsilon}{2}\right\} \\
\le & P\left\{\max_{t_1}\left|\left(\frac{f_Y^2(y_{t_1})}{\hat{f}_Y^2(y_{t_1})}-1\right)\widehat{\gamma}(y_{t_1})\right| \ge \frac{\epsilon}{2}, \min_{t_1}\hat{f}_Y(y_t) \ge \frac{c}{2}\right\} \\
& + P\left\{\min_{t_1}\hat{f}_Y(y_t) < \frac{c}{2}\right\} \\
\le & P\left\{\max_{t_1}\left|\widehat{\gamma}(y_{t_1})[f_Y^2(y_{t_1}) - \hat{f}_Y^2(y_{t_1})]\right| \ge \frac{c^2\epsilon}{8}\right\} \\
& + P\left\{\max_{t_1}\left|f_Y(y_t) - \hat{f}_Y(y_t)\right| \ge \frac{c\epsilon}{2}\right\} \\
:= & T_{21} + T_{22}.
\end{aligned}$$

By (R1), $T_{22} \le 2n\exp\{-c^2n\epsilon^2/2\}$. Let $\widehat{m}(y_{t_1}) := \widehat{\gamma}(y_{t_1})[f_Y^2(y_{t_1}) - \hat{f}_Y^2(y_{t_1})]$ and $\widehat{m}^U(y_{t_1})$ be the corresponding U-statistic. Similar to (R2), we can show that

$$\widehat{m}(y_{t_1}) = \widehat{m}^U(y_{t_1}) + o(1) \text{ and } E_{t_2t_3t_4t_5}\widehat{m}^U(y_{t_1}) = O(h^2).$$

Hence, for n sufficiently large,

$$\begin{aligned}
T_{21} &\le P\left\{\max_{t_1}\left|\widehat{m}^U(y_{t_1}) - E_{t_2t_3t_4t_5}\widehat{m}^U(y_{t_1})\right| \ge \frac{c^2\epsilon}{16}\right\} \\
&\le 2n\exp\left\{-\frac{c^4n\epsilon^2}{512}\right\}.
\end{aligned}$$

Finally, we have

$$P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \ge \epsilon\right\} \le 2n\exp\left\{-a_3n\epsilon^2\right\},$$

where a_3 is a constant depending on c. □

(III) If Y is discrete, then under conditions (C1)–(C5), for any $\epsilon \in (0,1)$,

$$P\{|\widehat{w}_j^C - w_j^C| \ge \epsilon\} \le 2Ln\exp\left\{-\frac{a_2n\epsilon^2}{L^3}\right\},$$

where $j = 1, \ldots, p$, and $a_2 > 0$ is a constant depending on c_0.

Proof. Let $\mathbf{Z}_j = (\mathbf{X}_{-j}, \mathbf{W})$, $j = 1, \ldots, p$. Note that

$$\begin{aligned}
\mathcal{H} :=& SSR_K(\mathbf{Z}_j|Y; X_j) \\
=& SSR_K(\mathbf{Z}_j|X_j; Y) + SSR_K(\mathbf{Z}_j|Y) - SSR_K(\mathbf{Z}_j|X_j) \\
:=& \mathcal{H}_1 + \mathcal{H}_2 + \mathcal{H}_3.
\end{aligned}$$

Denote $\mathcal{H}_1^{(l)} := SSR_K(\mathbf{Z}_j^{(l)}|X_j^{(l)})$ as the within-group SSR given $Y = y^{(l)}$, then $\mathcal{H}_1 = \sum_{l=1}^L P_l \mathcal{H}_1^{(l)}$. Let

$$\begin{aligned}\widehat{\mathcal{H}}_1 :=& SSR_{K,n}(\mathbf{Z}_j|X_j;Y)\\ =&\sum_{l=1}^L \frac{n_l}{n} SSR_{K,G,n}(\mathbf{Z}_j^{(l)}|X_j^{(l)}) := \sum_{l=1}^L \hat{P}_l \widehat{\mathcal{H}}_1^{(l)}.\end{aligned}$$

Then,

$$\begin{aligned}&P\left\{|\widehat{\mathcal{H}}_1 - \mathcal{H}_1| \geq \epsilon\right\}\\ =&P\left\{\left|\sum_{l=1}^L \hat{P}_l \widehat{\mathcal{H}}_1^{(l)} - \sum_{l=1}^L P_l \mathcal{H}_1^{(l)}\right| \geq \epsilon\right\}\\ \leq&P\left\{\sum_{l=1}^L \hat{P}_l \left|\widehat{\mathcal{H}}_1^{(l)} - \mathcal{H}_1^{(l)}\right| \geq \frac{\epsilon}{2}\right\} + P\left\{\sum_{l=1}^L \left|\hat{P}_l - P_l\right| \mathcal{H}_1^{(l)} \geq \frac{\epsilon}{2}\right\}\\ \leq&P\left\{\max_l \left|\widehat{\mathcal{H}}_1^{(l)} - \mathcal{H}_1^{(l)}\right| \geq \frac{\epsilon}{2L}\right\} + P\left\{\max_l \left|\hat{P}_l - P_l\right| \geq \frac{\epsilon}{4L}\right\}\\ \leq&P\left\{\max_l \left|\widehat{\mathcal{H}}_1^{(l)} - \mathcal{H}_1^{(l)}\right| \geq \frac{\epsilon}{2L}, \min_l \hat{P}_l \geq \frac{c_0}{L}\right\} + P\left\{\min_l \hat{P}_l < \frac{c_0}{L}\right\}\\ &+ P\left\{\max_l \left|\hat{P}_l - P_l\right| \geq \frac{\epsilon}{4L}\right\}\\ \leq&\sum_{l=1}^L P\left\{\left|\widehat{\mathcal{H}}_1^{(l)} - \mathcal{H}_1^{(l)}\right| \geq \frac{\epsilon}{2L}, n_l \geq \frac{c_0 n}{L}\right\} + P\left\{\min_l \hat{P}_l < \frac{c_0}{L}\right\}\\ &+ \sum_{l=1}^L P\left\{\left|\hat{P}_l - P_l\right| \geq \frac{\epsilon}{4L}\right\}\\ \leq&2L \max_l n_l \exp\left\{-\frac{a c_0 n\epsilon^2}{L^3}\right\} + 2L\exp\left\{-\frac{2c_0^2 n\epsilon^2}{L^2}\right\} + 2L\exp\left\{-\frac{n\epsilon^2}{8L^2}\right\}\\ &\text{by (II) and Lemma 1 for some constants } a > 0\\ \leq&2Ln\exp\left\{-\frac{\tilde{a}_2 n\epsilon^2}{L^3}\right\} \text{ for some constant } a_2 > 0 \text{ depending on } c_0.\end{aligned}$$

Similar to (I), it can be shown that

$$P\left\{|\widehat{\mathcal{H}}_2 - \mathcal{H}_2| \geq \epsilon\right\} \leq 2L\exp\left\{-\frac{\tilde{a}_1 n\epsilon^2}{L^3}\right\}$$

for some $\tilde{a}_1 > 0$ depending on c_0. Similar to (II),

$$P\left\{|\widehat{\mathcal{H}}_3 - \mathcal{H}_3| \geq \epsilon\right\} \leq 2n\exp\left\{-\tilde{a}_3 n\epsilon^2\right\}$$

for some $\tilde{a}_3 > 0$. Therefore,

$$P\left\{|\widehat{\mathcal{H}} - \mathcal{H}| \geq \epsilon\right\} \leq \sum_{i=1}^{3} P\left\{|\widehat{\mathcal{H}}_i - \mathcal{H}_i| \geq \epsilon/3\right\} \leq 2Ln \exp\left\{-\frac{a_2 n\epsilon^2}{L^3}\right\},$$

for some $a_2 > 0$ depending on c_0. □

(IV) If Y is continuous, then under conditions (C1)–(C4) and (C6), for any $\epsilon \in (0, 1)$,

$$P\{|\widehat{w}_j^C - w_j^C| > \epsilon\} \leq 2n \exp\left\{-a_4 n\epsilon^2\right\},$$

where $j = 1, \ldots, p$, and $a_4 > 0$ is a constant.

Proof. The proof is analogous to (II) since

$$SSR_{K,\widetilde{G},G,n}(\mathbf{Z}_j|Y; X_j) = SSR_{K,\widetilde{G},n}(\mathbf{Z}_j|(Y, X_j)) - SSR_{K,G,n}(\mathbf{Z}_j|X_j).$$ □

A.2. Proof of Theorem 2

Proof. Let $\mathcal{A}_M := \{j \in \mathcal{A} : X_j \not\perp\!\!\!\perp \mathbf{Y}\}$ with $|\mathcal{A}_M| = s_1$ and $\mathcal{A}_C := \mathcal{A} \setminus \mathcal{A}_M$ with $|\mathcal{A}_C| = s_2$. If Y is categorical, following from Theorem 1,

$$\begin{aligned} P\left\{\max_{j\in\mathcal{A}_M} |\widehat{w}_j^M - w_j^M| > c_1 n^{-\gamma_1}\right\} &\leq 2s_1 L \exp\left\{-\frac{a_1 c_1^2 n^{1-2\gamma_1}}{L^3}\right\} \\ &\leq O\left(s_1 \exp\left\{-b_1 n^{1-2\gamma_1-3\kappa} + \kappa \log n\right\}\right), \end{aligned}$$

where $b_1 > 0$ is a constant depending on c_0 and c_1. Similarly, by Theorem 1,

$$\begin{aligned} P\left\{\max_{j\in\mathcal{A}_C} |\widehat{w}_j^C - w_j^C| > c_2 n^{-\gamma_2}\right\} &\leq 2s_2 Ln \exp\left\{-\frac{a_2 c_2^2 n^{1-2\gamma_2}}{L^3}\right\} \\ &\leq O\left(s_2 \exp\left\{-b_2 n^{1-2\gamma_2-3\kappa} + (1+\kappa)\log n\right\}\right), \end{aligned}$$

for some constant $b_2 > 0$ depending on c_0 and c_2. Under condition (C7), if $\mathcal{A} \nsubseteq \widehat{\mathcal{A}}$, there must exist some $j \in \mathcal{A}_M$ such that $w_j^M \geq 2c_1 n^{-\gamma_1}$ but $\widehat{w}_j^M < c_1 n^{-\gamma_1}$, or some $j \in \mathcal{A}_C$ such that $w_j^C \geq 2c_2 n^{-\gamma_2}$ but $\widehat{w}_j^C < c_2 n^{-\gamma_2}$. Therefore,

$$\begin{aligned} P\left\{\mathcal{A} \nsubseteq \widehat{\mathcal{A}}\right\} \leq & P\left\{|\widehat{w}_j^M - w_j^M| > c_1 n^{-\gamma_1} \text{ for some } j \in \mathcal{A}_M\right\} \\ & + P\left\{|\widehat{w}_j^C - w_j^C| > c_2 n^{-\gamma_2} \text{ for some } j \in \mathcal{A}_C\right\} \\ \leq & P\left\{\max_{j\in\mathcal{A}_M} |\widehat{w}_j^M - w_j^M| > c_1 n^{-\gamma_1}\right\} \\ & + P\left\{\max_{j\in\mathcal{A}_C} |\widehat{w}_j^C - w_j^C| > c_2 n^{-\gamma_2}\right\} \\ \leq & O\left(s_1 \exp\left\{-b_1 n^{1-2\gamma_1-3\kappa} + \kappa \log n\right\}\right) \\ & + O\left(s_2 \exp\left\{-b_2 n^{1-2\gamma_2-3\kappa} + (1+\kappa)\log n\right\}\right) \\ \leq & O\left(s \exp\left\{-b n^{1-2\gamma-3\kappa} + (1+\kappa)\log n\right\}\right), \end{aligned}$$

where b is a constant depending on c_0, c_1 and c_2, and $\gamma = \max\{\gamma_1, \gamma_2\}$. In other words,

$$P\left\{\mathcal{A} \subseteq \widehat{\mathcal{A}}\right\} \geq 1 - O\left(s \exp\left\{-bn^{1-2\gamma-3\kappa} + (1+\kappa)\log n\right\}\right).$$

The proof for continuous Y is analogous. □

A.3. Proof of Theorem 3

Proof. If Y is categorical, by condition (C8) and Theorem 1,

$$\begin{aligned}
&P\left\{\left(\min_{j\in\mathcal{A}_M} \widehat{w}_j^M - \max_{j\notin\mathcal{A}_M} \widehat{w}_j^M\right) < c_3 n^{-\gamma_3}\right\} \\
\leq &P\left\{\left(\min_{j\in\mathcal{A}_M} \widehat{w}_j^M - \max_{j\notin\mathcal{A}_M} \widehat{w}_j^M\right) - \left(\min_{j\in\mathcal{A}_M} w_j^M - \max_{j\notin\mathcal{A}_M} w_j^M\right) < -c_3 n^{-\gamma_3}\right\} \\
\leq &P\left\{\left|\left(\min_{j\in\mathcal{A}_M} \widehat{w}_j^M - \max_{j\notin\mathcal{A}_M} \widehat{w}_j^M\right) - \left(\min_{j\in\mathcal{A}_M} w_j^M - \max_{j\notin\mathcal{A}_M} w_j^M\right)\right| > c_3 n^{-\gamma_3}\right\} \\
\leq &P\left\{\max_j \left|\widehat{w}_j^M - w_j^M\right| > c_3 n^{-\gamma_3}/2\right\} \\
\leq &2pL\exp\left\{-\frac{a_5 n^{1-2\gamma_3}}{L^3}\right\}
\end{aligned}$$

for some $a_5 > 0$ depending on c_3. Since $\log(p) = o(n^{1-2\gamma_3-3\kappa})$ and $L = O(n^\kappa)$, we have $\log(p) \leq a_5 n^{1-2\gamma_3}/(2L^3)$, $a_5 n^{1-2\gamma_3}/(2L^3) \geq 3\log(n)$, $\log(L) \leq \log(n)$ for n sufficiently large. For some n_0 sufficiently large,

$$\sum_{n=n_0}^{+\infty} pL\exp\{-a_5 n^{1-2\gamma_3}/L^3\} \leq \sum_{n=n_0}^{+\infty} n^{-2} < +\infty.$$

By Borel-Cantelli Lemma,

$$\liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_M} \widehat{w}_j^M - \max_{j\notin\mathcal{A}_M} \widehat{w}_j^M\right\} \geq c_3 n^{-\gamma_3} > 0$$

a.s. Similarly, by condition (C8) and Theorem 1,

$$P\left\{(\min_{j\in\mathcal{A}_C} \widehat{w}_j^C - \max_{j\notin\mathcal{A}_C} \widehat{w}_j^C) < c_4 n^{-\gamma_4}\right\} \leq 2npL\exp\left\{-\frac{a_6 n^{1-2\gamma_4}}{L^3}\right\},$$

for some $a_6 > 0$ depending on c_4. Since $\log(p) = o(n^{1-2\gamma_4-3\kappa})$ and $L = O(n^\kappa)$, we can derive similarly that

$$\liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_C} \widehat{w}_j^C - \max_{j\notin\mathcal{A}_C} \widehat{w}_j^C\right\} \geq c_4 n^{-\gamma_4} > 0$$

a.s. The proof for continuous Y is analogous. □

Appendix B: Computational issues

Computing efficiency is an important factor for variable screening procedures. In this section, we discuss the computational aspect of the proposed procedure.

Let $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{Z}$ be three random vectors on $\mathbb{R}^{p_1}$, $\mathbb{R}^{p_2}$ and $\mathbb{R}^{p_3}$, respectively. Since kernel learning typically requires computing and handling an $n \times n$ Gram matrix, the computational complexity for $R_n^2(\mathbf{U}|V)$ is $O(n^2(p_1 + p_2))$ if $\mathbf{V}$ is continuous or $O(n^2 p_1)$ if $\mathbf{V}$ is categorical, and the computational complexity for $R_n^2(\mathbf{U}|V;\mathbf{Z})$ is $O(n^2 p_1(p_2 + p_3))$ if $\mathbf{V}$ is continuous or $O(n^2(p_1 + p_3))$ if $\mathbf{V}$ is categorical. Low-rank approximation to the Gram matrix via incomplete Cholesky decomposition can be used to improve efficiency and the resulting computational complexity depends on the decaying spectrum of the kernel function (Bach and Jordan, 2002). In the univariate case with a Gaussian kernel, the decay is exponential, reducing the computational complexity to $O(n \log n)$.

As for the utility measures, the computational complexity is $O(n^2)$ for marginal utility and $O(n^2 p)$ for conditional utility per predictor due to the computation of the Gram matrix for X_j and for ultrahigh-dimensional vector $(\mathbf{X}_{-j}, \mathbf{W})$. A naive implementation of the screening procedure would therefore scale as $O(n^2 p^2)$, which can be a serious liability in applications to large data sets. However, the computation of the conditional utility measure can be easily reduced to $O(n^2)$. For example, if a Gaussian kernel is used, the distance matrix of $(\mathbf{X}, \mathbf{W})$ can be computed in $O(n^2 p)$, based on which the Gram matrix for $(\mathbf{X}_{-j}, \mathbf{W})$ can be obtained within $O(n^2)$. As a result, the computation complexity of KCSVS can be optimized to $O(n^2 p)$. In theory this is comparable to single-measure screening procedures that adopt similar reproducing-kernel-based indexes such as DCOR, HSIC and CDCOR, despite that KCSVS relies on two utility measures and calculating the conditional measures involves ultrahigh-dimensional vectors. Again, incomplete Cholesky decomposition may help further improve the efficiency.

We also compare KCSVS with other methods in terms of computational cost in the simulation studies. The average computing times are summarized in Table 4 based on 100 replicates in R on a laptop with i5 1.4 GHz processor and 16G RAM. We only report the results for the three submodels (c)–(e) of Model 2 (regression) and Model 3 (classification) because the computing time does not vary significantly between different models for given n, p and p_0. All the existing methods only compute one marginal measure for each predictor, yet KCSVS is shown to be more efficient than CDCSIS, a conditional screening procedure based on CDCOR that can be implemented using the *cdcsis* package in R. Note that methods that appear to be very fast such as NIS, CCSIS and CSIS are model-based procedures. Moreover, KCSVS can handle high-dimensional controls and remain equally efficient for a wide range of p_0.

Appendix C: Kernel-based univariate conditional screening

When it is believed that $\mathcal{A} = \mathcal{A}_1$ or when the main interest lies in recovering $\mathcal{A}_1$, we should evaluate the marginal importance of each predictor after adjusting for

TABLE 4
Average computing time (in seconds) based on 100 replicates. A cell displays a dash if the corresponding method is not applicable to the specific model.

Model	p_0	n	p	KCSVS	NIS	CCSIS	CSIS	CDCSIS	BKRSIS
2(c)	1	100	5000	21.42	1.76	0.41	4.54	47.50	9.03
			10000	41.34	3.61	0.80	9.09	95.22	17.77
		300	5000	262.67	2.37	3.22	4.79	1469.66	55.98
			10000	530.14	4.70	6.42	9.43	2892.08	112.08
2(d)	2	100	5000	20.34	–	–	4.68	48.35	9.27
			10000	41.89	–	–	9.57	95.78	18.63
		300	5000	261.12	–	–	4.91	1484.34	56.34
			10000	522.08	–	–	9.86	2941.17	111.85
2(e)	2000	100	5000	20.19	–	–	–	–	–
			10000	41.70	–	–	–	–	–
		300	5000	261.13	–	–	–	–	–
			10000	523.76	–	–	–	–	–
3(c)	1	100	5000	22.35	–	–	6.93	–	–
			10000	45.83	–	–	13.79	–	–
		300	5000	189.78	–	–	12.47	–	–
			10000	386.68	–	–	24.27	–	–
3(d)	2	100	5000	22.38	–	–	8.04	–	–
			10000	46.80	–	–	16.06	–	–
		300	5000	189.85	–	–	15.29	–	–
			10000	381.38	–	–	29.79	–	–
3(e)	2000	100	5000	22.60	–	–	–	–	–
			10000	45.35	–	–	–	–	–
		300	5000	190.97	–	–	–	–	–
			10000	382.07	–	–	–	–	–

the effect of the control variables. Leaving out $\mathbf{X}_{-j}$ in Proposition 1, conditions $(b1)$ $X_j \perp\!\!\!\perp Y$ and $(b2')$ $\mathbf{W} \perp\!\!\!\perp Y|X_j$ jointly imply condition (a') $X_j \perp\!\!\!\perp Y|\mathbf{W}$. Therefore, we turn to use $w_j^M = R_K^2(X_j|Y)$ and $w_j^{C^*} = R_{\widetilde{K}}^2(\mathbf{W}|Y; X_j)$ as the utility measures. Denote their estimators by $\widehat{w}_j^M = R_n^2(X_j|Y)$ and $\widehat{w}_j^{C^*} = R_n^2(\mathbf{W}|Y; X_j)$, respectively. Then $\mathcal{A}_1$ is estimated by

$$\widehat{\mathcal{A}}_1 = \left\{1 \le j \le p : \widehat{w}_j^M \ge c_1 n^{-\gamma_1} \text{ or } \widehat{w}_j^{C^*} \ge c_2^* n^{-\gamma_2^*}\right\},$$

where c_1, c_2^*, γ_1 and γ_2^* are pre-specified threshold values defined later. We refer to the above screening procedure as kernel-based univariate conditional screening (KUSC).

C.1. Theoretical properties

Following from Theorem 1, deviation bounds can be found for $\widehat{w}_j^{C^*}$ for different types of the response variable, regardless of the dimension of $\mathbf{W}$.

Corollary 1. *If Y is categorical, then under conditions (C1)–(C5),*

$$P\{|\widehat{w}_j^{C^*} - w_j^{C^*}| > \epsilon\} \le 2Ln \exp\left\{-\frac{a_2^* n\epsilon^2}{L^3}\right\},$$

for any $\epsilon > 0$, *where* $j = 1, \ldots, p$, *and* $a_2^* > 0$ *is some constant depending on* c_0. *If* Y *is continuous, then under conditions (C1)–(C4) and (C6),*

$$P\{|\widehat{w}_j^{C^*} - w_j^{C^*}| > \epsilon\} \le 2n \exp\left\{-a_4^* n\epsilon^2\right\}$$

for any $\epsilon > 0$, *where* $j = 1, \ldots, p$, *and* $a_4^* > 0$ *is some constant.*

With the above appealing property, it can be shown that KUSC asymptotically almost surely select $\mathcal{A}_1$. Let $\mathcal{A}_M = \{j \in \mathcal{A} : X_j \not\perp\!\!\!\perp \mathbf{Y}\}$, $\mathcal{A}_{C^*} = \mathcal{A}_1 \setminus \mathcal{A}_M$. The following condition is required to establish the sure screening property.

(C7*) There exist $c_1, c_2^* > 0$ and $\gamma_1, \gamma_2^* \in [0, 1/2)$, such that

$$\min_{j \in \mathcal{A}_M} w_j^M \ge 2c_1 n^{-\gamma_1} \text{ and } \min_{j \in \mathcal{A}_{C^*}} w_j^{C^*} \ge 2c_2^* n^{-\gamma_2^*}.$$

Corollary 2 (Sure Screening)**.** *If* Y *is categorical, then under conditions (C1)–(C5) and (C7*), we have*

$$P\left(\mathcal{A}_1 \subset \widehat{\mathcal{A}}_1\right) \ge 1 - O\left(s_1 \exp\left\{-b^* n^{1-2\gamma^*-3\kappa} + (1+\kappa)\log n\right\}\right)$$

for $\kappa \in [0, \frac{1}{3} - \frac{2\gamma^*}{3})$, *where* s_1 *is the cardinality of* $\mathcal{A}_1$, b^* *is a positive constant depending on* c_0, c_1 *and* c_2^*, *and* $\gamma^* = \max\{\gamma_1, \gamma_2^*\}$. *If* Y *is continuous, then under conditions (C1)–(C4), (C6) and (C7*), we have*

$$P\left(\mathcal{A}_1 \subset \widehat{\mathcal{A}}_1\right) \ge 1 - O\left(s_1 \exp\left\{-\tilde{b}^* n^{1-2\gamma^*} + \log n\right\}\right),$$

where s_1 *is the cardinality of* $\mathcal{A}_1$, $\tilde{b}^*$ *is a positive constant depending on* c_1 *and* c_2^*, *and* $\gamma^* = \max\{\gamma_1, \gamma_2^*\}$.

Moreover, the procedure ranks important predictors above irrelevant ones with high probability if a stronger condition is assumed as follows.

(C8*) There exist $c_3, c_4^* > 0$ and $\gamma_3, \gamma_4^* \in [0, 1/2)$, such that

$$\min_{j \in \mathcal{A}_M} w_j^M - \max_{j \notin \mathcal{A}_M} w_j^M \ge 2c_3 n^{-\gamma_3} \text{ and } \min_{j \in \mathcal{A}_{C^*}} w_j^{C^*} - \max_{j \notin \mathcal{A}_{C^*}} w_j^{C^*} \ge 2c_4^* n^{-\gamma_4^*}.$$

Corollary 3 (Rank Consistency)**.** *Let* $\tilde{\gamma}^* = \max\{\gamma_3, \gamma_4^*\}$. *If* $\mathbf{Y}$ *is categorical and* $\log p = o(n^{1-2\tilde{\gamma}^*-3\kappa})$ *for* $\kappa \in [0, \frac{1}{3} - \frac{2\tilde{\gamma}^*}{3})$, *then under conditions (C1)–(C5) and (C8*),*

$$\liminf_{n\to\infty} \left\{ \min_{j \in \mathcal{A}_M} \widehat{w}_j^M - \max_{j \notin \mathcal{A}_M} \widehat{w}_j^M \right\} > 0$$

$$\text{and } \liminf_{n\to\infty} \left\{ \min_{j \in \mathcal{A}_{C^*}} \widehat{w}_j^{C^*} - \max_{j \notin \mathcal{A}_{C^*}} \widehat{w}_j^{C^*} \right\} > 0$$

almost surely. If $\mathbf{Y}$ *is continuous and* $\log p = o(n^{1-2\tilde{\gamma}^*})$, *then under conditions (C1)–(C4), (C6) and (C8*),*

$$\liminf_{n\to\infty} \left\{ \min_{j \in \mathcal{A}_M} \widehat{w}_j^M - \max_{j \notin \mathcal{A}_M} \widehat{w}_j^M \right\} > 0$$

$$\textit{and } \liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_{C^*}}\widehat{w}_j^{C^*}-\max_{j\notin\mathcal{A}_{C^*}}\widehat{w}_j^{C^*}\right\}>0$$

almost surely.

In other words, the true marginal and conditional signal can be well separated from noise by some thresholds. Such thresholds, however, could be difficult to determined in practice. We can instead choose a relatively generous model size d and select $\widehat{\mathcal{A}}_1^* := \widehat{\mathcal{A}}_M^* \cup \widehat{\mathcal{A}}_{C^*}^*$, where

$$\widehat{\mathcal{A}}_M^*(d_1) := \{1 \le j \le p : \widehat{w}_j^M \text{ is among the first } d_1 \text{ largest of all}\}$$
$$\text{and } \widehat{\mathcal{A}}_{C^*}^*(d_2) := \{j \notin \widehat{\mathcal{A}}_M^*(d_1) : \widehat{w}_j^{C^*} \text{ is among the first } d_2 \text{ largest of all}\},$$

for $d_1+d_2 = d$. The sure screening property guarantees the coverage of $\mathcal{A}_1$ when d is large.

C.2. Numerical studies

We examine the performance of KUSC through the same 10 models considered for KCSVS and the results are reported in Table 5 with $n = 200$, $p = 2,000$ and $d_1 = d_2 = [n/\log n] = 38$. As expected, KUSC selects all important variables in $\mathcal{A}_1$ with high probability and results in small MMS (close to the true size) when $\mathcal{A} = \mathcal{A}_1$. In general, KUSC performs similarly to KCSVS except for that KUSC cannot detect important variables beyond $\mathcal{A}_1$. Compared with existing alternatives that also aim to recover $\mathcal{A}_1$, KUSC allows multiple or even high-dimensional control variables. Notice that unimportant control variables that are correlated with the response help improve the performance of KUSC in some models, which suggests that KUSC can utilize indirect or even inaccurate prior information to better select important predictors.

Appendix D: Kernel-based sufficient variable screening

In the absence of $\mathbf{W}$, or equivalently when $\mathbf{W}$ is empty, KCSVS performs kernel-based unconditional sufficient variable screening (KSVS) with the conditional utility measure being $w_j^C = R^2_{\widetilde{K}}(\mathbf{X}_{-j}|Y; X_j)$. The sure screening property and the rank consistency property naturally hold by Theorems 2 and 3. The most important merit of sufficient variable screening is the ability to select important predictors that are marginally independent with the response. There are different paths towards achieving sufficient variable screening for ultrahigh-dimensional data. In this section, we elaborate the advantage of KSVS over existing competitors. Yang, Yin and Zhang (2019) and Yuan et al. (2022) developed sufficient variable screening procedures for continuous and categorical responses, respectively. The following proposition summarizes the sufficient conditions adopted by KSVS and the two aforementioned methods for identifying redundant variables.

TABLE 5

Quantiles of MMS M_τ's and selection proportions P_j^s's and P_a's for KUSC across all models based on 200 replicates. A cell for P_j^s displays a dash if the corresponding variable X_j is assigned to the control set and thus protected in screening.

Model	p_0	s	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_2^s	P_3^s	P_4^s	P_5^s	P_a
1(a)	1	4	227.8	1223.2	1734.5	1940.8	1998.0	0.990	0.995	0.995	0.010	–	0.010
1(b)	2000	4	5.0	6.0	6.0	12.0	44.0	0.996	0.995	0.995	1.000	–	0.980
1(c)	1	3	3.0	4.0	4.0	6.0	32.3	–	0.995	0.995	1.000	–	0.990
1(d)	2	2	2.0	2.0	2.0	2.0	10.0	–	–	0.990	1.000	–	0.990
1(e)	2000	2	2.0	2.0	2.0	2.0	10.0	–	–	0.990	1.000	–	0.990
2(a)	1	4	65.9	368.8	936.5	1536.0	1933.8	0.990	1.000	0.990	0.070	–	0.070
2(b)	2000	4	5.0	6.0	6.0	8.0	30.3	0.985	1.000	0.990	1.000	–	0.975
2(c)	1	3	4.0	4.0	9.5	25.2	139.4	–	1.000	0.990	0.915	–	0.905
2(d)	2	2	2.0	2.0	2.0	2.0	14.2	–	–	0.985	1.000	–	0.985
2(e)	2000	2	2.0	2.0	2.0	2.0	14.2	–	–	0.985	1.000	–	0.985
3(a)	1	4	303.3	1043.5	1508.5	1829.5	1991.0	0.970	1.000	0.990	0.020	–	0.020
3(b)	2000	4	5.0	5.0	9.0	29.0	233.3	0.980	1.000	0.990	0.935	–	0.905
3(c)	1	3	16.0	66.8	313.0	873.2	1695.4	–	1.000	0.985	0.275	–	0.275
3(d)	2	2	2.0	2.0	2.0	7.0	48.2	–	–	0.980	1.000	–	0.980
3(e)	2000	2	2.0	2.0	2.0	9.0	50.2	–	–	0.975	1.000	–	0.975
4(a)	1	4	162.6	635.2	1045.0	1591.8	1869.3	0.965	0.950	0.965	0.015	–	0.015
4(b)	2000	4	4.0	5.0	8.0	27.0	114.4	0.960	0.945	0.960	0.990	–	0.865
4(c)	1	3	3.0	3.0	6.0	14.0	82.2	–	0.950	0.960	0.990	–	0.900
4(d)	2	2	2.0	2.0	2.0	5.0	35.4	–	–	0.960	1.000	–	0.960
4(e)	2000	2	2.0	2.0	2.0	6.0	47.4	–	–	0.960	0.990	–	0.950

Model	p_0	s	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_2^s	P_4^s	P_6^s	P_8^s	P_{10}^s	P_a
5(a)	5	5	9.0	10.0	12.0	30.2	320.7	0.995	1.000	1.000	1.000	0.880	0.875
5(b)	2000	5	9.0	10.0	12.0	30.0	296.5	0.995	1.000	1.000	1.000	0.885	0.880
6(a)	5	5	9.0	9.0	10.0	12.2	28.1	1.000	1.000	1.000	1.000	0.995	0.995
6(b)	2000	5	9.0	9.0	10.0	13.0	29.2	1.000	1.000	1.000	1.000	0.995	0.995

Model	p_0	s	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_3^s	P_5^s	P_7^s	P_9^s	P_a
5(c)	5	5	9.0	10.0	11.5	20.0	151.9	0.910	1.000	1.000	1.000	0.995	0.905
5(d)	2000	5	9.0	10.0	12.0	25.0	168.3	0.995	1.000	0.995	1.000	0.900	0.895
6(c)	5	5	9.0	9.0	10.0	12.0	26.4	1.000	1.000	1.000	1.000	1.000	1.000
6(d)	2000	5	9.0	9.0	10.0	12.0	28.0	1.000	1.000	1.000	1.000	1.000	1.000

Model	p_0	s	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_2^s	P_3^s	P_4^s	P_5^s	P_a
7(a)	1	5	624.0	1220.0	1610.5	1879.2	1991.0	1.000	1.000	0.975	0.060	0.045	0.000
7(b)	4	5	421.9	984.0	1473.5	1796.5	1991.0	1.000	1.000	0.955	0.075	0.055	0.000
7(c)	2000	5	6.0	6.0	8.0	18.0	105.9	1.000	1.000	0.955	0.990	0.980	0.925
8(a)	1	5	6.0	25.5	90.0	324.0	1025.1	1.000	1.000	0.960	0.655	0.640	0.455
8(b)	4	5	6.0	6.0	12.5	30.5	142.1	1.000	1.000	0.975	0.975	0.955	0.905
8(c)	2000	5	5.0	6.0	6.0	16.0	64.3	1.000	1.000	0.975	0.995	0.990	0.960
9(a)	1	5	97.6	635.5	1198.0	1706.8	1980.2	1.000	0.985	0.985	0.990	0.045	0.045
9(b)	4	5	88.8	442.8	1023.0	1472.5	1919.6	1.000	0.980	0.990	0.995	0.050	0.050
9(c)	2000	5	7.0	7.0	9.0	17.0	57.1	1.000	0.990	0.990	0.995	0.995	0.970

Model	p_0	s	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_2^s	P_{12}^s	P_{22}^s	P_{33}^s	P_a
10(a)	1	5	8.0	11.2	30.0	104.5	465.2	0.770	0.925	1.000	0.950	0.990	0.680
10(b)	4	5	9.0	10.0	11.0	25.5	135.6	0.965	0.975	1.000	0.980	0.985	0.910
10(c)	2000	5	9.0	10.0	12.0	26.0	150.5	0.965	0.975	1.000	0.985	0.985	0.915

Proposition 2. *Let* $\mathbf{X}_{-j}$ *denote the vector of all predictors excluding* X_j *(*$j = 1, \ldots, p$*), then*

1. $j \notin \mathcal{A}$ *if and only if condition*

$$(a)\ X_j \perp\!\!\!\perp Y | \mathbf{X}_{-j}$$

holds, $j = 1, \ldots, p$*; and*
2. *The following pair of conditions (b1) and (b2) implies condition (a):*

$$(b1)\ X_j \perp\!\!\!\perp Y;\ (b2)\ \mathbf{X}_{-j} \perp\!\!\!\perp Y | X_j.$$

3. *The pair of conditions (b1) and (b3) is equivalent to condition (b4), which implies condition (a):*

$$(b3)\ X_j \perp\!\!\!\perp \mathbf{X}_{-j} | Y;$$
$$(b4)\ X_j \perp\!\!\!\perp (Y, \mathbf{X}_{-j}).$$

Statements 1 and 2 follow immediately from Proposition 1. The pair of $(b1)$ and $(b3)$ is considered by Yang, Yin and Zhang (2019) for sliced continuous responses and by Yuan et al. (2022) for categorical response, in which case condition $(b3)$ can be more easily assessed than condition (a) given the conditional variable is categorical. Condition $(b4)$ is used in Yang, Yin and Zhang (2019) for continuous responses. Most existing unconditional screening methods only evaluate condition $(b1)$, which is not sufficient for condition (a). As a consequence, important predictors that are marginally independent but jointly correlated with the response could be falsely ruled out. Therefore, one should further check either one of conditions $(b2)$–$(b4)$ before removing a variable. However, we argue that a marginally silent predictor may not survive $(b3)$ or $(b4)$ if it does not have a stronger correlation with the rest variables compared to the other candidates. False discoveries of spurious variables are exacerbated in the meanwhile. The reason is that the relation between X_j and $\mathbf{X}_{-j}$ dominates the two conditions due to the ultrahigh dimensionality of $\mathbf{X}_{-j}$. In contrast, KSVS uses condition $(b2)$ and avoids such an issue. Taking a closer look at the conditional utility measure of KSVS,

$$R^2_{\tilde{K}}(\mathbf{X}_{-j} | Y; X_j) = \frac{SSR_{\tilde{K}}(\mathbf{X}_{-j} | Y; X_j)}{SSE_{\tilde{K}}(\mathbf{X}_{-j} | X_j)},$$

where

$$SSR_{\tilde{K}}(\mathbf{X}_{-j} | Y; X_j) = SSR_{\tilde{K}}(\mathbf{X}_{-j} | (Y, X_j)) - SSR_{\tilde{K}}(\mathbf{X}_{-j} | X_j)$$
$$\text{and } SSE_{\tilde{K}}(\mathbf{X}_{-j} | X_j) = SSTO_{\tilde{K}}(\mathbf{X}_{-j}) - SSR_{\tilde{K}}(\mathbf{X}_{-j} | X_j),$$

we notice that the correlation between X_j and $\mathbf{X}_{-j}$ is adjusted for in this kernel partial R^2, which solves the issue of the existing sufficient variable screening methods. In addition, another common issue of unconditional screening procedures is also alleviated. That is, important variables that are difficult to be

detected marginally because they are weakly correlated with the response (especially when spurious predictors that are highly correlated with the important predictors exist) may still be discovered through strong conditional signal.

To support the above argument with numerical evidence, we compare KSVS with DCSVS (Yang, Yin and Zhang, 2019) by revisiting Model 1 (linear) and Model 2 (heterogeneity) with $n = 200$, $p = 2,000$ and $d = 2[n/\log n] = 76$. The two methods are benchmarked against DCSIS (Li, Zhong and Zhu, 2012), a well-known marginal screening procedure based on DCOR, which is also the marginal measure adopted by Yang, Yin and Zhang (2019). The results are summarized in Table 6. Recall that for either model, $X_4 \perp\!\!\!\perp Y$ and X_5 is weakly correlated with Y although both variables are truly important. Moreover, unlike the other predictors, X_5 acts on the second-order moment of the response variable. Therefore, DCSIS as a marginal method fails to select X_4 all the time. As we can also observe from the table, KSVS outperforms DCSVS by large margins in terms of MMS quantiles and selection probabilities. In particular, KSVS is more powerful than DCSVS in detecting both X_4 and X_5.

TABLE 6

Quantiles of MMS M_τ's and selection proportions P_j^s's and P_a's for all Models 1 and 2 based on 200 replicates to compare KSVS, DCSVS and DCSIS. Note that DCSVS1 is based on the pair of conditions ($b1$) *and* ($b3$)*, while DCSVS2 is based on the pair of conditions* ($b1$) *and* ($b4$).

Model	s	Method	$M_{5\%}$	$M_{25\%}$	$M_{50\%}$	$M_{75\%}$	$M_{95\%}$	P_1^s	P_2^s	P_3^s	P_4^s	P_5^s	P_a
1	5	KSVS	6.0	8.0	16.0	45.2	223.4	0.990	0.995	0.995	1.000	0.860	0.845
1	5	DCSVS1	340.1	925.2	1436.0	1787.2	1988.0	1.000	1.000	1.000	0.000	0.615	0.000
1	5	DCSVS2	8.0	10.0	49.0	532.8	1716.1	1.000	1.000	1.000	0.945	0.615	0.575
1	5	DCSIS	1997.0	2000.0	2000.0	2000.0	2000.0	1.000	1.000	1.000	0.000	0.620	0.000
2	5	KSVS	6.0	6.0	8.0	22.0	220.3	0.985	1.000	0.990	1.000	0.925	0.900
2	5	DCSVS1	2000.0	2000.0	2000.0	2000.0	2000.0	0.995	1.000	1.000	0.000	0.000	0.000
2	5	DCSVS2	2000.0	2000.0	2000.0	2000.0	2000.0	0.995	1.000	1.000	0.930	0.000	0.000
2	5	DCSIS	2000.0	2000.0	2000.0	2000.0	2000.0	0.995	1.000	1.000	0.000	0.000	0.000

Acknowledgments

We sincerely thank the editor, the associate editor and two anonymous referees for their constructive comments, which led to a significant improvement of this article.

References

ALIZADEH, A. A., EISEN, M. B., DAVIS, R. E., MA, C., LOSSOS, I. S., ROSENWALD, A., BOLDRICK, J. C., SABET, H., TRAN, T., YU, X. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** 503–511.

BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research* **3** 1–48. MR1966051

Balasubramanian, K., Sriperumbudur, B. and Lebanon, G. (2013). Ultrahigh dimensional feature screening via RKHS embeddings. In *Artificial Intelligence and Statistics* 126–134.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085. MR3375876

Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics* **47** 2504–2537. MR3988764

Barut, E., Fan, J. and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association* **111** 1266–1277. MR3561948

Blenk, S., Engelmann, J., Weniger, M., Schultz, J., Dittrich, M., Rosenwald, A., Müller-Hermelink, H.-K., Müller, T. and Dandekar, T. (2007). Germinal center B cell-like (GCB) and activated B cell-like (ABC) type of diffuse large B cell lymphoma (DLBCL): analysis of molecular predictors, signatures, cell cycle state and patient survival. *Cancer Informatics* **3** 399–420.

Chen, X., Cook, R. D. and Zou, C. (2015). Diagnostic studies in sufficient dimension reduction. *Biometrika* **102** 545–558. MR3394274

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association* **86** 328–332. MR1137117

Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association* **110** 630–641. MR3367253

Dalla-Favera, R., Migliazza, A., Chang, C.-C., Niu, H., Pasqualucci, L., Butler, M., Shen, Q. and Cattoretti, G. (1999). Molecular pathogenesis of B cell malignancy: the role of BCL-6. In *Mechanisms of B Cell Neoplasia 1998* 257–265. Springer.

Dunleavy, K. and Wilson, W. H. (2014). Appropriate management of molecular subtypes of diffuse large B-cell lymphoma. *Oncology (Williston Park, NY)* **28** 326.

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** 849–911. MR2530322

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 101–148. MR2640659

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** 1270–1284. MR3265696

Fukumizu, K., Gretton, A., Lanckriet, G. R., Schölkopf, B. and Sriperumbudur, B. K. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1750–1758. Curran Associates, Inc. MR2825431

Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory* 63–77. Springer. MR2255909

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B. and Smola, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds.) **20** 585–592. MIT Press.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K. and Sriperumbudur, B. K. (2012). Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 1205–1213. Curran Associates, Inc.

Hans, C. P., Weisenburger, D. D., Greiner, T. C., Gascoyne, R. D., Delabie, J., Ott, G., Muller-Hermelink, H. K., Campo, E., Braziel, R. M., Jaffe, E. S. et al. (2004). Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood* **103** 275–282.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13–30. MR0144363

Ke, C. and Yin, X. (2020). Expected conditional characteristic function-based measures for testing independence. *Journal of the American Statistical Association* **115** 985–996. MR4107694

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327. MR1137117

Li, L. (2006). Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics* **22** 466–471.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102** 997–1008. MR2354409

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107** 1129–1139. MR3010900

Li, L., Zhu, L. and Zhu, L. (2011). Inference on the primary parameter of interest with the aid of dimension reduction estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73** 59–80. MR2797736

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** 266–274. MR3180562

Liu, W., Ke, Y., Liu, J. and Li, R. (2022). Model-free feature screening and FDR control with knockoff features. *Journal of the American Statistical Association* **117** 428–443. MR4399096

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100** 229–234. MR3034336

Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: a nonparametric model-free screening method. *The Annals of Statistics* **43** 1471–1497. MR3357868

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346** 1937–1947.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* 2263–2291. MR3127866

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* **48** 1514–1538. MR4124333

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association* **109** 1302–1318. MR3265698

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC Press. MR0848134

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769–2794. MR2382665

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association* **110** 1726–1734. MR3449068

Wen, C., Pan, W., Huang, M. and Wang, X. (2018). Sure independence screening adjusted for confounding covariates with ultrahigh dimensional data. *Statistica Sinica* **28** 293–317. MR3752262

Yang, G., Yang, S. and Li, R. (2020). Feature screening in ultrahigh dimensional generalized varying-coefficient models. *Statistica Sinica* **30** 1049–1067. MR4214173

Yang, B., Yin, X. and Zhang, N. (2019). Sufficient variable selection using independence measures for continuous response. *Journal of Multivariate Analysis* **173** 480–493. MR3948760

Yin, X. and Yuan, Q. (2020). A new class of measures for testing independence. *Statistica Sinica* **30** 2131–2154. MR4260758

Yuan, Q., Chen, X., Ke, C. and Yin, X. (2022). Independence index sufficient variable screening for categorical responses. *Computational Statistics & Data Analysis* **174** 107530. MR4432145

Zhou, Y., Liu, J. and Zhu, L. (2020). Test for conditional independence with application to conditional screening. *Journal of Multivariate Analysis* **175** 104557. MR4017960

Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106** 1464–1475. MR2896849