mathematics

Article

Quantile-Composited Feature Screening for Ultrahigh-Dimensional Data

Shuaishuai Chen [1] **and Jun Lu** [2,*]

[1] School of Mathematics, Shandong University, Jinan 250100, China
[2] School of Science, National University of Defense and Technology, Changsha 410000, China
* Correspondence: penguin1020@foxmail.com

Abstract: Ultrahigh-dimensional grouped data are frequently encountered by biostatisticians working on multi-class categorical problems. To rapidly screen out the null predictors, this paper proposes a quantile-composited feature screening procedure. The new method first transforms the continuous predictor to a Bernoulli variable, by thresholding the predictor at a certain quantile. Consequently, the independence between the response and each predictor is easy to judge, by employing the Pearson chi-square statistic. The newly proposed method has the following salient features: (1) it is robust against high-dimensional heterogeneous data; (2) it is model-free, without specifying any regression structure between the covariate and outcome variable; (3) it enjoys a low computational cost, with the computational complexity controlled at the sample size level. Under some mild conditions, the new method was shown to achieve the sure screening property without imposing any moment condition on the predictors. Numerical studies and real data analyses further confirmed the effectiveness of the new screening procedure.

Keywords: feature screening; discriminative analysis; quantile-composited

MSC: 62H20; 62H30

Citation: Chen, S.; Lu, J. Quantile-Composited Feature Screening for Ultrahigh-Dimensional Data. *Mathematics* **2023**, *11*, 2398. https://doi.org/10.3390/math11102398

Academic Editors: Xu Guo and Xuehu Zhu

Received: 24 April 2023
Revised: 12 May 2023
Accepted: 19 May 2023
Published: 22 May 2023

1. Introduction

With the rapid advancements in science and technology, ultrahigh-dimensional data are becoming increasingly common across various fields of scientific research: these include, but are not limited to, biomedical imaging, neuroscience, tomography, and tumor classifications, where the number of variables or parameters can exponentially increase with the sample size. In such a situation, an important task is to recover the important features from thousands or even millions of predictors.

In order to rapidly lower the huge dimensionality of data to an acceptable size, Fan and Lv [1] introduced the method of sure independence screening, which ranks the importance of predictors according to their marginal utilities. Since then, a series in the literature has been devoted to this issue, in various scenarios, which can basically be divided into two groups: the model-based and the model-free methods. For the former, the typical literature includes, but is not limited to, Wang [2], Chang et al. [3], and Wang and Leng [4] for linear models, Fan et al. [5] for additive models, and Fan et al. [6] and Liu et al. [7] for varying coefficients models, amongst others. Model-based methods are computationally efficient, but can suffer from the risk of model misspecification. To avoid such a risk, researchers developed the model-free methods, without specifying a concrete model. For example, Zhu et al. [8] proposed a screening procedure named SIRS for the multi-index model; Li et al. [9] introduced a sure screening procedure via the distance correlation called DCS; for the heterogeneous data, He et al. [10] developed a quantile-adaptive screening method; Lin et al. [11] proposed a novel approach, dubbed Nonparametric Ranking Feature Screening (NRS), leveraging the local information flows of the predictors;

Lu and Lin [12] developed a conditional model-free screening procedure, utilizing the conditional distance correlation; and Tong et al. [13] proposed a model-free conditional feature screening method with FDR control. Additionally, Ref. [14] recently introduced a data-adaptive threshold selection procedure with error rate control, which is applicable to most kinds of popular screening methods. Ref. [15] proposed a feature screening method for the interval-valued response.

The literature listed above mainly concentrated on the continuous response; however, ultrahigh-dimensional grouped data, in which the label of a sample can be seen as a categorical response, are also very frequently encountered in many scientific research fields—specifically, for biostatisticians who work on multi-class categorical problems. For example, in the diagnosis of tumor classification, researchers need to judge the type of tumor, according to the gene expression level. If we do not reduce the dimension of the predictors, the established classifier will behave as poorly as random guessing, due to the diverging spectra and accumulation of noise (Fan et al. [16]); therefore, it makes sense to screen out the null predictors before further analysis. The following are the existing works that have made some progress on this issue. Huang et al. [17] proposed a screening method based on Pearson chi-square statistics, for discrete predictors. Pan et al. [18] set the maximal mean difference for each pair of classes as a ranking index and, based on this, proposed a corresponding screening procedure. Mai and Zou [19] built a Kolmogorov–Smirnov type distance, to measure the dependence between two variables, and used it as a filter for screening out noise predictors. Cui et al. [20] proposed a screening method via measuring the distance of the distribution of the subgroup from the whole distribution. Recently, Xie et al. [21] established a category-adaptive screening procedure, by calculating the difference between the conditional distribution of the response and the marginal one. All these aforementioned methods were clearly motivated, and have been examined effectively for feature screening in different settings.

In this paper, we propose a new robust screening method for ultrahigh-dimensional grouped data. Our research was partly motivated by an empirical analysis of a leukemia dataset, consisting of 72 observations and 3571 genes, of which 47 were acute lymphocytic leukemia (ALL), and 25 were acute myelogenous leukemia (AML). Figure 1 plots the density function of the first 20 features selected from the 3571 genes of the 47 ALLs, from which it can be seen that all of them are far from being a regular distribution, most of them have sharp peaks and heavy tails (e.g., gene 9 and gene 12), and some of them are even multi-modal (e.g., gene 6 and gene 8), although these samples are from the same ALL group. This phenomenon challenges most of the existing methods. For example, the method in Pan et al. [18] might fail, if data are not normally distributed, and the method in Xie et al. [21] might lose efficiency when the distribution of a predictor is multi-modal. It is known that quantile-based statistics are not sensitive to outliers and heavy-tailed distributed data; thus, it was expected that the quantile-based screening method would be robust against heterogeneous data. Furthermore, compared to point estimation, quantile-based statistics can usually provide a more detailed picture of a predictor at different quantile levels. Motivated by the above discussion, we propose a quantile-composited screening approach, by aggregating the distributional information over many quantile levels. The basic idea of our method is straightforward. If X_j has no contribution to predicting the category of an outcome variable, denoted by Y, at the τ-th quantile level, the conditional quantile function of X_j given Y should be equal to the unconditional one, i.e, $q_{X_j|Y}(\tau) = q_{X_j}(\tau)$. Moreover, if X_j and Y are independent, we have $q_{X_j|Y}(\tau) = q_{X_j}(\tau)(a.s.)$ for all $\tau \in (0,1)$, where a.s. means 'almost surely'. Thus, the equality $q_{X_j|Y}(\tau) = q_{X_j}(\tau)$ plays a key role in measuring the independence between Y and X_j. To quantify this kind of independence, we show that $q_{X_j|Y}(\tau) = q_{X_j}(\tau)$ for a given τ is equivalent to the independence between the index variable $I(X_j - q_{X_j}(\tau) > 0)$ and the label variable Y. Then, the equality between $q_{X_j|Y}(\tau)$ and $q_{X_j}(\tau)$ is converted to testing the independence between two discrete variables, which can be easily checked by the Pearson chi-square test statistics. Finally, we aggregate all the discriminative information over the whole distribution in an efficient way, based on which, we establish the corresponding screening procedure. Our newly proposed screening method enjoys the following salient features. First of all, compared to the existing methods, it is robust against non-normal data, which are very common in high dimensions. Secondly, it is model-free, in the sense that we do not need to assume a specific statistical model, such as the linear or quadratic discriminant analysis model, between the predictor and the outcome variable. Thirdly, its ranking index has a very simple form, and the computational complexity is controlled at the sample size level, so that the proposed screening method can be implemented very quickly. In addition, as a by-product, our new method is invariant, in regard to the monotonic transformations of the data.

The rest of the paper is organized as follows. Section 2 gives the details of the quantile-composited screening procedure, including the methodological development, theoretical properties, and some extensions. Section 3 provides convincing numerical results and two real data analyses. Technical proofs of the main results are deferred to Appendix A.

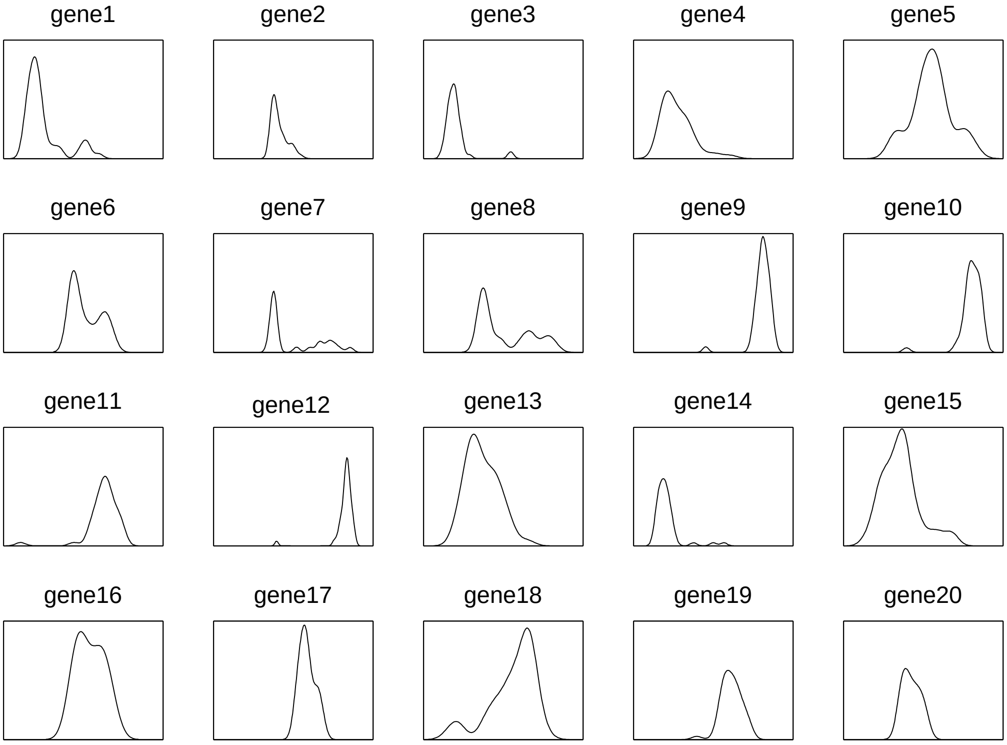

Figure 1. The sample histograms of the 47 ALLs corresponding to the first 20 features selected from 3571 genes.

2. A New Feature Screening Procedure

Let $\mathbf{X} = \{X_1, \cdots, X_p\}$ be the p-dimensional predictor, and without loss of generality, let $Y \in \{1, \cdots, K\}$ be the outcome variable indicating which group $\mathbf{X}$ belongs to, where K is allowed to grow with the sample size at some certain rate. Define the index set of active predictors corresponding to quantile level τ as

$$\mathcal{A}_\tau = \{1 \leq j \leq p : q_{X_j|Y}(\tau) \text{ functionally depends on } Y\}, \tag{1}$$

where $q_{X_j|Y}(\tau) = \inf\{t : P(X_j \leq t|Y) \geq \tau\}$. Denote by $|\mathcal{A}_\tau|$ the cardinality of $\mathcal{A}_\tau$; $|\mathcal{A}_\tau|$ is usually less than the sample size n under the sparsity assumption. Denote by $q_{X_j}(\tau) = \inf\{t : P(X_j \leq t) \geq \tau\}$ the τ-th quantile of X_j. Intuitively, if $q_{X_j|Y}(\tau)$ does not functionally depend on Y, it should be the case that $q_{X_j|Y}(\tau) = q_{X_j}(\tau)$ for all Y: in other words, X_j has no ability to predict its label Y at the quantile level τ. On the other hand, if $q_{X_j|Y}(\tau)$ is far away from $q_{X_j}(\tau)$ for some Y, X_j will be helpful for predicting the category of Y. Hence, the difference between $q_{X_j|Y}$ and $q_{X_j}(\tau)$ determines whether X_j is a contributive predictor at the τ-th quantile level. The following lemma was of central importance to our methodological development.

Lemma 1. *Let Y be the outcome variable, and let X be a continuous variable; then, we have two conclusions:*

(1) $q_{X_j|Y}(\tau) = q_{X_j}(\tau)$ a.s. if and only if the Bernoulli variable $I\{X > q_{X_j}(\tau)\}$ and Y are independent, where $I\{\cdot\}$ is the indicator function;
(2) $q_{X_j|Y}(\tau) = q_{X_j}(\tau)$ a.s. for $\forall\tau \in (0,1)$ if Y and X_j are independent.

The proof of this lemma is presented in Appendix A. Conclusions (1) and (2) imply that the independence between X_j and Y for $\forall\tau \in (0,1)$ is equivalent to the independence between $I\{X > q_{X_j}(\tau)\}$ and Y; consequently, it is natural to apply the Pearson chi-square statistics, to measure the independence between them. Let $Z_j(\tau) = I\{X_j > q_{X_j}(\tau)\}$, $\pi_{yk} = P(Y = k)$, $\pi_{jb}(\tau) = P(Z_j(\tau) = b)$, $\pi_{yk,jb}(\tau) = P(Y = k, Z_j(\tau) = b)$. Then, the dependence of X_j on the response Y, at quantile level τ, can be evaluated by

$$Q_j(\tau) = \sum_{k=1}^{K}\sum_{b=0}^{1} \frac{(\pi_{yk}\pi_{jb}(\tau) - \pi_{yk,jb}(\tau))^2}{\pi_{yk}\pi_{jb}(\tau)}. \tag{2}$$

Clearly, $Q_j(\tau) = 0$ iff $Z_j(\tau)$ and Y are independent.

$Q_j(\tau)$ provides a way to identify whether X_j is active at quantile level τ. However, it is not easy to determine the informative quantiles for every predictor. Moreover, the active predictors could be contributive at many quantiles instead of a single one. For these reasons, we propose a quantile-composited screening index, which makes an integration for $Q_j(\tau)$ at the interval $(0,1)$. More specifically, the ranking index is defined as

$$Q_j = \int_{\alpha}^{1-\alpha} Q_j(\tau)w_j(\tau)d\tau, \tag{3}$$

where $w_j(\tau)$ is some positive weight function, and α is a value tending to 0 at some certain rate related to the sample size, which will be specified in the next section. Note that Q_j avoids making integration at the endpoints 0 and 1, because $Q_j(\tau)$ could be ill-defined at the two points. Theoretically, $Q_j = 0$ if X is independent of Y, regardless of the choice of $w_j(\tau)$, which is easy to prove according to Lemma 1. According to the above analysis, $Q_j(\tau)$ is always non-negative for $\forall\tau \in (0,1)$, and will equal 0 if X_j is independent of Y.

For the choice of weight $w_j(\tau)$, the different settings will lead to different values of Q_j. For example, a naive setting is $w_j(\tau) = 1$ for $\tau \in (0,1)$, which means that all $Q_j(\tau)$ are treated equally. Clearly, this is not a good option. Intuitively, if X_j is active, $Q_j(\tau)$ should be large for some τ in $(0,1)$. Then, we should place more weight on these quantile levels. For this reason, we set $w_j(\tau) = Q_j(\tau)/\int_{\alpha}^{1-\alpha} Q_j(\tau)d\tau$; then, the resultant Q_j has the following form:

$$Q_j = \int_{\alpha}^{1-\alpha} Q_j^2(\tau)d\tau \Big/ \int_{\alpha}^{1-\alpha} Q_j(\tau)d\tau. \tag{4}$$

In addition, for the precise-definition of Q_j, we restrict $Q_j = 0$ when $Q_j(\tau) = 0$ for all $\tau \in (0,1)$.

In the following, we give the estimation of Q_j. Suppose $\{X_i, Y_i\}_{i=1}^{n}$ is a set of i.i.d samples from (X, Y), where i.i.d means independent and identically distributed. Let $\hat{q}_{X_j}(\tau)$ be the τth sample quantile of X_j and $Z_{ij}(\tau) = I\{X_{ij} > \hat{q}_{X_j}(\tau)\}$, π_{yk}, $\pi_{jb}(\tau)$ and $\pi_{yk,jb}(\tau)$ can be estimated as $\hat{\pi}_{yk} = n^{-1}\sum_{i=1}^{n} I\{Y_i = k\}$, $\hat{\pi}_{jb}(\tau) = n^{-1}\sum_{i=1}^{n} I\{Z_{ij}(\tau) = b\}$ and

$\hat{\pi}_{yk,jb}(\tau) = n^{-1}\sum_{i=1}^{n} I\{Y_i = k\}I\{Z_{ij}(\tau) = b\}$, respectively. Then, by plug-in method, $Q_j(\tau)$ is estimated as

$$\widehat{Q}_j(\tau) = \sum_{k=1}^{K}\sum_{b=0}^{1} \frac{(\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau) - \hat{\pi}_{yk,jb}(\tau))^2}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)}, \tag{5}$$

and Q_j is estimated as

$$\widehat{Q}_j = \int_{\alpha}^{1-\alpha} \widehat{Q}_j^2(\tau)d\tau \Big/ \int_{\alpha}^{1-\alpha} \widehat{Q}_j(\tau)d\tau.$$

Regarding $\widehat{Q}_j(\tau)$, we make the following remarks:

Remark 1. *1.* *If $q_{X_j|Y}(\tau) = q_{X_j}(\tau)$, $n\widehat{Q}_j(\tau)$ follows the χ^2 distribution with $K-1$ degrees of freedom [22].*

2. *$\widehat{Q}_j$ is invariant to any monotonic transformation on predictors, because $Z_j(\tau)$ is free of the monotonic transformation on X_j.*
3. *The computation of $Q_j(\tau)$ involves the integration of τ. We can calculate it by an approximate numerical method as*
$$\widehat{Q}_j = \sum_{i=1}^{s} \widehat{Q}_j^2(i/s) \Big/ \sum_{i=1}^{s} \widehat{Q}_j(i/s).$$
4. *The choice of s. Intuitively, a large s will make the approximation of integration more accurate. However, our method aims to efficiently separate the active predictors from the null ones, instead of getting an accurate estimate of Q_j. Figure 2 displays the density curves of marginal utilities of active and inactive predictors versus different choices of s with Example 2 in Section 3. It can be seen that the choice of s does not affect the distribution of either active predictors or inactive ones.*
5. *Figure 2 also shows that the gap between the indices of active predictors and inactive ones is clear, which means the proposed method is efficient at separating the influential predictors from the inactive ones well. Moreover, it can also be observed that the marginal utilities of active predictors are, with a smaller variance, comparable to those of inactive ones, which implies that the new method is sensitive to the active predictors.*

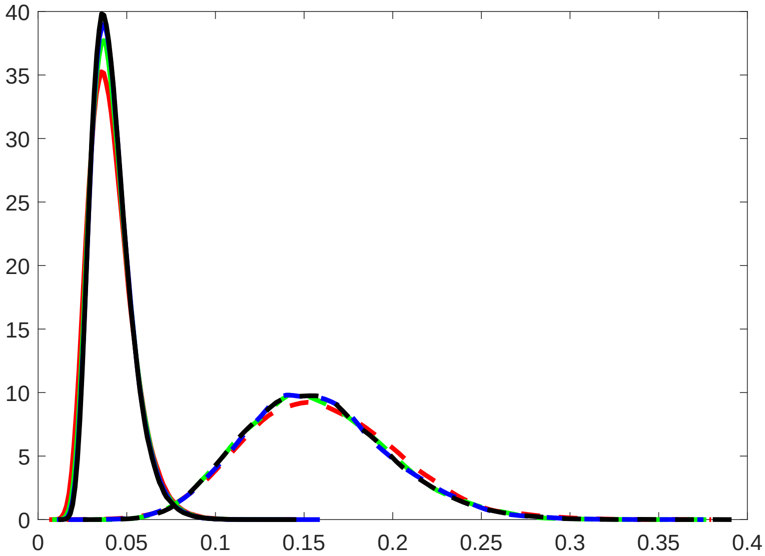

Figure 2. Density curves of marginal utilities of active predictors (solid line) and inactive ones (dashed line) for $s = 10$(red), 20(green), 50(blue), 100(black). The simulations were repeated 1000 times, using the model in Example 2 in Section 3 with a balanced response and $r = 0.05$.

With the estimation of $\widehat{Q}_j$, the index set of active predictors can be estimated as

$$\widehat{\mathcal{A}} = \{1 \leq j \leq p : \widehat{Q}_j \geq cn^{-\eta}\},$$

where c and η are two predetermined thresholding values. In practice, we usually take a hard threshold criterion, to determine the submodel as

$$\widehat{\mathcal{A}}^{\star} = \{1 \leq j \leq p : \widehat{Q}_j \text{ is among the top } d_n \text{ largest of all}\},$$

where d_n is a predetermined threshold value. We call the above quantile-composited screening procedure, based on $\widehat{Q}_j$ as QCS.

2.1. Theoretical Properties

This section provides the sure screening property of the newly proposed method, which guarantees the effectiveness of the newly proposed method. The corresponding technical details of the proof can be found in Appendix A.

We first prove the consistency of $\widehat{Q}_j(\tau)$. To this end, we require the following condition.

(C1): There exist two constants $c_1, c_2 (c_1 < c_2)$, such that $c_1/K < \pi_{yk} < c_2/K$ for $k \in \{1, 2, \ldots, K\}$ with $K = O(n^{\gamma})$.

Condition C1 requires that the sample size of each subgroup can be neither too small nor too large. The condition $K = O(n^{\gamma})$ allows that the number of categories can diverge to infinity at some certain rate, with the increase of sample size. The following theorem states the consistency of $\hat{Q}_j(\tau)$.

Theorem 1. *For a given quantile $\tau \in (\alpha, 1-\alpha)$, under condition (C1),*

$$P\big(|\hat{Q}_j(\tau) - Q_j(\tau)| \geq cn^{\eta}\big) = K\exp(\{-n^{1-2\gamma-2\eta} + n^{\eta-2\gamma}/\bar{\tau}\}), \tag{6}$$

where $\bar{\tau} = \min(\tau, 1-\tau)$.

This theorem shows that the consistency of $\widehat{Q}_j(\tau)$ can be guaranteed under suitable conditions. In addition, it reminds us that we cannot select the quantiles either very close to zero or to one, because the items $\bar{\tau}$ would collapse to zero, which would make the consistency of $\widehat{Q}_j(\tau)$ problematic. Based on the above theorem, the following theorem provides the consistency of $\widehat{Q}_j$.

Corollary 1. *According to the conditions in Theorem 1, if $\bar{\tau} = O(n^{\eta-1})$,*

$$P\left(\max_{1 \leq j \leq p} |\widehat{Q}_j - Q_j| > cn^{-\eta}\right) \leq pK\exp(-O\{n^{1-2\eta-2\gamma}\}). \tag{7}$$

This theorem states that the gap between $\widehat{Q}_j$ and Q_j will disappear with probability tending to 1 as $n \to \infty$. This theorem also shows that our method can address the dimensionality of order $o\left(\exp\left\{n^{(1-2\gamma-2\eta)}\right\}\right)$.

In the following, we provide the sure screening property of our method.

Theorem 2. *Sure screening property: let $\mathcal{A} = \{1 \leq j \leq p : Q_j > 0\}$; then, under condition (C1) and the following condition, $\min_{j \in \mathcal{A}} Q_j \geq 2cn^{-\eta}$,*

$$P\left(\mathcal{A} \subseteq \widehat{\mathcal{A}}\right) \geq 1 - s_n K\exp\left\{-O(n^{1-2\eta-2\gamma})\right\},$$

where s_n is the cardinality of $\mathcal{A}$.

2.2. Extensions

Up to this point, the new methods have been designed for ultrahigh-dimensional categorical data. In this section, to make the proposed methods applicable in more settings, we give two natural extensions for our method, and in the next section, we use some numerical simulation, to illustrate the effectiveness of these extensions.

Extension to Genome-Wide Association Studies. We first apply our method to the typical case of the genome-wide association studies (GWAS), where the predictors are single-nucleotide polymorphisms (SNPs) in three classes, denoted by $\{AA, Aa, aa\}$, and the response is continuous. Our strategy for this problem is straightforward: define the sample space $\{AA, Aa, aa\}$ as $\{-1, 0, 1\}$, respectively; then, the marginal utility of X_j at quantile level τ is defined as

$$\widehat{Q}_{1,j}(\tau) = \sum_{b=-1}^{1}\sum_{k=0}^{1}\frac{(\hat{\pi}^1_{yk}(\tau)\hat{\pi}^1_{jb} - \hat{\pi}^1_{yk,jb}(\tau))^2}{\hat{\pi}^1_{yk}(\tau)\hat{\pi}^1_{jb}}, j = 1, \cdots, p, \tag{8}$$

where $Y_i(\tau) = I(Y_i > \hat{q}_Y(\tau))$, $\hat{\pi}^1_{yk}(\tau) = n^{-1}\sum_{i=1}^n I(Y_i(\tau) = k)$, $\hat{\pi}^1_{jb} = n^{-1}\sum_{i=1}^n I(X_{ij} = b)$, $\hat{\pi}^1_{yk,jb}(\tau) = n^{-1}\sum_{i=1}^n I(Y_i(\tau) = k, X_{ij} = b)$ for $b = -1, 0, 1$.

Extension to additive models. We can extend our method to the model in which both the response and predictors are continuous. To make our method applicable, we first slice the predictors into several segments, according to some threshold values. For example, taking the quartiles of the predictor as the cut points, then the predictors are transformed to a balanced four-categorical variable. Specifically, let $(\widehat{Q}_{j1}, \cdots, \widehat{Q}_{jN-1})$ be N percentiles of X_j, and define $X^*_{ij} = bI\{\widehat{Q}_{jb} \leq X_{ij} < \widehat{Q}_{j(b+1)}\}$, where $b = 0, 1, \cdots, N-1$; here, we define $\widehat{Q}_{j0} = \min_i X_{ij}$ and $\widehat{Q}_{j(N)} = \max_i X_{ij}$. Then, similar to (9), we define the marginal utility of X_j at quantile level τ as

$$\widehat{Q}_{2,j}(\tau) = \sum_{b=0}^{N-1}\sum_{k=0}^{1}\frac{(\hat{\pi}^*_{yk}(\tau)\hat{\pi}^*_{jb} - \hat{\pi}^*_{yk,jb}(\tau))^2}{\hat{\pi}^*_{yk}(\tau)\hat{\pi}^*_{jb}}, j = 1, \cdots, p, \tag{9}$$

where $\hat{\pi}^*_{yk}(\tau) = \hat{\pi}^1_{yk}(\tau)$, $\hat{\pi}^*_{jb} = n^{-1}\sum_{i=1}^n I(X^*_{ij} = b)$, $\hat{\pi}^*_{yk,jb}(\tau) = n^{-1}\sum_{i=1}^n I(Y_i(\tau) = k, X^*_{ij} = b)$ for $b = 0, 1, \cdots, N-1$.

3. Numerical Studies

3.1. General Settings

For this section, we first conducted some Monte Carlo simulations, to compare our method to those of several competitors. Then, we applied our screening procedure to two real data examples.

We compared our method to: (1) MV-based sure independence screening (MVS) [20], which can be seen as the weighted average of the Cramér–von Mises distances between the conditional distribution function of X given $Y = k$ and the unconditional distribution function of X; (2) distance correlation–sure independence screening (DCS) [9], which employs distance correlation as a measure to evaluate the importance of each predictor; (3) category-adaptive variable screening (CAS) [21], which screens the inactive predictor, by comparing its marginal distribution to its marginal conditional one; (4) Kolmogorov filter screening (KFS) [19], which filters the inactive predictors, by comparing the Kolmogorov distance between the conditional distribution and the unconditional one. Note that DCS is not efficient for categorical variables. Thus, we transferred the categorical variable into a multivariate dummy variable, with the i-th coordinate equal to 1, and other coordinates equal to 0, where i was the category of a sample, e.g., we transformed $Y = 3$ into $(0, 0, 1, 0, 0)$ if Y was five-category.

Throughout the simulation, we repeated each experiment 1000 times, and always set $s = 50$. To fairly evaluate the performances of the different methods, the following criteria were employed: (1) MS: the minimum model size of the selected models that are required to have a sure screening; (2) $\mathcal{P}_s$: the percentage of submodels that contain all active predictors under a predetermined model size d_n over 1000 replications. We let MS(t) be the result of the t-th numerical experiment, and denoted by MS$_\alpha$ the α-level quantile of

{MS(1),$\cdots$,MS(1000)}; then, we reported the median of MS (MMS), the interquartile range (IQR) of MS, and the extreme percentile range (EPR) of MS, namely:

$$\begin{aligned} &\text{MMS} = \text{MS}_{0.5}, \text{IQR} = \text{MS}_{0.75} - \text{MS}_{0.25}, \text{EPR} = \text{MS}_{0.95} - \text{MS}_{0.05}, \\ &\mathcal{P}_s = \frac{1}{1000}\sum_{t=1}^{1000} I(\text{MS}(t) \le s) \times 100\%. \end{aligned}$$

We considered $d_n = [n/\log n]$ and $s = 2[n/\log n]$ for a small and large model size, respectively, where $[a]$ was the integer part of a. By the two criteria, a well-behaved screening method should have small MS, but with $\mathcal{P}_a$ close to 1.

3.2. Monte Carlo Simulations

Example 1. *Data were generated in the following manner. For a given $Y = k$, the p-dimensional random vector of $X|\{Y = k\}$ was generated from a mixture distribution $(1 - r)Z + rW$, where $X \sim N(\mu_k, I_p)$, with I_p being the identity matrix and $\mu_k = (\mu_{k1}, \cdots, \mu_{kp})^\top$; W was a random vector, with each component being an independent student's t-distribution with one degree of freedom. Here, r was used to check the robustness of our method against the heavy-tailed distribution. We considered $r = 0.05$ and 0.15, representing, respectively, a low and high proportion of the heavy-tailed samples in the data. The categorical variable Y was set to be binary and multi-category, with both balanced and imbalanced design, by the following scenarios:*

Case 1. *$P(Y = 1) = P(Y = 2) = 0.5$, $\mu_1 = (1.5, 0, \cdots, 0)^\top$ and $\mu_2 = (0, 1.5, 0, \cdots, 0)^\top$.*

Case 2. *The same setup as Case 1, except that $P(Y = 1) = 1/3$ and $P(Y = 2) = 2/3$.*

Case 3. *$P(Y = k) = 1/K$ for $k = 1, \cdots, 8$, and $\mu_k = (0_{k-1}^\top, 2, 0_{p-k}^\top)^\top$ for $k = 1, \cdots, 8$, where 0_d represented a d-dimensional zero-valued vector.*

Case 4. *The same setup as Case 1, except that $P(Y = k) = 2[1 + (k - 1)/(K - 1)]/3K$.*

The numerical results are reported in Table 1, by setting $(n, p) = (50, 1000)$ for $K = 2$, and $(n, p) = (160, 2000)$ for $K = 8$. From this table, it can be seen that the QCS, MVS, CAS, and KFS performed comparably well with both $\mathcal{P}_{d_n}$ and $\mathcal{P}_{2d_n}$ equal to 100%. However, the performance of DCS was unsatisfactory, in that it was sensitive to heavy-tailed data, and was easily affected by the imbalanced response.

Example 2. *In this example, we used a more complex setting to check the effectiveness of the proposed methods. This example was similar to Example 2 in Xie et al. [21]. For a given $Y = k$, the p-dimensional random vector of $X|\{Y = k\}$ was generated in the same way as in Example 1, but the correlation structure among the predictors was set as $\text{Corr}(X_i, X_j) = 0.5^{|i-j|}$. We considered a five-categorical response; the mean shifts μ_k for each class were $\mu_1 = (1.5, 1.5, 0_{p-2}^\top)^\top$, $\mu_2 = (0_5^\top, 1.5, 1.5, 1.5, 0_{p-8}^\top)^\top$, $\mu_3 = (0_{10}^\top, 1.5, 1.5, 1.5, 1.5, 0_{p-14}^\top)^\top$, $\mu_4 = (0_{20}^\top, 1.5, 1.5, 1.5, 1.5, 1.5, 0_{p-25}^\top)^\top$, $\mu_5 = (0_{30}^\top, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 0_{p-36}^\top)^\top$, so the corresponding active sets were $\mathcal{A} = \{1, 2, 6, 7, 8, 11, \cdots, 14, 21, \cdots, 25, 31, \cdots, 36\}$. Y was also generated in a balanced way, with $P(Y = k) = 0.2$ for $k = 1, \cdots, 5$, and in an imbalanced way, with $P(Y = k) = 0.1$ for $k = 1, 2, 3$ and $P(Y = k) = 0.35$ for $k = 4.5$. We considered $n = 200$ and $p = 1000$ or 3000.*

Table 2 presents the simulation results. In this example, we can see that QCS performed better than its competitors: it had the smallest MMS, IQR, and EPR. Secondly, it can be seen that the increase of dimensionality p had a negative effect on all methods, but that our method suffered the least.

Table 1. Simulation results of Example 1.

Method	$r = 0.05$					$r = 0.15$				
	MMS	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$	**MMS**	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$
Case 1										
QCS	2	1	12	94.2	97.3	2	2	21	91.2	95.4
MVS	2	1	6	97.1	98.6	2	1	16.5	93.4	97.0
DCS	7	5	14	86.4	98.0	20	13	44	12.0	65.2
CAS	2	0	5	97.5	98.9	2	1	15	92.1	95.9
KFS	2	2	16	91.2	96.1	2	3	29	86.2	93.1
Case 2										
QCS	2	2	23	90.7	94.7	3	4	34	85.8	92.6
MVS	2	1	15	91.6	96.0	2	2	22	90.4	93.7
DCS	7	6	23.5	78.8	94.8	22.5	19	62	13.3	56.1
CAS	2	1	16	93.3	96.0	2	3	24	90.1	93.2
KFS	3	4	35	85.0	91.1	3	5	53	83.0	90.2
Case 3										
QCS	8	0	3	98.8	99.3	8	0	10	98	99.3
MVS	8	2	15	96.4	98.3	8.5	3	45	92.3	96.0
DCS	68.5	31	126	0.2	41.2	190.5	103	334	0	0
CAS	8	0	2	99.7	99.9	8	1	6	98.2	98.9
KFS	9.5	4	21	96.0	99.2	11	7	43	90.0	96.3
Case 4										
QCS	8	0	11	96.9	98.7	8	2	24.5	94.9	97.0
MVS	9	4	39.5	92.6	96.0	11	14	84	83.5	92.0
DCS	91	63	270.5	0.0	19.2	272	213	630	0.0	0.0
CAS	8	1	6	98.4	99.4	9	4	19	93.1	96.3
KFS	12	10	47	86.6	96.2	14.5	20	98	77.0	91.6

Table 2. Simulation results of Example 2.

Method	$r = 0.05$					$r = 0.15$				
	MMS	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$	**MMS**	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$
Balanced response, $p = 1000$										
QCS	20	0	1	100	100	20	0	1	99.2	100
MVS	20	0	0	100	100	20	0	1	99.5	100
DCS	33	8	19	74.0	99.7	68.5	21	58	0	66.3
CAS	20	0	0	100	100	20	0	0	100	100
KFS	20	0	2	100	100	20	1	6	99.2	100
Imbalanced response, $p = 1000$										
QCS	20	3	24	92.5	98.0	22	8	51	85.5	95.3
MVS	21	7	31	88.8	97.6	23	14	66	78.4	92.4
DCS	83	83	287	1.7	41.6	203.5	143	420	0	0.7
CAS	26	13	38	80.2	97.3	32	19	55	63.5	94.0
KFS	31	20	64	64.7	91.6	36	29	137	52.4	83.1
Balanced response, $p = 3000$										
QCS	20	0	1	100	100	20	0	2	99.2	100
MVS	20	0	1	100	100	20	0	3	99.2	100
DCS	61	16	56	0	80.4	159.5	50	162	0	0
CAS	20	0	0	100	100	20	0	1	100	100
KFS	20	2	8	96.8	100	21	2	17	95.1	98.8
Imbalanced response, $p = 3000$										
QCS	21	8	80	84.4	93.2	24	27	230	69.9	83.2
MVS	23	15	106	77.1	90.3	29	46	220	60.2	78.2
DCS	228.5	237	908	0	1.2	581	423	1148	0	0
CAS	37	34	109	50.9	81.7	57	62	161	20.5	63.3
KFS	48.5	56	257	36.6	69.2	68.5	80	306	19.1	54.8

Example 3. *This example mimicked the scenario that the samples in the same group had multi-modals. Given $Y = k$, the random vector of X was generated in the same way as in Example 1,*

except that we fixed $r = 0.05$, and generated Z from a mixture normal distribution, designed as follows:

Case 1. $Z|\{Y = k\} \sim 0.2N(\mu_k, I_p) + 0.8N(-\mu_k, I_p)$;

Case 2. $Z|\{Y = k\} \sim 0.3N(\mu_k, I_p) + 0.7N(-\mu_k, I_p)$;

Case 3. $Z|\{Y = k\} \sim 0.4N(\mu_k, I_p) + 0.6N(-\mu_k, I_p)$,

where $\mu_{kk} = 2.5$ for $k = 1, 2, \cdots, K$ and 0 for other components in μ_k. In this example, we only considered a balanced setting for Y. Similarly, we considered $(K, n, p) = (2, 50, 1000)$ and $(8, 160, 2000)$, respectively.

The simulation results are shown in Table 3. This table shows that the category K of Y had a greatly negative effect on all the competitors, in that they suffered much efficiency loss for the screening when we increased K from 2 to 8. In particular, in case 3, where the distribution of data had two comparable modals, all methods except ours missed the active predictors completely, even under a large model size $2d_n$. The above results show that the newly proposed method is very robust.

Table 3. Simulation results of Example 3.

Method	**$K = 2$**					**$K = 8$**				
	MMS	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$	**MMS**	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$
					Case 1					
QCS	2	0	3	98.9	99.7	8	0	4	98.9	99.8
MVS	2	0	2	100	100	21	29	141	64.5	83.4
DCS	10	4.5	12	75.0	99.0	147.5	79	238	0	0
CAS	3	3	12.5	92.9	97.4	31	37.5	169.5	50.5	79.5
KFS	2	0	2	99.6	100	22	23	102	67.0	87.2
					Case 2					
QCS	3	2	14	93.6	97.3	9	3	27.5	94.3	96.9
MVS	4	5	12.5	93.3	98.4	103	126.5	366	3.0	25.5
DCS	13	5	16	49.5	96.5	192.5	92	420	0	0
CAS	35	41	125	11.3	33.0	467	385	1014	0	0
KFS	2	2	11	95.0	97.6	103.5	108	344.5	8.0	28.4
					Case 3					
QCS	6	10	44.5	74.5	87.6	13	16	86.5	78.1	89.7
MVS	12	13	49	52.1	79.4	346.5	277.5	841	0	0
DCS	15	8	21	28.7	88.2	271	153.5	545.5	0	0
CAS	297.5	207.5	550.5	0	0	1644	317	696.5	0	0
KFS	9	18	55	60.9	79.3	411	332.5	738	0	0

Example 4. *This example considered a K-categorical logistic model with*

$$P(Y = k|X) = \frac{\exp(X^\top \beta_k)}{1 + \sum_{i=1}^{K} \exp(X^\top \beta_k)}, k = 1, 2, \cdots, K$$

where the model settings were configured as follows:

Case 1. $K = 2$, $\beta_2 = 0_p$ *and* $\beta_1 = (\beta_1, \cdots, \beta_{10}, 0_{p-10}^\top)^\top$ *with* $\beta_j \sim Uniform(1, 2)$;

Case 2. $K = 5$, $\beta_1 = 0_p$ $\beta_2 = (\beta_1, 0_{p-1}^\top)^\top$, $\beta_3 = (0, \beta_2, \beta_3, 0_{p-3}^\top)^\top$; $\beta_4 = (0_3^\top, \beta_4, \beta_5, \beta_6, 0_{p-6}^\top)^\top$ *and* $\beta_5 = (0_6^\top, \beta_7, \beta_8, \beta_9, \beta_{10}, 0_{p-10}^\top)^\top$ *with* $\beta_j \sim Uniform(1, 2)$.

We considered the multivariate normal distribution $X_j \sim N(0, 1)$ and the student t-distribution $X_j \sim t_3$. The correlation structure among the predictors was equal to

$\mathrm{Corr}(X_i, X_j) = 0.5^{|i-j|}$. We set $(n, p) = (150, 1000)$. The corresponding simulation results are shown in Table 4, which shows that all the methods performed similarly, but that our methods behaved slightly better under the t-distribution. In addition, it seems that the t-distribution led to a more accurate screening result for all methods. The reason may be attributed to the structure of the logistic model. Consider the simplest case, $P(Y = 1|X) = 1/(1 + \exp(-X))$ and $P(Y = 0|X) = 1/(1 + \exp(X))$: clearly, a larger $|X|$ will make the classification between positive and negative easier. Consequently, under logistic function, the t-distributed data will result in a more accurate result, because the t-distribution has a higher probability of generating predictors with large values.

Table 4. Simulation results of Example 4.

Method	$K = 2$					$K = 5$				
	MMS	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$	**MMS**	**IQR**	**EPR**	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$
					$X_j \sim N(0,1)$					
QCS	12.5	15	112	77.7	88.3	12	12	101	81.6	90.4
MVS	13	18	165	75.2	84.8	12	14	143	76.6	88.1
DCS	13	18	177	76.0	85.6	18	37	210	65.1	80.0
CAS	12.5	13	153	78.1	87.1	31.5	46.0	172	47.2	72.8
KFS	21	36	281	62.2	78.1	32.5	53	230	48.0	70.4
					$X_j \sim t_3$					
QCS	10	15	137	85.6	92.2	10	12	49	92.0	95.2
MVS	10	15	150	85.3	92.4	10	12	67	90.1	94.4
DCS	10	15	133	84.7	91.2	11	23	112	79.0	88.6
CAS	10	14	106	85.2	92.8	14	30	136	72.6	83.6
KFS	12	24	139	79.2	88.5	13	34	174	72.8	82.8

Example 5. *This example aimed to check the effectiveness of the two extensions of the new method in Section 2.2. We considered the following three models:*

1. $Y = \sum_{i=1}^{5} X_i + \exp\left(\sum_{i=6}^{10} X_i\right) + \varepsilon$, *where* $X_j \sim N(0,1)$ *with* $\mathrm{Corr}(X_i, X_j) = 0.5^{|i-j|}$ *and* $\varepsilon \sim N(0,1)$*;*
2. $Y = 3f_1(X_1) + f_2(X_2) - 1.5f_3(X_3) + f_4(X_4) + \varepsilon$, *where* $f_1(x) = -\sin(2x)$, $f_2(x) = x^2 - 25/12$, $f_3(x) = x$, $f_4(x) = \exp(-x) - 0.4\sinh(2.5)$, *where* X_j *was independent of* $Uniform(-2.5, 2.5)$*;*
3. $Y = 1.5\log(n)/\sqrt{n}(X_1 + X_2 - 2X_{10} + 2X_{20} - 2|X_{100}|) + \varepsilon$, *where* X_j *was equal to* -1 *if* $Z_j < q_1$*, 1 if* $Z_j \geq q_3$*, and 0 otherwise, and where* $Z_j \sim N(0,1)$ *with* $\mathrm{Corr}(Z_j, Z_k) = 0.5^{|j-k|}$*, and* q_1 *and* q_3 *were the first and third quartiles, respectively, of a standard normal distribution.*

Model 1 was an index model from Zhu et al. [8]. Model 2 was an additive model from Meier et al. [23]. Model 3 mimicked the SNPs, with equal allele frequencies $\{-1, 0, 1\}$ representing $\{AA, Aa, aa\}$, respectively; this model has been analyzed in Cui et al. [20]. We report the simulation results in Table 5. It is clear that the proposed method always demonstrated a superior performance under the three models. More specifically, in Models 1 and 2, DCS did not work, though the predictor was not heavy-tailed. In Model 3, the performance of DCS and CAS were unsatisfying, with large MS and less probability of including the active predictors.

Overall, through the above simulations, we can summarize that QCS was the most robust method: compared to its competitors, it had a very stable performance within different model settings.

Computational complexity. Before the end of this subsection, we discuss the computational complexity of our method. Theoretically, the computational complexity of our method is $O(np)$, which is restricted at the sample size level. To obtain a clearer view of the computational complexity of our method, we conducted some simulations, to compare the computing time of our method to its competitors (see Figure 3). This figure showed that the computing time of our method linearly increased with the sample size, while the computing times of the other methods had a quadratic form against n. The simulations were conducted using Matlab 2013a in a Dell OptiPlex 7060 SFF, equipped with eight 3.20 GHz Intel(R) Core(TM) i7-8700 CPUs 3.20 Ghz and 16.0 GB RAM.

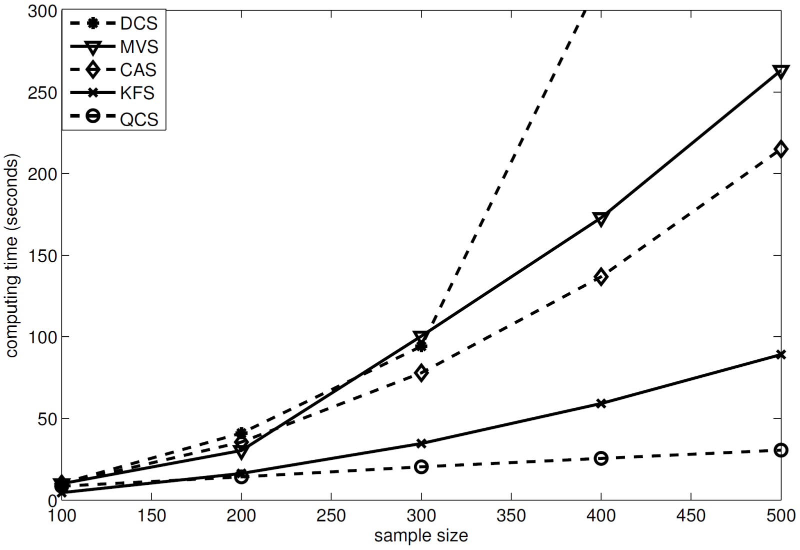

Figure 3. Computing time of different methods based on 100 replications, where QCS is our method, MVS is the MV-based sure independence screening method in [20], DCS is the distance correlation–sure independence screening procedure in [9], CAS is the category-adaptive variable screening in [21], and KFS is the Kolmogorov filter method in [19]. This simulation used Example 1, with $(K, p) = (8, 2000)$.

Table 5. Simulation results of Example 5.

	Method	MMS	IQR	EPR	$\mathcal{P}_{d_n}$	$\mathcal{P}_{2d_n}$
Model 1	QCS	11	10	111	86.2	92.1
	MVS	11.5	14	132	82.4	89.5
	DCS	957	713	1513	0.4	0.8
	CAS	30	78	338	56.7	71.2
	KFS	14	18	165	83.2	90.3
Model 2	QCS	4	1	7	98.1	100
	MVS	4	1	9	98.0	100
	DCS	49	58	195	34.6	64.5
	CAS	6	10	35	92.9	98.3
	KFS	5	13	72	92.4	94.1
Model 3	QCS	6.5	14	109	88.7	94.2
	MVS	7	16	114	86.4	90.8
	DCS	13	26	215	76.5	86.6
	CAS	17	54	258	60.8	78.9
	KFS	25	87	316	58.2	68.4

3.3. Real Data Analyses

For this section, we applied our new screening methods to two cancer datasets. One was leukemia data, consisting of 72 samples and 3571 genes, of which 47 were acute lymphocytic leukemia (ALL) and the rest 25 were acute myelogenous leukemia (AML). Note that the original leukemia data had 73 samples and 7129 genes. The data we analyzed here had been pre-feature-selected (see details in Dettling M. [24]). The other cancer dataset comprised small-round-blue-cell tumors (SRBCT) data, consisting of 63 samples and 2308 genes. Among the 63 subjects, there were four types of tumors, including Burkitt lymphoma (BL), having 23 cases, Ewing sarcoma (EWS), having 20 cases, neuroblastoma (NB), having 12 cases, and rhabdomyosarcoma (RMS), having 8 cases, respectively, so the data was four-categorical. The two datasets are available on the website http://www.stat.cmu.edu/~jiashun/Research/software/GenomicsData/, accessed on 5 March 2022.

The purpose of the two datasets was to identify the key genes that have a dominant effect on predicting the diagnostic category of a tumor. We first employed the screening methods, to reduce the large p to a suitable scale s. Then, we invoked the penalized linear discriminant analysis (penLDA) [25], to further select the discriminative predictors from the s_n predictors. The above procedure is the popular two-stage method that is commonly used in the analysis of ultrahigh-dimensional data. Note that we could also replace the penLDA in the second stage with other penalized methods, such as sparse discriminant analysis, as proposed by Clemmensen et al. [26].

We randomly extracted 70% of the samples from each class, as the training data, and set the rest of the samples as the testing data, in which the training data were used both to implement the screening procedure and to build the classifier, while the testing data were used to check the performance of the trained classifier. We repeated the above procedure for 500 replications, and we report both the training errors and testing errors for different methods. Note that in the screening stage, we set $d_n = [n/\log n]$ and $2[n/\log n]$, respectively: thus, $d_n = 16$ and 32 in the leukemia dataset, and $d_n = 15$ and 30 in the SRBCT dataset. In the second stage, the tuning parameter of the penLDA method was determined according to the 5-fold cross-validation method. Table 6 displays the corresponding results, where QCS–penLDA denotes the two-stage method of QCS followed by penLDA; a similar definition applies to MQS–penLDA, MVS–penLDA, etc.

The numerical results are summarized in Table 6, from which the following conclusion can be obtained. For the leukemia dataset, all methods except DCS performed reasonably well, such that all of them could control the testing errors below 1. However, for the SRBCT data, our method performed significantly better than the other methods: it achieved the smallest training errors, and testing errors closer to 0. The CAS-based two-stage method yielded bad results for both the training error and the testing error. The reason may be that the distribution behind the data was not unimodal.

Table 6. Numerical results of the real data analyses.

Data	d_n	Method	No. of Training Errors		No. of Testing Errors	
			Mean	Std	Mean	Std
Leukemia	16	QCS-penLDA	0.176	0.401	0.794	0.865
		MVS-penLDA	0.166	0.383	0.828	0.864
		DCS-penLDA	1.188	0.783	1.334	1.082
		CAS-penLDA	0.140	0.353	0.814	0.867
		KFS-penLDA	0.260	0.478	0.896	0.924
	32	QCS-penLDA	0.152	0.359	0.670	0.813
		MVS-penLDA	0.130	0.336	0.696	0.808
		DCS-penLDA	0.898	0.732	0.974	0.920
		CAS-penLDA	0.128	0.334	0.682	0.801
		KFS-penLDA	0.210	0.417	0.792	0.873
SRBCT	15	QCS-penLDA	0.100	0.319	0.574	0.818
		MVS-penLDA	0.436	1.777	1.180	1.399
		DCS-penLDA	1.366	2.753	1.740	1.701
		CAS-penLDA	7.236	2.246	5.744	2.174
		KFS-penLDA	2.850	1.665	2.852	1.744
	30	QCS-penLDA	0.088	1.343	0.206	0.872
		MVS-penLDA	0.130	1.710	0.470	1.021
		DCS-penLDA	0.320	2.693	0.604	1.458
		CAS-penLDA	3.360	1.601	3.864	1.787
		KFS-penLDA	0.594	0.831	0.860	0.964

4. Conclusions

This paper proposes a new quantile-composited feature screening (QCS) procedure, to rapidly screen out the null predictors. Compared to the existing methods, QCS sheds light on the following aspects. Firstly, the ranking index is a simple structure, so that the implementation of the screening procedure is computationally easy. Secondly, QCS is a quantile-composited method: it can utilize much distributional information, so as to significantly improve the screening efficiency, but retains the computational cost at a low level. The simulation and real data analysis also demonstrated the effectiveness of QCS.

In addition, it is worth mentioning that QCS can be further improved. For example, the selection of the number s of the quantiles is still a problem, which could be the focus of future work, based on this article.

Author Contributions: Methodology, S.C. and J.L.; Formal analysis, S.C.; Writing—review & editing, S.C. and J.L.; Funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: Jun Lu's research was partly supported by National Natural Science Foundation of China (No.12001486), Natural Science Foundation of Zhejiang Province (No.LQ20A010001), and Advance Research Plan of National University of Defense Technology.

Data Availability Statement: The datasets used in this paper can be accessed freely on the website http://www.stat.cmu.edu/~jiashun/Research/software/GenomicsData/.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Main Results

Proof of Lemma 1. For Lemma 1(1), we only prove the sufficient; the necessity can be proved similarly. If $q_X(\tau) = q_{X|Y}(\tau)$, then for any $Y = 1, \cdots, K$,

$$\begin{aligned} & P(I\{X > q_X(\tau)\} = 0, Y = k) \\ = \;& P(X \le q_X(\tau)|Y = k)P(Y = k) \\ = \;& P(X \le q_{X|Y}(\tau)|Y = k)P(Y = k) \\ = \;& P(X \le q_X(\tau))P(Y = k) \\ = \;& P(I\{X > q_X(\tau)\} = 0)P(Y = k). \end{aligned}$$

For $I\{X > q_X(\tau)\} = 1$, the proof is the same.

To prove Lemma 1(2), we have $\forall x_0 \in R$, $\exists \tau_0$ $s.t.$ $x_0 = q_X(\tau_0)$; then,

$$\begin{aligned} & P(X \le x_0, Y = k) \\ = \;& P(X \le q_X(\tau_0), Y = k) \\ = \;& P(X \le q_X(\tau_0)|Y = k)P(Y = k) \\ = \;& P(X \le q_{X|Y}(\tau_0)|Y = k)P(Y = k) \\ = \;& P(X \le q_{X|Y}(\tau_0))P(Y = k) \\ = \;& P(X \le q_X(\tau_0))P(Y = k) \\ = \;& P(X \le x_0)P(Y = k). \end{aligned}$$

□

Proof of Theorem 1. We prove this theorem in two steps.

Firstly, we prove the consistency of $\hat{\pi}_{jb}$, $\hat{\pi}_{yk}$, $\hat{\pi}_{yk,jb}(\tau)$ and $\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)$.

(1) If $b = 0$, then

$$\begin{aligned}\left|\hat{\pi}_{j0}(\tau)-\pi_{j0}(\tau)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n} I(X_{ij}\le \hat{q}_{X_j}(\tau)) - P(X_j \le q_{X_j}(\tau))\right| \\ &= \left|\frac{[n\tau]+I(n\tau>[n\tau])}{n}-\frac{n\tau}{n}\right| \\ &= \left|\frac{I(n\tau>[n\tau])-(n\tau-[n\tau])}{n}\right| \le \frac{1}{n}.\end{aligned} \tag{A1}$$

The conclusion for $b=1$ can be proved similarly.

(2) By using Hoeffding's inequality, we obtain

$$P\left(\left|\hat{\pi}_{yk}-\pi_{yk}\right| > \varepsilon\right) \le 2\exp\left\{-2n\varepsilon^2\right\}. \tag{A2}$$

(3) Define $\widetilde{Z}_{ij}(\tau) = I(X_{ij} > q_{X_j}(\tau))$, such that

$$\begin{aligned} & P\left(\left|\hat{\pi}_{yk,jb}(\tau)-\pi_{yk,jb}(\tau)\right| \ge \varepsilon\right) \\ = \; & P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\{Y_i=k\}I\{Z_{ij}(\tau)=b\} - P(Y=k, Z_j(\tau)=b)\right| \ge \varepsilon\right) \\ \le \; & P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\{Y_i=k\}(I\{Z_{ij}(\tau)=b\} - I\{\widetilde{Z}_{ij}(\tau)=b\})\right| \ge \frac{\varepsilon}{2}\right) \qquad \text{(A3)} \\ & +P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\{Y_i=k\}I\{\widetilde{Z}_{ij}(\tau)=b\} - P(Y=k, Z_j(\tau)=b)\right| \ge \frac{\varepsilon}{2}\right) \qquad \text{(A4)}\end{aligned}$$

For (A3), for each j, $I\{Z_{ij}=b\} - I\{\widetilde{Z}_{ij}(\tau)=b\} \le 0$ for any i, or $I\{Z_{ij}=b\} - I\{\widetilde{Z}_{ij}(\tau)=b\} > 0$ for any i. Using Hoeffding's inequality, (A3) can be deduced, such that

$$\begin{aligned} & P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\{Y_i=k\}(I\{Z_{ij}(\tau)=b\} - I\{\widetilde{Z}_{ij}(\tau)=b\})\right| \ge \frac{\varepsilon}{2}\right) \\ = \; & P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\{Y_i=k\}(I\{X_{ij}\le \hat{q}_{X_j}(\tau)\} - I\{X_{ij}\le q_{X_j}(\tau)\})\right| \ge \frac{\varepsilon}{2}\right) \\ \le \; & P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(I\{X_{ij}\le \hat{q}_{X_j}(\tau)\} - I\{X_{ij}\le q_{X_j}(\tau)\})\right| \ge \frac{\varepsilon}{2}\right) \\ = \; & P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\{X_{ij}\le \hat{q}_{X_j}(\tau)\} - \frac{1}{n}\sum_{i=1}^{n} I\{X_{ij}\le q_{X_j}(\tau)\}\right| \ge \frac{\varepsilon}{2}\right) \\ = \; & P\left(\left|\hat{\pi}_{j0}(\tau) - \frac{1}{n}\sum_{i=1}^{n} I\{X_{ij}\le q_{X_j}(\tau)\}\right| \ge \frac{\varepsilon}{2}\right) \\ = \; & P\left(\left|\hat{\pi}_{j0}(\tau)-\pi_{j0}(\tau)\right| + \left|\pi_{j0}(\tau) - \frac{1}{n}\sum_{i=1}^{n} I\{X_{ij}\le q_{X_j}(\tau)\}\right| \ge \frac{\varepsilon}{2}\right) \\ \le \; & P\left(\frac{1}{n} + \left|\frac{1}{n}\sum_{i=1}^{n} I\{X_{ij}\le q_{X_j}(\tau)\} - \tau\right| \ge \frac{\varepsilon}{2}\right) \\ = \; & P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\{X_{ij}\le q_{X_j}(\tau)\} - \tau\right| \ge \frac{\varepsilon}{2} - \frac{1}{n}\right) \\ \le \; & 2\exp\left\{-\frac{n(\varepsilon-\frac{2}{n})^2}{2}\right\}.\end{aligned} \tag{A5}$$

For (A4), using Hoeffding's inequality,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(y_i = k)I(\widetilde{Z}_{ij}(\tau) = b) - P(Y = k, Z_j(\tau) = b)\right| > \varepsilon/2\right) \leq 2\exp\left\{\frac{-n\varepsilon^2}{2}\right\}. \quad \text{(A6)}$$

Consequently, combining the results of (A5) and (A6), it is simple to establish that

$$P\left(\left|\hat{\pi}_{yk,jb}(\tau) - \pi_{yk,jb}(\tau)\right| > \varepsilon\right) \leq 2\exp\left\{-\frac{n(\varepsilon - \frac{2}{n})^2}{2}\right\} + 2\exp\left\{\frac{-n\varepsilon^2}{2}\right\}. \quad \text{(A7)}$$

(4) By employing a similar argument, $\left|\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau) - \pi_{yk}\pi_{jb}(\tau)\right|$ can be bounded easily, as

$$\begin{aligned}
& P\left(\left|\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau) - \pi_{yk}\pi_{jb}(\tau)\right| \geq \varepsilon\right) \\
\leq\ & P\left(\left|\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau) - \hat{\pi}_{yk}\pi_{jb}(\tau)\right| + \left|\hat{\pi}_{yk}\pi_{jb}(\tau) - \pi_{yk}\pi_{jb}(\tau)\right| \geq \varepsilon\right) \\
=\ & P\left(\hat{\pi}_{yk}\cdot\left|\hat{\pi}_{jb}(\tau) - \pi_{jb}(\tau)\right| + \pi_{jb}(\tau)\cdot\left|\hat{\pi}_{yk} - \pi_{yk}\right| \geq \varepsilon\right) \\
\leq\ & P\left(\frac{1}{n} + \left|\hat{\pi}_{yk} - \pi_{yk}\right| \geq \varepsilon\right) \\
\leq\ & P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(Y_i = k) - P(Y = k)\right| \geq \varepsilon - \frac{1}{n}\right) \\
\leq\ & 2\exp\left\{-2n\left(\varepsilon - \frac{1}{n}\right)^2\right\}, \qquad \text{(A8)}
\end{aligned}$$

where the second inequality holds because $\left|\hat{\pi}_{jb}(\tau) - \pi_{jb}(\tau)\right| \leq \frac{1}{n}$, and where the last inequality holds due to Hoeffding's inequality.

Secondly, we prove the consistency of $\widehat{Q}_j(\tau) - Q_j(\tau)$. Because

$$\begin{aligned}
& \left|\hat{Q}_j(\tau) - Q_j(\tau)\right| \\
=\ & \left|\sum_{k=0}^{K}\sum_{b=0}^{1}\frac{(\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau) - \hat{\pi}_{yk,jb}(\tau))^2}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)} - \sum_{k=0}^{K}\sum_{b=0}^{1}\frac{(\pi_{yk}\pi_{jb}(\tau) - \pi_{yk,jb}(\tau))^2}{\pi_{yk}\pi_{jb}(\tau)}\right| \\
=\ & \left|\sum_{k=0}^{K}\sum_{b=0}^{1}\left(\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau) - 2\hat{\pi}_{yk,jb}(\tau) + \frac{\hat{\pi}^2_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)}\right)\right. \\
& \left. - \sum_{k=0}^{K}\sum_{b=0}^{1}\left(\pi_{yk}\pi_{jb}(\tau) - 2\pi_{yk,jb}(\tau) + \frac{\pi^2_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right| \\
=\ & \left|\sum_{k=0}^{K}\sum_{b=0}^{1}(\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau) - \pi_{yk}\pi_{jb}(\tau)) + 2\sum_{k=0}^{K}\sum_{b=0}^{1}(\hat{\pi}_{yk,jb}(\tau) - \pi_{yk,jb}(\tau))\right. \\
& \left. + \sum_{k=0}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^2_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)} - \frac{\pi^2_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right| \\
=\ & \left|0 + 0 + \sum_{k=0}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^2_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)} - \frac{\pi^2_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right| \\
=\ & \left|\sum_{k=0}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^2_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)} - \frac{\pi^2_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right|,
\end{aligned}$$

we only need to prove the consistency of $\sum_{k=0}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^2_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)} - \frac{\pi^2_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)$.

When $0 < \tau \leq \frac{1}{2}$,

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^2_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)} - \frac{\pi^2_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right| \\
= &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)} + \frac{\hat{\pi}^2_{yk,j1}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j1}(\tau)} - \frac{\pi^2_{yk,j1}(\tau)}{\pi_{yk}\pi_{j1}(\tau)}\right)\right| \\
= &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)} + \frac{[\hat{\pi}_{yk} - \hat{\pi}_{yk,j0}(\tau)]^2}{\hat{\pi}_{yk}\hat{\pi}_{j1}(\tau)} - \frac{[\pi_{yk} - \pi_{yk,j0}(\tau)]^2}{\pi_{yk}\pi_{j1}(\tau)}\right)\right| \\
= &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)} + \frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j1}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j1}(\tau)}\right.\right. \\
&\left.\left.+\frac{\hat{\pi}_{yk} - 2\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{j1}(\tau)} - \frac{\pi_{yk} - 2\pi_{yk,j0}(\tau)}{\pi_{j1}(\tau)}\right)\right| \\
\leq &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)} + \frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j1}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j1}(\tau)}\right)\right| \qquad \text{(A9)} \\
&+\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}_{yk} - 2\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{j1}(\tau)} - \frac{\pi_{yk} - 2\pi_{yk,j0}(\tau)}{\pi_{j1}(\tau)}\right)\right|. \qquad \text{(A10)}
\end{aligned}$$

For (A9), combining the results of (A1), (A2), (A7) and (A8), we have :

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)} + \frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j1}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j1}(\tau)}\right)\right| \\
= &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}}\left(\frac{1}{\hat{\pi}_{j0}(\tau)} + \frac{1}{\hat{\pi}_{j1}(\tau)}\right) - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}}\left(\frac{1}{\pi_{j0}(\tau)} + \frac{1}{\pi_{j1}(\tau)}\right)\right)\right| \\
= &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)\hat{\pi}_{j1}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)\pi_{j1}(\tau)}\right)\right| \\
= &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)\hat{\pi}_{j1}(\tau)} - \frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)\pi_{j1}(\tau)} + \frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)\pi_{j1}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)\pi_{j1}(\tau)}\right)\right| \\
\leq &\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} \cdot \frac{\pi_{j1}(\tau) - \hat{\pi}_{j1}(\tau)}{\hat{\pi}_{j1}(\tau)\pi_{j1}(\tau)}\right)\right| + \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)}\right) \cdot \frac{1}{\pi_{j1}(\tau)}\right| \\
= &\frac{\pi_{j1}(\tau) - \hat{\pi}_{j1}(\tau)}{\hat{\pi}_{j1}(\tau)\pi_{j1}(\tau)} \cdot \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)}\right)\right| + \frac{1}{\pi_{j1}(\tau)} \cdot \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)}\right)\right| \\
\leq &\frac{1}{n(1-\tau)\cdot(1-\tau-\frac{1}{n})} \cdot \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{j0}(\tau)}\right)\right| + 2\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)}\right)\right| \\
\leq &\frac{1}{n(1-\tau)\cdot(1-\tau-\frac{1}{n})} + 2\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)}\right)\right|, \qquad \text{(A11)}
\end{aligned}$$

and

$$
\begin{aligned}
& \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)}\right)\right| \\
= & \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{j0}(\tau)} - \frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\pi_{j0}(\tau)} + \frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\pi_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)}\right)\right| \\
= & \left|\sum_{k=1}^{K}\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}}\left(\frac{1}{\hat{\pi}_{j0}(\tau)} - \frac{1}{\pi_{j0}(\tau)}\right) + \sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}\pi_{j0}(\tau)} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}\pi_{j0}(\tau)}\right)\right| \\
\leq & \left|\left(\frac{1}{\hat{\pi}_{j0}(\tau)} - \frac{1}{\pi_{j0}(\tau)}\right)\sum_{k=1}^{K}\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}}\right| \\
& \quad + \frac{1}{\pi_{j0}(\tau)}\left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}^2_{yk,j0}(\tau)}{\hat{\pi}_{yk}} - \frac{\hat{\pi}_{yk,j0}(\tau)\pi_{yk,j0}(\tau)}{\hat{\pi}_{yk}} + \frac{\hat{\pi}_{yk,j0}(\tau)\pi_{yk,j0}(\tau)}{\hat{\pi}_{yk}} - \frac{\pi^2_{yk,j0}(\tau)}{\pi_{yk}}\right)\right| \\
\leq & \left|\left(\frac{\pi_{j0}(\tau) - \hat{\pi}_{j0}(\tau)}{\hat{\pi}_{j0}(\tau)\pi_{j0}(\tau)}\right)\sum_{k=1}^{K}\hat{\pi}_{yk,j0}(\tau)\right| \\
& \quad + \frac{1}{\pi_{j0}(\tau)}\left|\sum_{k=1}^{K}\left[\frac{\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{yk}}\Big(\hat{\pi}_{yk,j0}(\tau) - \pi_{yk,j0}(\tau)\Big) + \frac{\pi_{yk,j0}(\tau)}{\hat{\pi}_{yk}}\Big(\hat{\pi}_{yk,j0}(\tau) - \pi_{yk,j0}(\tau)\Big)\right]\right| \\
\leq & \left|\frac{\pi_{j0}(\tau) - \hat{\pi}_{j0}(\tau)}{\hat{\pi}_{j0}(\tau)\pi_{j0}(\tau)} \cdot \hat{\pi}_{j0}(\tau)\right| + \frac{1}{\pi_{j0}(\tau)}\left|\sum_{k=1}^{K}\frac{\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{yk}}\Big(\hat{\pi}_{yk,j0}(\tau) - \pi_{yk,j0}(\tau)\Big)\right| \\
& \quad + \frac{1}{\pi_{j0}(\tau)}\left|\sum_{k=1}^{K}\frac{\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{yk}}\Big(\hat{\pi}_{yk,j0}(\tau) - \pi_{yk,j0}(\tau)\Big)\right| \\
\leq & \left|\frac{\pi_{j0}(\tau) - \hat{\pi}_{j0}(\tau)}{\pi_{j0}(\tau)}\right| + \frac{1}{\pi_{j0}(\tau)}\sum_{k=1}^{K}\Big|\hat{\pi}_{yk,j0}(\tau) - \pi_{yk,j0}(\tau)\Big| + \frac{1}{\pi_{j0}(\tau)}\sum_{k=1}^{K}\Big|\hat{\pi}_{yk,j0}(\tau) - \pi_{yk,j0}(\tau)\Big| \\
= & \frac{1}{n\tau} + \frac{2}{1-\tau}\sum_{k=1}^{K}\Big|\hat{\pi}_{yk,j0}(\tau) - \pi_{yk,j0}(\tau)\Big|.
\end{aligned} \tag{A12}
$$

For (A10),

$$
\begin{aligned}
& \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}_{yk} - 2\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{j1}(\tau)} - \frac{\pi_{yk} - 2\pi_{yk,j0}(\tau)}{\pi_{j1}(\tau)}\right)\right| \\
= & \left|\sum_{k=1}^{K}\left(\frac{\hat{\pi}_{yk}}{\hat{\pi}_{j1}(\tau)} - \frac{\pi_{yk}}{\pi_{j1}(\tau)} + \frac{2\pi_{yk,j0}(\tau)}{\pi_{j1}(\tau)} - \frac{2\hat{\pi}_{yk,j0}(\tau)}{\hat{\pi}_{j1}(\tau)}\right)\right| \\
= & \left|\frac{1}{\hat{\pi}_{j1}(\tau)} - \frac{1}{\pi_{j1}(\tau)} + 2\left(\frac{\pi_{j0}(\tau)}{\pi_{j1}(\tau)} - \frac{\hat{\pi}_{j0}(\tau)}{\hat{\pi}_{j1}(\tau)}\right)\right| \\
= & \left|\frac{1}{\hat{\pi}_{j1}(\tau)} - \frac{1}{\pi_{j1}(\tau)} + 2\left(\frac{1-\pi_{j1}(\tau)}{\pi_{j1}(\tau)} - \frac{1-\hat{\pi}_{j1}(\tau)}{\hat{\pi}_{j1}(\tau)}\right)\right| \\
= & \left|\frac{1}{\hat{\pi}_{j1}(\tau)} - \frac{1}{\pi_{j1}(\tau)} + 2\left(\frac{1}{\pi_{j1}(\tau)} - \frac{1}{\hat{\pi}_{j1}(\tau)}\right)\right| \\
= & \left|\frac{1}{\hat{\pi}_{j1}(\tau)} - \frac{1}{\pi_{j1}(\tau)}\right| \\
\leq & \frac{1}{n(1-\tau)\cdot(1-\tau-\frac{1}{n})}.
\end{aligned} \tag{A13}
$$

Combining the results of (A9)–(A13),

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^{2}_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)}-\frac{\pi^{2}_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right| \\
\leq\ &\frac{2}{(1-\tau)\cdot[n(1-\tau)-1]}+\frac{2}{n\tau}+\frac{4}{1-\tau}\sum_{k=1}^{K}\left|\hat{\pi}_{yk,j0}(\tau)-\pi_{yk,j0}(\tau)\right|. \qquad \text{(A14)}
\end{aligned}$$

For $\frac{1}{2}<\tau<1$, by employing a similar argument, it can be proved that

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^{2}_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)}-\frac{\pi^{2}_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right| \\
\leq\ &\frac{2}{\tau\cdot(n\tau-1)}+\frac{2}{n(1-\tau)}+\frac{4}{\tau}\sum_{k=1}^{K}\left|\hat{\pi}_{yk,j1}(\tau)-\pi_{yk,j1}(\tau)\right|. \qquad \text{(A15)}
\end{aligned}$$

For any $\tau\in(0,1)$, by (A14) and (A15), it holds that

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\sum_{b=0}^{1}\left(\frac{\hat{\pi}^{2}_{yk,jb}(\tau)}{\hat{\pi}_{yk}\hat{\pi}_{jb}(\tau)}-\frac{\pi^{2}_{yk,jb}(\tau)}{\pi_{yk}\pi_{jb}(\tau)}\right)\right| \\
\leq\ &\frac{2}{\widetilde{\tau}\cdot(n\widetilde{\tau}-1)}+\frac{2}{n(\overline{\tau})}+\frac{4}{\widetilde{\tau}}\sum_{k=1}^{K}\left|\hat{\pi}_{yk,ja}(\tau)-\pi_{yk,ja}(\tau)\right|, \qquad \text{(A16)}
\end{aligned}$$

where $\widetilde{\tau}=\max\{\tau,1-\tau\}$, $\overline{\tau}=\min\{\tau,1-\tau\}$, and $\overline{b}=I\{\tau>1-\tau\}$. For (A7) and (A16), we can obtain

$$\begin{aligned}
&P\left(\left|\hat{Q}_{j}(\tau)-Q_{j}(\tau)\right|\geq\varepsilon\right) \\
\leq\ &P\left(\frac{2}{\widetilde{\tau}\cdot(n\widetilde{\tau}-1)}+\frac{2}{n(\overline{\tau})}+\frac{4}{\widetilde{\tau}}\sum_{k=1}^{K}\left|\hat{\pi}_{yk,j\overline{b}}(\tau)-\pi_{yk,j\overline{b}}(\tau)\right|\geq\varepsilon\right) \\
=\ &P\left(\sum_{k=1}^{K}\left|\hat{\pi}_{yk,j\overline{b}}(\tau)-\pi_{yk,j\overline{b}}(\tau)\right|\geq\frac{\widetilde{\tau}}{4}\left(\varepsilon-\frac{2}{\widetilde{\tau}\cdot(n\widetilde{\tau}-1)}-\frac{2}{n(\overline{\tau})}\right)\right) \\
\leq\ &2KP\left(\left|\hat{\pi}_{yk,jb}(\tau)-\pi_{yk,jb}(\tau)\right|\geq\frac{\widetilde{\tau}}{4K}\left(\varepsilon-\frac{2}{\widetilde{\tau}\cdot(n\widetilde{\tau}-1)}-\frac{2}{n(\overline{\tau})}\right)\right) \\
\leq\ &4K\exp\left\{-\frac{n\left[\frac{\widetilde{\tau}}{4K}\left(\varepsilon-\frac{2}{\widetilde{\tau}\cdot(n\widetilde{\tau}-1)}-\frac{2}{n(\overline{\tau})}\right)-\frac{2}{n}\right]^{2}}{2}\right\} \\
&+4K\exp\left\{\frac{-n\left[\frac{\widetilde{\tau}}{4K}\left(\varepsilon-\frac{2}{\widetilde{\tau}\cdot(n\widetilde{\tau}-1)}-\frac{2}{n(\overline{\tau})}\right)\right]^{2}}{2}\right\}. \qquad \text{(A17)}
\end{aligned}$$

Let $\tau\in(\alpha,1-\alpha)$, and by condition (C1), it can be derived that

$$P\left(\left|\hat{Q}_{j}(\tau)-Q_{j}(\tau)\right|\geq\varepsilon\right)=K\exp(\{-nK^{-2}\varepsilon^{2}+\varepsilon K^{-2}/\bar{\tau}\}). \qquad \text{(A18)}$$

Let $K=O(n^{\gamma})$ and $\varepsilon=cn^{-\eta}$, if $\bar{\tau}=o(n^{\eta-1})$; then

$$P\left(\left|\hat{Q}_{j}(\tau)-Q_{j}(\tau)\right|\geq cn^{\eta}\right)=K\exp(-O\{n^{1-2\eta-2\gamma}\}). \qquad \text{(A19)}$$

The proof of Corollary 1. Under the conditions in Theorem 1, following (A17) and (A19), we obtain

$$
\begin{aligned}
& P\left(\left|\hat{Q}_j - Q_j\right| \geq \varepsilon\right) \\
= \;& P\left(\left|\frac{\int_a^{1-a} \hat{Q}_j^2(\tau)d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau} - \frac{\int_a^{1-a} Q_j^2(\tau)d\tau}{\int_a^{1-a} Q_j(\tau)d\tau}\right| \geq \varepsilon\right) \\
= \;& P\left(\left|\frac{\int_a^{1-a} \hat{Q}_j^2(\tau)d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau} - \frac{\int_a^{1-a} \hat{Q}_j(\tau)Q_j(\tau)d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau} + \frac{\int_a^{1-a} \hat{Q}_j(\tau)Q_j(\tau)d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau}\right.\right. \\
& \left.\left. - \frac{\int_a^{1-a} \hat{Q}_j(\tau)Q_j(\tau)d\tau}{\int_a^{1-a} Q_j(\tau)d\tau} + \frac{\int_a^{1-a} \hat{Q}_j(\tau)Q_j(\tau)d\tau}{\int_a^{1-a} Q_j(\tau)d\tau} - \frac{\int_a^{1-a} Q_j^2(\tau)d\tau}{\int_a^{1-a} Q_j(\tau)d\tau}\right| \geq \varepsilon\right) \\
= \;& P\left(\left|\frac{\int_a^{1-a} \hat{Q}j(\tau)\left[\hat{Q}_j(\tau) - Q_j(\tau)\right]d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau} + \frac{\int_a^{1-a} \hat{Q}_j(\tau)Q_j(\tau)d\tau \int_a^{1-a}\left[\hat{Q}_j(\tau) - Q_j(\tau)\right]d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau \int_a^{1-a} Q_j(\tau)d\tau}\right.\right. \\
& \left.\left. + \frac{\int_a^{1-a} \hat{Q}_j(\tau)\left[\hat{Q}_j(\tau) - Q_j(\tau)\right]d\tau}{\int_a^{1-a} Q_j(\tau)d\tau}\right| \geq \varepsilon\right) \\
\leq \;& P\left(\frac{\int_a^{1-a} \hat{Q}j(\tau)\left|\hat{Q}_j(\tau) - Q_j(\tau)\right|d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau} + \frac{\int_a^{1-a} \hat{Q}_j(\tau)Q_j(\tau)d\tau \int_a^{1-a}\left|\hat{Q}_j(\tau) - Q_j(\tau)\right|d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau \int_a^{1-a} Q_j(\tau)d\tau}\right. \\
& \left. + \frac{\int_a^{1-a} Q_j(\tau)\left|\hat{Q}_j(\tau) - Q_j(\tau)\right|d\tau}{\int_a^{1-a} Q_j(\tau)d\tau} \geq \varepsilon, \left|\hat{Q}_j(\tau) - Q_j(\tau)\right| < \frac{\varepsilon}{4}\right) \\
& + P\left(\left|\hat{Q}_j(\tau) - Q_j(\tau)\right| \geq \frac{\varepsilon}{4}\right) \\
\leq \;& P\left(\frac{\varepsilon}{4} + \frac{\varepsilon}{4} \cdot \frac{\int_a^{1-a} \hat{Q}_j(\tau)Q_j(\tau)d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau \int_a^{1-a} Q_j(\tau)d\tau} + \frac{\varepsilon}{4} \geq \varepsilon, \left|\hat{Q}_j(\tau) - Q_j(\tau)\right| < \frac{\varepsilon}{4}\right) \\
& + P\left(\left|\hat{Q}_j(\tau) - Q_j(\tau)\right| \geq \frac{\varepsilon}{4}\right) \\
\leq \;& P\left(\frac{\varepsilon}{2} + \frac{\varepsilon}{4} \cdot \frac{\sqrt{\int_a^{1-a} \hat{Q}_j^2(\tau)d\tau \int_a^{1-a} Q_j^2(\tau)d\tau}}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau \int_a^{1-a} Q_j(\tau)d\tau} \geq \varepsilon, \left|\hat{Q}_j(\tau) - Q_j(\tau)\right| < \frac{\varepsilon}{4}\right) \\
& \quad + P\left(\left|\hat{Q}_j(\tau) - Q_j(\tau)\right| \geq \frac{\varepsilon}{4}\right) \\
\leq \;& P\left(\frac{\varepsilon}{2} + \frac{\varepsilon}{4} \cdot \frac{\int_a^{1-a} \hat{Q}_j(\tau)d\tau \int_a^{1-a} Q_j(\tau)d\tau}{\int_a^{1-a} \hat{Q}_j(\tau)d\tau \int_a^{1-a} Q_j(\tau)d\tau} \geq \varepsilon, \left|\hat{Q}_j(\tau) - Q_j(\tau)\right| < \frac{\varepsilon}{4}\right) \\
& + P\left(\left|\hat{Q}_j(\tau) - Q_j(\tau)\right| \geq \frac{\varepsilon}{4}\right) \\
= \;& P\left(\frac{\varepsilon}{2} + \frac{\varepsilon}{4} \cdot 1 \geq \varepsilon, \left|\hat{Q}_j(\tau) - Q_j(\tau)\right| < \frac{\varepsilon}{4}\right) + P\left(\left|\hat{Q}_j(\tau) - Q_j(\tau)\right| \geq \frac{\varepsilon}{4}\right) \\
= \;& P\left(\left|\hat{Q}_j(\tau) - Q_j(\tau)\right| \geq \frac{\varepsilon}{4}\right) \\
\leq \;& K\exp(-O\{n^{1-2\eta-2\gamma}\}).
\end{aligned}
$$

□

Proof of Theorem 2. If $\mathcal{A} \not\subseteq \widehat{\mathcal{A}}$, then there must exist some $k \in \mathcal{A}$, such that $\widehat{Q}_k < cn^{-\eta}$. It follows from condition (C2) that $\left|\widehat{Q}_k - Q_k\right| > cn^{-\eta}$ for some $k \in \mathcal{A}$, indicating that the events satisfy $\left\{\mathcal{A} \not\subseteq \widehat{\mathcal{A}}\right\} \subseteq \left\{\left|\widehat{Q}_k - Q_k\right| > cn^{-\kappa}\text{, for some } k \in \mathcal{A}\right\}$, and hence $\mathcal{E}_n = \left\{\max_{k\in\mathcal{A}}\left|\widehat{Q}_k - Qk\right| \le cn^{-\eta}\right\} \subseteq \left\{\mathcal{A} \subseteq \widehat{\mathcal{A}}\right\}$. Consequently,

$$\begin{aligned} P\left(\mathcal{A} \subseteq \widehat{\mathcal{A}}\right) &\ge \Pr(\mathcal{E}_n) = 1 - P(\mathcal{E}_n^c) = 1 - P\left(\min_{k\in\mathcal{A}}|\widehat{\omega}_k - \omega_k| \ge cn^{-\eta}\right) \\ &= 1 - s_n P\left\{|\widehat{\omega}_k - \omega_k| \ge cn^{-\eta}\right\} \ge 1 - s_n K \exp\left\{-O(n^{1-2\eta-2\gamma})\right\}, \end{aligned} \tag{A20}$$

where s_n is the cardinality of $\mathcal{A}$. □

References

1. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B* **2008**, *70*, 849–911. [CrossRef] [PubMed]
2. Wang, H. Forward regression for ultra-high dimensional variable screening. *J. Am. Statist. Assoc.* **2009**, *104*, 1512–1524. [CrossRef]
3. Chang, J.; Tang, C.; Wu, Y. Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.* **2013**, *41*, 2123–2148. [CrossRef] [PubMed]
4. Wang, X.; Leng, C. High dimensional ordinary least squares projection for screening variables. *J. R. Statist. Soc. B* **2016**, *78*, 589–611. [CrossRef]
5. Fan, J.; Feng, Y.; Song, R. Nonparametric independence screening in sparse ultrahigh dimensional additive models. *J. Am. Statist. Assoc.* **2011**, *106*, 544–557. [CrossRef] [PubMed]
6. Fan, J.; Ma, Y.; Dai, W. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Am. Statist. Assoc.* **2014**, *109*, 1270–1284. [CrossRef] [PubMed]
7. Liu, J.; Li, R.; Wu, R. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Am. Statist. Assoc.* **2014**, *109*, 266–274. [CrossRef]
8. Zhu, L.; Li, L.; Li, R.; Zhu, L. Model-free feature screening for ultrahigh-dimensional data. *J. Am. Statist. Assoc.* **2011**, *106*, 1464–1475. [CrossRef]
9. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Statist. Assoc.* **2012**, *107*, 1129–1139. [CrossRef]
10. He, X.; Wang, L.; Hong, H. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **2013**, *41*, 342–369. [CrossRef]
11. Lin, L.; Sun, J.; Zhu, L. Nonparametric feature screening. *Comput. Statist. Data Anal.* **2013**, *67*, 162–174. [CrossRef]
12. Lu, J.; Lin, L. Model-free conditional screening via conditional distance correlation. *Statist. Pap.* **2020**, *61*, 225–244. [CrossRef]
13. Tong, Z.; Cai, Z.; Yang, S.; Li, R. Model-Free Conditional Feature Screening with FDR Control. *J. Am. Statist. Assoc.* **2002**. [CrossRef]
14. Guo, X.; Ren, H.; Zou, C.; Li, R. Threshold Selection in Feature Screening for Error Rate Control. *J. Am. Statist. Assoc.* **2022**, *36*, 1–13. [CrossRef]
15. Zhong, W.; Qian, C.; Liu, W.; Zhu, L.; Li, R. Feature Screening for Interval-Valued Response with Application to Study Association between Posted Salary and Required Skills. *J. Am. Statist. Assoc.* **2023**. [CrossRef]
16. Fan, J.; Feng, Y.; Tong, X. A road to classification in high dimensional space: The regularized optimal affine discriminant. *J. R. Statist. Soc. B* **2012**, *74*, 745–771. [CrossRef]
17. Huang, D.; Li, R.; Wang, H. Feature screening for ultrahigh dimensional categorical data with applications. *J. Bus. Econ. Stat.* **2014**, *32*, 237–244. [CrossRef]
18. Pan, R.; Wang, H.; Li, R. Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *J. Am. Statist. Assoc.* **2016**, *111*, 169–179. [CrossRef]
19. Mai, Q.; Zou, H. The fused kolmogorov filter: A nonparametric model-free feature screening. *Ann. Statist.* **2015**, *43*, 1471–1497. [CrossRef]
20. Cui, H.; Li, R.; Zhong, W. Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Am. Statist. Assoc.* **2015**, *110*, 630–641. [CrossRef]
21. Xie, J.; Lin, Y.; Yan, X.; Tang, N. Category-Adaptive Variable Screening for Ultra-high Dimensional Heterogeneous Categorical Data. *J. Am. Statist. Assoc.* **2019**, *36*, 747–760. [CrossRef]
22. Shao, J. *Mathematical Statistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2003.
23. Meier, L.; van de Geer, S.; Buhlmann, P. High-Dimensional Additive Modeling. *Ann. Statist.* **2009**, *37*, 3779–3821. [CrossRef]
24. Dettling, M. Bagboosting for tumor classification with gene expression data. *Bioinformatics* **2004**, *20*, 3583–3593. [CrossRef] [PubMed]
25. Witten, D.M.; Tibshirani, R. Penalized classification using fisher's linear discriminant. *J. R. Statist. Soc. B* **2011**, *73*, 753–772. [CrossRef]
26. Clemmensen, L.; Hastie, T.; Witten, D.; Ersbøll, B. Sparse discriminant analysis. *Technometrics* **2011**, *53*, 406–413. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.