Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Category-Adaptive Variable Screening for Ultra-high Dimensional Heterogeneous Categorical Data

Jinhan Xie, Yuanyuan Lin, Xiaodong Yan & Niansheng Tang

To cite this article: Jinhan Xie, Yuanyuan Lin, Xiaodong Yan & Niansheng Tang (2019): Category-Adaptive Variable Screening for Ultra-high Dimensional Heterogeneous Categorical Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1573734

To link to this article: https://doi.org/10.1080/01621459.2019.1573734

Category-Adaptive Variable Screening for Ultra-high Dimensional Heterogeneous Categorical Data

Jinhan Xie[a] Yuanyuan Lin[b *] Xiaodong Yan[c] and Niansheng Tang[a]

[a] Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming 650091, P. R. of China

[b] Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China

[c] School of Economics, Shandong University, Jinan 250100, China

[*]Correspondence to: Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China (E-mail: ylin@sta.cuhk.edu.hk)

February 6, 2019

Abstract The populations of interest in modern studies are very often heterogeneous. The population heterogeneity, the qualitative nature of the outcome variable and the high dimensionality of the predictors pose significant challenge in statistical analysis. In this article, we introduce a category-adaptive screening procedure with high-dimensional heterogeneous data, which is to detect category-specific important covariates. The proposal is a model-free approach without any specification of a regression model and an adaptive procedure in the sense that the set of active variables is allowed to vary across different categories, thus making it more flexible to accommodate heterogeneity. For response-selective sampling data, another main discovery of this paper is that the proposed method works directly without any modification. Under mild regularity conditions, the newly procedure is shown to possess the sure screening and ranking consistency properties. Simulation studies contain supportive evidence that the proposed method performs well under various

settings and it is effective to extract category-specific information. Applications are illustrated with two real data sets.

Keywords: Categorical response; Ultrahigh dimensional data; Heterogeneity; Feature screening; Response-selective sampling

1 Introduction

Recent development in science and technology enables the growth of raw data to occur at an explosive rate. The high dimensional data in which the number of predictors is larger than the sample size pose unprecedent challenge for statistical analysis and numerical computation (Fan et al., 2009). Feature screening methods have been examined to be effective in dimension reduction to retain important variables with high probability. Fan and Lv (2008) proposed the sure independence screening (SIS) for linear regression by using marginal Pearson's correlation coefficient. Other popular advances on feature screening for the generalized linear model, the additive models and a general model framework can be found in Fan and Song (2010), Fan et al. (2011), Li et al. (2012), Chang et al. (2013), among many others. The aforementioned approaches are model-based methods. Recently, significant findings on model-free feature screening are reported in the literature, see, for example, Zhu et al. (2011), Li et al. (2012), He et al. (2013), among others. These screening methods are originally designed for continuous outcome variable.

High dimensional covariates coupled with categorical response are common in many fields of modern science, for example, economic, genomics, bioinformatics, social and biological studies, etc. In the analysis of high dimensional settings with categorical response, classical approaches such as the linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA) may not be directly applicable due to the singularity of the covariance matrix. To circumvent the problem, novel solutions have been proposed for classification problems with qualitative outcome variable. The Kolmogorov-Smirnov statistic (Mai and Zou,

2013) was developed for binary classification problems and was further extended to handle continuous response in Mai and Zou (2015). Huang et al. (2014) studied a feature screening procedure based on the Pearson's Chi-Square statistics (PC-SIS) when all covariates are categorical. Cui et al. (2015) advocated a marginal feature screening procedure for ultrahigh dimensional discriminant analysis based on the empirical conditional distribution function. Pan et al. (2016) proposed a pairwise sure screening for linear discriminant analysis with ultrahigh dimensional data. To the best of our knowledge, all these existing methodologies are clearly motivated and have examined to be effective for dimension reduction with prospective samples or i.i.d. samples from the underlying population.

However, when it comes to imbalanced design in which one of the classes/categories is rather rare, for instance, data for diagnosing rare diseases in clinical studies, the response-selective sampling would be useful in terms of its effectiveness and its saving in costs and time (Chen, 2001; Chen et al., 2017). The response-selective sampling is popular and widely used in biomedical, epidemiological, social studies, etc. There is a rich literature on response-selective or response-biased sampling, see, for example, Prentice and Pyke (1979), Scott and Wild (1986, 1997), Manski (1993), Lawless et al. (1999), Chen (2001a, 2001b), Tsai (2009), Luo and Tsai (2009), Luo et al. (2009), Lin (2000), Ning et al. (2010), Huang and Qin (2011), Kim et al. (2013), Kim et al. (2016), Xu et al. (2017), Qin (2017), Sun et al. (2018), among others. Generally speaking, most existing screening methods cannot be directly used to handle high dimensional categorical data subject to response-selective sampling, as the zero mean property of the ranking statistic under statistical independence of the response and a given predictor no longer holds. In addition, as an important feature of big data, variety or heterogeneity is ubiquitous. Typically, medical data of the same disease can be collected from various sources. With the availability of enormous high dimensional data in various disciplines, it is common that the sets of important predictors relevant to different categorical levels of the

response can be rather different. In this regard, directly applying the aforementioned methods without accounting for category-specific information may lead to inaccurate results. Hence, detecting important predictors of a heterogeneous categorical dataset is a crucial step for developing category-targeted strategies.

The aim of this paper is to develop a category-adaptive screening approach for analyzing ultrahigh dimensional heterogeneous data. By defining dummy variables pertaining to each categorical level respectively, one appealing feature of the newly procedure is that it is able to provide a complete picture of the heterogeneous nature of the categorical response given predictors. Moreover, it's a model-free approach without specifying a regression model of the explanatory variables and the response variable, which allows us to discover the nonlinear relationship between the response and predictors. Its corresponding marginal utility may be easily computed without any numerical optimization. We show the sure screening and ranking consistency properties of our proposed procedure under regularity conditions. Perhaps more importantly, another major finding of this paper is that our proposed procedure can be directly applied to the response-biased sampling data without any modification.

The rest of this paper is organized as follows. In Section 2, we present our main results and the theoretical properties of the proposed method. We evaluate the performance of the proposed procedure through extensive simulation studies. Real data examples are presented in Section 4. A few concluding remarks and discussions are given in Section 5. All technical details are deferred to Appendix. A supplementary material contains some additional simulation studies and theoretical properties of an extension of the proposed method.

2 Main results and methodologies

2.1 Category-adaptive variable screening with prospective samples

We consider the problem of feature screening in high dimensional settings, where we observe a categorical response variable Y with R ($R > 2$) classes $\{y_1, y_2, \ldots, y_R\}$, $p_r = \Pr(Y = y_r) > 0$ for all $r = 1, \ldots, R$ and $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is a p-vector of continuous covariates with a support $\mathbb{R}^p$. In big data era, data of the same categorical variable may be collected from different sources during different time periods, or under distinct experimental methods. In other words, the high dimensional data with categorical response is usually heterogeneous. To explore the heterogeneous nature and provide a complete picture of the conditional distribution of the outcome variable given the predictor vector, we assume that

(i) (Heterogeneity) the set of important predictors $\mathcal{A}_r \equiv \{1 \le j \le p : \Pr(W_r \le w_r \mid \boldsymbol{X})$ functionally depends on $X_j\}$, where $W_r \equiv I(Y = y_r)$, may be different for different $r = 1, \ldots, R$;

(ii) (Sparsity) the dimensionality $p = o\{\exp(n^\alpha)\}$ for some constant $\alpha > 0$, but $|\mathcal{A}_r| = s_r = o(n)$ where $|\mathcal{A}_r|$ is the cardinality of $\mathcal{A}_r$ and n is the sample size.

In practical situations, it is usually hard to specify a proper statistical model for ultrahigh dimensional data. In this paper, we impose nearly no requirement on the actual model structure except for the sparsity assumption in part (ii). In addition, we also denote $\mathcal{I}_r = \{1, \ldots, p\} \setminus \mathcal{A}_r$ as the inactive set pertaining to W_r, $r = 1, \ldots, R$. If $j \in \mathcal{A}_r$, X_j is referred as an active predictor related to W_r, a dummy variable relevant to the rth category. In this general framework, our goal is to pursue feature screening by investigating the dependence between $I(Y = y_r)$ for given $r = 1, \ldots, R$ and each X_j, $j = 1, \ldots, p$. To this end, we consider $F_{j,r}(x) = \Pr(X_j \le x \mid Y = y_r)$, the conditional distribution function of X_j given $Y = y_r$, and propose the following marginal screening utility

$$\tau_{j,r} = E_X\{F_{j,r}(X)\} - \frac{1}{2}. \qquad (1.1)$$

Let $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$ be independent and identically distributed random samples of size n of $(\boldsymbol{X}, Y)$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$. Denote $\hat{p}_r = (1/n)\sum_{i=1}^{n} I(Y_i = y_r)$, $r = 1, \ldots, R$. Then, a sample estimate of $\tau_{j,r}$, $j = 1, 2, \ldots, p$, is

$$\hat{\tau}_{j,r} = \frac{1}{n+1}\sum_{k=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{I(X_{ij} \le X_{kj}, Y_i = y_r)}{\hat{p}_r}\right\} - \frac{1}{2}. \qquad (1.2)$$

Remark 1. When X_j and W_r are statistically independent, a direct calculation yields

$$\begin{aligned}
&E\left\{\frac{1}{n+1}\sum_{k=1}^{n}\frac{1}{n}\sum_{i=1}^{n}\frac{I(X_{ij} \le X_{kj}, Y_i = y_r)}{\hat{p}_r}\right\} \\
&= \frac{1}{n(n+1)}\sum_{k=1}^{n}\sum_{i=1}^{n} E\left\{\frac{I(Y_i = y_r)}{\hat{p}_r}\right\} E\left\{I(X_{ij} \le X_{kj})\right\} \\
&= \frac{1}{n(n+1)}\left[\sum_{k=1}^{n}\sum_{i\neq k}^{n} E\left\{\frac{I(Y_i = y_r)}{\hat{p}_r}\right\} \Pr(X_{ij} \le X_{kj}) + \sum_{k=1}^{n} E\left\{\frac{I(Y_k = y_r)}{\hat{p}_r}\right\}\right] \\
&= \frac{1}{n(n+1)} E\left\{\frac{n}{2\hat{p}_r}\sum_{i=1}^{n} I(Y_i = y_r) + \frac{1}{2\hat{p}_r}\sum_{k=1}^{n} I(Y_k = y_r)\right\} \\
&= \frac{1}{2}.
\end{aligned} \qquad (1.3)$$

The denominator $(n+1)$ in (1.2) ensures that $\hat{\tau}_{j,r}$ is an unbiased estimate of $\tau_{j,r}$ when X_j and W_r are statistically independent. In other words, $\tau_{j,r} = 0$ if X_j and W_r are statistically independent. Moreover, in the absence of independence of X_j and W_r, it can be checked in a straightforward way that $\hat{\tau}_{j,r}$ is asymptotically unbiased for $\tau_{j,r}$. Thus, $\tau_{j,r}$ can serve as a utility to characterize the linear or nonlinear relationships between each predictor X_j and W_r, $r = 1, \ldots, R$.

Remark 2. Note that 1/2 in (1.2) and (1.3) is from $\Pr(X_{ij} \le X_{kj}) = 1/2$ for $i \neq k$, under the continuously distributed assumption of X_j. One may simply extend our proposed method to accommodate discrete X_j and replace 1/2 by the empirical counterpart of $\Pr(X_{ij} \le X_{kj}), i \neq k$.

In this paper, we propose to rank all the candidate predictors X_j, $j = 1, \ldots, p$ according to $|\hat{\tau}_{j,r}|$ from the largest to smallest. To be specific, we select the set of variables

$$\mathcal{A}_r = \{1 \le j \le p : |\hat{\tau}_{j,r}| > \gamma_n\},$$

where γ_n is pre-specified threshold value. Following popular feature screening methods (Li et al. 2012; Cui et al. 2015), we select a reduced model

$$\tilde{\mathcal{A}}_r = \{1 \le j \le p : |\hat{\tau}_{j,r}| \text{ is among the top } d_n \text{ largest of all}\},$$

where d_n is pre-determined size, chosen as $d_n = \lfloor n/\log(n) \rfloor$ in practice, and $\lfloor a \rfloor$ denotes the integer part of a.

Remark 3. When the categorical variable Y is binary with only two classes y_1 and y_2 ($R = 2$), by the definition of $\mathcal{A}_1$ and $\mathcal{A}_2$, $\mathcal{A}_1 = \mathcal{A}_2$, which reduces to the active set defined in the literature, see Cui et al. (2015). Nevertheless, for $R > 2$, $\mathcal{A}_r$ could be different for $r = 1, \ldots, R$. Hence, our proposed method is able to identify category-specific $\mathcal{A}_r$, $r = 1, \ldots, R$.

Next, we study the theoretical properties of the proposed procedure. The following conditions are needed to establish the sure screening and ranking consistency properties of the proposed procedure.

(C1). There exist two positive constants c_1 and c_2 such that $c_1/R \le \min_{1 \le r \le R} p_r \le \max_{1 \le r \le R} p_r \le c_2/R$.

(C2). There exist positive constants $c > 0$ and $0 \le \kappa < 1/2$ such that $\min_{j \in \mathcal{A}_r} |\tau_{j,r}| \ge 2cn^{-\kappa}$.

(C3). The number of classes $R = O(n^{\xi}), \xi > 0$ satisfying $2\kappa + \xi < 1$.

(C4). $\liminf_{p \to \infty} \{\min_{j \in \mathcal{A}_r} |\tau_{j,r}| - \max_{j \in \mathcal{I}_r} |\tau_{j,r}|\} \ge m_0$ for some $m_0 > 0$.

Similar to Condition (C1) in Cui et al. (2015), Condition (C1) requires that the proportion of each class of the response should not be either too small or too large. Condition (C2) allows the minimum true signal is at the order of $n^{-\kappa}$. Condition (C3) allows the number of classes for the response to diverge as *n* increases. Condition (C4) guarantees the active and inactive predictors can be well separated at the population level.

THEOREM 1

1. (Sure screening property) Suppose Conditions (C1) and (C3) hold. Then, for any constant $c_4 > 0$, there exists $c_5 > 0$ such that

$$\Pr(\max_{1 \le j \le p} |\hat{\tau}_{j,r} - \tau_{j,r}| \ge c_4 n^{-\kappa}) \le 2(2n+3)p \exp(-c_5 n^{1-2\kappa-\xi}),$$

where $r = 1, \ldots, R$. *Moreover, under Condition (C2), we have*

$$\Pr(\mathcal{A}_r \subset \hat{\mathcal{A}}_r) \ge 1 - 2(2n+3)s_r \exp(-c_6 n^{1-2\kappa-\xi}),$$

where c_6 *is some positive constant and* $s_r = |\mathcal{A}_r|$ *is the true model size,* $r = 1, \ldots, R$.

2. (Ranking consistency property) *Suppose Conditions (C1) and (C4) hold. If* $R\log(n) = o(nm_0^2)$ *and* $R\log(p) = o(nm_0^2)$, *then*

$$\liminf_{n \to \infty} \left\{ \min_{j \in \mathcal{A}_r} |\hat{\tau}_{j,r}| - \max_{j \in \mathcal{I}_r} |\hat{\tau}_{j,r}| \right\} > 0$$

almost surely, for n sufficiently large.

3. (Controlling false discovery rate) *Furthermore, under Conditions (C1)-(C3) and* $\gamma_n = c_8 n^{-\kappa}$ *, there exists positive constant* c_9 *such that*

$$\Pr\{|\mathcal{A}_r| \le (c_8/2)^{-1} n^{\kappa} \sum_j |\tau_{j,r}|\} \ge 1 - 2(2n+3)p\exp(-c_9 n^{1-2\kappa-\xi}).$$

Theorem 1 indicates that the proposed method can handle the NP-dimensionality problem when $\log(p) = o(n^{\alpha})$ with $\alpha < 1 - 2\kappa - \xi, 0 \le \kappa < 1/2$ and $2\kappa + \xi < 1$. The sure screening property holds under Conditions (C1)-(C3) here. Since we do not require any regression models and moment assumptions of the predictors, our method is robust to heavy-tailed distributions of the predictors and the presence of potential outliers. The ranking consistency property indicates that the values of $\hat{\tau}_{j,r}$ of active predictors can be ranked above that of the inactive ones with high probability, which implies that we can separate the active and inactive predictors with certain thresholding value. Property (iii) implies that the number of selected variables is bounded by $O(n^{\kappa}\sum_j |\tau_{j,r}|)$ with high probability. In particular, when the true model size $\sum_j |\tau_{j,r}| = O(n^{\lambda})$, the size of the selected variables is at the order of $O(n^{\kappa+\lambda})$ and $|\mathcal{A}|$ is of order $O(n^{\kappa+\lambda})$ according to Theorem 1, which confirms that the number of the selected variables can be effectively controlled. In addition, when $\kappa + \lambda < 1$, the hard thresholding rule with the top $a\lfloor n/\log(n)\rfloor$ variables selected in numerical studies can select the true active predictors with high probability, where *a* is some constant and the value of *a* may reflect researchers' prior knowledge of the number of susceptible predictors.

Remark 4. Alternatively, if one focuses on selecting active predictors relevant to certain categories $\mathcal{G}$, a given subset of $\{1, \ldots, R\}$, one can define $W_{\mathcal{G}} \equiv I(Y = y_r, r \in \mathcal{G})$ and consider a refined version

$$\hat{\tau}_{j,\mathcal{G}} \equiv \sup_{r \in \mathcal{G}} |\hat{\tau}_{j,r}|$$

for measuring the dependence between each X_j, $j = 1, \ldots, p$, and $W_{\mathcal{G}}$. As a special case,

$$\hat{\tau}_j \equiv \sup_{r \in \{1,2,\ldots,R\}} |\hat{\tau}_{j,r}| \qquad (1.4)$$

measures the dependence between X_j and the categorical response Y, the topic of interest in Mai and Zou (2013), Cui et al. (2015), etc. Based on the magnitude of $\hat{\tau}_j$, we can select a set of variables that are most relevant to Y. For such a variation, under similar conditions to Conditions (C1)-(C4), the asymptotic properties analogous to Theorem 1 can be established without further difficulty. As suggested by an anonymous reviewer, we present the technical conditions and the main results for such a variation in the supplementary materials for completion.

2.2 Category-adaptive variable screening with response-selective samples

Recall that existing popular screening methods in the literatures are effective with prospective samples or i.i.d samples of the underlying population. However, in imbalanced design where one of the classes/categories of Y is rather rare, e.g, in the studies of rare diseases, the imbalanced nature of case and control could be at the order of $1:10^3$ or $1:10^4$, the response-biased or response-selective sampling is particularly designed to collect samples containing more information relevant to one's interest by adjusting the mixture of the categories to enrich the rare class. Nonetheless, there exists very limited amount of work on variable screening with response-selective sampling. Another main finding of this paper, the proposed category-adaptive screening method in Section 2.1 can be directly applied to response-selective sampling data without any modification.

To facilitate presentation, we let $(\boldsymbol{X}, Y)$ be the pair of covariates and response in the population and let $(\boldsymbol{X}^*, Y^*)$ be a pair of covariates and response in the sample collected by a response-biased sampling design. The observations $(\boldsymbol{X}_i^*, Y_i^*), i = 1, \ldots, n$ are independent and identically distributed random samples of $(\boldsymbol{X}^*, Y^*)$. For instance, the ratio of case and control in the study of ovarian cancer at the population level is about $1:10^3$. With case-control sampling, a special case of response-selective sampling, one could arbitrarily set the proportion of the cases and controls when taking samples and their respective predictors. The resulting samples $(\boldsymbol{X}^*, Y^*)$ are regarded as response-biased sampling data. It is known that the joint distribution of $(\boldsymbol{X}^*, Y^*)$ is generally not the same as that of $(\boldsymbol{X}, Y)$ in the population. Nevertheless, according to Chen (2001a) and Chen et al. (2017), the response-selective sampling assumes that, the conditional distribution of $\boldsymbol{X}^*$ given Y^* is the same as that of $\boldsymbol{X}$ given Y. For example, in case-control studies, the conditional distribution of $\boldsymbol{X}$ given $Y = 1(0)$ is the population distribution of the covariates for all cases (controls), which is the same as the distribution of the covariates of cases (controls) in the case-control sample.

Our proposed marginal screening utility with response-selective sampling data $(\boldsymbol{X}_i^*, Y_i^*), i = 1, \ldots, n$ is

$$\tau_{j,r}^* \equiv E_{X^*}\{F_{j,r}^*(X^*)\} - \frac{1}{2},$$

where $F_{j,r}^*(x) \equiv \Pr(X_j^* \le x \mid Y^* = y_r)$ for each $j = 1, \ldots, p$. Similar to Section 2.2, a sample estimate of $\tau_{j,r}^*, r = 1, \ldots, R$, is

$$\hat{\tau}_{j,r}^* = \frac{1}{n+1}\sum_{k=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{I(X_{ij}^* \le X_{kj}^*, Y_i^* = y_r)}{\hat{p}_r^*}\right\} - \frac{1}{2}, \qquad (1.5)$$

where $\hat{p}_r^* \equiv \sum_{i=1}^{n} I(Y_i^* = y_r)/n$. The rationale of using $\tau_{j,r}^*$ to measure the importance of X_j on W_r is that, when X_j and W_r are statistical independent in the population,

$$
\begin{aligned}
& E\left\{\frac{1}{n+1}\sum_{k=1}^{n}\frac{1}{n}\sum_{i=1}^{n}\frac{I(X_{ij}^* \le X_{kj}^*, Y_i^* = y_r)}{\hat{p}_r^*}\right\} \\
&= \frac{1}{n(n+1)}\sum_{k=1}^{n}\sum_{i=1}^{n} E\left[\frac{I(Y_i^* = y_r)}{\hat{p}_r^*} E\left\{I(X_{ij}^* \le X_{kj}^*) \mid I(Y_1^* = y_r), I(Y_2^* = y_r), \ldots, I(Y_n^* = y_r)\right\}\right] \\
&= \frac{1}{n(n+1)}\sum_{k=1}^{n}\sum_{i=1}^{n} E\left[\frac{I(Y_i^* = y_r)}{\hat{p}_r^*} E\left\{I(X_{ij} \le X_{kj}) \mid I(Y_1 = y_r), I(Y_2 = y_r), \ldots, I(Y_n = y_r)\right\}\right] \\
&= \frac{1}{n(n+1)}\sum_{k=1}^{n}\sum_{i=1}^{n} E\left[\frac{I(Y_i^* = y_r)}{\hat{p}_r^*} E\left\{I(X_{ij} \le X_{kj})\right\}\right] \\
&= \frac{1}{n(n+1)}\left[\sum_{k=1}^{n}\sum_{i\ne k}^{n} E\left\{\frac{I(Y_i^* = y_r)}{\hat{p}_r^*}\right\}\Pr(X_{ij} \le X_{kj}) + \sum_{k=1}^{n} E\left\{\frac{I(Y_k^* = y_r)}{\hat{p}_r^*}\right\}\right] \\
&= \frac{1}{n(n+1)} E\left\{\frac{n}{2\hat{p}_r^*}\sum_{i=1}^{n} I(Y_i^* = y_r) + \frac{1}{2\hat{p}_r^*}\sum_{k=1}^{n} I(Y_k^* = y_r)\right\} \\
&= \frac{1}{2},
\end{aligned}
$$

where $\hat{p}_r^* = \sum_{i=1}^{n} I(Y_i^* = y_r)/n$ and $r = 1, \ldots, R$. The second equality above holds due to the nature of response-respective sampling design, while the third equality follows from the independence of X_j and W_r. The above calculation reveals that $\tau_{j,r}^* = 0$ if X_j and W_r are statistical independent at the population level. In other words, our proposed screening method in Section 2.2 is valid and directly applicable to handle response-biased sampling data. Therefore, with response-selective sampling data $(\boldsymbol{X}_i^*, Y_i^*), i = 1, \ldots, n$, we propose to rank $X_j, j = 1, \ldots, p$ according to $|\hat{\tau}_{j,r}^*|$ and select the set of variables

$$\mathcal{A}_r^* = \{1 \le j \le p : |\hat{\tau}_{j,r}^*| > \gamma_n\},$$

where γ_n is a pre-specified threshold value.

We next present the theoretical properties of the proposed screening measure with response-selective sampling data. Define $p_r^* \equiv \Pr(Y^* = y_r), r = 1, \ldots, R$. In fact, p_r^* is the proportion of samples from the rth category of Y in the response-selective sampling, which is usually specified before sampling. When important variables pertaining to the rth category of Y is of concern, an additional condition on p_r^* is needed for such a given r.

(C5). There exist two positive constants $\tilde{c}_1$ and $\tilde{c}_2$ such that $\tilde{c}_1 / R \le p_r^* \le \tilde{c}_2 / R$.

Condition (C5) implies that, when important variables pertaining to the rth category of Y is of interest, the proportion of samples from the rth category in response-biased sampling design should not be too small or too large. In other words, the response-selective sampling data cannot consist of all samples or no sample from the rth category.

The following theorem presents the asymptotic properties of the proposed method with response-selective sampling, when feature screening pertaining to the rth category of Y is of concern, $r = 1, \ldots, R$.

THEOREM 2

(i) (Sure screening property) *Suppose Conditions (C3) and (C5) hold. Then, for any constant $\tilde{c}_4 > 0$, there exists $\tilde{c}_5 > 0$ such that*

$$\Pr(\max_{1 \le j \le p} |\hat{\tau}_{j,r}^* - \tau_{j,r}| \ge \tilde{c}_4 n^{-\kappa}) \le 2(2n+3)\, p \exp(-\tilde{c}_5 n^{1-2\kappa-\xi}).$$

Moreover, together with Condition (C2), we have

$$\Pr(\mathcal{A}_r \subset \mathcal{A}_r^*) \ge 1 - 2(2n+3)s_r \exp(-\tilde{c}_6 n^{1-2\kappa-\xi}),$$

where $\tilde{c}_6$ *is some positive constant and* $s_r = |\mathcal{A}_r|$ *is the true model size for the rth category.*

(ii) (Ranking consistency property) *Suppose Conditions (C4) and (C5) hold. If* $R\log(n) = o(nm_0^2)$ *and* $R\log(p) = o(nm_0^2)$*, then*

$$\liminf_{n\to\infty}\left\{\min_{j\in\mathcal{A}_r}|\hat{\tau}^*_{j,r}| - \max_{j\in\mathcal{I}_r}|\hat{\tau}^*_{j,r}|\right\} > 0$$

almost surely, for sufficiently large n.

(iii) (Controlling false discovery rate) *Furthermore, under Conditions (C2), (C3) and (C5),* $\gamma_n = \tilde{c}_8 n^{-\kappa}$*, there exists positive constant* $\tilde{c}_9$ *such that*

$$\Pr\{|\mathcal{A}^*_r| \le (\tilde{c}_8/2)^{-1} n^{\kappa}\sum_j |\tau_{j,r}|\} \ge 1 - 2(2n+3)p\exp(-\tilde{c}_9 n^{1-2\kappa-\xi}).$$

3 Simulations

Extensive simulation studies are conducted to investigate the performance of the proposed category-adaptive screening method in terms of the following criteria: $\mathcal{A}_r$, the minimum model size needed to include all the true predictors, we report the standard deviation (SD), the median and the interquartile range (IQR) of $\mathcal{A}_r$; R_a, the average of the ranks of all active predictors among all candidate variables sorted by the screening procedure; P_a, the proportion of all active predictors being selected into the submodel with size $\lfloor n/\log(n)\rfloor$. For comparison, we compare the performance with well-known screening procedures including the Kolmogorov filter (KF) method (Mai and Zou 2013), the mean-variance screening (MV-SIS) (Cui et al. 2015) and the pairwise sure independence screening (PSIS) (Pan et al. 2016). For brevity, we refer to our proposed category-specific screening approach pertaining to the *r*th category in (1.2) as CAS$_r$, and refer to the refined version in (1.4) as CAS$^{\text{sup}}$. The simulations of each scenario are based on 200 replications.

Example 1. The first example is conducted to examine the performance of the CAS^{sup} procedure to screen important predictors for the categorical response Y compared with existing screening methods. The simulated data are generated from a linear discriminant analysis model with ultrahigh dimensional predictors by letting the categorical response Y_i follow a distribution with $p_r = \Pr(Y_i = r), r = 1, \ldots, R, i = 1, \ldots, n$. Given $Y_i = r$, the ith sample vector of predictors $\boldsymbol{X}_i$ is generated from a mixture distribution $0.9\boldsymbol{X}_i + 0.1\boldsymbol{Z}_i$, where $\boldsymbol{X}_i = \boldsymbol{\mu}_r + \boldsymbol{\epsilon}_i$ with $\boldsymbol{\mu}_r = (\mu_{r1}, \ldots, \mu_{rp})^\top$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^\top$ follows $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, and $\boldsymbol{Z}_i$ is a random vector with each component being independent Student's t distribution with 1 degree of freedom, $i = 1, \ldots, n$. We set the categorical response Y to be binary or multi-category with both balanced or imbalanced design by considering the following scenario:

Case 1: $R = 2$, $p_1 = p_2 = 0.5$, $n = 100$, $p = 1000$ or 3000, $\mu_{rr} = 1.5$ and $\mu_{rl} = 0$ for $r \neq l$;

Case 2: the same setup as in Case 1 except that $p_1 = 0.1$ and $p_2 = 0.9$;

Case 3: $R = 8$, $p_r = 1/8$ for $r = 1, \ldots, 8$, $n = 200$, $p = 1000$ or 3000, $\mu_{rr} = 2.4$ and $\mu_{rl} = 0$ for $r \neq l$;

Case 4: the same setup as in Case 3 except that $p_1 = p_2 = 0.05$ and $p_r = 0.15$ for $r = 3, \ldots, 8$.

In sum, the true active set for the categorical response Y is $\{X_1, X_2, X_3, X_4\}$ for Cases 1-2, while the true active set is $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8\}$ for Cases 3-4. The results are summarized in Table 1, which reveals three observations. First, when the covariates are contaminated with possible outliers, the proposed CAS^{sup} is quite robust and performs comparably to MV-SIS and KF, whilst the PSIS breaks down in terms of R_a and mean of the minimum model size. Second,

the four methods work similarly well in balanced design, but in contrast, the CAS^{sup} with i.i.d samples performs the best among the four methods in imbalanced design in Case 4. Third, as the number of categories increases in Case 4, the proposed CAS^{sup} shows superior performance in terms of P_a.

Example 2. In this example, we conduct simulations with the proposed CAS$_r$ for screening category-specific active predictors in a more complex setting. The data are generated from the same model as in Example 1, except for $R = 5$ and $\boldsymbol{\mu}_1 = (1.5,1.5,\boldsymbol{\theta}_{p-2}^\top)$, $\boldsymbol{\mu}_2 = (\boldsymbol{\theta}_7^\top,1.5,1.5,1.5,\boldsymbol{\theta}_{p-10}^\top)$, $\boldsymbol{\mu}_3 = (\boldsymbol{\theta}_3^\top,1.5,1.5,1.5,\boldsymbol{\theta}_{p-6}^\top)$, $\boldsymbol{\mu}_4 = (\boldsymbol{\theta}_{15}^\top,1.5,1.5,1.5,1.5,1.5,\boldsymbol{\theta}_{p-20}^\top)$, . Accordingly, the true active sets are $\mathcal{A}_1 = \{X_1, X_2\}$, $\mathcal{A}_2 = \{X_8, X_9, X_{10}\}$, $\mathcal{A}_3 = \{X_4, X_5, X_6\}$, $\mathcal{A}_4 = \{X_{16}, X_{17}, X_{18}, X_{19}, X_{20}\}$ and $\mathcal{A}_5 = \{X_{31}, X_{32}, X_{33}, X_{34}, X_{35}\}$, respectively. We consider the balanced and imbalanced design as follows:

Case 1: $p_i = 0.2, i = 1,\ldots,5$, n = 200, p = 1000 or 3000;

Case 2: $p_1 = p_2 = p_3 = 0.1, p_4 = p_5 = 0.35$, n = 200, p = 1000 or 3000.

Note that the active sets pertaining to each category are different in Example 2. For comparison, following the constructive suggestion from the anonymous reviewers, we also compute a modified version of MV-SIS, KF and PSIS by redefining the response variable as $I(Y = y_r)$ in their methods, denoted by MV-SIS$_r$, KF$_r$ and PSIS$_r$, respectively. The simulation results of Example 2 are summarized in Table 2 indicating that: (i) When the covariates are contaminated with possible outliers in both cases, the screening performance of PSIS$_r$ is largely discounted; (ii) In Case 1 with p = 1000 and balanced design, MV-SIS$_r$, KF$_r$ and our proposed CAS$_r$ perform comparably in selecting category-specific active variables, while the CAS$_r$ performs slightly better than the other two for p = 3000; Note that with i.i.d samples, MV-SIS$_r$ and our proposed CAS$_r$ work comparably well in terms of median of $\mathcal{A}_r$ in balanced cases; (iii) In Case 2 with relatively

imbalanced design, the CAS$_r$ shows superior screening performance and robustness in terms of larger coverage proportions P_a.

Example 3. In this example, we conduct simulation studies to check the performance of our proposed method with case-control sampling data. First, we generate a large pool of data from a linear discriminant analysis model with the categorical response Y following a distribution with $p_r = \Pr(Y = r), r = 1, \ldots, R = 4$. We set $p_1 = p_4 = 0.05$ and $p_2 = p_3 = 0.45$. This is a relatively imbalanced design. Given $Y_i = r$, the ith sample vector of predictors $\boldsymbol{X}_i$ is generated from $\boldsymbol{X}_i = \boldsymbol{\mu}_r + \boldsymbol{\epsilon}_i$ with $\boldsymbol{\mu}_r = (\mu_{r1}, \ldots, \mu_{rp})^\top$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^\top$, $i = 1, \ldots, N$. Here $N = 10^4$. We consider two distributions of ϵ: (1) each component of ϵ follows independent $\mathcal{N}(0,1)$; (2) each component of ϵ follows independent Student's t distribution with 2 degrees of freedom. We consider the following two settings:

Case 1:
$\boldsymbol{\mu}_1 = (1.5, 1.5, \boldsymbol{0}_{p-2}^\top)$, $\boldsymbol{\mu}_2 = (\boldsymbol{0}_7^\top, 1.5, 1.5, 1.5, \boldsymbol{0}_{p-10}^\top)$, $\boldsymbol{\mu}_3 = (\boldsymbol{0}_3^\top, 1.5, 1.5, \boldsymbol{0}_{p-5}^\top)$, $\boldsymbol{\mu}_4 = (\boldsymbol{0}_{15}^\top, 1.5, 1.5, 1.5, 1.5, 1.5, \boldsymbol{0}_{p-20}^\top)$, p = 1000 or 3000. Accordingly, the true active sets are $\mathcal{A}_1 = \{X_1, X_2\}$, $\mathcal{A}_2 = \{X_8, X_9, X_{10}\}$, $\mathcal{A}_3 = \{X_4, X_5\}$ and $\mathcal{A}_4 = \{X_{16}, X_{17}, X_{18}, X_{19}, X_{20}\}$, respectively.

Case 2: $\boldsymbol{\mu}_1 = (1.5, 1.5, \boldsymbol{0}_{p-2}^\top)$, $\boldsymbol{\mu}_2 = (\boldsymbol{0}_5^\top, 1.5, 1.5, 1.5, 1.5, 1.5, \boldsymbol{0}_{p-10}^\top)$, $\boldsymbol{\mu}_3 = (\boldsymbol{0}_5^\top, 1.5, 1.5, 1.5, 1.5, 1.5, \boldsymbol{0}_{p-10}^\top)$, $\boldsymbol{\mu}_4 = (\boldsymbol{0}_{15}^\top, 1.5, 1.5, 1.5, 1.5, 1.5, \boldsymbol{0}_{p-20}^\top)$, p = 1000 or 3000. Accordingly, the true active sets are $\mathcal{A}_1 = \{X_1, X_2\}$, $\mathcal{A}_2 = \mathcal{A}_3 = \{X_6, X_7, X_8, X_9, X_{10}\}$ and $\mathcal{A}_4 = \{X_{16}, X_{17}, X_{18}, X_{19}, X_{20}\}$, respectively.

Next, we take case-control sampling with sample size 200 from this imbalanced data set by setting the mixture of the four categories in the case-control samples as $1:1:1:1$. We apply our proposed method to the case-control samples. For comparison, we also take 200 random samples out of the pool and compute the MV-SIS$_r$, PSIS$_r$ and KF$_r$ defined in Example 2.

The results are reported in Tables 3-4. It can be seen that, compared with existing methods with i.i.d sampling, our proposed method with case-control sampling data are more favorable in terms of the median, mean, IQR of $\mathcal{A}_r$, R_a and P_a. Our method with case-control sampling data outperforms MV-SIS$_r$, PSIS$_r$ and KF$_r$ with heavy-tailed error distributions. In our view, the reason accounting for the improvement is that the case-control sampling data is supposed to contain more information according to one's interest in heavy-tailed case, compared with simple random sampling data. Note that the results of CAS$_r$ in Tables 3-4 are based on one single case-control sampling data in each replication. In the simulation, we also compute a refined version of our method by repeatedly taking case-control samples with proportion of each category $1:1:1:1$ for L times with replacement. The resulting fused ranking statistic is calculated by averaging the values of CAS$_r$ based on each case-control samples. The results are presented in Table 5, which contains supportive evidence that our proposed method based on multiple case-control sampling provides stable and robust results in variable screening.

4 Applications

In this section, we illustrate the usefulness of the proposed method using the US post-office handwritten digits data and the lung carcinomas data.

4.1 Handwritten digits data

The first real example is to analyze the US post-office handwritten digits data (Le Cun et al. 1990), which contains 7291 training images of the digits $\{0,1,\ldots,9\}$ and 2007 testing samples. Each image is digitized to a 16 × 16 gray-scale with 16 × 16 = 256 pixels as features. Here p = 256 and the response Y is a categorical variable with R = 10 classes $\{y_1, y_2, \ldots, y_{10}\}$ and $y_1 = 0, y_2 = 1, \ldots, y_{10} = 9$. Some handwritten samples are shown in Figure 1. For detecting heterogeneous important features for each digit, we apply the proposed category-adaptive screening method defined in (1.2), denoted by CAS$_r$, $r = 1,\ldots,10$ to the training

dataset. We also compute the MV-SIS$_r$, PSIS$_r$ and KF$_r$ defined in Section 3 for comparison.

The implementation of our proposed method is as follows. Since our method works directly for response-selective sampling data, we first take case-control sampling to screen category-specific features for each digit/category. The case-control sampling is drawn for appropriate contrast and savings in computational cost. For instance, in order to screen features pertaining to the category $Y = y_1$, a simple design of case-control sampling one can draw from the training data is to take all the training samples with $Y = y_1$ and $Y = y_2$, for some $y_2 \in \{0,1,\ldots,9\}$ but $y_2 \neq y_1$. There are total 9 different case-control sample. For each case-control samples, we apply our proposed CAS$_r$ method and retain the top 100 features, respectively. Considering that some unimportant predictors might be retained after the screening procedure, to further remove those irrelevant features, we next fit a regularized logistic regression for the binary variable $W_1 = I(Y = y_1)$ with lasso penalty to all the training data. For the fitted logistic models of $W_1 = I(Y = y_1)$ with different case-control samples respectively, we calculate their predictive power in terms of classification accuracy with the testing data, respectively. Hence, for each digit $Y = y_1$, $y_1 \in \{0,1,\ldots,9\}$, the respective final model is determined to be the one with largest classification accuracy. The selected features for each digit are presented in Figure 2, from which one can observe that the sets of important predictors are generally different for different digits. Consequently, with the availability of 10 fitted logistic regression models for $W_r = I(Y = y_r), r = 1,\ldots,10$ respectively, we adopt the one-versus-rest (OvR) strategy in multi-category classification to classify the testing data and evaluate the predictive power. In other words, based on the 10 fitted logistic regression models, a testing sample will be classified into category m, where $m \equiv \operatorname{argmax}_{r\in\{1,2,\ldots,10\}} \Pr(W_r = 1)$, and $\Pr(W_r = 1)$ is the fitted value of $\Pr(W_r = 1)$. The predictive power can be thus evaluated on the entire testing data and reported in Table 6 in terms of classification accuracy (CA), the proportion of corrected-

classified samples among all testing data. For comparison, we also implement the MV-SIS$_r$, PSIS$_r$ and KF$_r$ in the screening step with all training data and then fit the penalized logistic regression model for $W_r = I(Y = y_r)$, $r \in \{1, \ldots, 10\}$, respectively. Similarly, their classification accuracies are evaluated on the testing data. One can see from Table 6 that our proposed method performs the best with $\mathrm{CA} = 91.73\%$, followed by 90.43% of MV-SIS$_r$, 90.53% of PSIS$_r$ and 90.18% of KF$_r$.

4.2 Lung carcinomas data

The human lung carcinomas data was studied by using mRNA expression profiling (Bhattacharjee et al. 2001) and was previously analyzed by Cui et al. (2015). In the original lung carcinomas data set, there are total 203 snap-frozen lung tumors and normal lungs and there are 12600 mRNA expression levels. Here, the categorical response Y is the tumor type. The 203 specimens/samples were classified into five subclasses: 139 in lung adenocarcinomas (ADEN), 21 in squamous cell lung carcinomas (SQUA), 6 in small cell lung carcinomas (SCLC), 20 in pulmonary carcinoid tumors (COID) and the remaining 17 normal lung samples (NORMAL).

To check the performance of various methods, we randomly split the 203 samples into training and testing data. Since there are very few samples in the SCLC class, we include the 6 samples of SCLC in the training set. For other four classes, we randomly partition $100\gamma\%$ samples as training data and the rest $100(1-\gamma)\%$ samples as testing data, where $\gamma \in (0,1)$. We apply the proposed CAS$_r$, as well as the MV-SIS$_r$, PSIS$_r$ and KF$_r$, to the training data and retain the top $\lfloor n / \log(n) \rfloor$ features for $r = 1, \ldots, 5$. Here n is the size of the training set. Following the previous example in subsection 4.1, to further remove some irrelevant predictors after screening, we next fit penalized logistic regression with lasso penalty for each binary variable $W_r = I(Y = y_r), r = 1, \ldots, 5$, respectively, with all training data. With the 5 fitted binary logistic regression models, we classify

the testing samples by adopting the one-versus-rest (OvR) strategy. In addition, we also use the proposed $\mathrm{CAS}^{\mathrm{sup}}$ defined in (1.4), the original MV-SIS, PSIS and KF methods to screen features relevant to the response Y. Similar to Cui et al. (2015), we apply the sparse discriminant analysis (SDA) subsequent to the dimension reduction with different screening methods; see Clemmensen et al. (2011). We refer these two-stage approaches as $\mathrm{CAS}^{\mathrm{sup}}$ + SDA, MV-SIS + SDA, PSIS + SDA, KF + SDA.

We set $\gamma = 0.6$, 0.7 or 0.8 and repeat the above procedure with 100 random splits of the data set. The mean and standard deviation of classification accuracy (CA), the proportion of corrected-classified samples among all testing data, are computed. The analysis results of different methods are reported in Table 7. One can observe that, all methods work reasonably well and our proposed method CAS_r performs slightly better in terms of the classification accuracy compared with other methods. On the contrary, the classification accuracies of $\mathrm{CAS}^{\mathrm{sup}}$ + SDA, MV-SIS + SDA, PSIS + SDA and KF + SDA are smaller than those of CAS_r, $\mathrm{MV\text{-}SIS}_r$, PSIS_r and KF_r approaches, which provides supportive evidence that considering category-adaptive information improves predictive power.

5 Conclusion

In this paper, we advocate a screening procedure that is a model-free method and a nonparametric measure of statistical dependence between a candidate predictor and the dummy variable that the response belongs to certain level of a qualitative variable. Our proposed procedure has some desirable features: it is capable of identifying the active variables that is allowed to vary across different categories; it is a model-free framework without specifying any regression form of predictors and response variable; it is most useful to handle imbalanced data subject to response-selective sampling without any modification, as evidenced in the simulation studies and real data analysis; moreover, it can be directly applied to data collected in retrospective studies. Nonetheless, we have to admit that our

proposed method and MV-SIS perform similarly in many balanced cases. And the proposed procedure cannot handle the case that there is interaction effects among predictors in ultrahigh dimensional discriminant analysis. We leave space here for future research.

Acknowledgment

The authors are indebted to the Editor, the Associate Editor and four anonymous reviewers for their professional review and insightful comments that lead to significant improvements in the paper. Yuanyuan Lin's research is supported by the Hong Kong Research Grants Council (Grants No. 509413 and 14311916), the National Natural Science Foundation of China (Grant No. 71874028) and Direct Grants for Research, The Chinese University of Hong Kong. Xiaodong Yan's research is supported by the Young Scholars Program of Shandong University (YSPSDU: 11020088964008). Niansheng Tang's research is supported by the National Natural Science Foundation of China (Grant No. 11671349) and the Key Projects of the National Natural Science Foundation of China (Grant No. 11731101).

Appendix

Some notations are needed. Let $p_r = \Pr(Y = y_r)$, $f_j(r,x) = I(X_j \le x, Y = y_r)$ and $W_r = I(Y = y_r)$. For some $j = 1, \ldots, p$, let $\{(X_{ij}, Y_i) : i = 1, \ldots, n\}$ be a random sample of (X_j, Y). Write $W_{ri} = I(Y_i = y_r)$, $f_{ij}(r,x) = I(X_{ij} \le x, Y_i = y_r)$, $\zeta_j(r) = E_X\{F_{j,r}(X)\}$,

$$\tilde{\zeta}_j(r) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{i=1}^{n} \frac{I(X_{ij} \le X_{kj}, Y_i = y_r)}{\hat{p}_r}$$

and

$$\hat{\zeta}_j(r) = \frac{1}{n(n+1)} \sum_{k=1}^{n} \sum_{i=1}^{n} \frac{I(X_{ij} \le X_{kj}, Y_i = y_r)}{\hat{p}_r}.$$

LEMMA 1 *For given r and j*, $r = 1, \ldots, R,\ j = 1, \ldots, p$,

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} W_{ri} - E(W_r)\right| \geq \epsilon\right\} \leq 2\exp\left\{-\frac{n\epsilon^2}{2(p_r + \epsilon/3)}\right\}, \qquad \text{(A.1)}$$

$$\Pr\left[\sup_{x \in \mathbb{R}_{X_j}} \left|\frac{1}{n}\sum_{i=1}^{n} I(X_{ij} \leq x) - E\{I(X_j \leq x)\}\right| \geq \epsilon\right] \leq 2(n+1)\exp\left(-2n\epsilon^2\right), \qquad \text{(A.2)}$$

$$\Pr\left[\sup_{x \in \mathbb{R}_{X_j}} \left|\frac{1}{n}\sum_{i=1}^{n} f_{ij}(r,x) - E\{f_j(r,x)\}\right| \geq \epsilon\right] \leq 2(n+1)\exp\left\{-\frac{n\epsilon^2}{2(p_r + \epsilon/3)}\right\}, \qquad \text{(A.3)}$$

for any $\epsilon \in (0,1)$ *, where* $\mathbb{R}_{X_j}$ *denotes the support of a continuous predictor* X_j.

Proof of Lemma 1. Note that $E(W_{ri}) = p_r$, W_{ri} follows Bernoulli(p_r) and $\sum_{i=1}^{n} W_{ri}$ follows Binomial(n, p_r). To prove (A.1), we use the Bernstein's inequality and get

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} W_{ri} - E(W_r)\right| \geq \epsilon\right\} = \Pr\left\{\left|\sum_{i=1}^{n}(W_{ri} - p_r)\right| \geq n\epsilon\right\}$$

$$\leq 2\exp\left\{-\frac{n^2\epsilon^2}{2(np_r + n\epsilon/3)}\right\}$$

$$= 2\exp\left\{-\frac{n\epsilon^2}{2(p_r + \epsilon/3)}\right\}.$$

Since $|I(X_{ij} \leq x)| \leq 1$, it can be proved straightforwardly by Hoeffding's inequality and the empirical process theory (Pollard 1984) that (A.2) holds. On the other hand, in view of

$$|f_{ij}(r,x) - E\{f_j(r,x)\}| = |I(X_{ij} \leq x, Y_i = y_r) - F_j(x \mid Y = y_r)p_r| \leq 1,$$

we can prove (A.3) by Bernstein's inequality and the empirical process theory (Pollard 1984). Hence, we complete the proof of Lemma 1.

LEMMA 2 *Under Conditions (C1) and (C3), for any* $\epsilon \in (0, 1/2)$ *and* $j = 1, \ldots, p$, *there exists a positive constant* c_3 *such that*

$$\Pr\{|\tilde{\zeta}_j(r) - \zeta_j(r)| \geq \epsilon\} \leq 2(2n+3)\exp(-c_3 n\epsilon^2 / R).$$

Proof of Lemma 2. Note that

$$\begin{aligned}
&|\tilde{\zeta}_j(r) - \zeta_j(r)| \\
&= \left| \int \frac{1}{n}\sum_{i=1}^{n} \frac{f_{ij}(r,x)}{\hat{p}_r} d\hat{F}_j(x) - E_X\{F_{j,r}(X)\} \right| \\
&\leq \left| \int \frac{1}{n}\sum_{i=1}^{n} \frac{f_{ij}(r,x)}{\hat{p}_r} d\{\hat{F}_j(x) - F_j(x)\} \right| + \sup_{x \in \mathbb{R}^{x_j}} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{f_{ij}(r,x)}{\hat{p}_r} - \frac{E\{f_j(r,x)\}}{p_r} \right| \\
&\equiv I_1 + I_2,
\end{aligned}$$

where $\hat{F}_j(x) = (1/n)\sum_{i=1}^{n} I(X_{ij} \leq x)$ and $F_j(x) = \Pr(X_j \leq x)$. We first consider I_1. Since

$$\left| \int \frac{1}{n}\sum_{i=1}^{n} \frac{f_{ij}(r,x)}{\hat{p}_r} d\{\hat{F}_j(x) - F_j(x)\} \right| \leq \sup_{x \in \mathbb{R}^{x_j}} \left| \hat{F}_j(x) - F_j(x) \right|$$

and (A.2), we have

$$\Pr(I_1 > \epsilon) \leq \Pr\left\{ \sup_{x \in \mathbb{R}^{x_j}} \left| \hat{F}_j(x) - F_j(x) \right| > \epsilon \right\} \leq 2(n+1)\exp(-2n\epsilon^2).$$

We next consider I_2. Write

$$\sup_{x\in\mathbb{R}_{X_j}}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{f_{ij}(r,x)}{\hat{p}_r}-\frac{E\{f_j(r,x)\}}{p_r}\right|$$

$$\leq \sup_{x\in\mathbb{R}_{X_j}}\left[\frac{|\sum_{i=1}^{n} f_{ij}(r,x)/n - E\{f_j(r,x)\}|}{\hat{p}_r}+\frac{E\{f_j(r,x)\}\,|\,\hat{p}_r - p_r\,|}{p_r\hat{p}_r}\right]$$

$$= \sup_{x\in\mathbb{R}_{X_j}}\left[\frac{|\sum_{i=1}^{n} f_{ij}(r,x)/n - E\{f_j(r,x)\}|}{\hat{p}_r}\right]+\frac{|\,\hat{p}_r - p_r\,|}{\hat{p}_r},$$

where the last equality holds due to $\sup_{x\in\mathbb{R}_{X_j}} E\{f_j(r,x)\} = \sup_{x\in\mathbb{R}_{X_j}} \Pr(X_j \leq x, Y = y_r) = p_r$.

Thereby, it follows from Lemma 1 that, for any $\epsilon > 0$,

for some constant $c_0 > 0$. In particular, under Condition (C1), the fifth inequality above holds because of

$$\left| \frac{1}{n} \sum_{i=1}^{n} W_{ri} - E(W_r) \right| = | \hat{p}_r - p_r | \geq p_r - \hat{p}_r > \frac{c_1}{R} - \frac{c_1}{2R} = \frac{c_1}{2R}.$$

On the other hand, under Condition (C3), $c_0 n^{\epsilon^2} / R < 2n^{\epsilon^2}$. Therefore,

$$\begin{aligned} &\Pr\{|\tilde{\zeta}_j(r) - \zeta_j(r)| \geq \epsilon\} \leq \Pr(I_1 + I_2 \geq \epsilon) \\ &\leq \Pr(I_1 \geq \epsilon / 2) + \Pr(I_2 \geq \epsilon / 2) \\ &\leq 2(2n+3)\exp(-c_3 n^{\epsilon^2} / R) \end{aligned}$$

for some constant $c_3 > 0$. We complete the proof of Lemma 2.

Proof of Theorem 1. We first prove part (i) of Theorem 1. By the definitions of $\hat{\zeta}_j(r)$ and $\tilde{\zeta}_j(r)$, we write

$$\hat{\zeta}_j(r) - \zeta_j(r) \equiv \frac{n}{n+1}\{\tilde{\zeta}_j(r) - \zeta_j(r)\} + \Delta,$$

where $\Delta = -\zeta_j(r)/(n+1)$. It follows from Condition (C3) and Lemma 2 that $R = O(n^{\xi})$ for $\xi + 2\kappa < 1$. By letting $\epsilon = c_4 n^{-\kappa}$ for $0 \leq \kappa < 1/2$ and some constant $c_4 > 0$, we have

$$\Pr\Big(\max_{1\le j\le p} |\hat{\tau}_{j,r}-\tau_{j,r}|\ge c_4 n^{-\kappa}\Big) \le p\Pr(|\hat{\tau}_{j,r}-\tau_{j,r}|\ge c_4 n^{-\kappa})$$

$$= p\Pr\{|\hat{\zeta}_j(r)-\zeta_j(r)|\ge c_4 n^{-\kappa}\}$$

$$= p\Pr\left\{|\tilde{\zeta}_j(r)-\zeta_j(r)|\ge \frac{(n+1)(c_4 n^{-\kappa}-|\Delta|)}{n}\right\}$$

$$\le 2(2n+3)p\exp(-c_5 n^{1-2\kappa-\xi})$$

for some constant c_5. The last inequality above holds due to $\Delta = O(1/n)$. Furthermore, in view of Condition (C2) and $\max_{j\in\mathcal{A}_r}|\hat{\tau}_{j,r}-\tau_{j,r}|\le cn^{-\kappa}$, we have

$$\min_{j\in\mathcal{A}_r}|\hat{\tau}_{j,r}|\ge \min_{j\in\mathcal{A}_r}(|\tau_{j,r}|-|\hat{\tau}_{j,r}-\tau_{j,r}|)\ge \min_{j\in\mathcal{A}_r}|\tau_{j,r}|-\max_{j\in\mathcal{A}_r}|\tau_{j,r}-\hat{\tau}_{j,r}|\ge cn^{-\kappa}.$$

Therefore,

$$\Pr(\mathcal{A}_r\subset\mathcal{A}_r)\ge \Pr\Big(\max_{j\in\mathcal{A}_r}|\hat{\tau}_{j,r}-\tau_{j,r}|\le cn^{-\kappa}\Big)\ge 1-2(2n+3)s_r\exp(-c_6 n^{1-2\kappa-\xi})$$

for some constant $c_6>0$.

Next, we prove the ranking consistency property in part (ii) of Theorem 1. It follows from Condition (C4) and Lemma 2 that

$$\Pr\left(\min_{j\in\mathcal{A}_r}|\hat{\tau}_{j,r}|-\max_{j\in\mathcal{I}_r}|\hat{\tau}_{j,r}|<\frac{m_0}{2}\right)$$

$$\le \Pr\left\{\left(\min_{j\in\mathcal{A}_r}|\hat{\tau}_{j,r}|-\max_{j\in\mathcal{I}_r}|\hat{\tau}_{j,r}|\right)-\left(\min_{j\in\mathcal{A}_r}|\tau_{j,r}|-\max_{j\in\mathcal{I}_r}|\tau_{j,r}|\right)<-\frac{m_0}{2}\right\}$$

$$\le \Pr\left\{\left|\left(\min_{j\in\mathcal{A}_r}|\hat{\tau}_{j,r}|-\max_{j\in\mathcal{I}_r}|\hat{\tau}_{j,r}|\right)-\left(\min_{j\in\mathcal{A}_r}|\tau_{j,r}|-\max_{j\in\mathcal{I}_r}|\tau_{j,r}|\right)\right|>\frac{m_0}{2}\right\}$$

$$\le \Pr\left(2\max_{1\le j\le p}|\hat{\tau}_{j,r}-\tau_{j,r}|>\frac{m_0}{2}\right)$$

$$\le 2(2n+3)p\exp(-c_7 n m_0^2/R),$$

for some constant $c_7 > 0$. The additional conditions $R\log(n) = o(nm_0^2)$ and $R\log(p) = o(nm_0^2)$ in part (ii) of Theorem 1 ensures that, there exists some $n_0 > 0$, for $n > n_0$, $p \le \exp(c_7 nm_0^2 / 2R)$ and $c_7 nm_0^2 / 2R \ge 3\log\{2(2n+3)\}$. As a result,

$$\sum_{n=n_0}^{\infty} 2(2n+3)p\exp(-c_7 nm_0^2 / R) \le \sum_{n=n_0}^{\infty} \exp[-c_7 nm_0^2 / R + c_7 nm_0^2 / 2R + \log\{2(2n+3)\}]$$

$$\le \sum_{n=n_0}^{\infty} \exp[-3\log\{2(2n+3)\} + \log\{2(2n+3)\}]$$

$$= \sum_{n=n_0}^{\infty} \{2(2n+3)\}^{-2} < \infty.$$

By Borel-Contelli Lemma,

$$\liminf_{n\to\infty} \left\{ \min_{j\in\mathcal{A}_r} \hat{\tau}_{j,r} - \max_{j\notin\mathcal{A}_r} \hat{\tau}_{j,r} \right\} > 0$$

almost surely.

The last step is to prove part (iii) of Theorem 1. Take $\gamma_n = c_8 n^{-\kappa}$ and define the event $\mathcal{M}_r = \left\{ \max_{1\le j\le p} |\hat{\tau}_{j,r} - \tau_{j,r}| \ge c_8 n^{-\kappa} \right\}$. On this event, the number of $\{j : |\hat{\tau}_{j,r}| \ge c_8 n^{-\kappa}\}$ cannot exceed the number of $\{j : |\tau_{j,r}| \ge c_8 n^{-\kappa} / 2\}$, which is bounded by $(c_8/2)^{-1} n^{\kappa} \sum_j |\tau_{j,r}|$. Therefore,

$$\Pr\left\{ |\mathcal{A}_r| \le (c_8/2)^{-1} n^{\kappa} \sum_{j=1}^{p} |\tau_{j,r}| \right\} \ge \Pr(\mathcal{M}_r) \ge 1 - 2(2n+3)p\exp(-c_9 n^{1-2\kappa} / R)$$

for some constant $c_9 > 0$. We have completed the proof of Theorem 1.

Proof of Theorem 2. Define

$$\tilde{\zeta}_j^*(r) \equiv \frac{1}{n^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\frac{I(X_{ij}^* \le X_{kj}^*, Y_i^* = y_r)}{\hat{p}_r^*},$$

$$\hat{\zeta}_j^*(r) \equiv \frac{1}{n(n+1)}\sum_{k=1}^{n}\sum_{i=1}^{n}\frac{I(X_{ij}^* \le X_{kj}^*, Y_i^* = y_r)}{\hat{p}_r^*}$$

and $\zeta_j^*(r) \equiv E_{X^*}\{F_{j,r}^*(X^*)\}$. Here $F_{j,r}^*(x) = \Pr(X_j^* \le x \mid Y^* = y_r)$. With response-selective sampling data, we notice that $\{(X_{ij}^*, Y_i^*), i = 1, \ldots, n\}$ are still independent and identically distributed copies of (X_j^*, Y^*). Thus, Lemma 1 and Lemma 2 together imply that, for any $\epsilon > 0$,

$$\Pr\{|\tilde{\zeta}_j^*(r) - \zeta_j^*(r)| \ge \epsilon\} \le 2(2n+3)\exp(-\tilde{c}_3 n\epsilon^2 / R).$$

Most importantly, we recall that the nature of response-selective sampling design, for any y in the support of Y, the conditional distribution of X given $Y = y$ is the same as that of X^* given $Y^* = y$. Then,

$$\zeta_j^*(r) = \zeta_j(r) = E_X\{F_{j,r}(X)\}.$$

As a result,

$$\begin{aligned}
&\Pr(\max_{1\le j\le p}|\hat{\tau}_{j,r}^* - \tau_{j,r}| \ge \tilde{c}_4 n^{-\kappa}) \\
&\le p\Pr(|\hat{\tau}_{j,r}^* - \tau_{j,r}| \ge \tilde{c}_4 n^{-\kappa}) \\
&= p\Pr\{|\hat{\zeta}_j^*(r) - \zeta_j(r)| \ge \tilde{c}_4 n^{-\kappa}\} \\
&= p\Pr\left\{|\tilde{\zeta}_j^*(r) - \zeta_j(r)| \ge \frac{(n+1)(\tilde{c}_4 n^{-\kappa} - |\Delta|)}{n}\right\} \\
&= p\Pr\left\{|\tilde{\zeta}_j^*(r) - \zeta_j^*(r)| \ge \frac{(n+1)(\tilde{c}_4 n^{-\kappa} - |\Delta|)}{n}\right\} \\
&\le 2(2n+3)p\exp(-\tilde{c}_5 n^{1-2\kappa-\xi}),
\end{aligned}$$

where $\Delta = -\zeta_j(r)/(n+1) = O(1/n)$ and $\tilde{c}_5$ is some positive constant. The remaining steps are similar to those of Theorem 1. We omit them here. Hence, we complete the proof of Theorem 2.

References

Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa1, P., Ladd, C., Beheshti, J., Bueno R., Gillette, M., Loda1, M., Weber, G., Mark, E., Lander3, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D., and Meyerson, M. (2001), "Classifcation of Human Lung Carcinomas by mRNA Expression Profling Reveals Distinct Adenocarcinoma Subclasses, " PNAS, 98, 13790-13795.

Chang, J., Tang, C. Y., and Wu, Y. (2013), "Marginal Empirical Likelihood and Sure Independence Feature Screening," The Annals of Statistics, 41, 2123-2148.

Chen, K. (2001a), "Parametric Models for Response-Biased Sampling," Journal of the Royal Statistical Society, Series B, 63, 775-789.

Chen, K. (2001b), "Generalized Case-Cohort Sampling," Journal of the Royal Statistical Society, Series B, 63, 791-809.

Chen, K., Lin, Y., Yao, Y., and Zhou, C. (2017), "Regression Analysis with Response-Biased Sampling," Statistica Sinica, 27, 1699-1714.

Clemmensen, L., Hastie, T., Witten, D., and Ersboll, B. (2011), "Sparse Discriminant Analysis," Technometrics, 53, 406-415.

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.

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultrahigh-Dimensional Additive Models," Journal of the American Statistical Association, 106, 544-557.

Fan, J. and Lv, J. (2008), “Sure Independence Screening for Ultrahigh Dimensional Feature Space,” Journal of the Royal Statistical Society, Series B, 70, 849-911.
Fan, J., Samworth, R., and Wu, Y. (2009), “Ultrahigh Dimensional Feature Selection: Beyond the Linear Model,” The Journal of Machine Learning Research, 10, 2013-2038.
Fan, J. and Song, R. (2010), “Sure Independence Screening in Generalized Linear Models with NP-Dimensionality,” The Annals of Statistics, 38, 3567-3604.
Golub, T. R., Slonim, D. K., and Tamayo, P. et al. (1999), “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, 286, 531-537.
He, X., Wang, L., and Hong, H. G. (2013), “Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data,” The Annals of Statistics, 41, 342-369.
Huang, D., Li, R., and Wang, H. (2014), “Feature Screening for Ultrahigh Dimensional Categorical Data with Applications,” Journal of Business & Economic Statistics, 32, 237-244.
Huang, C. Y. and Qin, J. (2010), “Nonparametric Estimation for Length-Biased and Right-Censored Data,” Biometrika, 98, 177-186.
Kim, J. P., Lu, W., Sit, T., and Ying, Z. (2013), “A Unified Approach to Semiparametric Transformation Models under General Biased Sampling Schemes,” Journal of the American Statistical Association, 108, 217-227.
Kim, J. P., Sit, T., and Ying, Z. (2016), “Accelerated Failure Time Model under General Biased Sampling Scheme,” Biostatistics, 17, 576-588.
Lawless, J.F., Wild, C.J., and Kalbfleisch, J.D. (1999), “Semiparametric Methods for Response-Selective and Missing Data Problems in Regression,” Journal of the Royal Statistical Society, Series B, 61, 413-438.

Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1990), “Handwritten Digit Recognition with A Back-Propogation Betwork in D. Touretzky, ed,” Advances in Neural Information Processing Systems, 2, 386-404.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012), “Robust Rank Correlation Based Screening,” The Annals of Statistics, 40, 1846-1877.

Li, R., Zhong, W., and Zhu, L. (2012), “Feature Screening via Distance Correlation Learning,” Journal of the American Statistical Association, 107, 1129-1139.

Lin, D. Y. (2000), “On Fitting Coxs Proportional Hazards Models to Survey Data,” Biometrika, 87, 37-47.

Luo, X. and Tsai, W. Y. (2009), “Nonparametric Estimation for Right-Censored Length-Biased Data: A Pseudo-Partial Likelihood Approach,” Biometrika, 96, 873-886.

Luo, X., Tsai, W. Y., and Xu, Q. (2009), “Pseudo-Partial Likelihood Estimators for The Cox Regression Model with Missing Covariates,” Biometrika, 96, 617-633.

Mai, Q. and Zou, H. (2013), “The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification,” Biometrika, 100, 229-234.

Mai, Q. and Zou, H. (2015), “The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method,” The Annals of Statistics, 43, 1471-1497.

Ning, J., Qin, J., and Shen, Y. (2010), “Nonparametric Tests for Right-Censored Data with Biased Sampling,” Journal of the Royal Statistical Society, Series B, 5, 609-630.

Pan, R., Wang, H., and Li, R. (2016), “Ultrahigh Dimensional Multi-Class Linear Discriminant Analysis by Pairwise Sure Independence Screening,” Journal of the American Statistical Association, 111, 169-179.

Pollard, D. (1984), Convergence of Stochastic Processes, Springer-Verlag New York Inc.

Prentice, R. L. and Pyke, R. (1979), “Logistic Disease Incidence Models with Case-Control Studies,” Biometrika, 66, 403-411.
Qin, J. (2017), Biased Sampling, Over-Identified Parameter Problems and Beyond, Springer, New York.
Scott, A. J. and Wild, C. J. (1986), “Fitting Logistic Models under Case-Control or Choice Based Sampling,” Journal of the Royal Statistical Society, Series B, 48, 170-182.
Scott, A. J. and Wild, C. J. (1997), “Fitting Regression Models to Case-Control Data by Maximum Likelihood,” Biometrika, 84, 57-71.
Sun, Y., Chan, K.C.G., and Qin, J. (2018), “Simple and Fast Overidentified Rank Estimation for Right-Censored Length-Biased Data and Backward Recurrence Time,” Biometrics, 74, 77-85.
Tsai, W. Y. (2009), “Pseudo-Partial Likelihood for Proportional Hazards Models with Biased-Sampling Data,” Biometrika, 96, 601-615.
Xu, G., Sit, T., Wang, L., and Huang, C. Y. (2017), “Estimation and Inference of Quantile Regression for Survival Data under Biased Sampling,” Journal of the American Statistical Association, 112, 1571-1586.
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.

Table 1. Screening results with different methods for Example 1.

Case	(n, p)	Method	Median	IQR	Mean(SD)	R_a	P_a
Case 1	(100,1000)	CAS^{sup}	2.00	0.00	2.04(0.28)	1.53	1.00
		MV-SIS	2.00	0.00	2.05(0.29)	1.53	1.00
		PSIS	110.50	62.00	159.34(144.92)	120.60	0.00
		KF	2.00	0.00	2.09(0.51)	1.55	1.00
	(100,3000)	CAS^{sup}	2.00	0.00	2.11(0.86)	1.56	1.00
		MV-SIS	2.00	0.00	2.20(1.49)	1.60	1.00
		PSIS	329.00	149.00	449.71(401.42)	346.95	0.00
		KF	2.00	0.00	3.48(10.92)	2.23	0.99
Case 2	(100,1000)	CAS^{sup}	6.00	22.00	32.11(83.48)	17.72	0.73
		MV-SIS	6.00	25.00	31.61(70.72)	17.49	0.73
		PSIS	132.00	100.50	199.13(180.75)	141.95	0.00
		KF	9.00	35.25	44.62(88.29)	24.49	0.64
	(100,3000)	CAS^{sup}	14.50	81.25	130.68(302.76)	72.95	0.57
		MV-SIS	14.50	109.50	132.81(317.55)	75.31	0.56
		PSIS	388.00	376.50	631.37(586.08)	454.95	0.00
		KF	29.50	128.25	174.13(366.91)	97.58	0.47
Case 3	(200,1000)	CAS^{sup}	8.00	0.00	8.08(0.53)	4.51	1.00
		MV-SIS	8.00	0.00	8.39(2.44)	4.55	1.00
		PSIS	295.50	74.75	329.46(119.62)	201.45	1.00
		KF	8.00	0.00	8.38(1.07)	4.56	1.00
	(200,3000)	CAS^{sup}	8.00	0.00	9.23(6.14)	4.66	1.00
		MV-SIS	8.00	0.00	9.21(15.23)	4.66	0.99
		PSIS	857.50	233.25	963.69(348.15)	602.35	0.00

Case	(n, p)	Method	Median	IQR	Mean(SD)	R_a	P_a
		KF	8.00	1.00	9.45(7.99)	4.71	1.00
Case 4	(200,1000)	CAS^{sup}	11.00	8.00	24.12(46.57)	7.89	0.89
		MV-SIS	13.00	25.75	45.68(93.96)	9.58	0.76
		PSIS	290.00	64.25	328.21(132.59)	199.77	0.00
		KF	16.00	21.25	31.05(67.28)	9.86	0.83
	(200,3000)	CAS^{sup}	18.00	24.00	54.83(130.22)	15.93	0.77
		MV-SIS	17.50	73.00	134.68(346.91)	20.90	0.64
		PSIS	871.00	237.00	976.59(331.95)	604.41	0.00
		KF	30.50	63.75	92.65(175.91)	23.68	0.60

Note: CAS^{sup}, our proposed method defined in (1.4); KF, Kolmogorov filter method (Mai and Zou 2013); MV-SIS, mean variance screening (Cui et al. 2015); PSIS, pairwise sure independence (Pan et al. 2016). Median, IQR, Mean and SD denote the median, the interquartile range, mean and standard deviation of minimum model size needed to include all the true predictors; R_a is the average of the ranks of all active predictors; P_a denotes the proportion of all active predictors being selected into the submodel with size $\lfloor n / \log(n) \rfloor$.

Table 2. Screening results with different methods for Example 2.

	$p = 1000$					$p = 3000$				
Method	Median	IQR	Mean(SD)	R_a	P_a	Median	IQR	Mean(SD)	R_a	P_a
Case 1										
					1.00					1.00
CAS_1	2.00	0.00	2.00(0.00)	1.50	0	2.00	0.00	2.01(0.07)	1.50	0
CAS_2	3.00	0.00	3.00(0.00)	2.00	1.00	3.00	0.00	3.05(0.57)	2.02	1.00

					0					0
					1.0					1.0
CAS_3	3.00	0.00	3.02(0.22)	2.01	0	3.00	0.00	3.06(0.52)	2.02	0
					1.0					1.0
CAS_4	5.00	0.00	5.03(0.20)	3.01	0	5.00	0.00	5.05(0.37)	3.01	0
					1.0					1.0
CAS_5	5.00	0.00	5.05(0.28)	3.01	0	5.00	0.00	5.02(0.22)	3.00	0
MV-					1.0					1.0
SIS_1	2.00	0.00	2.01(0.07)	1.50	0	2.00	0.00	2.01(0.10)	1.51	0
MV-					1.0					1.0
SIS_2	3.00	0.00	3.00(0.00)	2.00	0	3.00	0.00	3.06(0.78)	2.02	0
MV-					1.0					1.0
SIS_3	3.00	0.00	3.02(0.21)	2.01	0	3.00	0.00	3.06(0.57)	2.02	0
MV-					1.0					1.0
SIS_4	5.00	0.00	5.03(0.19)	3.01	0	5.00	0.00	5.07(0.50)	3.01	0
MV-					1.0					1.0
SIS_5	5.00	0.00	5.02(0.14)	3.00	0	5.00	0.00	5.08(0.54)	3.02	0
	110.0		173.97(181.	126.4	0.0	325.0	161.0	477.96(484.	356.2	0.0
$PSIS_1$	0	53.00	36)	6	0	0	0	06)	3	0
	124.5		184.39(158.	118.0	1.0	373.0	188.2	583.27(539.	362.5	0.0
$PSIS_2$	0	72.75	41)	7	0	0	5	17)	2	0
	119.0		184.30(177.	116.8	0.0	366.0	219.2	558.55(498.	356.2	0.0
$PSIS_3$	0	70.00	76)	2	0	0	5	63)	4	0
	154.5	123.0	248.60(226.	121.3	0.0	447.5	310.7	689.67(603.	359.0	0.0
$PSIS_4$	0	0	72)	1	0	0	5	77)	1	0
	149.0	105.5	231.88(190.	117.7	0.0	466.5	365.0	720.22(608.	372.6	0.0
$PSIS_5$	0	0	44)	8	0	0	0	42)	4	0
KF_1	2.00	0.00	2.03(0.35)	1.51	1.0	2.00	0.00	2.08(0.51)	1.54	1.0

					0					0
					1.0					1.0
KF_2	3.00	0.00	3.02(0.17)	2.01	0	3.00	0.00	3.10(0.63)	2.03	0
					1.0					1.0
KF_3	3.00	0.00	3.18(1.30)	2.06	0	3.00	0.00	3.32(2.98)	2.10	0
					1.0					1.0
KF_4	5.00	0.00	5.12(0.75)	3.03	0	5.00	0.00	5.18(1.06)	3.04	0
					1.0					1.0
KF_5	5.00	0.00	5.25(1.13)	3.05	0	5.00	0.00	5.13(1.04)	3.02	0
Case 2										
					0.9					0.9
CAS_1	2.00	0.00	3.60(5.78)	2.21	9	2.00	1.00	6.18(17.82)	3.63	8
					0.9			21.83(139.8		0.9
CAS_2	3.00	1.00	5.00(6.54)	2.72	9	3.00	1.25	0)	8.40	6
					0.9					0.9
CAS_3	3.00	1.00	5.87(10.27)	3.02	8	3.00	1.00	11.73(43.91)	5.34	6
					1.0					1.0
CAS_4	5.00	0.00	5.03(0.17)	3.01	0	5.00	0.00	5.03(0.17)	3.01	0
					1.0					1.0
CAS_5	5.00	0.00	5.04(0.21)	3.01	0	5.00	0.00	5.05(0.29)	3.01	0
MV-					0.9					0.9
SIS_1	2.00	0.00	3.35(5.47)	2.19	9	2.00	1.00	2.64(20.70)	3.86	7
MV-					0.9			26.61(183.0		0.9
SIS_2	3.00	1.00	5.47(8.04)	2.90	9	3.00	2.00	0)	9.96	5
MV-					0.9					0.9
SIS_3	3.00	1.00	6.23(11.89)	3.14	7	3.00	1.00	13.11(54.44)	5.73	5
MV-	5.00	0.00	5.03(0.17)	3.01	1.0	5.00	0.00	5.04(0.20)	3.01	1.0

SIS_4					0					0
MV-SIS_5	5.00	0.00	5.04(0.24)	3.01	1.00	5.00	0.00	5.04(0.26)	3.01	1.00
$PSIS_1$	114.00	55.50	171.88(161.49)	126.01	0.00	351.50	194.25	543.98(548.09)	390.14	0.00
$PSIS_2$	134.00	89.00	196.63(170.07)	121.64	0.00	394.50	285.25	582.03(473.90)	366.71	0.00
$PSIS_3$	133.50	103.50	210.99(196.88)	126.67	0.00	390.00	294.00	638.96(593.83)	389.02	0.00
$PSIS_4$	151.00	134.50	223.78(178.62)	117.51	0.00	419.00	263.25	635.42(560.81)	345.54	0.00
$PSIS_5$	140.00	84.75	210.27(182.78)	113.68	0.00	424.50	320.75	703.03(653.47)	352.89	0.00
KF_1	2.00	1.00	5.30(12.26)	3.22	0.97	2.00	3.00	12.05(36.17)	6.67	0.94
KF_2	3.00	3.00	7.36(12.44)	3.55	0.98	4.00	4.25	30.02(155.06)	11.27	0.93
KF_3	3.00	2.00	9.93(26.78)	4.40	0.96	4.00	5.00	18.06(50.34)	7.73	0.90
KF_4	5.00	0.00	5.08(0.28)	3.02	1.00	5.00	0.00	5.12(0.38)	3.03	1.00
KF_5	5.00	0.00	5.10(0.41)	3.03	1.00	5.00	0.00	5.09(0.41)	3.02	1.00

Note: CAS_r, our proposed category-adaptive screening method defined in (1.2); MV-SIS_r, $PSIS_r$ and KF_r denote the modified version of MV-SIS (Cui et al. 2015), PSIS (Pan et al. 2015) and KF (Mai and Zou 2013), respectively; Other notations are the same as in Table 1.

Table 3. Screening results with different methods for Case 1 in Example 3.

			$p = 1000$					$p = 3000$		
Method	Median	IQR	Mean(SD)	R_a	P_a	Median	IQR	Mean(SD)	R_a	P_a
$\epsilon_{ij} \sim \mathcal{N}(0,1)$										
CAS_1	2.00	0.00	2.00(0.00)	1.50	1.00	2.00	0.00	2.00(0.00)	1.50	1.00
CAS_2	3.00	0.00	3.00(0.00)	2.00	1.00	3.00	0.00	3.00(0.00)	2.00	1.00
CAS_3	2.00	0.00	2.00(0.00)	1.51	1.00	2.00	0.00	2.00(0.00)	1.50	1.00
CAS_4	5.00	0.00	5.00(0.00)	3.00	1.00	5.00	0.00	5.00(0.00)	3.00	1.00
$MV\text{-}SIS_1$	3.00	5.00	20.45(69.14)	11.32	0.91	3.00	12.00	32.86(151.21)	17.30	0.87
$MV\text{-}SIS_2$	3.00	0.25	3.31(0.57)	2.14	1.00	3.00	0.00	3.31(0.60)	2.14	1.00
$MV\text{-}SIS_3$	2.00	0.00	2.31(0.61)	1.69	1.00	2.00	1.00	2.40(0.72)	1.74	1.00
$MV\text{-}SIS_4$	7.00	14.75	28.55(71.58)	9.58	0.87	17.00	70.50	107.59(280.91)	27.43	0.64
$PSIS_1$	2.00	2.00	12.54(46.57)	7.02	0.95	2.00	3.00	19.24(113.41)	10.38	0.95
$PSIS_2$	3.00	1.00	3.41(0.64)	2.20	1.00	3.00	1.00	3.42(0.69)	2.19	1.00
$PSIS_3$	2.00	1.00	2.44(0.79)	1.77	1.00	2.00	1.00	2.49(0.81)	1.79	1.00
$PSIS_4$	6.00	5.00	18.75(50.69)	6.73	0.92	9.00	30.25	48.86(108.12)	13.95	0.76
KF_1	4.50	12	25.64(71.29)	14.48	0.87	6.00	32.25	45.77(129.35)	24.57	0.77

KF_2	3.00	1.00	3.44(0.66)	2.22	1.00	3.00	1.00	3.53(0.72)	2.25	1.00
KF_3	2.00	1.00	2.51(0.78)	1.81	1.00	2.00	1.00	2.58(0.84)	1.87	1.00
KF_4	13.00	26.50	42.30(86.73)	13.52	0.77	34.50	121.50	157.87(329.93)	40.41	0.52
$\epsilon_{ij} \sim t(2)$										
CAS_1	2.00	0.00	2.04(0.26)	1.52	1.00	2.00	0.00	2.22(1.91)	1.61	1.00
CAS_2	3.00	0.00	3.07(0.38)	2.02	1.00	3.00	0.00	3.18(0.99)	2.06	1.00
CAS_3	2.00	0.00	2.04(0.31)	1.52	1.00	2.00	0.00	2.06(0.29)	1.53	1.00
CAS_4	5.00	0.00	5.24(1.73)	3.05	1.00	5.00	0.00	5.22(0.69)	3.05	1.00
$MV\text{-}SIS_1$	11.00	47.25	61.04(137.81)	34.71	0.73	34.00	264.25	269.15(554.28)	155.28	0.52
$MV\text{-}SIS_2$	3.00	1.00	3.64(0.75)	2.31	1.00	3.00	1.00	3.65(0.74)	2.30	1.00
$MV\text{-}SIS_3$	2.00	0.00	2.31(0.61)	2.02	1.00	2.00	1.00	2.67(0.87)	1.92	1.00
$MV\text{-}SIS_4$	61.50	117.75	133.20(187.36)	40.21	0.38	114.50	299.50	347.47(571.14)	101.63	0.23
$PSIS_1$	125.00	173.50	199.01(198.56)	130.45	0.08	455.00	795.25	701.38(669.83)	460.69	0.00
$PSIS_2$	26.00	23.50	66.99(140.56)	31.72	0.72	76.50	66.25	128.16(217.67)	67.29	0.10
$PSIS_3$	21.00	15.00	38.19(66.18)	25.43	0.81	57.00	49.25	92.13(131.78)	61.94	0.22
$PSIS_4$	267.00	408.50	371.00(269.93)	147.95	0.01	838.50	949.25	1055.04(734.61)	421.58	0.00
KF_1	12.50	66.50	66.01(138.09)	37.43	0.67	54.00	218.50	281.48(565.98)	162.19	0.45
KF_2	3.00	1.00	3.65(0.75)	2.32	1.00	4.00	1.00	3.71(0.76)	2.36	1.00

KF_3	3.00	1.25	2.88(1.01)	2.07	1.00	2.00	1.00	2.73(0.92)	1.94	1.00
KF_4	70.00	173.00	150.10(192.69)	44.89	0.35	154.50	377.00	392.32(600.09)	114.50	0.20

Note: CAS_r, our proposed category-adaptive screening method defined in (1.2); MV-SIS_r, $PSIS_r$ and KF_r denote the modified version of MV-SIS (Cui et al. 2015), PSIS (Pan et al. 2015) and KF (Mai and Zou 2013), respectively; Other notations are the same as in Table 1.

Table 4. Screening results with different methods for Case 2 in Example 3.

	p = 1000					p = 3000				
Method $\epsilon_{ij} \sim \mathcal{N}(0,1)$	Median	IQR	Mean(SD)	R_a	P_a	Median	IQR	Mean(SD)	R_a	P_a
CAS_1	2.00	0.00	2.09(0.46)	1.55	1.00	2.00	0.00	2.04(0.20)	1.52	1.00
CAS_2	5.00	1.00	5.60(1.63)	3.17	1.00	5.00	1.00	6.94(7.27)	3.43	0.99
CAS_3	5.00	1.00	5.82(2.95)	3.19	1.00	5.00	1.00	6.54(5.17)	3.34	1.00
CAS_4	5.00	0.00	5.12(0.36)	3.03	1.00	5.00	0.00	5.10(0.39)	3.02	1.00
$MV\text{-}SIS_1$	6.00	7.00	22.93(68.94)	13.20	0.91	6.00	14.00	35.35(151.11)	19.12	0.86
$MV\text{-}SIS_2$	570.00	428.25	537.14(263.17)	232.07	0.02	1640.00	1355.75	1606.54(792.48)	682.05	0.00
$MV\text{-}SIS_3$	499.00	480.25	515.14(263.63)	215.79	0.01	1386.00	1052.75	1497.84(729.57)	633.17	0.00
$MV\text{-}SIS_4$	11.00	13.75	31.78(71.28)	11.45	0.85	20.00	69.50	110.76(280.39)	29.35	0.62
$PSIS_1$	6.00	3.00	14.97(46.51)	8.96	0.95	6.00	5.00	21.76(113.28)	12.30	0.95
$PSIS_2$	437.00	558.75	466.22(295.96)	179.41	0.04	1480.50	1567.25	1464.88(896.32)	534.17	0.03
$PSIS_3$	429.00	453.00	442.70(264.06)	165.18	0.03	1199.50	1462.75	1377.90(870.38)	509.25	0.00

$PSIS_4$	10.00	5.00	21.97(50.33)	8.60	0.91	12.00	29.00	51.97(107.84)	15.97	0.75
KF_1	8.00	13.25	28.13(71.15)	16.39	0.85	9.00	34.25	48.27(129.30)	26.41	0.76
KF_2	573.00	458.00	565.05(253.09)	249.16	0.00	1716.50	1219.50	1699.52(782.03)	735.24	0.00
KF_3	562.00	487.50	541.71(266.46)	243.98	0.00	1588.50	1194.50	1602.13(745.39)	696.45	0.00
KF_4	16.00	25.75	45.20(86.29)	15.31	0.76	38.00	120.50	160.77(329.34)	42.29	0.50
$\epsilon_{ij} \sim t(2)$										
CAS_1	2.00	1.00	2.77(1.51)	1.92	1.00	2.00	1.00	2.54(1.14)	1.79	1.00
CAS_2	12.00	19.00	32.68(62.72)	9.42	0.83	23.50	61.50	73.28(207.17)	19.23	0.62
CAS_3	10.00	15.50	28.32(55.97)	8.44	0.85	19.50	44.25	60.13(127.47)	15.82	0.70
CAS_4	5.50	2.00	5.94(1.20)	3.30	1.00	6.00	2.00	6.35(1.91)	3.38	1.00
$MV\text{-}SIS_1$	16.00	47.00	58.12(109.83)	33.92	0.64	45.00	237.25	249.23(484.22)	137.03	0.47
$MV\text{-}SIS_2$	692.50	398.25	661.39(248.62)	303.30	0.00	2113.00	1055.50	1954.31(728.78)	911.62	0.00
$MV\text{-}SIS_3$	701.50	349.00	663.33(232.70)	299.41	0.01	1996.50	1091.75	1891.19(736.37)	898.44	0.00
$MV\text{-}SIS_4$	58.00	154.25	146.83(208.66)	41.99	0.40	143.00	502.25	430.04(587.72)	121.14	0.21
$PSIS_1$	123.5	193.	213.41(227	138.	0.0	462.5	768.0	728.67(662.	469.6	0.0

	0	00	.00)	10	8	0	0	01)	1	0
	835.5	258.	790.41(175	422.	0.0	2489.	756.0	2287.17(580	1249.	0.0
$PSIS_2$	0	75	.43)	31	0	00	0	.95)	15	0
	814.5	287.	769.44(177	425.	0.0	2453.	716.0	2324.71(525	1303.	0.0
$PSIS_3$	0	25	.80)	89	0	00	0	.19)	32	0
	282.5	345.	357.78(253	83.9	0.2	1032.	1276.	1202.43(781	448.3	0.0
$PSIS_4$	0	25	.74)	5	3	00	75	.53)	3	0
		65.0	63.64(112.	37.2	0.6		262.7	262.75(493.	144.7	0.4
KF_1	19.50	0	29)	8	1	51.50	5	16)	8	3
	708.5	346.	668.62(238	311.	0.0	2098.	973.2	1965.69(715	935.7	0.0
KF_2	0	25	.19)	54	0	00	5	.97)	8	0
	727.0	317.	688.93(230	316.	0.0	2111.	1017.	1945.26(713	926.3	0.0
KF_3	0	75	.11)	70	0	50	50	.94)	3	0
		172.	153.99(204	44.3	0.3	175.0	608.0	453.50(610.	130.2	0.1
KF_4	67.50	00	.94)	9	7	0	0	48)	7	9

Note: CAS_r, our proposed category-adaptive screening method defined in (1.2); $MV\text{-}SIS_r$, $PSIS_r$ and KF_r denote the modified version of MV-SIS (Cui et al. 2015), PSIS (Pan et al. 2015) and KF (Mai and Zou 2013), respectively; Other notations are the same as in Table 1.

Table 5. Screening results with the proposed method under repeated case-control samplings in Example 3.

		$p = 1000$				$p = 3000$			
	Method	Median	Mean(SD IQR)	R_a	P_a	Median	Mean(SD IQR)	R_a	P_a
L									
Case 1: $\epsilon_{ij} \sim \mathcal{N}(0,1)$									

			p = 1000					p = 3000				
3	CAS_1	2.00	0.00	2.00(0.00)	1.50	1.00	2.00	0.00	2.00(0.00)	1.50	1.00
	CAS_2	3.00	0.00	3.00(0.00)	2.00	1.00	3.00	0.00	3.00(0.00)	2.00	1.00
	CAS_3	2.00	0.00	2.00(0.00)	1.50	1.00	2.00	0.00	2.00(0.00)	1.50	1.00
	CAS_4	5.00	0.00	5.00(0.00)	3.00	1.00	5.00	0.00	5.00(0.00)	3.00	1.00
5	CAS_1	2.00	0.00	2.00(0.00)	1.50	1.00	2.00	0.00	2.00(0.00)	1.50	1.00
	CAS_2	3.00	0.00	3.00(0.00)	2.00	1.00	3.00	0.00	3.00(0.00)	2.00	1.00
	CAS_3	2.00	0.00	2.00(0.00)	1.50	1.00	2.00	0.00	2.00(0.00)	1.50	1.00
	CAS_4	5.00	0.00	5.00(0.00)	3.00	1.00	5.00	0.00	5.00(0.00)	3.00	1.00
Case 1: $\epsilon_{ij} \sim t(2)$											
3	CAS_1	2.00	0.00	2.00(0.00)	1.50	1.00	2.00	0.00	2.00(0.00)	1.50	1.00
	CAS_2	3.00	0.00	3.00(0.00)	2.00	1.00	3.00	0.00	3.00(0.00)	2.00	1.00
	CAS_3	2.00	0.00	2.00(0.00)	1.50	1.00	2.00	0.00	3.00(0.00)	1.50	1.00
	CAS_4	5.00	0.00	5.00(0.00)	3.00	1.00	5.00	0.00	5.00(0.00)	3.00	1.00
5	CAS_1	2.00	0.00	2.00(0.00	1.50	1.00	2.00	0.00	2.00(0.00	1.50	1.00

			p = 1000				p = 3000			
			0	)	0	0		0	)	0	0
			0.0	3.00(0.00	2.0	1.0		0.0	3.00(0.00	2.0	1.0
	CAS_2	3.00	0	)	0	0	3.00	0	)	0	0
			0.0	2.00(0.00	1.5	1.0		0.0	2.00(0.00	1.5	1.0
	CAS_3	2.00	0	)	0	0	2.00	0	)	0	0
			0.0	5.00(0.00	3.0	1.0		0.0	5.00(0.00	3.0	1.0
	CAS_4	5.00	0	)	0	0	5.00	0	)	0	0
Case 2: $\epsilon_{ij} \sim \mathcal{N}(0,1)$											
			0.0	2.00(0.00	1.5	1.0		0.0	2.00(0.00	1.5	1.0
3	CAS_1	2.00	0	)	0	0	2.00	0	)	0	0
			0.0	5.00(0.00	3.0	1.0		0.0	5.01(0.14	3.0	1.0
	CAS_2	5.00	0	)	0	0	5.00	0	)	0	0
			0.0	5.01(0.07	3.0	1.0		0.0	5.00(0.00	3.0	1.0
	CAS_3	5.00	0	)	0	0	5.00	0	)	0	0
			0.0	5.00(0.00	3.0	1.0		0.0	5.00(0.00	3.0	1.0
	CAS_4	5.00	0	)	0	0	5.00	0	)	0	0
			0.0	2.00(0.00	1.5	1.0		0.0	2.00(0.00	1.5	1.0
5	CAS_1	2.00	0	)	0	0	2.00	0	)	0	0
			0.0	5.00(0.00	3.0	1.0		0.0	5.01(0.14	3.0	1.0
	CAS_2	5.00	0	)	0	0	5.00	0	)	0	0
			0.0	5.00(0.00	3.0	1.0		0.0	5.00(0.00	3.0	1.0
	CAS_3	5.00	0	)	0	0	5.00	0	)	0	0
			0.0	5.00(0.00	3.0	1.0		0.0	5.00(0.00	3.0	1.0
	CAS_4	5.00	0	)	0	0	5.00	0	)	0	0
Case 2:											

			p = 1000				p = 3000			
$\epsilon_{ij} \sim t(2)$										
3	CAS_1	2.00	0.00	2.07(0.32)	1.54	1.00	2.00	0.00	2.05(0.25)	1.53	1.00
	CAS_2	5.00	0.00	5.25(0.57)	3.06	1.00	5.00	0.00	5.26(1.00)	3.06	1.00
	CAS_3	5.00	0.00	5.23(0.91)	3.05	1.00	5.00	0.00	5.20(0.75)	3.04	1.00
	CAS_4	5.00	0.00	5.14(0.40)	3.03	1.00	5.00	0.00	5.18(0.56)	3.05	1.00
5	CAS_1	2.00	0.00	2.01(0.10)	1.51	1.00	2.00	0.00	2.02(0.14)	1.51	1.00
	CAS_2	5.00	0.00	5.03(0.17)	3.01	1.00	5.00	0.00	5.05(0.27)	3.01	1.00
	CAS_3	5.00	0.00	5.05(0.32)	3.01	1.00	5.00	0.00	5.04(0.21)	3.01	1.00
	CAS_4	5.00	0.00	5.02(0.12)	3.00	1.00	5.00	0.00	5.07(0.36)	3.02	1.00

Note: L is the repeated times of case-control sampling. Other notations are the same as in Table 2.

Table 6. Classification accuracy of different methods for US post-office handwritten digits data.

Method	CAS,	MV-SIS,	PSIS,	KF,
CA (%)	91.73	90.43	90.53	90.18

Table 7. Classification accuracy of different methods for the lung carcinomas data.

	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$
Method	CA (%)	CA (%)	CA (%)
CAS_r	93.61(2.48)	93.83(3.00)	93.90(3.34)
$MV\text{-}SIS_r$	93.47(2.51)	93.61(2.98)	93.45(3.46)
$PSIS_r$	90.77(2.82)	91.45(3.64)	91.70(5.01)
KF_r	93.47(2.35)	93.70(3.25)	93.53(3.85)
CAS^{sup} + SDA	87.35(3.69)	87.45(3.79)	89.35(5.13)
MV-SIS + SDA	91.78(2.96)	92.53(3.46)	92.65(3.94)
PSIS + SDA	86.43(3.54)	87.28(3.45)	87.85(5.31)
KF + SDA	81.46(6.64)	84.62(4.99)	86.48(5.37)

Note: The standard deviations of the classification accuracy are presented in brackets.

Fig. 1 US post-office handwritten digits

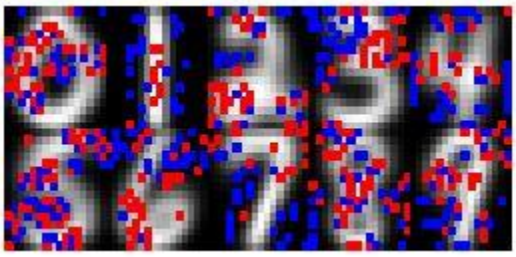

Fig. 2 Coefficients of the ten l_1-regularized binary logistic regression models by the proposed method CAS_r, displayed as images for each digit class. The gray background image is the average training example for that class. Superimposed in two colors (red for positive, blue for negative) are the nonzero coefficients for each class.