PREPRINT

Robust Feature Screening Procedures for Single and Mixed types of Data

Jinhui Sun[a], Pang Du[b], Hongyu Miao[c], and Hua Liang[d]

[a]Ping An Technology Co., Ltd., Beijing 100028, China; [b]Department of Statistics, Virginia Tech, Blacksburg, Virginia 24061, U. S. A.; [c]Department of Biostatistics, University of Texas Health Science Center at Houston, Houston, Texas 77030, U. S. A.; [d]Department of Statistics, George Washington University, Washington, D.C. 20052, U. S. A.;

ARTICLE HISTORY
Compiled January 30, 2020

ABSTRACT
Feature screening procedures aim to reducing the dimensionality of data with exponentially-growing dimensions. Existing procedures all focused on a single type of predictors, which are either all continuous or all discrete. They cannot address mixed types of variables, outliers, or nonlinear trends. In this paper we first propose new feature screening procedure(s) for different continuous/discrete combinations of response and predictor variables. They are respectively based on marginal Spearman correlation, marginal ANOVA test, marginal Kruskal-Wallis test, Kolmogorov-Smirnov test, Mann-Whitney test, and smoothing splines modeling. Extensive simulation studies are performed to compare the new and existing procedures, with the aim of identifying a best robust screening procedure for each single type of data. Then we combine these best screening procedures to form the robust feature screening procedure for mixed type of data. We demonstrate its robustness against outliers and model misspecification through simulation studies and a real example.

KEYWORDS
Feature screening procedures; ultra-high dimensions; mixed types of data; robust

CONTACT Pang Du. Email: pangdu@vt.edu

1. Introduction

High dimensional data analysis has become increasingly frequent and important in many areas such as economics, finance, health sciences and machine learning. Variable selection and feature extraction play a crucial role in knowledge discovery in all of these areas. Classical model selection methods have been developed and applied to different areas for many decades. Traditional variable selection, for example, by AIC[1],BIC[2],Mallow's C_p[3], RIC[4] and GCV[5], involves an NP-hard combinatorial optimization problem. It is natural that these classical variable selection methods use penalized L_0 regularization, which gives a nice interpretation of best subset selection and admits nice sampling properties[6]. However, the expensive computational cost makes classical procedures infeasible for high dimensional data analysis. Therefore, in the past decades a number of penalization methods have been developed to exploit the sparse nature of such high dimensional data. Some well-known examples are the bridge penalty [7], the least absolute shrinkage and selection operator (Lasso) penalty [8], the elastic net penalty [9], the adaptive Lasso penalty [10], the smoothly clipped absolute deviation (SCAD) penalty [11], and the minimax concave penalty [12]. A more comprehensive list of such methods can be found at the review [13].

The aforementioned regularization methods can comfortably deal with high dimensional cases when the number of predictors p is almost as large as the sample size n. But they may have difficulty when p can increase in an exponential order $\exp\{O(n^\alpha)\}$ of the sample size n, where $\alpha > 0$. For example, bioinformatic studies often see data with a sample size of a few hundreds and predictors of tens of thousand dimensions. To deal with the ultra high dimensionality, one appealing idea is to first use a fast, reliable and efficient method to reduce the dimensionality p from an ultra-high scale to a relatively large scale d (e.g., $O(n^b)$ for some $b > 0$), then the regularization methods can be applied to the reduced feature space. This motivates the sure independence screening (SIS) procedure introduced in [14] which ranks the predictors according to the magnitudes of their individual sample correlations with the response and keeps only the top-ranked predictors in the model. They have shown that SIS possesses the *sure screening property*, that is, it can detect a subset of covariates which contains the important ones and its size is much smaller than p.

Since then, a number of extensions have emerged to refine the procedure or generalize this idea to various settings. For example, Fan and Lv [14] provided an iterative SIS procedure (ISIS) by iteratively replacing the response with the residual obtained from the regression of the response on selected covariates in the previous step. Wang [15] studied the property of forward regression with ultrahigh-dimensional predictors and proposed using the extended BIC [16] to determine the size of the active predictor set. Hall and Miller [17] proposed using the generalized correlation as a marginal screening utility and ranking all predictors based on the magnitude of estimated generalized correlation. Li et al. [18] proposed a robust rank correlation screening(RRCS) procedure based on the Kendall rank correlation to deal with the heavy-tail distributions. And the RRCS procedure is robust to outliers and influence points in the observations, which is not the case for the Pearson correlation in the SIS procedure. Besides the linear model, [19] and [20] also considered the SIS procedure for generalized linear models. And [21] proposed a nonparametric independence screening(NIS) procedure for an ultra-high dimensional additive nonparametric regression model. Furthermore, [22] proposed a model-free variable screening procedure called the sure independent ranking screening(SIRS) procedure.

All the aforementioned screening procedures only deal with continuous predictors.

There is also some work focusing on the screening procedures for discrete predictors only. For example, when the response is also categorical, Huang et al. [23] employed the Pearson χ^2 test statistic as a marginal utility for feature screening. Screening procedures have also been studied under the scenario of classification. For example, Fan and Fan [24] proposed using two sample t-statistic as the marginal utility for feature screening in high dimensional binary classification. However, this procedure may break down for heavy-tailed distributions or data with outliers. To overcome this drawback, Mai and Zou [25] proposed a feature screening method for binary classification based on the Kolmogorov-Smirnov statistic. Besides establishing the sure screening property, they also showed that this method is almost as fast as the t-test screening [24] and is ten times faster than nonparametric maximum marginal likelihood screening [21]. Cui et al. [26] proposed a model-free feature screening procedure using mean variance index for ultra high dimensional discriminant analysis. It is not only robust to heavy-tailed distributions of predictors and the presence of potential outliers, but also allows the categorical response having a diverging number of classes in the order of $\mathbf{O}(n^k)$ with some $k \geq 0$.

One common drawback of these existing methods is that they all focus on single type of predictors, which means the predictors are all continuous or all discrete. However, in practice, we often collect mixed type of data, which contains both continuous and discrete predictors. For example, in genetic studies, researchers can collect information on both gene expression profiles and single nucleotide polymorphisms (SNPs) genotypes. Numerous gene expression(continuous variables) based strategies have been developed ([27], [28], and [29]) and many methods have been developed for pathway analyses using SNP data([30], [31], and [32]). As discussed by Xiong et al.[33], valuable associations may be discarded in single data type analyses. For instance, genes with only genetic alterations are not considered in gene set analyses based solely on expression data. Similarly, genes with only expression changes cannot be captured by a purely SNP-based approach. These issues create a need to integrate both gene expression and SNPs into the association analysis of gene sets. This motivates us to develop a feature screening procedure for mixed type of data. Furthermore, data with ultra-high dimensions are often contaminated with outliers. Many existing screening procedures may suffer from such contamination. And many procedures assume strict parametric models that might not be realistic for most practical data. Therefore, in this paper we are interested in developing screening procedures for mixed type of data that are robust against outliers and model misspecification.

Our new development is divided in two parts. In Section 2, we first focus on feature screening procedures for single type of data. For each type of data, we propose a new robust procedure and conduct simulation studies to assess the performance of the proposed procedure and compare them with existing procedures. The goal of this part is to identify a candidate robust screening procedure for each type of data which will be combined together to form the robust screening procedure for mixed type of data in the next part. Our contribution and findings in this section can be summarized below. For models with a continuous response and continuous predictors, we propose a robust screening method by the marginal Spearman correlation and our simulations show that our Spearman correlation screening procedure and the RRCS [18] procedure are the most robust procedures against all the types of outliers considered in our simulations. For models with a continuous response and categorical predictors, we propose two screening procedures respectively by the marginal ANOVA and Kruskal-Wallis tests, and our simulations show that the ANOVA screening procedure is the best when there is no outlier but the Kruskal-Wallis screening and the RRCS are better when

there are outliers. For models with a categorical response and continuous predictors, we propose two screening procedures respectively by the Kolmogorov-Smirnov and Mann-Whitney tests, and our simulations show that the Mann-Whitney test outperforms all the other competitors. For nonparametric models with continuous predictors, we propose a screening procedure by smoothing splines modeling of the predictor effects and our simulations show that the smoothing spline screening procedure is the best for such models together with the NIS procedure [21]. In summary, all the new screening procedures we propose for models with a single type of predictor variables are either competitive or better than the existing methods in general, and especially so when there are outliers. Our extensive simulations also provide important numerical comparisons of all the existing methods that are not available anywhere else.

In Section 3, we propose robust screening procedures for mixed type of data based on our findings in Section 2. When the response is continuous, we propose a screening procedure combining the B-spline modeling of continuous predictors as in the NIS method and the ANOVA/Kruskal-Wallis screening for categorical predictors. When the response is categorical, we propose a screening procedure combining the Chi-square test for continuous predictors with the ANOVA/Kruskal-Wallis test for categorical predictors. We conduct extensive simulation studies to evaluate the performance of the proposed procedure options for both types of responses. We further illustrate the procedure using a real-life data example in Section 5.

2. Screen procedures for single type of data

In this section, we focus on feature screening procedures for single type of data and aim to identify a best robust candidate screening procedure for each type of data, which will be combined together to form the screening procedure for mixed type of data. For models with a continuous response and continuous predictors, we introduce the Spearman correlation screening procedure and conduct simulation studies to compare the performance with SIS [14], RRCS [18], CQC-SIS [34] and DC-SIS [35]. For models with a continuous response and categorical predictors, we introduce the screening procedures respectively by the ANOVA and Kruskal-Wallis test and conduct simulation studies to compare their performances with SIS and RRCS. For categorical response and continuous predictors, we introduce the screening procedures respectively by the Kolmogorov-Smirnov and Mann-Whitney tests and conduct simulation studies to compare their performances with NIS [21] and SIRS [22]. For models with a categorical response and categorical predictors, screening with χ^2 test statistic[23] seems to be the only option. These studies for single type of data prepare us for developing a robust procedure for data of mixed types.

2.1. *Models with a continuous response and continuous predictors*

2.1.1. Robust Screening by Spearman Correlation

Consider the random vectors (X_i, Y_i), $i = 1, \ldots, n$. After converting the raw values X_i, Y_i to ranks rgX_i, rgY_i, the Spearman's ρ rank correlation between X_i and Y_i is defined as $\rho = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}$, where $cov(rg_X, rg_Y)$, σ_{rg_X} and σ_{rg_Y} are respectively the covariance and standard deviations of the rank variables. If there are no ties, it can be computed as $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where $d_i = rg(X_i) - rg(Y_i)$ is the difference between

the two ranks at the ith observation.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be an n-vector of response, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be an $n \times p$ design matrix. Define $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T$ with components $\omega_k = 1 - \frac{6\sum_{i=1}^{n} d_{ik}^2}{n(n^2-1)}$, $k = 1, \ldots, p$, where $d_{ik} = rg(X_{ik}) - rg(Y_i)$. Then ω_k is essentially the marginal rank correlation coefficient between $\mathbf{Y}$ and $\mathbf{X}_k$. We can sort the magnitudes of all the components of $\boldsymbol{\omega}$ in a decreasing order and select a submodel

$$\mathcal{M}_{d_n} = \{1 \leq k \leq p : |\omega_k| \text{ is among the first } d_n \text{ largest of all}\},$$

where d_n is a predefined threshold value. This reduces the full model of size p to a submodel with the size d_n.

The Spearman rank correlation between two variables is the nonparametric version of the Pearson correlation and equal to the Pearson correlation between the rank values of those two variables. Because of the robustness of the Spearman correlation against heavy-tailed distributions and outliers, a screening method using Spearman correlation is expected to be more robust than the SIS.

2.1.2. Numerical Studies

In this section, we present simulations to compare the performances of the Spearman correlation screening procedure with the existing methods, such as SIS([14]), RRCS([18]), CQC-SIS([34]) and DC-SIS([35]).

We used the linear model with multivariate normal predictors and the noise ε was generated from three different distributions: the standard normal distribution, the standard normal distribution with 10% of the outliers following the Cauchy distribution, and all the errors from the $t(1)$ distribution. We considered two settings with $(n, p) = (100, 1000)$ and $(200, 1000)$, respectively. The sizes s of the true models, i.e., the numbers of nonzero coefficients, were chosen to be 5 and 8, respectively, and the nonzero values of the p-vectors β were all set to be 5. We considered three designs for the predictor covariance matrix: (1) $\Sigma_1 = I_{p\times p}$; (2) $\Sigma_2 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.5$; (3) $\Sigma_3 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.8$. To examine robustness against outliers in the predictors, we also considered the scenarios of having 10% of outlier predictors in each design that were generated from multivariate t-distribution with $t(1)$ marginal distributions and the same covariance matrix as the multivariate normal distribution in the design, using the `rmvt` function in the R pakcage `mvtnorm`. We chose $d = [n/\log n]$. For each model we simulated 500 data sets.

We used the median number of correctly selected predictors and the proportion of times that the screened predictor set contained the true model to evaluate the performances of the procedures. Tables 1 and 2 summarized the simulation results and we can draw the following conclusions:

(1) When there were no outliers, SIS and CQC-SIS performed better than others, yielding higher proportions of predictors containing the true model selected. The difference became smaller with a larger sample size. But when outliers were present in data, Spearman, RRCS and CQC-SIS performed much better than others. SIS was very sensitive to outliers.

(2) Spearman, RRCS and CQC-SIS could outperform DC-SIS with or without outliers in response. When outliers were in predictors, Spearman and RRCS performed much better than the others. Generally speaking, the performance of Spearman and RRCS were the best.

Table 1. Results of simulation comparisons for the methods SIS, Spearman, RRCS, CQC-SIS, DC-SIS with $n = 100$, $p = 1000$ and $s = 5$ in Section 2.1.2. Four outlier types are: no outliers (None), 10% Cauchy errors (ϵ : 10% Cauchy), 100% $t(1)$ errors ($\epsilon \sim t(1)$), and 10% multivariate $t(1)$ predictors ($\mathbf{x}$: 10% MVT). The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

ρ	Outlier Type	SIS	Spearman	RRCS	CQC-SIS	DC-SIS
0	None	5	5	5	5	5
0	ϵ : 10% Cauchy	5	5	5	5	5
0	$\epsilon \sim t(1)$	2	5	5	5	5
0	$\mathbf{x}$: 10% MVT	1	5	5	3	3
0.5	None	5	5	5	5	5
0.5	ϵ : 10% Cauchy	5	5	5	5	5
0.5	$\epsilon \sim t(1)$	2	5	5	5	5
0.5	$\mathbf{x}$: 10% MVT	1	5	5	3	4
0.8	None	5	5	5	5	5
0.8	ϵ : 10% Cauchy	5	5	5	5	5
0.8	$\epsilon \sim t(1)$	2	4	4	4	4
0.8	$\mathbf{x}$: 10% MVT	1	5	5	3	3
0	None	0.940	0.924	0.930	0.944	0.898
0	ϵ : 10% Cauchy	0.848	0.904	0.896	0.936	0.888
0	$\epsilon \sim t(1)$	0.144	0.768	0.776	0.802	0.642
0	$\mathbf{x}$: 10% MVT	0.040	0.856	0.864	0.248	0.296
0.5	None	0.910	0.902	0.904	0.922	0.894
0.5	ϵ : 10% Cauchy	0.704	0.824	0.848	0.832	0.800
0.5	$\epsilon \sim t(1)$	0.104	0.704	0.712	0.768	0.576
0.5	$\mathbf{x}$: 10% MVT	0.032	0.832	0.832	0.248	0.328
0.8	None	0.744	0.712	0.718	0.744	0.676
0.8	ϵ : 10% Cauchy	0.664	0.706	0.712	0.726	0.654
0.8	$\epsilon \sim t(1)$	0.062	0.428	0.398	0.446	0.312
0.8	$\mathbf{x}$: 10% MVT	0.012	0.708	0.712	0.208	0.284

(3) With the increase of the sample size, they all had improved performances.

2.2. *Models with a continuous response and categorical predictors*

2.2.1. Screening by the ANOVA and Kruskal-Wallis Tests

Given observations $(X_i, Y_i), i = 1, \ldots, n$, of a continuous variable Y and a categorical variable X, where $X_i \in \{1, \ldots, K\}$ is the observed class label. We can divide the n-vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ into K groups according to the corresponding class label X_i. Then we can perform a one-way ANOVA to test whether the means of the K groups are all the same. The p-value of the test indicates the level of association between Y and X.

The ANOVA model assumes that $\mathbf{Y}$ is normally distributed. When this assumption does not hold, we can use the Kruskal-Wallis test [36], which is the nonparametric equivalent of ANOVA. Let n_i represent the sample size for the ith group, $i = 1, \ldots, K$. Rank all the observations and compute R_i, the sum of the ranks for group i. Then the

Table 2. Results of simulation comparisons for the methods SIS, Spearman, RRCS, CQC-SIS, DC-SIS with $n = 200$, $p = 1000$ and $s = 8$ in Section 2.1.2. Four outlier types are: no outliers (None), 10% Cauchy errors (ϵ : 10% Cauchy), 100% $t(1)$ errors ($\epsilon \sim t(1)$), and 10% multivariate $t(1)$ predictors ($\mathbf{x}$: 10% MVT). The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

ρ	Outlier Type	SIS	Spear-man	RRCS	CQC-SIS	DC-SIS
0	None	8	8	8	8	8
0	ϵ : 10% Cauchy	8	8	8	8	8
0	$\epsilon \sim t(1)$	5	8	8	8	8
0	$\mathbf{x}$: 10% MVT	1	8	8	5	5
0.5	None	8	8	8	8	8
0.5	ϵ : 10% Cauchy	8	8	8	8	8
0.5	$\epsilon \sim t(1)$	3	8	8	8	8
0.5	$\mathbf{x}$: 10% MVT	1	8	8	5	5
0.8	None	8	8	8	8	8
0.8	ϵ : 10% Cauchy	8	8	8	8	8
0.8	$\epsilon \sim t(1)$	3	8	8	8	7
0.8	$\mathbf{x}$: 10% MVT	1	8	8	5	5
0	None	0.988	0.982	0.982	0.986	0.974
0	ϵ : 10% Cauchy	0.804	0.966	0.966	0.978	0.952
0	$\epsilon \sim t(1)$	0.124	0.858	0.862	0.812	0.762
0	$\mathbf{x}$: 10% MVT	0.008	0.944	0.944	0.112	0.246
0.5	None	0.938	0.926	0.926	0.946	0.904
0.5	ϵ : 10% Cauchy	0.782	0.930	0.938	0.948	0.902
0.5	$\epsilon \sim t(1)$	0.088	0.832	0.866	0.822	0.714
0.5	$\mathbf{x}$: 10% MVT	0.000	0.914	0.922	0.066	0.190
0.8	None	0.738	0.706	0.710	0.756	0.688
0.8	ϵ : 10% Cauchy	0.570	0.682	0.682	0.716	0.658
0.8	$\epsilon \sim t(1)$	0.018	0.572	0.554	0.552	0.428
0.8	$\mathbf{x}$: 10% MVT	0.000	0.692	0.706	0.092	0.172

Kruskal-Wallis test statistic is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1). \tag{1}$$

This statistic approximately follows a χ^2 distribution with $K-1$ degrees of freedom if the null hypothesis is true.

Let $\mathbf{X}_j = (\mathbf{X}_{1j}, \ldots, \mathbf{X}_{nj})^T$ be the vector of observed values for the jth categorical predictor and $\omega = (\omega_1, \ldots, \omega_p)^T$ be the vector of p-values of tests on the marginal association between $\mathbf{Y}$ and $\mathbf{X}_j$. We can then sort the magnitudes of all the components of ω in an increasing order and select a submodel

$$\mathcal{M}_{d_n} = \{1 \leq k \leq p : \omega_k \text{ is among the first } d_n \text{ smallest of all}\},$$

where d_n is a predefined threshold value. This reduces the full model of size p to a submodel with size d_n.

2.2.2. Numerical Studies

In this section, we present several simulations to compare the performances of four methods: screening by the ANOVA test, screening by the Kruskal-Wallis test, SIS([14]) and RRCS([18]).

The true model was the linear model with p binary predictors where only s of them had nonzero coefficients. The random error was generated from two different distributions: the standard normal distribution and the standard t distribution with one degree of freedom. We considered two settings with $(n, p) = (100, 1000)$ and $(200, 1000)$, respectively. The true size s of the model was chosen to be 5 or 8, with all the nonzero components of the coefficient vector β equal to 5. We considered the same three designs for the predictor covariance matrix as in Section 2.1.2: (1) $\mathbf{\Sigma}_1 = \mathbf{I}_{p\times p}$ (2) $\mathbf{\Sigma}_2 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.5$; (3) $\mathbf{\Sigma}_3 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.8$. We chose $d = [n/\log n]$. For each model we simulated 500 data sets. Table 3 and Table 4 summarized the simulation results and we can draw the following conclusions:

(1) With the standard normal noise, the ANOVA test performed better than the others, yielding higher proportions of predictors containing the true model selected. The difference became smaller with a larger sample size. But with the t distribution noise, the Kruskal-Wallis test and RRCS performed much better than the others.
(2) Generally speaking, the performance of the Kruskal-Wallis test and RRCS were the best.
(3) With the increase of the sample size, they all had improved performances.
(4) An interesting finding was that: the performance of the Kruskal-Wallis test and RRCS were the same in almost all the settings. This may be due to their common nonparametric nature.

Table 3. Results of simulation comparisons for the ANOVA, Kruskal-Wallis, SIS and RRCS methods with $n = 100$, $p = 1000$ and $s = 5$ in Section 2.2.2. Two error distributions, $N(0, 1)$ and $t(1)$, are considered. The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

ρ	ε	ANOVA	K-W	SIS	RRCS
0	$N(0,1)$	5	5	2	5
0	$t(1)$	2	5	1	5
0.5	$N(0,1)$	5	5	2	5
0.5	$t(1)$	3	5	1	5
0.8	$N(0,1)$	5	5	2	5
0.8	$t(1)$	2	4	1	4
0	$N(0,1)$	0.972	0.928	0.008	0.928
0	$t(1)$	0.136	0.626	0.004	0.626
0.5	$N(0,1)$	0.946	0.880	0.004	0.880
0.5	$t(1)$	0.112	0.610	0.000	0.610
0.8	$N(0,1)$	0.912	0.826	0.002	0.826
0.8	$t(1)$	0.076	0.478	0.002	0.478

Table 4. Results of simulation comparisons for the ANOVA, Kruskal-Wallis, SIS and RRCS methods with $n = 200$, $p = 1000$ and $s = 8$ in Section 2.2.2. Two error distributions, $N(0, 1)$ and $t(1)$, are considered. The top half contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

ρ	ε	ANOVA	K-W	SIS	RRCS
0	$N(0,1)$	8	8	3	8
0	$t(1)$	4	8	2	8
0.5	$N(0,1)$	8	8	3	8
0.5	$t(1)$	4	8	2	8
0.8	$N(0,1)$	8	8	3	8
0.8	$t(1)$	5	8	2	8
0	$N(0,1)$	0.998	0.996	0.000	0.996
0	$t(1)$	0.148	0.870	0.000	0.870
0.5	$N(0,1)$	0.976	0.970	0.002	0.970
0.5	$t(1)$	0.130	0.872	0.000	0.872
0.8	$N(0,1)$	0.960	0.932	0.004	0.934
0.8	$t(1)$	0.134	0.764	0.000	0.764

2.3. *Models with a categorical response and continuous predictors*

2.3.1. Screening by the Kolmogorov-Smirnov and Mann-Whitney Tests

We first review the Kolmogorov-Smirnov and Mann-Whitney tests for testing whether two samples come from the same distribution. The Kolmogorov-Smirnov test is a nonparametric hypothesis test that evaluates the difference between the cumulative distribution functions (c.d.f.) of the two sample data vectors over the data range. Suppose that the first sample $X_1, \ldots, X_m$ of size m has a distribution with c.d.f. $\mathbf{F}_1(x)$ and the second sample $Y_1, \ldots, Y_n$ of size n has a distribution with c.d.f. $\mathbf{F}_2(x)$. The Kolmogorov-Smirnov statistic is $D = \max_x |\mathbf{F}_1(x) - \mathbf{F}_2(x)|$, which is the maximum absolute value of the differences between the two c.d.f.s. A natural estimator for D is

$$\hat{D}_{mn} = \max_x |\hat{\mathbf{F}}_1(x) - \hat{\mathbf{F}}_2(x)|, \tag{2}$$

where $\hat{\mathbf{F}}_1$ and $\hat{\mathbf{F}}_1$ are the sample c.d.f.s. The null hypothesis of two samples having the same distribution is rejected at level α if $\hat{D}_{mn} > c(\alpha)\sqrt{\frac{m+n}{mn}}$, where $c(\alpha)$ is the critical value for the Kolmogorov-Smirnov distribution.

The Mann-Whitney test is another non-parametric test that can be used to test whether two samples come from the same distribution. It is based on a comparison of every observation in the first sample with every observation in the other sample. Suppose we have a sample $X_1, \ldots, X_m$ of size m and another sample $Z_1, \ldots, Z_n$ of size n. To calculate the test statistic, one first ranks all the pooled observations and let S_j be the rank of Z_j in this joint ordering. When there are ties, S_j is computed as the average rank of all the observations that are tied. Then the Mann-Whitney test statistic is defined as $U = \sum_{j=1}^{n} S_j - n(n+1)/2$. Note that if the number of observations is large enough, a normal approximation can be used with $\mu_U = \frac{mn}{2}$, $\sigma_U = \sqrt{\frac{mn(m+n+1)}{12}}$.

Both the Kolmogorov-Smirnov and Mann-Whitney tests are nonparametric tests to compare two unpaired groups of data. Both compute p-values for testing the null hy-

pothesis that the two groups have the same distribution. The Kolmogorov-Smirnov test is sensitive to any distributional differences. Substantial differences in shape, spread or median will result in a small p-value. In contrast, the Mann-Whitney test is mostly sensitive to changes in the median. Both tests can be used when we have two groups. When we have three or more groups, we can use the Kruskal-Wallis test as described in Section 2.2.1.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be an n-vector of categorical responses where $Y_i \in \{1, \ldots, K\}$ is the ith class label, and $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})^T$ be the jth continuous predictor. For each pair of $\mathbf{Y}$ and $\mathbf{X}_j$, we can divide $\mathbf{X}_j$ into K groups according to the class label Y_i and perform a test to see whether the K groups come from the same distribution. Let $\omega = (\omega_1, \ldots, \omega_p)^T$ be a p-vector each being the p-value of the selected test. We can then sort the magnitudes of all the components of ω in an increasing order and select a submodel

$$\mathcal{M}_{d_n} = \{1 \leq k \leq p : \omega_k \text{ is among the first } d_n \text{ smallest of all}\},$$

where d_n is a predefined threshold value. This reduces the full model of size p to a submodel with the size d_n.

2.3.2. Numerical Studies

In this section, we present simulations to compare the performances of four methods: NIS([21]), SIRS([22]), screening with the Kolmogorov-Smirnov test (K-S) and screening with the Mann-Whitney test (M-W).

In this example, the observations were independently generated such that $Y|\mathbf{X} = \mathbf{x}$ is distributed as Binomial$(1, p(\mathbf{x}))$, with $\log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$ where the additional noise variable ε was inserted into the model to create outliers. The noise variable ε was generated from three different distributions: the standard normal distribution, 90% from standard normal distribution and 10% from the Cauchy distribution, and all from the $t(1)$ distribution. We chose $n = 200, p = 1000$. The true size s of the model was chosen to be 8 and the nonzero components of the coefficient vector β were all equal to 5. We considered three designs for the predictor covariance matrix: (1) $\boldsymbol{\Sigma}_1 = \mathbf{I}_{p\times p}$ (2) $\boldsymbol{\Sigma}_2 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.5$; (3) $\boldsymbol{\Sigma}_3 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.8$. We chose $d = [n/\log n]$. For each model we simulated 500 data sets. Table 5 summarized the simulation results and we can draw the following conclusions:

(1) The Mann-Whitney test outperformed the other three methods.
(2) With the increase of ρ, the performances all became worse.

2.4. *Nonparametric screening for models with continuous predictors*

When continuous predictors are involved, the underlying form of their effects may not necessarily be linear as considered in the previous sections. This motivates screening procedures under a nonparametric regression model. The NIS procedure developed in [21] is such an example where each marginal nonparametric regression model is fitted by B-splines.

Table 5. Results of simulation comparisons for the NIS, SIRS, Mann-Whitney test and Kolmogorov-Smirnov test methods with $n = 200$, $p = 1000$ and $s = 8$ in Section 2.3.2. Three noise types are: all from $N(0,1)$, 90% from $N(0,1)$ + 10% from Cauchy, and all from $t(1)$. The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

ρ	Noise Type	NIS	SIRS	M-W	K-S
0	$N(0,1)$	8	6	8	7
0	90% $N(0,1)+10\%$ Cauchy	7	6	8	7
0	$t(1)$	7	5	8	7
0.5	$N(0,1)$	7	6	8	7
0.5	90% $N(0,1)+10\%$ Cauchy	7	6	8	7
0.5	$t(1)$	7	5	8	7
0.8	$N(0,1)$	7	5	7	7
0.8	90% $N(0,1)+10\%$ Cauchy	7	5	7	6
0.8	$t(1)$	6	5	7	6
0	$N(0,1)$	0.518	0.042	0.714	0.436
0	90% $N(0,1)+10\%$ Cauchy	0.368	0.048	0.728	0.408
0	$t(1)$	0.332	0.000	0.632	0.404
0.5	$N(0,1)$	0.452	0.054	0.664	0.366
0.5	90% $N(0,1)+10\%$ Cauchy	0.384	0.024	0.642	0.328
0.5	$t(1)$	0.350	0.036	0.610	0.260
0.8	$N(0,1)$	0.256	0.008	0.304	0.144
0.8	90% $N(0,1)+10\%$ Cauchy	0.202	0.006	0.326	0.132
0.8	$t(1)$	0.116	0.004	0.242	0.080

2.4.1. Screening by smoothing splines modeling of predictor effects

B-splines are good for modeling simple nonlinear trends but may suffer when the nonlinear trend becomes more complicated. In this section, we consider using smoothing splines instead of B-splines for the marginal regression models. For responses generated from exponential family distributions, we assume the following marginal model for response Y given the jth predictor $X_j = x$

$$f(y|X_j = x) = \exp\left(y\eta_j(x) - b(\eta_j(x))\right)/a(\phi_j) + c(y, \phi_j), \tag{3}$$

where $a > 0$, b and c are known functions, $\eta_j(\cdot)$ is the marginal regression function of the jth predictor, and ϕ_j is either known or considered as a nuisance parameter. We use the smoothing splines in Chapter 5 of [37] to estimate η_j.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an n-vector of observed responses and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)^T$ be an $n \times p$ design matrix. For each pair of $\mathbf{Y}$ and $\mathbf{X}_j$, we can fit a marginal regression model by smoothing splines and get an estimate $\hat{\eta}_j$ for η_j. We then test the significance of the relationship by examining whether $\hat{\eta}_j$ is a constant function, or equivalently, $\hat{\eta}_j' \equiv 0$. Define $\omega_j = \sum_{i=1}^n \{\hat{\eta}_j'(X_{ij})\}^2$ and let $\omega = (\omega_1, \ldots, \omega_p)^T$. We can then sort the magnitudes of all the components of ω in an decreasing order and select a submodel

$$\mathcal{M}_{d_n} = \{1 \le k \le p : \omega_k \text{ is among the first } d_n \text{ largest of all}\},$$

where d_n is a predefined threshold value. This reduces the full model of size p to a submodel with the size d_n. Here we opt for examining ω_j instead of the p-values of a nonparametric marginal test since p-values for hypothesis tests involving nonpara-

metric smoothing tend to be conservative and may not be as accurate as those in parametric tests [38].

2.4.2. Numerical Studies

Continuous Response:

For continuous response, we compared the performance of screening by smoothing spline with SIS([14]), CQC-SIS([34]) and NIS([21]). We set $n = 400$ and $p = 1000$. For NIS, the number of basis is set to be 5 as suggested by Fan et al.[21]. For smoothing spline, the number of basis was set to be $\max(30, 10n^{2/9})$ and the modified GCV with $a = 1.4$ was used for smoothing parameter selection as suggested by Kim and Gu[39]. For each model we simulated 500 data sets.

Example 2.1. This example was adapted from [21]. Let $g_1(x) = x$, $g_2(x) = (2x-1)^2$, $g_3(x) = \sin(2\pi x)/(2-\sin(2\pi x))$ and $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$. The data were generated from the following model:

$$Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\varepsilon.$$

The covariates $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_P)^T$ are simulated according to the random-effects model

$$\mathbf{X}_j = \frac{\mathbf{W}_j + t\mathbf{U}}{1+t}, j = 1, \ldots, p,$$

where $\mathbf{W}_1, \ldots, \mathbf{W}_p$ and $\mathbf{U}$ are iid Uniform$(0,1)$, and $\varepsilon \sim N(0,1)$. When $t = 0$, the covariates are all independent, and when $t = 1$, the pairwise correlation of covariates is 0.5.

Example 2.2. The settings and model were the same as Example 2.1 except that the covariates $\mathbf{X} = (X_1, \ldots, X_p)^T$ were generated from the multivariate normal distribution with mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$ for $i \neq j$. We considered three cases: $\rho = 0.5$, $\varepsilon \sim N(0,1)$; $\rho = 0.8$, $\varepsilon \sim N(0,1)$; $\rho = 0.8$, $\varepsilon \sim t(1)$.

Based on the summary results in Table 6 we can draw the following conclusions:

(1) Generally speaking, the performance of NIS and smoothing splines were the best.
(2) When $\rho = 0.8$, the performances for all procedures became worse while both NIS and smoothing splines gave good performance.

Discrete Response from Exponential Family:

For discrete response from exponential family, we compare the performance of screening by smoothing spline with NIS([21]), SIRS([22]) and screening by p-values from the Kruskal-Wallis test. We set $n = 400$ and $p = 1000$. The choices of the number of basis functions for NIS and smoothing splines screening, as well as the smoothing parameter selection criterion for smoothing splines screening, were the same as those in the previous simulations for continuous responses. For each model we simulated 500 data sets.

Example 2.3. Let $g_1(x) = x^2$, $g_2(x) = x^3$ and $g_3(x) = exp(x)$. Y is distributed, conditional on $\mathbf{X} = \mathbf{x}$, as Binomial$(1, p(\mathbf{x}))$, with $\log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = 5g_1(\mathbf{X}_1) + 5g_2(\mathbf{X}_2) +$

Table 6. Results of simulation comparisons for the SIS, CQC-SIS, NIS and smoothing spline methods with data generated from models in Examples 2.1 and 2.2 of Section 2.4.2. The true number of active predictors is $s = 4$. For Example 2.1, two values (0 and 1) for the constant t are considered. For Example 2.2, the distribution for ε is always Uniform(0,1) and two values (0.5 and 0.8) of the constant ρ are considered. The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

Model Specification	SIS	CQC	NIS	SS
Example 2.1 ($t = 0$)	1	3	4	4
Example 2.1 ($t = 1$)	4	4	3	4
Example 2.2 ($\rho = 0.5$)	1	3	4	4
Example 2.2 ($\rho = 0.8$)	0	3	4	4
Example 2.1 ($t = 0$)	0.000	0.074	0.962	0.968
Example 2.1 ($t = 1$)	0.578	0.336	0.426	0.524
Example 2.2 ($\rho = 0.5$)	0.000	0.120	0.960	0.960
Example 2.2 ($\rho = 0.8$)	0.000	0.055	0.924	0.854

$5g_3(\mathbf{X}_3)$. The covariates $\mathbf{X}$ were generated from the multivariate normal distribution with mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$ for $i \neq j$. We considered two cases: $\rho = 0$ and $\rho = 0.8$.

Example 2.4. Let $g_1(x) = x^2$, $g_2(x) = x^3$ and $g_3(x) = \exp(x)$. Y is distributed, conditional on $\mathbf{X} = \mathbf{x}$, as $Poisson(\mu(\mathbf{x}))$, with $\log(\mu(\mathbf{x})) = 5g_1(\mathbf{X}_1) + 5g_2(\mathbf{X}_2) + 5g_3(\mathbf{X}_3)$. The covariates $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_P)^T$ are generated from the multivariate normal distribution with mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma} = (\sigma)_{p\times p}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho^{|i-j|}$ for $i \neq j$. We considered two cases: $\rho = 0$ and $\rho = 0.8$.

Based on the summary results in Table 7 we can draw the following conclusions:

(1) Generally speaking, the performance of NIS and smoothing spline are the best.
(2) The performance of NIS and smoothing splines were not much affected by the increase of correlations between predictors.

Table 7. Results of simulation comparisons for the NIS, SIRS, Kruskal-Wallis test and smoothing spline methods with data generated from models in Examples 2.3 and 2.4 of Section 2.4.2. The true number of active predictors is $s = 4$. For each example, three values (0, 0.5, and 0.8) for the constant ρ are considered. The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

Model Specification	NIS	SIRS	K-W	SS
Example 2.3 ($\rho = 0$)	4	3	3	4
Example 2.3 ($\rho = 0.5$)	4	3	3	4
Example 2.3 ($\rho = 0.8$)	4	3	3	4
Example 2.4 ($\rho = 0$)	4	3	3	4
Example 2.4 ($\rho = 0.5$)	4	3	3	4
Example 2.4 ($\rho = 0.8$)	4	3	3	4
Example 2.3 ($\rho = 0$)	0.876	0.012	0.018	0.886
Example 2.3 ($\rho = 0.5$)	0.774	0.032	0.026	0.868
Example 2.3 ($\rho = 0.8$)	0.850	0.000	0.000	0.870
Example 2.4 ($\rho = 0$)	0.732	0.062	0.088	0.786
Example 2.4 ($\rho = 0.5$)	0.712	0.032	0.048	0.736
Example 2.4 ($\rho = 0.8$)	0.748	0.018	0.026	0.728

We note that in both the continuous and discrete response cases, the performance of the NIS with B-splines is similar to that of smoothing spline screening. Therefore, we opt to use the NIS with B-splines for the development in the next section.

3. Screening procedure for mixed types of data

The studies for single type of data in Section 2 have prepared us to define a robust screening procedure for mixed type of ultra-high dimensional data. The best robust screening procedure for each type of data has been identified. We will combine these best screening procedures to form the robust feature screening procedure for mixed type of data.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be an n-vector response and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)^T$ be the $n \times p$ design matrix. For each pair of $\mathbf{Y}$ and $\mathbf{X}_j$, we want to perform a marginal test and the p-value of the test indicates the significance of the marginal relationship between the response and the predictor.

Case 1: Response Y is continuous:

If the predictor $\mathbf{X}_j$ is also continuous, we use the B-splines to estimate the marginal regression function similar to the NIS procedure. Consider the marginal model $Y = f_j(X_j) + \epsilon$. We estimate the marginal regression function $f_j(x)$ by the B-splines expansion $\hat{f}_j(x) = \hat{\beta}_j^T \mathbf{B}_j(x)$ where $\mathbf{B}_j(x) = \{\mathbf{B}_{j1}(x), \ldots, \mathbf{B}_{jd}(x)\}^T$ is the vector of B-spline basis functions. The coefficients $\hat{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})^T$ is computed as

$$\hat{\beta}_j = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - \beta_j^T \mathbf{B}_j(X_{ij})\}^2.$$

Then we can test whether $\hat{f}_j$ is a constant function and get a p-value.

When the predictor $\mathbf{X}_j$ is discrete, we can treat different values of the predictor as group labels. Then we perform a one-way ANOVA test or Kruskal-Wallis test. Suppose we have K groups, let $n_i (i = 1, \ldots, K)$ represent the sample sizes for each of the K groups. If we choose one-way ANOVA test, our test statistic would be:

$$F = \frac{\sum_{i=1}^{K} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 / (K-1)}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2 / (n-K)}.$$

This test statistics follows an F distribution with degrees of freedom $K-1$ and $n-K$. And we can get a p-value from the ANOVA test. If we choose the Kruskal-Wallis test, we need to rank the response, and compute $R_i =$ the sum of the ranks for group i. Then the Kruskal-Wallis test statistic is:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1).$$

This statistic approximately follows a χ^2 distribution with $K-1$ degrees of freedom and we can get a p-value from the K-W test.

Case 2: Response Y is discrete:

If the predictor $\mathbf{X}_j$ is continuous, we can treat different values of the response as group labels, then we can perform a one-way ANOVA test or Kruskal-Wallis test simi-

lar to the previous case but with the roles of X and Y switched. Both the corresponding test statistics and their approximate distributions are the same as above.

If the predictor $\mathbf{X}_j$ is also discrete, we can perform a Chi-square test. Suppose $Y_i \in \{1, \ldots, K_1\}$ and $X_{ij} \in \{1, \ldots, K_2\}$. Define $P(Y_i = k) = \pi_{yk}$, $P(X_{ij} = k) = \pi_{jk}$, and $P(Y_i = k_1, X_{ij} = k_2) = \pi_{yj,k_1k_2}$. Those quantities can be estimated by $\hat{\pi}_{yk} = n^{-1}\sum \mathbf{I}(Y_i = k)$, $\hat{\pi}_{jk} = n^{-1}\sum \mathbf{I}(X_{ij} = k)$, and $\hat{\pi}_{yj,k_1k_2} = n^{-1}\sum \mathbf{I}(Y_i = k_1)\mathbf{I}(X_{ij} = k_2)$. The Chi-square test statistic is

$$\hat{\triangle}_j = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \frac{(\hat{\pi}_{yk_1}\hat{\pi}_{jk_2} - \hat{\pi}_{yj,k_1k_2})^2}{\hat{\pi}_{yk_1}\hat{\pi}_{jk_2}}.$$

This test statistics follows a χ^2 distribution with $(K_1 - 1)(K_2 - 1)$ degrees of freedom and we can get a p-value from the test.

In either case for the type of the response variable, we combine the p-values of all the marginal tests for both continuous and discrete predictors to form the vector $\omega = (\omega_1, \ldots, \omega_p)^T$. We can then sort the magnitudes of all the components of ω in an decreasing order and select a submodel

$$\mathcal{M}_{d_n} = \{1 \le k \le p : \omega_k \text{ is among the first } d_n \text{ smallest of all}\},$$

where d_n is a predefined threshold value. This reduces the full model of size p to a submodel with the size d_n. Then the regularization methods, such as SCAD and MCP, can be applied to the reduced feature space.

4. Simulation Studies

In this section we perform simulations to study the empirical performance for the proposed screening procedure for mixed types of data. When the response is continuous, the marginal model for a continuous predictor is fitted by B-splines and there are two options for assessing the marginal effect of a discrete predictor: the ANOVA or the Kruskal-Wallis test. We shall denote them respectively by B-sp & ANOVA and B-sp & K-W. Similarly, when the response is discrete, the marginal effect of a discrete predictor is assessed by the Chi-square test and there are two options for assessing the marginal effect of a continuous predictor: the ANOVA or Kruskal-Wallis test. We shall denote them respectively by ANOVA & Chi-sq and K-W & Chi-sq.

Example 4.1. We considered the linear model $Y = X\boldsymbol{\beta} + \epsilon$. One half of the predictors were generated from multivariate normal distribution whose covariance matrix had one of the following two designs: (1) $\mathbf{\Sigma}_1 = \mathbf{I}_{p\times p}$; (2) $\mathbf{\Sigma}_3 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.8$. The other half of predictors were binary predictors generated from the Bernoulli(0.5) distribution. The random error ϵ was generated from three different distributions: all from the standard normal distribution, 90% from the standard normal and 10% from the Cauchy distributions, and all from the t distribution with three degrees of freedom. We chose $(n, p) = (400, 1000)$, $s = 8$, $d = [n/\log n]$ and the nonzero components of the p-vectors β were all equal to 5. For each model we simulated 500 data sets.

Example 4.2. Same as Example 4.1 except that $Y = X^2\boldsymbol{\beta} + \epsilon$.

Example 4.3. Same as Example 4.1 except that $Y = \sin(X)\boldsymbol{\beta} + \epsilon$.

Example 4.4. Let $g_1(x) = x$, $g_2(x) = x^2$ and $g_3(x) = \sin(x)$. The true model was $Y = 5g_1(X_1)+5g_2(X_2)+5g_3(X_3)+5g_1(X_4)+5g_2(X_5)+5g_3(X_6)+\epsilon$, where X_1, X_2, X_3 are continuous predictors and X_4, X_5, X_6 are binary predictors. The other settings are the same with Example 4.1.

Based on the summary results in Tables 8 and 9, we draw the following conclusions:

(1) Both options, B-sp & ANOVA and B-sp & K-W, performed better with standard normal noise and independent predictors shown by the higher proportions of selected predictors containing the true model.
(2) Generally both options performed well and the performances of the two options were comparable.

Table 8. Results of simulation comparisons for the B-sp & ANOVA and B-sp & K-W methods with data generated from models in Examples 4.1-4.3 of Section 4. The true number of active predictors is $s = 8$. In each example, two values (0 and 0.8) are used for the constant ρ and three distributions are used for simulating ε. The errors are generated respectively all from the standard normal distribution ($N(0,1)$), 90% from the standard normal and 10% from the Cauchy (10%), and all from the t distribution with 3 degrees of freedom ($t(3)$). The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

ρ & Errors ϵ	0&$N(0,1)$	0&10%	0&$t(3)$	0.8&$N(0,1)$	0.8&10%	0.8&$t(3)$
Ex 4.1/anova	8	8	8	8	8	8
Ex 4.1/K-W	8	8	8	8	8	8
Ex 4.2/anova	8	8	8	8	7	8
Ex 4.2/K-W	8	8	8	8	8	8
Ex 4.3/anova	8	8	8	8	8	8
Ex 4.3/K-W	8	8	8	8	8	8
Ex 4.1/anova	0.984	0.902	0.976	0.900	0.766	0.894
Ex 4.1/K-W	0.982	0.938	0.962	0.880	0.834	0.876
Ex 4.2/anova	0.748	0.656	0.738	0.636	0.488	0.620
Ex 4.2/K-W	0.882	0.864	0.888	0.824	0.768	0.820
Ex 4.3/anova	1.000	0.998	0.998	0.996	0.896	0.992
Ex 4.3/K-W	0.998	0.994	0.998	0.990	0.938	0.990

Example 4.5. Repeat Example 4.1 with $(n, p) = (100, 200)$ and $s = 6$.

Example 4.6. Repeat Example 4.2 with $(n, p) = (100, 200)$ and $s = 6$.

Example 4.7. Repeat Example 4.3 with $(n, p) = (100, 200)$ and $s = 6$.

Example 4.8. Repeat Example 4.4 with $(n, p) = (100, 200)$.

Example 4.9. To make the simulation mimic the arrhythmia application in the next section, we chose $(n, p, s) = (450, 250, 6)$ and Y is distributed as Binomial$(1, p(\mathbf{x}))$ conditional on $\mathbf{X} = \mathbf{x}$, with $\log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = \mathbf{x}^T\boldsymbol{\beta}$. The other settings were the same as Example 4.1.

Example 4.10. Same as Example 4.9 except that $\log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = (\mathbf{x}^2)^T\boldsymbol{\beta}$.

Example 4.11. Same as Example 4.9 except that $\log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = \sin(\mathbf{x})^T\boldsymbol{\beta}$.

Table 9. Results of simulation comparisons for the B-sp & ANOVA and B-sp & K-W methods with data generated from models in Examples 4.4-4.8 of Section 4. The true number of active predictors is $s = 6$. In each example, two values (0 and 0.8) are used for the constant ρ and three distributions are used for simulating ε. The errors are generated respectively all from the standard normal distribution ($N(0,1)$), 90% from the standard normal and 10% from the Cauchy (10%), and all from the t distribution with 3 degrees of freedom ($t(3)$). The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the half the proportions of times that the screened predictor set contained the true model.

ρ & Errors ϵ	0&$N(0,1)$	0&10%	0&$t(3)$	0.8&$N(0,1)$	0.8&10%	0.8&$t(3)$
Ex 4.4/anova	6	6	6	6	6	6
Ex 4.4/K-W	6	6	6	6	6	6
Ex 4.5/anova	6	6	6	6	6	6
Ex 4.5/K-W	6	6	6	6	6	6
Ex 4.6/anova	6	6	6	6	6	6
Ex 4.6/K-W	6	6	6	6	6	6
Ex 4.7/anova	6	6	6	6	6	6
Ex 4.7/K-W	6	6	6	6	6	6
Ex 4.8/anova	6	6	6	6	6	6
Ex 4.8/K-W	6	6	6	6	6	6
Ex 4.4/anova	0.986	0.912	0.978	1.000	0.866	1.000
Ex 4.4/K-W	0.996	0.958	0.996	1.000	0.906	1.000
Ex 4.5/anova	0.946	0.812	0.938	0.876	0.672	0.876
Ex 4.5/K-W	0.938	0.916	0.932	0.876	0.728	0.872
Ex 4.6/anova	0.846	0.608	0.826	0.756	0.504	0.740
Ex 4.6/K-W	0.822	0.842	0.778	0.718	0.782	0.684
Ex 4.7/anova	0.936	0.826	0.948	0.902	0.816	0.872
Ex 4.7/K-W	0.952	0.908	0.966	0.920	0.848	0.900
Ex 4.8/anova	0.724	0.696	0.696	0.644	0.648	0.642
Ex 4.8/K-W	0.774	0.812	0.726	0.674	0.776	0.674

Example 4.12. $(n, p) = (450, 250)$ and Y is distributed as Binomial$(1, p(\mathbf{x}))$ conditional on $\mathbf{X} = \mathbf{x}$, with $\log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = 5g_1(x_1) + 5g_2(x_2) + 5g_3(x_3) + 5g_1(x_4) + 5g_2(x_5) + 5g_3(x_6)$.

The summary results in Table 10 lead us to the following conclusions similar to the case with a continuous response:

(1) Both options, ANOVA & Chi-sq and K-W & Chi-sq, performed better with standard normal noise and independent predictors shown by the higher proportions of selected predictors containing the true model.
(2) Generally both options performed well and the performances of the two options were comparable.

5. Application: Arrhythmia

We apply the proposed screening procedure to the Arrhythmia data set downloaded from the UC-Irvine Machine Learning Respository *https* : *//archive.ics.uci.edu/ml/datasets/Arrhythmia*. The original data contained 452 patient records and 279 attributes, such as age, sex, height, weight and patients' ECG related measurements. The patients were divided into 16 different classes, with

Table 10. Results of simulation comparisons for the ANOVA & Chi-sq and K-W & Chi-sq methods with data generated from models in Examples 4.9-4.12 of Section 4. The true number of active predictors is $s = 6$. In each example, two values (0 and 0.8) are used for the constant ρ and three distributions are used for simulating ε. The errors are generated respectively all from the standard normal distribution ($N(0,1)$), 90% from the standard normal and 10% from the Cauchy (10%), and all from the t distribution with 3 degrees of freedom ($t(3)$). The top half of the table contains the empirical medians for the numbers of correctly selected variables, and the bottom half the proportions of times that the screened predictor set contained the true model.

ρ & Errors ϵ	0&$N(0,1)$	0&10%	0&$t(3)$	0.8&$N(0,1)$	0.8&10%	0.8&$t(3)$
Ex 4.9/anova	6	6	6	6	5	6
Ex 4.9/K-W	6	6	6	5	5	5
Ex 4.10/anova	5	5	5	5	5	5
Ex 4.10/K-W	6	5	6	6	5	5
Ex 4.11/anova	6	6	6	6	6	6
Ex 4.11/K-W	6	6	6	6	6	6
Ex 4.12/anova	6	5	6	5	5	5
Ex 4.12/K-W	6	6	6	6	5	5
Ex 4.9/anova	0.740	0.602	0.720	0.530	0.442	0.528
Ex 4.9/K-W	0.718	0.584	0.706	0.484	0.424	0.482
Ex 4.10/anova	0.416	0.454	0.406	0.318	0.434	0.306
Ex 4.10/K-W	0.606	0.476	0.588	0.506	0.442	0.486
Ex 4.11/anova	0.934	0.896	0.914	0.792	0.772	0.754
Ex 4.11/K-W	0.892	0.876	0.886	0.748	0.804	0.730
Ex 4.12/anova	0.586	0.486	0.580	0.494	0.456	0.404
Ex 4.12/K-W	0.692	0.538	0.648	0.522	0.496	0.468

class 1 corresponding to the normal ECG with no arrhythmia and classes 2 to 15 corresponding to different types of arrhythmia. We removed the patient records belonging to class 16, the unlabeled class. And we removed single-valued attributes and attributes with missing values. Therefore, we were left with 430 patients and 257 attributes. Among the attributes, 206 were continuous variables and the rest were nominal. We noticed that the data set had 245 patients belonging to the normal class (class 1) and many of the rest of the classes had very few patients. Therefore, we pooled all the patients in classes 2 to 15 into one abnormal group and aimed to distinguish normal from abnormal heartbeat behavior based on the 257 attributes.

We applied our screening procedure to the data set and used 10-fold cross validation to measure the classification accuracy. For continuous features, we used the ANOVA test (the results for the Kruskal-Wallis test option were similar and not presented here). For categorical features, we used the Chi-square test. The features were selected based on the p-value of the selected tests. The number of features selected was $d_t = [n_t/\log n_t]$, where n_t was the sample size of the training set. We applied the generalized linear model with SCAD penalty to further reduce the feature space. The fitted model from the training set was then used to obtain estimated classes for the test set. The classification accuracy was calculated using the estimated and true classes of the test set. We repeated the whole procedure 100 times.

From the study by Gupta et al.[40], we know that the performance of random forest is quite well compared with other classification methods. Therefore we compared the performance of our method with random forest. The results are summarized in Table 11. From the table, we can see that, with a much smaller model size and less computational time, the mean classification accuracy of our method was competitive to that of the random forest.

We also applied our screening procedure to the whole data set, which reduced the feature space to contain 73 features. After further application of the generalized linear model with SCAD penalty to the reduced feature space, we got 12 features in the final model: QRS duration, $DII90$, $DII91$, $DII93$, $DII100$, $DII103$, $DII112$, and $DI167$, $DI169$, $DII199$, $DII211$, $DII277$. We also applied the random forest method to the whole data set. The top 12 important features selected by model accuracy and Gini index were as follows. RF with model accuracy: $DII224$, $DII91$, $DII277$, $DII93$, $DII228$, $DII234$, $DII199$, $DII103$, $DII179$, $DII76$, QRS duration, and $DII250$. RF with Gini index: $DII224$, $DII277$, QRS duration, $DII199$, $DII197$, $DII91$, $DII179$, $DII93$, $DII228$, $DI167$, $DII177$, and $DI169$. Mitra and Samanta[41] also studied the Arrhythmia data set by neural network. Their final model contained 18 features: Sex, QRS duration, $DII49$, $DII76$, $DII91$, $DII103$, $DII112$, $DI163$, $DI167$, $DI169$, $DII173$, $DII199$, $DII207$, $DII211$, $DII261$, $DII267$, $DII271$, and $DII277$. Table 12 listed the important features selected by at least two methods mentioned above. From the table we can see that, 11 out of 12 features selected by our method were also selected by at least one different method. Only one feature, $DII76$, was selected by two other methods and was not selected by our method. Only one feature, $DII100$, was selected by our method and was not selected by other methods.

Table 11. Cross-validation results of the Arrhythmia data set: Mean values of the model size, classification accuracy and time.

Method	Model Size	Classification Accuracy	Time (seconds)
Screening	13.70	76.47%	19.58
Random Forest	257	80.09%	31.54

Table 12. Features selected by at least two methods

Attribute	Type	Screening +SCAD	RF-Gini	RF-Accuracy	Neural Networks
QRS	continuous	Y	Y	Y	Y
DII76	discrete	N	N	Y	Y
DII90	discrete	Y	N	Y	N
DII91	discrete	Y	Y	Y	Y
DII93	discrete	Y	Y	Y	N
DII103	discrete	Y	N	Y	Y
DII112	discrete	Y	N	N	Y
DI167	continuous	Y	Y	N	Y
DI169	continuous	Y	Y	N	Y
DII199	continuous	Y	Y	Y	Y
DII211	continuous	Y	N	N	Y
DII277	continuous	Y	Y	Y	Y

6. Conclusion

In this paper we have studied feature screening procedures for ultra-high dimensional data with various combinations of single-type responses and predictors. Our intensive simulations compare the screening properties of these procedures and identified a best

procedure for each combination of response and predictors. Based on these findings, we have developed a screening procedure for mixed types of data through the integration of these single-type variable procedures. Its use is demonstrated in the analysis of a high dimensional data set on arrhythmia with mixed types of predictors.

The sure screening property has been theoretically verified for data with single-type variables in numerious papers. For the new screening procedures proposed in this paper for data with single-type variables, their sure screening properties can be derived similar to the existing literature with a simple replacement of the corresponding metric. For example, such a property for the Spearman correlation screening in Section 2.1.1 can be established similar to [14] with the Pearson correlation replaced by the Spearman correlation in their proof. However, such a rigorous proof for the screening procedure for mixed types of data is very challenging. One possible detour is to assume the discrete data are actually observations from some latent continuous variables and then use the characterization of the joint distribution of these latent continuous variables and the continuous variables in the original data to derive the property. Such an investigation is beyond the scope of the paper and left as a future research direction.

Acknowledgements

The authors are grateful to the AE and two reviewers for their insightful comments that have significantly improved the paper. The authors would also like to thank the support from the U.S. National Science Foundation under grants DMS1620898, DMS-1620945, DMS-1620957, and DMS-1916174.

References

[1] Akaike H. Information theory and an extension of the maximum likelihood principle. in: Second international symposium on information theory, vol. 1, 267-281. Budapest: Akademinai Kiado; 1973.

[2] Schwarz H. Estimating the dimension of a model. The Annals of Statistics. 1978;6:461–464.

[3] Mallows C. Some comments on cp. Technometrics. 1973;15:661–675.

[4] Foster D, George E. The risk inflation criterion for multiple regression. The Annals of Statistics. 1994;22:1947–1975.

[5] Stone M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B. 1974;36:111–147.

[6] Barron A, Birge L, Massart P. Risk bounds for model selection via penalization. Probability Theory and Related Fields. 1999;113:301–413.

[7] Frank L, Friedman J. A statistical view of some chemometrics regression tools. Technometrics. 1993;35:109–135.

[8] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B. 1996;58:267–288.

[9] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B. 2005;67:301–320.

[10] Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006;101:1418–1429.

[11] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001;96:1348–1360.

[12] Zhang C. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics. 2010;38:894–942.
[13] Chen Y, Du P, Wang Y. Variable selection in linear models. WIREs Computational Statistics. 2014;6:1–9.
[14] Fan J, Lv J. Sure independence screening for ultra-high dimensional feature space. Journal of the Royal Statistical Society, Series B. 2008;70:849–911.
[15] Wang H. Forward regression for ultra-high dimensional variable screening. Journal of the American Statistical Association. 2009;104:1512–1524.
[16] Chen J, Chen Z. Extended bayesian information criteria for model selection with large model spaces. Biometrika. 2008;95:759–771.
[17] Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. Journal of Computational and Graphical Statistics. 2009;18:533–550.
[18] Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. The Annals of Statistics. 2012;40:1846–1877.
[19] Fan J, Samworth R, Wu Y. Ultra-high dimensional feature selection: beyond the linear model. Journal of Machine Learning Research. 2009;10:2013–2038.
[20] Fan J, Song R. Sure independence screening in generalized linear models with np-dimensionality. The Annals of Statistics. 2010;38:3567–3604.
[21] Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high dimensional additive models. Journal of the American Statistical Association. 2011;106:544–557.
[22] Zhu L, Li L, Li R, et al. Model-free feature screening for ultrahigh-dimensional data. Journal of the American Statistical Association. 2011;106:1464–1475.
[23] Huang D, Li R, Wang H. Feature screening for ultrahigh-dimensional categorical data with applications. Journal of Business & Economic Statistics. 2014;32:237–244.
[24] Fan J, Fan Y. High dimensional classification using features annealed independence rules. The Annals of Statistics. 2008;36:2605–2637.
[25] Mai Q, Zou H. The kolmogorov filter for variable screening in high-dimensional binary classification. Biometrika. 2013;100:229–234.
[26] Cui H, Li R, Zhong W. Model-free feature screening for ultra-high dimensional discriminant analysis. Journal of the American Statistical Association. 2015;110:630–641.
[27] Geoman J, van de Geer S, de Kort F, et al. A global test for groups of genes: Testing association with a clinical outcome. Bioinformatics. 2004;20:93–99.
[28] Kim S, Volsky D. Page: Parametric analysis of gene set enrichment. Bioinformatics. 2005; 6:144.
[29] Mansmann U, Meister R. Testing differential gene expression in functional groups. goeman's global test versus an ancova approach. Methods of Information in Medicine. 2005; 44:449–453.
[30] Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. The American Journal of Human Genetics. 2007;81:1278–1283.
[31] Holden M, Deng S, Wojnowski L, et al. Gsea-snp: Applying gene set enrichment analysis to snp data from genome-wide association studies. Bioinformatics. 2008;24:2784–2785.
[32] Zhong H, Yang X, Kaplan L, et al. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. The American Journal of Human Genetics. 2010;86:581–591.
[33] Xiong Q, ancona N, Hauser E, et al. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Research. 2012;22:386397.
[34] Ma X, Zhang J. Robust model-free feature screening via quantile correlation. Journal of Multivariate Analysis. 2016;143:472–480.
[35] Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. Journal of the American Statistical Association. 2012;107:1129–1139.
[36] Kruskal W, Wallis W. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association. 1952;47:583–621.
[37] Gu C. Smoothing spline ANOVA models (2nd ed.). New York: Springer-Verlag; 2013.

[38] Wood SN. Generalized additive models: An introduction with r (2nd ed.). Boca Raton: Chapman and Hall/CRC; 2017.
[39] Kim Y, Gu C. Smoothing spline gaussian regression: more scalable computation via efficient approximation. Journal of the Royal Statistical Society: Series B. 2004;66:337–356.
[40] Gupta V, Srinivasan S, Kudli S. Prediction and classification of cardiac arrhythmia. Department of Statistics, Stanford University; 2014. Report No.: CS229-Project2014.
[41] Mitra M, Samanta R. Cardiac arrhythmia classification using neural networks with selected features. Procedia Technology. 2013;10:76–84.