Robust Feature Screening Procedures for Mixed Type of Data

Jinhui Sun

Dissertation Submitted to the Faculty of the Virginia Polytechnic Institute and State University in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in Statistics

Pang Du, Chair

Xinwei Deng

Yili Hong

Inyoung Kim

Virginia Polytechnic Institute and State University

Robust Feature Screening Procedures for Mixed Type of Data

Jinhui Sun

Abstract

High dimensional data have been frequently collected in many fields of scientific research and technological development. The traditional idea of best subset selection methods, which use penalized L_0 regularization, is computationally too expensive for many modern statistical applications. A large number of variable selection approaches via various forms of penalized least squares or likelihood have been developed to select significant variables and estimate their effects simultaneously in high dimensional statistical inference. However, in modern applications in areas such as genomics and proteomics, ultra-high dimensional data are often collected, where the dimension of data may grow exponentially with the sample size. In such problems, the regularization methods can become computationally unstable or even infeasible. To deal with the ultra-high dimensionality, Fan and Lv (2008) proposed a variable screening procedure via correlation learning to reduce dimensionality in sparse ultra-high dimensional models. Since then many authors further developed the procedure and applied to various statistical models. However, they all focused on single type of predictors, that is, the predictors are either all continuous or all discrete. In practice, we often collect mixed type of data, which contains both continuous and discrete predictors. For example, in genetic studies, we can collect information on both gene

expression profiles and single nucleotide polymorphism (SNP) genotypes. Furthermore, outliers are often present in the observations due to experimental errors and other reasons. And the true trend underlying the data might not follow the parametric models assumed in many existing screening procedures. Hence a robust screening procedure against outliers and model misspecification is desired. In my dissertation, I shall propose a robust feature screening procedure for mixed type of data. To gain insights on screening for individual types of data, I first studied feature screening procedures for single type of data in Chapter 2 based on marginal quantities. For each type of data, new feature screening procedures are proposed and simulation studies are performed to compare their performances with existing procedures. The aim is to identify a best robust screening procedure for each type of data. In Chapter 3, I combine these best screening procedures to form the robust feature screening procedure for mixed type of data. Its performance will be assessed by simulation studies. I shall further illustrate the proposed procedure by the analysis of a real example.

General Audience Abstract

In modern applications in areas such as genomics and proteomics, ultra-high dimensional data are often collected, where the dimension of data may grow exponentially with the sample size. To deal with the ultra-high dimensionality, Fan and Lv (2008) proposed a variable screening procedure via correlation learning to reduce dimensionality in sparse ultra-high dimensional models. Since then many authors further developed the procedure and applied to various statistical models. However, they all focused on single type of predictors, that is, the predictors are either all continuous or all discrete. In practice, we often collect mixed type of data, which contains both continuous and discrete predictors. Furthermore, outliers are often present in the observations due to experimental errors and other reasons. Hence a robust screening procedure against outliers and model misspecification is desired. In my dissertation, I shall propose a robust feature screening procedure for mixed type of data. I first studied feature screening procedures for single type of data based on marginal quantities. For each type of data, new feature screening procedures are proposed and simulation studies are performed to compare their performances with existing procedures. The aim is to identify a best robust screening procedure for each type of data. Then i combined these best screening procedures to form the robust feature screening procedure for mixed type of data. Its performance will be assessed by simulation studies and the analysis of real examples.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Professor Pang Du for the continuous support of my Ph.D. study, for his guidance, encouragement and many valuable suggestions. In addition to providing a very fruitful environment, his endless patience and continued interest helped me a lot. I consider myself extremely fortunate to have had the opportunity to work under his direction.

I would also like to thank Professor Xinwei Deng, Professor Yili Hong and Professor Inyoung Kim for their insightful comments and encouragement. During my study at Virginia Tech, they provided a lot of help and taught me a lot.

At last, I would like to thank my family: my parents, my wife and my parents-in-law for supporting me in the low moments. I could not imagine my life without them. No words can express my deep gratitude and love to them.

Contents

List of Tables

Chapter 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[53],Mallow's C_p[43], RIC[21] and GCV[55], 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[3]. However, the expensive computational cost makes classical procedures infeasible for high dimensional data analysis. Other variable selection procedures should be used.

1.1 High Dimensional Variable Selection

A considerable amount of existing variable selection techniques for high dimensional data are based on the assumption that the relationship between the response and covariates is linear. Consider the linear model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{y}$ is a vector of n observations from a response variable Y, X is an $n \times p$ design matrix from p predictors $X_1, X_2, \ldots, X_p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of p unknown coefficients, and $\boldsymbol{\epsilon} \sim \mathbb{N}(0, \sigma^2 I_n)$ is a vector of n independent and identically distributed random errors. A generalized form of the penalized least squares is

$$||\boldsymbol{y} - X\boldsymbol{\beta}||^2 + \sum_{j=1}^{p} p_\lambda(\beta_j), \tag{1.2}$$

where p_λ is a penalty function on individual coefficient. A natural generalization of L_0 regularization is penalized L_q regularization, referred as the bridge regression[22], in which $p_\lambda(t) = \lambda|t|^q$ for $0 < q \leq 2$. This bridges the best subset selection(penalized L_0 regularization) and ridge regression(penalized L_2 regularization). In particular, the well-known Lasso[56] is the penalized L_1 regression.

Fan and Li[12] considered three properties for the penalty function[14]:

1. Sparsity: The resulting estimator automatically sets small estimated coefficients to zero to accomplish variable selection and reduce model complexity.

2. Unbiasedness: The resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, to reduce model bias.

3. Continuity: The resulting estimator is continuous in the data to reduce instability in model prediction.

When $q > 1$, the convex L_q penalty does not satisfy the sparsity condition. When $0 \leq q < 1$, the concave L_q penalty does not satisfy the continuity condition. In particular, the Lasso produces biased estimates for large coefficients. This has motivated Fan and Li[12] to propose the smoothly clipped absolute deviation(SCAD) penalty

$$p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\} for some\ a > 2, \tag{1.3}$$

where $p_\lambda(0) = 0$ and, often, $a = 3.7$ is used. The SCAD penalty satisfies all aforementioned three properties and the resulting estimator possesses the oracle property[12]. With similar spirit, Zhang[68] proposed the following minimax concave penalty(MCP),

$$p'_\lambda(t) = (a\lambda - t)_+ / a, \tag{1.4}$$

and showed that the resulting procedure possesses the oracle property. However, since the SCAD penalty is nonconcave and it is challenging to optimize nonconcave penalized likelihood. Fan and Li[12] proposed a unified and effective local quadratic approximation(LQA) algorithm. Their idea is to locally approximate the objective function by a quardratic function. Zou and Li[75] proposed a better approximation

achieved by using the local linear approximation(LLA). Fan and Lv[17] are able to give the conditions under which the penalized likelihood estimator exists and is unique.

Zou[72] proposed a weighted version of L_1 penalty, called the adaptive Lasso, to overcome the lack of oracle property of the Lasso.The adaptive Lasso has the oracle property under some regularity conditions[31]. However, the penalty at zero is infinite. While, the penalty function such as SCAD and MCP do not have this undesired property. The adaptive Lasso estimates can be calculated using the same algorithms for Lasso. Efron et al.[9] proposed a fast and efficient least angle regression(LARS) algorithm for computing the whole solution path of the Lasso. The computation is based on the fact that the Lasso solution path is piecewise linear. The idea of the LARS algorithm can be expanded to solve the penalized least squares problem. Fu[23], Daubechies et al.[8],Osborne et al.[52] and Wu and Lang[64] proposed the coordinate descent algorithm, which is very efficient for large Lasso problems. This algorithm can also be applied to optimize the group Lasso[67] as shown in Meier et al.[45]. Asymptotic properties of the group Lasso have been studied by Bach[2] and Nardi and Rinaldo[49]. Other group level procedures were developed by Kim et al.[35], Wang et al.[61] and Zhao et al.[69].

Zou and Hastie[73] proposed the elastic net(ENet) which is a combination of the L_1 and L_2 penalties:

$$p_\lambda(t) = \lambda_1|t| + \lambda_2 t^2, \tag{1.5}$$

where the L_1 penalty encourages sparsity in the coefficients and the L_2 penalty en-

courages some grouping effects. Liu and Wu[40] proposed the L0L1 penalty which is a combination of the L_0 and L_1 penalties:

$$p_\lambda(t) = (1 - \lambda_1) \min\{|t|/\lambda_2, 1\} + \lambda_1|t|. \tag{1.6}$$

The L0L1 penalty overcome disadvantages of the L_0 and L_1 penalties. Wu et al. [63] proposed a procedure that combines the L_1 and L_∞ penalties:

$$J_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_\infty ||\boldsymbol{\beta}||_\infty, \tag{1.7}$$

where $||\boldsymbol{\beta}||_\infty = \max_{1 \le j \le p} |\beta_j|$. While the L_1 penalty leads to sparsity, the L_∞ penalty encourages grouping among highly correlated predictors.

1.1.1 Choice of Penalty Parameters

The performance of penalized likelihood methods depends on the choice of the tuning parameters, which control the trade-off between the bias and variance in resulting estimators[28]. Traditionally, cross-validation and information criteria, including AIC and BIC, are widely applied. Nishii[50] proposed generalized information criterion(GIC):

$$n \log ||\boldsymbol{y} - X\hat{\boldsymbol{\beta}}(\lambda)||^2 + \xi d(\lambda), \tag{1.8}$$

where $\hat{\boldsymbol{\beta}}(\lambda)$ is an estimate of $\boldsymbol{\beta}$, $d(\lambda)$ measures the model complexity, and ξ controls the trade-off between goodness of fit and model complexity. AIC and BIC correspond

to the special cases when $\xi = 2$ and $\xi = \log n$. Another criterion is the generalized cross validation(GCV):

$$\frac{||\boldsymbol{y} - X\hat{\boldsymbol{\beta}}(\lambda)||^2}{(n - d(\lambda))^2}. \tag{1.9}$$

However, we need an appropriate measure of model complexity $d(\lambda)$ to be able to use these criteria.

Existing model selection criteria are naturally incorporated to select the tuning parameter for some regularization methods. For the Lasso procedure, Zou et al. [74] showed that the number of nonzero coefficients is an unbiased and consistent estimator of $d(\lambda)$. For the SCAD approach, Wang et al. [59] showed that the model selected by GCV contains all important variables and the BIC can identify the true model consistently. Wang et al. [58] showed that a modified BIC continues to work in the setting of linear regression with diverging dimensionality. We also refer to the work of Chen and Chen[5] and Wang and Zhu[62]. New statistical methodologies as well as theories are needed on the choice of penalty parameters[13].

1.2 Ultra-high Dimensional Variable Selection

The regularization methods in the last section can comfortably deal with high dimensional cases when p is almost as large as n but may have difficulty when p can increase in an exponential order $\exp\{O(n^\alpha)\}$, $\alpha > 0$ of the sample size n. To deal with the ultra high dimensionality, one appealing idea is that a fast,reliable and efficient method is first used to reduce the dimensionality p from a large or huge 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 suggests a two-scale method: a crude large scale screening followed by a moderate scale selection[15]. Many screening techniques can be chosen in the first step, as long as the sure screening property introduced in Fan and Lv[15] is satisfied, such that all the important variables can be selected with asymptotic probability one.

1.2.1 Sure Independence Screening

The main theme behind independence screening is as follows: each feature is used independently as a predictor for predicting the response and, subsequently, those features which appear highly related to the response are selected. In the linear model, the marginal correlation coefficient serves as an example of a measure of association between the covariates and the response. Fan and Lv[15] proposed sure independence screening(SIS) that ranks features according to the magnitude of its sample correlation with the response variable. Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T = X^T\boldsymbol{y}$ be a p-vector obtained by componentwise regression, where we assume that each column of the $n \times p$ design matrix X has been standardized with mean zero and variance one. For any given d_n, take the selected submodel to be

$$\mathcal{M}_{d_n} = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } d_n \text{ largest of all}\}.$$

where a conservative practical choice of d_n suggested in the paper is $[n/\log n]$. Such correlation learning screens those variables that have weak marginal correlations with the response. SIS has been shown to possess the sure screening property, which means that with high probability(tending to 1) it is able to detect a subset of covariates which contains the important ones and its size is much smaller than p. After the dimensionality is reduced from a large scale p to a moderate scale d, a variable selection procedure such as SCAD and adaptive Lasso can be applied to the selected variables from the screening procedure. The sampling properties of the sure screening methods can be obtained by combining the theory of SIS and penalization methods[16]. However, the screening procedure may fail when some key conditions are not valid. For example, the SIS may fail to select the important variable which is jointly correlated but marginal uncorrelated with the response and tends to select the unimportant variable which is jointly uncorrelated but highly marginally correlated with the response. To refine the screening performance, Fan and Lv[15] 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; see Section 1.2.2 for more details. Wang[57] studied the property of forward regression with ultrahigh-dimensional predictors and proposed using the extended BIC[4] to determine the size of the active predictor set. Hall and Miller[27] 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.[37] 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 SIS procedure.

Since Fan and Lv[15] proposed SIS for ultra high dimensional variable selection, many authors further developed the SIS method and applied to various statistical models.

1.2.2 Iterative Sure Independence Screening

Fan and Lv[15] point out three potential problems with SIS: an important predictor that is marginally uncorrelated but jointly correlated with the response cannot be picked; unimportant predictors that are highly correlated with the important predictors can have higher priority to be selected by SIS than important predictors that are relatively weakly related to the response; the issue of collinearity among the predictors adds difficulty to the problem of variable selection. Fan and Lv[15] address these issues by proposing an iterative SIS(ISIS).

Fan et al.[19] extended and improved the idea of ISIS and proposed an iterative feature screening procedure under the more general statistical framework. Suppose that our objective is to find a sparse β to minimize

$$n^{-1}\sum_{i=1}^{n}\mathbf{L}(\mathbf{Y}_i,\mathbf{x}_i^T\beta)+\sum_{j=1}^{p}p_\lambda(|\beta_j|). \tag{1.10}$$

The proposed iterative procedure consists of the following steps.

1. Apply an SIS procedure to pick a set $\mathbf{A}_1$ of indices of size k_1, and then employ a penalized likelihood method such as Lasso, SCAD or MCP to select a subset $\mathcal{M}_1$ of these indices.

2. Instead of computing residuals as in Fan and Lv[15], compute

$$\mathbf{L}_j^{(2)} = \min_{\beta_0, \beta_{\mathcal{M}_1}, \beta_j} n^{-1} \sum_{i=1}^{n} \mathbf{L}(\mathbf{Y_i}, \beta_0 + \mathbf{x_{i,\mathcal{M}_1}}^T \beta_{\mathcal{M}_1} + \mathbf{X}_{ij}\beta_j), \tag{1.11}$$

for $j \notin \mathcal{M}_1$, where $\mathbf{x_{i,\mathcal{M}_1}}$ is the sub-vector of $\mathbf{x}_i$ consisting of those elements in $\mathcal{M}_1$. This measures the additional contribution of variable $\mathbf{X_j}$ in the presence of variables $\mathbf{x}_{\mathcal{M}_1}$. Pick k_2 variables with the smallest $\{\mathbf{L}_j^{(2)}, j \notin \mathcal{M}_1\}$ and let $\mathbf{A_2}$ be the resulting set.

3. Use penalized likelihood to obtain

$$\hat{\beta}_2 = \underset{\beta_0, \beta_{\mathcal{M}_1}, \beta_{A_2}}{argmin}\ n^{-1} \sum_{i=1}^{n} \mathbf{L}(\mathbf{Y_i}, \beta_0 + \mathbf{x}_{i,\mathcal{M}_1}^T \beta_{\mathcal{M}_1} + \mathbf{x}_{i,A_2}^T \beta_{\mathbf{A_2}}) + \sum_{j \in \mathcal{M}_1 \cup A_2} p_\lambda(|\beta_j|). \tag{1.12}$$

This gives a new set $\mathcal{M}_2$ of active indices consisting of nonvanishing elements of $\hat{\beta}_2$. This step also deviates importantly from the approach in Fan and Lv[15] even in the least squares case. It allows the procedure to delete variables from the previous selected variables $\mathcal{M}_1$.

4. Iterate the above two steps until d (a prescribed number) variables are recruited or $\mathcal{M}_l = \mathcal{M}_{l-1}$.

The final estimate is then $\hat{\beta}_{\mathcal{M}_l}$. This iterative procedure extends the ISIS to a general statistical framework. It can be easily extended to many procedures. It also allows the procedure to delete predictors from the previously selected set.

Xu and Chen[66] claimed that the gain of the iterative procedure is built on higher computational cost and increased complexity. They proposed the sparsity restricted MLE(SRMLE) method for the generalized linear models and demonstrated the SRMLE is conceptually simpler and computationally cheaper. They further demonstrated that the SRMLE procedure enjoys the sure screening property.

1.2.3 SIS for Generalized Linear Models

Consider the generalized linear model with a canonical link. The conditional density is given by

$$f(y|\mathbf{x}) = exp\{y\theta(\mathbf{x}) - b(\theta(\mathbf{x})) + c(y)\}, \tag{1.13}$$

for some known functions $b(\cdot)$, $c(\cdot)$, and $\theta(\mathbf{x}) = \mathbf{x}^T\beta$. Without loss of generality, we assume the dispersion parameter $\phi = 1$. As before, we assume that each variable has been standardized to have mean 0 and variance 1. The penalized likelihood is

$$-n^{-1}\sum_{i=1}^{n}\ell((\mathbf{x_i})^T\beta, y_i) - \sum_{j=1}^{p}p_\lambda(|\beta_j|), \tag{1.14}$$

where $\ell(\theta, y) = b(\theta) - y\theta$. Define the maximum marginal likelihood estimator(MMLE) $\hat{\beta}_j^M$ as the minimizer of the componentwise regression

$$\hat{\beta}_j^{\ M} = (\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = argmin_{\beta_0,\beta_j} \sum_{i=1}^{n} \ell(\beta_0 + \beta_j \mathbf{X}_{ij}, \mathbf{Y}_i), \tag{1.15}$$

where $\mathbf{X}_{ij}$ is the ith observation of the jth variable. Similar to the marginal least squares estimate, it is reasonable to consider the magnitude of $\hat{\beta}_j^M$ to rank the importance of the features. Fan and Song[20] select a set of variables whose marginal magnitude exceeds a predefined threshold value γ_n:

$$\mathcal{M}_{\gamma_n} = \{1 \leq j \leq p : |\hat{\beta}_j^{\ M}| \geq \gamma_n\}. \tag{1.16}$$

Fan and Song[20] further proved that, under some technical assumptions, the MMLEs are uniformly convergent to the population values and established the sure screening property of the MMLE screening procedure. They also discussed the size of the selected model.

1.2.4 Nonparametric Independence Screening

In practice, there is often little prior information indicating that the effects of the covariates take a linear form or belong to any other finite-dimensional parametric family. Substantial improvements are sometimes possible by using a more flexible class of nonparametric models, such as the additive model[54]. Fan et al.[11] proposed

a nonparametric independence screening(NIS) for the ultra high dimensional additive model

$$\mathbf{Y} = \sum_{j=1}^{p} m_j(\mathbf{X}_j) + \epsilon, \tag{1.17}$$

where $\{m_j(\mathbf{X}_j)\}_{j=1}^{p}$ have mean 0. They rank the utility of covariates according to $\mathbf{E}(f_j^2(\mathbf{X}_j))$, where $f_j(\mathbf{X}_j) = \mathbf{E}(\mathbf{Y}|\mathbf{X}_j)$ is the projection of $\mathbf{Y}$ onto $\mathbf{X}_j$. $f_j(x)$ can be estimated via a normalized B-spline basis $\mathbf{B}_j(x) = \{\mathbf{B}_{j1}(x), \ldots, \mathbf{B}_{jd_n}(x)\}^T$:

$$\hat{f}_{nj}(x) = \hat{\beta}_j^T \mathbf{B}_j(x), 1 \leq j \leq p, \tag{1.18}$$

where $\hat{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_n})^T$ is obtained through the componentwise least squares regression:

$$\hat{\beta}_j = \underset{\beta_j \in \mathbb{R}^{d_n}}{argmin} \sum_{i=1}^{n} (\mathbf{Y}_i - \beta_j^T \mathbf{B}_j(\mathbf{X}_{ij})). \tag{1.19}$$

Thus the screened model index set is

$$\mathcal{M}_\nu = \{1 \leq j \leq p : ||\hat{f}_{nj}||_n^2 \geq \nu_n\}, \tag{1.20}$$

where $||\hat{f}_{nj}||_n^2 = n^{-1} \sum_{i=1}^{n} \hat{f}_{nj}(\mathbf{X}_{ij})^2$ and ν_n is a predefined threshold value. Such an independence screening ranks the importance of features according to the marginal strength of the marginal nonparametric regression. Fan et al.[11] further introduced INIS-penGAM procedure to decrease the false selection rate and greedy INIS (g-INIS) algorithm to deal with the highly correlated covariates.

Fan et al.[18], and Liu et al.[39] extended the NIS to sparse ultra high dimensional

varying coefficient models. Fan et al.[18] proposed a conditional correlation screening procedure based on the kernel regression approach. From a different point of view, Liu et al.[39] proposed another sure independent screening procedure based on the conditional correlation learning(CC-SIS).

1.2.5 Model-free Feature Screening for Continuous Variables

Most of the screening procedures we reviewed above focuses on a class of specific model and its performance is based upon the belief that the imposed working model is close to the true model. However, it may be very challenging to specify the model structure on the regression function in ultra high dimensional modelling.

Zhu et al.[71] proposed a sure independent ranking screening(SIRS) procedure, which is a model-free variable screening procedure. Let $\mathbf{Y}$ be the response variable with support Ψ_y, and $\mathbf{Y}$ can be both univariate and multivariate. Let $\mathbf{x} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)^T$ be a covariate vector. Zhu et al.[71] first developed the notion of active predictors and inactive predictors without specifying a regression model. Consider the conditional distribution function of $\mathbf{Y}$ given $\mathbf{x}$, denoted by $\mathbf{F}(y|\mathbf{x}) = \mathbf{P}(\mathbf{Y} < y|\mathbf{x})$. Define the true model

$$\mathcal{M}_* = \{k : \mathbf{F}(y|\mathbf{x}) \text{ functionally depends on } \mathbf{X}_k \text{ for some } y \in \Psi_y\}, \tag{1.21}$$

if $k \in \mathcal{M}_*$, $\mathbf{X}_k$ is referred to as an active predictor, otherwise it is referred to as an inactive predictor. Zhu et al.[71] considered a general model framework under which

a unified screening approach was developed. Assume that

$$\mathbf{F}(y|\mathbf{x}) = \mathbf{F}_0(y|\mathbf{B}^T\mathbf{x}_{\mathcal{M}_*}), \tag{1.22}$$

where $\mathbf{F}_0(\cdot|\mathbf{B}^T\mathbf{x}_{\mathcal{M}_*})$ is an unknown distribution function for a given $\mathbf{B}^T\mathbf{x}_{\mathcal{M}_*}$. Assume that $\mathbf{E}(\mathbf{X}_k) = 0$ and $\mathbf{V}ar(\mathbf{X}_k) = 1$ for $k = 1, \ldots, p$. Define $\Omega(y) = \mathbf{E}[\mathbf{x}\mathbf{F}(y|\mathbf{x})]$. It then follows by the law of iterated expectations that $\mathbf{\Omega}(y) = \mathbf{E}[\mathbf{x}\mathbf{E}\{\mathbf{1}(\mathbf{Y} < y)|\mathbf{x}\}] = cov\{\mathbf{x}, \mathbf{1}(\mathbf{Y} < y)\}$. Let $\Omega_k(y)$ be the kth element of $\mathbf{\Omega}(y)$, and define

$$\omega_k = \mathbf{E}\{\Omega_k^2(\mathbf{Y})\}, k = 1, \ldots, p. \tag{1.23}$$

Then ω_k is to serve as the marginal utility measure for predictor ranking. A natural estimator for ω_k is

$$\hat{\omega}_k = \frac{1}{n}\sum_{j=1}^{n}\widehat{\Omega}_k^2(\mathbf{Y}_j) = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_{ik}\mathbf{1}(\mathbf{Y}_i < \mathbf{Y}_j)\right\}^2, k = 1, \ldots, p, \tag{1.24}$$

where $\mathbf{X}_{ik}$ denotes the kth element of $\mathbf{x}_i$. Zhu et al.[71] proposed ranking all the candidate predictor $\mathbf{X}_k$ according to $\hat{\omega}_k$ from the largest to smallest, and then selecting the top ones as the active predictors. And they empirically demonstrated that the combination of the soft cutoff and hard cutoff by setting $d = [n/\log(n)]$ works quite well in their simulation studies.

Several other model-free screening procedures have been proposed. Li et al.[37] proposed a robust rank correlation screening (RRCS) procedure based on the Kendall

rank correlation. RRCS can be used against outliers and influence points in the observations and the sure independence screening property can hold only under the existence of a second order moment of predictor variable. Li et al.[38] developed the sure independence screening procedure based on the distance correlation (DC-SIS) under general parametric models. The DC-SIS can be used directly to screen grouped predictor variables and for multivariate response variables.

1.2.6 Model-free Feature Screening for Categorical Data

The aforementioned methods implicitly assume that predictor variables are continuous. Ultra high dimensional data with categorical predictors and categorical responses are frequently encountered in practice.

To deal with the cases when the predictors and the responses are all categorical, Huang et al.[30] employed the Pearson χ^2 test statistic as a marginal utility for feature screening. Let $\mathbf{Y}_i \in \{\mathbf{1}, \ldots, \mathbf{K}\}$ be the corresponding class label, and $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})^T \in \mathbb{R}^p$ be the associated categorical predictor. 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 $\hat{\pi}_{yj,k_1k_2} = n^{-1}\sum \mathbf{I}(Y_i = k_1)\mathbf{I}(X_{ij} = k_2)$. Subsequently, a chi-square type statistic can be defined as

$$\hat{\triangle}_j = \sum_{k_1=1}^{K}\sum_{k_2=1}^{2} \frac{(\hat{\pi}_{yk_1}\hat{\pi}_{jk_2} - \hat{\pi}_{yj,k_1k_2})^2}{\hat{\pi}_{yk_1}\hat{\pi}_{jk_2}}, \tag{1.25}$$

which is a nature estimator of

$$\triangle_j = \sum_{k_1=1}^{K} \sum_{k_2=1}^{2} \frac{(\pi_{yk_1}\pi_{jk_2} - \pi_{yj,k_1k_2})^2}{\pi_{yk_1}\pi_{jk_2}}. \tag{1.26}$$

Huang et al.[30] proposed estimating the true model by $\hat{\mathbf{S}} = \{j : \hat{\triangle}_j > c\}$, where $c > 0$ is some prespecified constant. They further established the sure screening property under mild conditions.

1.2.7 SIS for Classification

Classification and discriminant analysis are useful for analysis of categorical response data. Traditional methods of classification and discriminant analysis may break down when the dimensionality is extremely large.

Let $\mathbf{Y}$ be a categorical response with $\mathbf{K}$ classes $\{y_1, \ldots, y_K\}$. If an individual covariate $\mathbf{X}_j$ is associated with the response $\mathbf{Y}$, then $\mu_{jk} = \mathbf{E}(\mathbf{X}_j|\mathbf{Y} = y_k)$ are likely different from the population mean $\mu_j = \mathbf{E}(\mathbf{X}_j)$. It is intuitive to use the test statistic for multi-sample mean problem as a marginal utility for feature screening. Fan and Fan[10] proposed using the two sample t-statistic as marginal utility for feature screening in high dimensional binary classification. They further showed that the t-statistic does not miss any important features with probability 1 under some technical conditions.

Although the variable screening based on two-sample t-statistic performs generally well in the high-dimensional classification problems, it may break down for

heavy-tailed distributions or data with outliers. To overcome this drawback, Mai and Zou[42] proposed a feature screening method for binary classification based on the Kolmogorov-Smirnov statistic. Let $F_{+j}(x)$ and $F_{-j}(x)$ denote the conditional cumulative probability functions of K_j given $Y = 1, -1$, respectively. Define $K_j = sup_{-\infty<x<\infty}|F_{+j}(x) - F_{-j}(x)|$. The sample version of K_j is defined as $K_{nj} = sup_{-\infty<x<\infty}|\hat{F}_{+j}(x) - \hat{F}_{-j}(x)|$. Mai and Zou[42] proposed ranking all the variables by the K_{nj} statistics, which is the Kolmogorov-Smirnov test statistic for testing the equivalence of two distributions. Mai and Zou[42] further established the sure screening property and showed that this method is almost as fast as t-test screening[10] and is ten times faster than nonparametric maximum marginal likelihood screening[11]. However, it is limited to the binary classification. Cui et al.[7] 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 $k \geq 0$.

1.3 New Challenges for Ultra-high Dimensional Variable Selection

In the previous section, we have provided a selective overview on feature screening for ultra-high dimensional data. We briefly described a variety of feature screen-

ing procedures for linear models, generalized linear models, nonparametric regression models, several model-free feature screening procedures and classification. 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, we can collect information on both gene expression profiles and single nucleotide polymorphisms(SNPs) genotypes. Numerous gene expression(continuous variables) based strategies have been developed(Goeman et al.[24]; Kim and Volsky[33]; Mansmann and Meister[44]) and many methods have been developed for pathway analyses using SNP data(Wang et al. [60]; Holden et al. [29]; Zhong et al.[70]). As discussed by Xiong et al.[65], 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. We first studied the performances of the procedures for single type of data and it prepared us to find a procedure for mixed type of data.

Furthermore, data with ultra-high dimensions are often contaminated with outliers. Many existing screening procedures can suffer from such contamination. And many procedures assume strict parametric models that might not be realistic for most practical data. Therefore, in this dissertation we are particularly interested in

developing screening procedures that are robust against outliers and model misspecifications.

The remainder of this dissertation is organized as follows. In Chapter 2, we 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 chapter 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 chapter. In Chapter 3, we propose a robust screening procedure for mixed type of data. We conduct simulation studies to evaluate the performance of the proposed procedure and we further illustrate the procedure using a real-life data example.

Chapter 2

SIS for Single Type of Data

In this chapter, 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 continuous response and continuous predictors, we introduce the Spearman correlation screening procedure and conduct simulation studies to compare the performance with SIS([15]), RRCS([37]), CQC-SIS([41]) and DC-SIS([38]). For 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([11]) and SIRS([71]). We also introduce the nonparametric screening procedure by smoothing spline: for continuous response,

we conduct simulation studies to compare its performance with SIS, CQC-SIS and NIS; for discrete response from exponential family, we conduct simulation studies to compare its performance with NIS, SIRS and the screening procedure by the Kruskal-Wallis test. For categorical response and categorical predictors, screening with χ^2 test statistic[30] seems to be the only option. These studies for single type of data prepared us to find a robust procedure for mixed type of data.

2.1 Continuous Response, Continuous Predictors

2.1.1 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

$$\rho = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}, \tag{2.1}$$

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

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{2.2}$$

where $d_i = rg(X_i) - rg(Y_i)$ is the difference between the two ranks at the ith observation.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an n-vector of response, $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ be an $n \times p$

design matrix and $\omega = (\omega_1, \ldots, \omega_p)^T$ be a p-vector with components

$$\omega_k = 1 - \frac{6\sum_{i=1}^{n} d_{ik}^2}{n(n^2-1)}, k = 1, \ldots, p, \tag{2.3}$$

where $d_{ik} = rg(X_{ik}) - rg(Y_i)$ and ω_k is essentially the marginal rank correlation coefficient between $\mathbf{Y}$ and $\mathbf{X}_{\cdot k}$. We can then sort the magnitudes of all the components of ω 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 several simulation settings to compare the performances of the Spearman correlation screening procedure with the existing methods, such as SIS([15]), RRCS([37]), CQC-SIS([41]) and DC-SIS([38]).

We used the linear model (1.1) with standard Gaussian predictors and the noise ε is generated from two different distributions: the standard normal distribution and the standard normal distribution with 10% of the outliers following the Cauchy distribution. We considered two such models 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 coefficients of the nonzero components of the p-vectors β were chosen to be 5. We consider three designs for the covariance matrix of X as follows: (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 $d = [n/\log n]$ and $d = [\frac{3}{2}n/\log n]$, respectively. For each model we simulated 500 data sets.

(n,p)	ρ	Outliers	SIS	Spear-man	RRCS	CQC-SIS	DC-SIS
(100,1000)	0	0	5	5	5	5	5
			0.940	0.924	0.930	0.944	0.898
(100,1000)	0	10%	4	5	5	5	5
			0.356	0.820	0.818	0.842	0.694
(100,1000)	0.5	0	5	5	5	5	5
			0.890	0.874	0.880	0.930	0.892
(100,1000)	0.5	10%	4	5	5	5	5
			0.296	0.772	0.784	0.798	0.680
(100,1000)	0.8	0	5	5	5	5	5
			0.658	0.622	0.642	0.724	0.658
(100,1000)	0.8	10%	3	5	5	5	4
			0.230	0.550	0.570	0.606	0.490

Table 2.1: Results of simulation with $s = 5$ in Section 2.1.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for SIS,Spearman,RRCS,CQC-SIS,DC-SIS with $n = 100$

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

(n,p)	ρ	Outliers	SIS	Spear-man	RRCS	CQC-SIS	DC-SIS
(200,1000)	0	0	8	8	8	8	8
			0.988	0.976	0.976	0.990	0.974
(200,1000)	0	10%	7	8	8	8	8
			0.404	0.934	0.940	0.946	0.862
(200,1000)	0.5	0	8	8	8	8	8
			0.970	0.952	0.948	0.980	0.946
(200,1000)	0.5	10%	7	5	5	5	5
			0.414	0.930	0.930	0.952	0.864
(200,1000)	0.8	0	8	8	8	8	8
			0.722	0.706	0.710	0.756	0.712
(200,1000)	0.8	10%	6	8	8	8	8
			0.206	0.618	0.616	0.616	0.558

Table 2.2: Results of simulation with $s = 8$ in Section 2.1.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for SIS,Spearman,RRCS,CQC-SIS,DC-SIS with $n = 200$

performances of the procedures. Table 2.1 and Table 2.2 summarized the simulation results and we can draw the following conclusions:

1. When there is no outlier, SIS and CQC-SIS performed better than others according to 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. Generally speaking, the performance of Spearman, RRCS and CQC-SIS were the best.

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

2.2 Continuous Response, 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. And we can get a p-value of the test, this p-value indicates the association between Y and X. The ANOVA 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 the combined sample and compute R_i, the sum of the ranks for group i. Then the Kruskal-Wallis test statistic is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1). \tag{2.4}$$

This statistic approximately follows a χ^2 distribution with $K-1$ degrees of freedom if the null hypothesis is true. Each of the n_i should be at least 5 for the approximation to be valid.

Let $\mathbf{X}_j = (\mathbf{X}_{1j}, \ldots, \mathbf{X}_{nj})$ 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([15]) and RRCS([37]).

We used the linear model (1.1) with binary predictors and the noise ε is generated from two different distributions: the standard normal distribution and the standard t distribution with one degree of freedom. We considered two such models 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 all the nonzero components of the coefficient vector β were chosen to be 5. We consider three designs for the covariance matrix of X as follows: (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 $d = [n/\log n]$ and $d = [\frac{3}{2}n/\log n]$, respectively. For each model we simulated 500 data sets.

We use 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. Table 2.3 and Table 2.4 summarized the simulation results and we can draw the following conclusions:

1. With the standard normal noise, the ANOVA test performed better than others according to 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 others.

2. Generally speaking, the performance of the Kruskal-Wallis test and RRCS are the best.

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

4. An interesting finding is 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.

2.3 Categorical Response, Continuous Predictors

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

(n,p)	ρ	ε	ANOVA	K-W	SIS	RRCS
(100,1000)	0	$\mathbf{N}(0,1)$	5	5	2	5
			0.972	0.928	0.008	0.928
(100,1000)	0	$t(1)$	2	5	1	5
			0.136	0.626	0.004	0.626
(100,1000)	0.5	$\mathbf{N}(0,1)$	5	5	2	5
			0.946	0.880	0.004	0.880
(100,1000)	0.5	$t(1)$	3	5	1	5
			0.112	0.610	0.000	0.610
(100,1000)	0.8	$\mathbf{N}(0,1)$	5	5	2	5
			0.912	0.826	0.002	0.826
(100,1000)	0.8	$t(1)$	2	4	1	4
			0.076	0.478	0.002	0.478

Table 2.3: Results of simulation with $s = 5$ in Section 2.2.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for the ANOVA, Kruskal-Wallis, SIS and RRCS with $n = 100$

(n,p)	ρ	ε	ANOVA	K-W	SIS	RRCS
(200,1000)	0	$\mathbf{N}(0,1)$	8	8	3	8
			0.998	0.996	0.000	0.996
(200,1000)	0	$t(1)$	4	8	2	8
			0.148	0.870	0.000	0.870
(200,1000)	0.5	$\mathbf{N}(0,1)$	8	8	3	8
			0.976	0.970	0.002	0.970
(200,1000)	0.5	$t(1)$	4	8	2	8
			0.130	0.872	0.000	0.872
(200,1000)	0.8	$\mathbf{N}(0,1)$	8	8	3	8
			0.960	0.932	0.004	0.934
(200,1000)	0.8	$t(1)$	5	8	2	8
			0.134	0.764	0.000	0.764

Table 2.4: Results of simulation with $s = 8$ in Section 2.2.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for the ANOVA, Kruskal-Wallis, SIS and RRCS with $n = 200$

The two sample Kolmogorov-Smirnov test is used to test whether two samples come from the same distribution. It 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

$$D_{mn} = \max_x |\mathbf{F}_1(x) - \mathbf{F}_2(x)|. \tag{2.5}$$

The statistic is calculated by finding the maximum absolute value of the differences between the two distribution c.d.f.s. The null hypothesis is H_0: both samples come from a population with the same distribution. A natural estimator for D_{mn} is

$$\hat{D}_{mn} = \max_x |\hat{\mathbf{F}}_1(x) - \hat{\mathbf{F}}_2(x)|, \tag{2.6}$$

where $\hat{\mathbf{F}}_1$ and $\hat{\mathbf{F}}_1$ are the sample c.d.f.s. The null hypothesis is rejected at level α if

$$\hat{D}_{mn} > c(\alpha)\sqrt{\frac{m+n}{mn}}, \tag{2.7}$$

where $c(\alpha)$ is given in the Kolmogorov-Smirnov Table.

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 $Y_1, \ldots, Y_n$ of size n. We can carry out the test by the following procedure:

1. Arrange all the observation in order of magnitude.

2. Under each observation, write down **X** or **Y** to indicate which sample they are from.

3. Under each X_i write down the number of Ys which are to the left of it, which indicates $X_i > Y_j$. Under each Y_j write down the number of Xs which are to the left of it, which indicates $Y_j > X_i$.

4. Add up the total number of times $X_i > Y_j$, denote by U_x. Add up the total number of times $Y_j > X_i$, denote by U_y. Check that $U_x + U_y = mn$.

5. Calculate $U = min(U_x, U_y)$.

6. Use statistical tables for the Mann-Whitney test to find the probability of observing a value of U or lower. If the test is one-sided, this is the p-value; if the test is two-sided, double this probability to obtain the p-value.

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 hypothesis 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 $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an n-vector of categorical response where $Y_i \in \{1, \ldots, K\}$ is the ith class label, and $\mathbf{X}_j = (\mathbf{X}_{1j}, \ldots, \mathbf{X}_{nj})$ be the jth continuous predictor. For each pair of $\mathbf{Y}$ and $\mathbf{X}_j$, we can devide $\mathbf{X}_j$ 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 two examples to compare the performances of four methods: NIS([11]), SIRS([71]), screening with the Kolmogorov-Smirnov test (K-S) and screening with the Mann-Whitney test (M-W).

Logistic Regression

In this example, the data $(\mathbf{x}_1^T, Y_1), \ldots, (\mathbf{x}_n^T, Y_n)$ are independent copies of a pair $(\mathbf{x}^T, Y)$, where Y is distributed, conditional on $\mathbf{X} = \mathbf{x}$, as $Bin(1, p(\mathbf{x}))$, with $log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$ and the noise ε is generated from two different distributions: the standard

normal distribution and the standard t distribution with one degree of freedom. We choose $n = 200, p = 1000$. The sizes s of the true models, i.e., the numbers of nonzero coefficients, was chosen to be 8 and the nonzero components of the coefficient vector β were chosen to be 5. We consider three designs for the covariance matrix of X as follows: (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.

We use the median number of correctly selected predictors and proportion of times that the screened predictor set contained the true model to evaluate the performances of the procedures. Table 2.5 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.

Poisson Regression

In the second example, the response Y was distributed, conditional on $\mathbf{X} = \mathbf{x}$, as $Poisson(\mu(\mathbf{x}))$, where $log(\mu(\mathbf{x})) = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$ and the noise ε was generated from two different distributions: the standard normal distribution and the standard t distribution with one degree of freedom. We chose $n = 200, p = 1000$. The sizes s of the true models, i.e., the numbers of nonzero coefficients, was chosen to be 3 and the nonzero components of the coefficient vector β were chosen to be 5. We considered three

(n,p)	ρ	ε	NIS	SIRS	M-W	K-S
(200,1000)	0	$\mathbf{N}(0,1)$	8	6	8	7
			0.518	0.042	0.714	0.436
(200,1000)	0	$t(1)$	7	5	8	7
			0.332	0.000	0.632	0.404
(200,1000)	0.5	$\mathbf{N}(0,1)$	7	6	8	7
			0.452	0.054	0.664	0.366
(200,1000)	0.5	$t(1)$	7	5	8	7
			0.350	0.036	0.610	0.260
(200,1000)	0.8	$\mathbf{N}(0,1)$	7	5	7	7
			0.256	0.008	0.304	0.144
(200,1000)	0.8	$t(1)$	6	5	7	6
			0.116	0.004	0.242	0.080

Table 2.5: Results of simulation with logistic regression in Section 2.3.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for the NIS, SIRS, Mann-Whitney test and Kolmogorov-Smirnov test with $s = 8$

designs for the covariance matrix of X as follows: (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.

We use median of the correctly selected predictors and proportion of predictors containing the true model to evaluate the performances of the procedures. Table 2.6 summarized the simulation results and we can draw the following conclusions:

1. The Kruskal-Wallis test outperforms the other three methods.

2. An interesting finding is that: When $\rho = 0.5$, the performance of the Kruskal-Wallis test was even better than when $\rho = 0$. This may be because, with large p, the sample correlation is non-negligible even with iid standard normal predictors.

(n,p)	ρ	ε	NIS	SIRS	K-W	K-S
(200,1000)	0	$\mathbf{N}(0,1)$	1	0	3	3
			0.002	0.320	0.860	0.630
(200,1000)	0	$t(1)$	1	0	3	2
			0.000	0.015	0.568	0.448
(200,1000)	0.5	$\mathbf{N}(0,1)$	1	0	3	3
			0.005	0.250	0.930	0.725
(200,1000)	0.5	$t(1)$	1	0	3	3
			0.000	0.000	0.730	0.538
(200,1000)	0.8	$\mathbf{N}(0,1)$	1	0	3	3
			0.008	0.230	0.802	0.596
(200,1000)	0.8	$t(1)$	1	0	3	2
			0.000	0.000	0.500	0.334

Table 2.6: Results of simulation with poisson regression in Section 2.3.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for NIS, SIRS, Kruskal-Wallis test and Kolmogorov-Smirnov test with $s = 3$

2.4 Nonparametric Screening with Continuous Predictors

2.4.1 Screening by Smoothing Spline with Continuous Response

Given $Y_i = \eta(x_i) + \epsilon_i$, with $\epsilon_i \sim \mathbf{N}(0, \sigma^2)$, the minus log likelihood function $L(f)$ reduces to the least squares functional proportional to $\sum_{i=1}^{n}(Y_i - f(x_i))^2$. Then, the general form of penalized least squares functional in a reproducing kernel Hilbert space $\mathcal{H} = \oplus_{\beta=0}^{p}\mathcal{H}_\beta$ can be written as

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \eta(x_i))^2 + \lambda J(\eta), \tag{2.8}$$

where $J(f) = J(f,f) = \sum_{\beta=1}^{p} \theta_\beta^{-1}(f,f)_\beta$ and $(f,g)_\beta$ are inner products in $\mathcal{H}_\beta$ with reproducing kernel $R_\beta(x,y)$. The penalty is seen to be

$$\lambda J(\eta) = \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1}(f,f)_\beta, \tag{2.9}$$

with λ and θ_β as smoothing parameters. The bilinear form $J(f,g) = \sum_{\beta=1}^{p} \theta_\beta^{-1}(f,g)_\beta$ is an inner product in $\oplus_{\beta=1}^{p}\mathcal{H}_\beta$, with a reproducing kernel $R_J(x,y) = \sum_{\beta=1}^{p} \theta_\beta R_\beta(x,y)$ and a null space $\mathcal{N}_J = \mathcal{H}_0$ of finite dimension, say m. The minimizer η_λ has the expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i R_J(x_i, x) = \phi^T \mathbf{d} + \xi^T \mathbf{c}, \tag{2.10}$$

where $\{\phi_\nu\}_{\nu=1}^{m}$ is a basis of $\mathcal{N}_J = \mathcal{H}_0$, ϕ and ξ are vectors of functions, and $\mathbf{c}$ and $\mathbf{d}$ are vectors of real coefficients. The estimation then reduces to the minimization of

$$(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c}, \tag{2.11}$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where S is $n \times m$ with the (i,ν)th entry $\phi_\nu(x_i)$ and Q is $n \times n$ with the (i,j)th entry $R_J(x_i, x_j)$. Suppose S is of full column rank. Let

$$S = FR^* = (F_1, F_2) \begin{pmatrix} \tilde{R} \\ \mathcal{O} \end{pmatrix} = F_1 \tilde{R}, \tag{2.12}$$

be the QR-decomposition of S with F orthogonal and $\tilde{R}$ upper-triangular. From $S^T\mathbf{c} = 0$, one has $F_1^T\mathbf{c} = 0$, so $\mathbf{c} = F_2F_2^T\mathbf{c}$. Some algebra leads to

$$\mathbf{c} = F_2(F_2^TQF_2 + n\lambda I)^{-1}F_2^T\mathbf{Y}, \mathbf{d} = \tilde{R}^{-1}(F_1^T\mathbf{Y} - F_1^TQ\mathbf{c}). \tag{2.13}$$

Denote the fitted values by $\hat{Y}$, some algebra yields

$$\hat{Y} = Q\mathbf{c} + S\mathbf{d} = (I - n\lambda F_2)(F_2^TQF_2 + n\lambda I)^{-1}F_2^T\mathbf{Y} = A(\lambda)\mathbf{Y}, \tag{2.14}$$

where $A(\lambda)$ is known as the smoothing matrix.

With varying smoothing parameter λ, the minimizer η_λ defines a family of possible estimates. We can use the method of cross-validation to choose the smoothing parameter λ. If an independent validation data set were available with $Y_i^* = \eta(x_i) + \epsilon_i^*$, then an intuitive strategy for the selection of λ would be to minimize $n^{-1}\sum_{i=1}^n(\eta_\lambda(x_i) - Y_i^*)^2$. Lacking an independent validation data set, an alternative strategy is to minimize

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^n(\eta_\lambda^{[i]}(x_i) - Y_i)^2, \tag{2.15}$$

where $\eta_\lambda^{[i]}$ is the minimizer of the "delete-one" functional

$$\frac{1}{n}\sum_{i\neq k}(Y_i - \eta(x_i))^2 + \lambda J(\eta). \tag{2.16}$$

Some algebra yields

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \eta_\lambda(x_i))^2}{(1-a_{i,i})^2}, \tag{2.17}$$

where $a_{i,i}$ is the (i,i)th entry of $A(\lambda)$. Craven and Wahba[6] substituted $a_{i,i}$ by its average $n^{-1}\sum_1^n a_{i,i}$ and obtained the generalized cross-validation(GCV) score

$$V(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I-A(\lambda))^2\mathbf{Y}}{n^{-1}tr(I-A(\lambda))^2}. \tag{2.18}$$

A desirable property of the GCV score is its invariance to an orthogonal transform of $\mathbf{Y}$. Despite its asymptotic optimality, the GCV score is known to occasionally deliver severe undersmoothing. Kim and Gu[34] proposed a modified version,

$$V(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I-A(\lambda))^2\mathbf{Y}}{n^{-1}tr(I-\alpha A(\lambda))^2}, \tag{2.19}$$

with a fudge factor $\alpha > 1$ proves rather effective in curbing undersmoothing while maintaining the otherwise good performance of GCV. And $\alpha = 1.4$ was found to be adequate in the simulation studies.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an n-vector of continuous response, $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_p})^T$ be an $n \times p$ design matrix.For each pair of $\mathbf{Y}$ and $\mathbf{X}_j$, we can fit a soonthing spline model and get an estimate $\hat{\eta}_j$ for η_j, we choose the soothing parameter λ using modified GCV score described above. Then we can test the significance of the relationship by examining whether $\hat{\eta}_j$ is a constant function, which means $\hat{\eta}_j' \equiv 0$. For some arbitrary points $(x_1, \ldots, x_m)$, let $\omega = (\omega_1, \ldots, \omega_p)^T$ be a p-vector each being $\sum_{i=1}^m [\hat{\eta}_j'(x_i)]^2$.

We can then sort the magnitudes of all the components of ω in an decreasing order and select a submodel

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

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

2.4.2 Screening by Smoothing Spline with Discrete Response from Exponential Families

Consider exponential family distributions with densities of the form

$$f(y|x) = exp(y\eta(x) - b(\eta(x)))/a(\phi) + c(y, \phi), \tag{2.20}$$

where $a > 0$, b and c are known functions, $\eta(x)$ is the parameter of interest dependent on a covariate x, and ϕ is either known or considered as a nuisance parameter that is independent of x. Observing $Y_i|x_i \sim f(y|x_i)$, one is to estimate the regression function $\eta(x)$. One has the penalized likelihood functional

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(x_i) - b(\eta(x_i))\} + \frac{\lambda}{2}J(\eta) \tag{2.21}$$

for $\eta \in \mathcal{H} = \oplus_{\beta=0}^{p}\mathcal{H}_\beta$, where $J(f) = J(f, f) = \sum_{\beta=1}^{p}\theta_\beta^{-1}(f, f)_\beta$ and $(f, g)_\beta$ are inner products in $\mathcal{H}_\beta$ with reproducing kernels $\mathbf{R}_\beta(x, y)$. The terms $c(Y_i, \phi)$ are independent

of $\eta(x)$ and, hence, are dropped, and the dispersion parameter $a(\phi)$ is absorbed into λ. The bilinear form $J(f,g)$ is an inner product in $\oplus_{\beta=1}^{p}\mathcal{H}_\beta$ with a reproducing kernel $\mathbf{R}_J(x,y) = \sum_{\beta=1}^{p}\theta_\beta\mathbf{R}_\beta(x,y)$ and a null space $\mathcal{N}_J = \mathcal{H}_0$. The first term of (2.21) depends on η only through the evaluations $[x_i]\eta = \eta(x_i)$, and the minimizer η_λ of (2.21) has an expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu\phi_\nu(x) + \sum_{i=1}^{n} c_i\mathbf{R}_J(x_i,x) = \phi^T\mathbf{d} + \xi^T\mathbf{c}, \tag{2.22}$$

where $\{\phi_\nu\}_{\nu=1}^{m}$ is a basis of $\mathcal{N}_J = \mathcal{H}_0$, ξ and ϕ are vectors of functions, and $\mathbf{c}$ and $\mathbf{d}$ are vectors of coefficients. Fixing the smoothing parameters λ, the minimizer η_λ may be computed via the Newton iteration. Write $\tilde{u}_i = -Y_i + \dot{b}(\tilde{\eta}(x_i)) = -Y_i + \tilde{\mu}(x_i)$ and $\tilde{w}_i = \ddot{b}(\tilde{\eta}(x_i)) = \tilde{v}(x_i)$. The quadratic approximation of $-Y_i\eta(x_i) + b(\eta(x_i))$ at $\tilde{\eta}(x_i)$ is

$$-Y_i\tilde{\eta}(x_i)+b(\tilde{\eta}(x_i))+\tilde{u}_i\{\eta(x_i)-\tilde{\eta}(x_i)\}+\frac{1}{2}\tilde{w}_i\left\{\eta(x_i)-\tilde{\eta}(x_i)\right\}^2 = \frac{1}{2}\tilde{w}_i\{\eta(x_i)-\tilde{\eta}(x_i)+\frac{\tilde{u}_i}{\tilde{w}_i}\}^2+C_i, \tag{2.23}$$

where C_i is independent of $\eta(x_i)$. The Newton iteration updates $\tilde{\eta}$ by the minimizer of the penalized weighted least squares functional

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2 + \lambda J(\eta) \tag{2.24}$$

where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$.

Smoothing parameter selection remains an important practical issue. Without

loss of generality, assume $a(\phi) = 1$. Consider the Kullback-Leibler distance

$$KL(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{\mu(x_i)(\eta(x_i) - \eta_\lambda(x_i)) - (b(\eta(x_i)) - b(\eta_\lambda(x_i)))\}. \tag{2.25}$$

Dropping terms that do not involve η_λ, one gets the relative Kullback-Leibler distance

$$RKL(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{-\mu(x_i)\eta_\lambda(x_i) + b(\eta_\lambda(x_i))\}. \tag{2.26}$$

Replacing $\mu(x_i)\eta_\lambda(x_i)$ by $Y_i\eta_\lambda^{[i]}(x_i)$, one obtains a cross-validation estimate of $RKL(\eta, \eta_\lambda)$,

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{-Y_i\eta_\lambda^{[i]}(x_i) + b(\eta_\lambda(x_i))\}, \tag{2.27}$$

where $\eta_\lambda^{[i]}$ minimizes the "delete-one" version of (2.21),

$$-\frac{1}{n}\sum_{i\neq k}\{Y_i\eta(x_i) - b(\eta(x_i))\} + \frac{\lambda}{2}J(\eta). \tag{2.28}$$

Note that $E[Y_i] = \mu(x_i)$ and that $\eta_\lambda^{[i]}$ is independent of Y_i. Write

$$V_0(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta_\lambda(x_i) - b(\eta_\lambda(x_i))\} + \frac{1}{n}\sum_{i=1}^{n}Y_i(\eta_\lambda(x_i) - \eta_\lambda^{[i]}(x_i)), \tag{2.29}$$

where the first term is readily available, but the second term is impractical to compute. We need computationally practical approximations of the second term. Gu and Xiang[25] substitute $\eta_{\lambda,\eta_\lambda}^{[i]}(x_i)$ for $\eta_\lambda^{[i]}(x_i)$, where $\eta_{\lambda,\eta_\lambda}^{[k]}(x_i)$ minimizes the "delete-one"

version of (2.24),

$$\frac{1}{n}\sum_{i\neq k}\tilde{w}_i(\tilde{Y}_i-\eta(x_i))^2+\lambda J(\eta), \tag{2.30}$$

for $\tilde{\eta}=\eta_\lambda$. Remember that $\eta_\lambda=\eta_{\lambda,\eta_\lambda}$. Some algebra yields

$$V(\lambda)=-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta_\lambda(x_i)-b(\eta_\lambda(x_i))\}+\frac{1}{n}\sum_{i=1}^{n}\frac{a_{i,i}}{1-a_{i,i}}\frac{Y_i(Y_i-\mu_\lambda(x_i))}{\tilde{w}_i}, \tag{2.31}$$

where $a_{i,i}$ is the ith diagonal of the smoothing matrix $A_w(\lambda)$. We can obtain a generalized approximate cross-validation(GACV) score by replacing $a_{i,i}/\tilde{w}_i$ by $n^{-1}\sum_{i=1}^{n}a_{i,i}/\tilde{w}_i$ and $1-a_{i,i}$ by $1-n^{-1}trA_w$:

$$V_g(\lambda)=-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta_\lambda(x_i)-b(\eta_\lambda(x_i))\}+\frac{tr(A_wW^{-1})}{n-trA_w}\frac{1}{n}\sum_{i=1}^{n}Y_i(Y_i-\mu_\lambda(x_i)), \tag{2.32}$$

where $W=diag(\tilde{w}_i)$.

Let $\mathbf{Y}=(Y_1,\ldots,Y_n)$ be an n-vector of discrete response, $\mathbf{X}=(\mathbf{X_1},\ldots,\mathbf{X_p})^T$ be an $n\times p$ design matrix.For each pair of $\mathbf{Y}$ and $\mathbf{X}_j$, we can choose a discrete distribution from exponential family, e.g. Binomial distribution or Poisson distribution, for $\mathbf{Y}$. Then we can fit a smoothing spline model and get an estimate $\hat{\eta}_j$ for η_j, we choose the smoothing parameter λ using GACV score described above. We then test the significance of the relationship by examining whether $\hat{\eta}_j$ is a constant function, which means $\hat{\eta}_j'\equiv 0$. For some arbitrary points $(x_1,\ldots,x_m)$,let $\omega=(\omega_1,\ldots,\omega_p)^T$ be a p-vector each being $\sum_{i=1}^{m}[\hat{\eta}_j'(x_i)]^2$. We can then sort the magnitudes of all the

components of ω in an decreasing order and select a submodel

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

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

2.4.3 Numerical Studies

Continuous Response

For continuous response, we compared the performance of screening by smoothing spline with SIS([15]), CQC-SIS([41]) and NIS([11]). We set $n = 400$ and $p = 1000$. For NIS, the number of basis is set to be 5 as suggested by Fan et al.[11]. For smoothing spline, the number of basis is set to be $max(30, 10n^{2/9})$ and $a = 1.4$ in modified GCV as suggested by Kim and Gu[34]. For each model we simulated 500 data sets.

*Example*1: This example is adapted from Fan, Feng and Song[11]. Let $g_1(x) = x$, $g_2(x) = (2x-1)^2$, $g_3(x) = sin(2\pi x)/(2 - sin(2\pi x))$ and $g_4(x) = 0.1sin(2\pi x) + 0.2cos(2\pi x) + 0.3sin(2\pi x)^2 + 0.4cos(2\pi x)^3 + 0.5sin(2\pi x)^3$. The data is generated from the following model:

$$\mathbf{Y} = 5g_1(\mathbf{X}_1) + 3g_2(\mathbf{X}_2) + 4g_3(\mathbf{X}_3) + 6g_4(\mathbf{X}_4) + \sqrt{1.74}\varepsilon. \tag{2.33}$$

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

$$\mathbf{X}_j = \frac{\mathbf{W}_j + t\mathbf{U}}{1+t}, j = 1, \ldots, p, \tag{2.34}$$

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

*Example*2: The settings and model are the same as *Example*1 except that 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 three cases: $\rho = 0.5$, $\varepsilon \sim \mathbf{N}(0,1)$; $\rho = 0.8$, $\varepsilon \sim \mathbf{N}(0,1)$; $\rho = 0.8$, $\varepsilon \sim t(1)$.

We used the median number of correctly selected predictors and proportion of times that the screened predictor set contained the true model to evaluate the performances of the procedures. Table 2.7 summarized the simulation results and 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 were the best. This may be because, when the correlation is weak, the signals work against the marginal effect estimation as accumulated noise, thus masking the relatively weak signals from $\mathbf{X}_3$ and $\mathbf{X}_4$ in the example.

Model	t	SIS	CQC	NIS	SS
*Example*1	0	1	3	4	4
		0.000	0.074	0.962	0.968
*Example*1	1	4	4	3	4
		0.578	0.336	0.426	0.524
Model	ρ & ε	SIS	CQC	NIS	SS
*Example*2	0.5	1	3	4	4
	$\mathbf{U}(0,1)$	0.000	0.120	0.960	0.960
*Example*2	0.8	0	3	4	4
	$\mathbf{U}(0,1)$	0.000	0.055	0.924	0.854

Table 2.7: Results of simulation with continuous response in Section 2.4.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for SIS, CQC-SIS, NIS and smoothing spline with 4 truly active predictors

Discrete Response from Exponential Family

For discrete response from exponential family, we compare the performance of screening by smoothing spline with NIS([11]), SIRS([71]) and screening with p-value of Kruskal-Wallis test. We set $n = 400$ and $p = 1000$. For NIS, the number of basis is set to be 5 as suggested by Fan et al.[11]. For smoothing spline, the number of basis is set to be $max(30, 10n^{2/9})$ and $a = 1.4$ in modified GCV as suggested by Kim and Gu[34]. For each model we simulated 500 data sets.

*Example*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 $Bin(1, p(\mathbf{x}))$, with $log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = 5g_1(\mathbf{X}_1) + 5g_2(\mathbf{X}_2) + 5g_3(\mathbf{X}_3)$. 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 $\rho = 0.8$.

*Example*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 $\rho = 0.8$.

We used the median number of correctly selected predictors and proportion of times that the screened predictor set contained the true model to evaluate the performances of the procedures. Table 2.8 summarized the simulation results and we can draw the following conclusions:

1. Generally speaking, the performance of NIS and smoothing spline are the best.

2. When $\rho = 0.8$, the performances of all methods are better than when $\rho = 0$. This may be because, when $\rho = 0$, the independent signals work against the marginal effect estimation as accumulated noise, thus masking the relatively weak signals.

2.5 Categorical Response, Categorical Predictors

When the predictors and the responses are all categorical, Huang et al.[30] employed the Pearson χ^2 test statistic as a marginal utility for feature screening. We described the details of this screening procedure in *section*1.2.6. It seems this procedure is the best option we have and we have not found a competitive procedure so far.

Model	ρ	NIS	SIRS	K-W	SS
*Example*3	0	4	3	3	4
		0.876	0.012	0.018	0.886
*Example*3	0.5	4	3	3	4
		0.774	0.032	0.026	0.868
*Example*3	0.8	4	3	3	4
		0.850	0.000	0.000	0.870
*Example*4	0	4	3	3	4
		0.732	0.062	0.088	0.786
*Example*4	0.5	4	3	3	4
		0.712	0.032	0.048	0.736
*Example*4	0.8	4	3	3	4
		0.748	0.018	0.026	0.728

Table 2.8: Results of simulation with discrete response in Section 2.4.3: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for NIS, SIRS, Kruskal-Wallis test and smoothing spline with 3 truly active predictors

2.6 Ordinal Response, Continuous Predictors

2.6.1 Screening by Polyserial Correlation

When the response is ordinal variable and the predictors are continuous variables. We propose to use the Polyserial correlation. Polyserial correlation measures the correlation between two continuous variables with a bi-variate normal distribution, where one variable is observed directly, and the other is unobserved. Information about the unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values[51].

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an n-vector of ordinal response, $Y_i \in 1, \ldots, K$ be the corresponding class label, $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ be an $n \times p$ design matrix and $\omega =$

$(\omega_1, \ldots, \omega_p)^T$ be a p-vector each being the marginal Polyserial correlation coefficient between $\mathbf{Y}$ and $\mathbf{X}_j$. We can then sort the magnitudes of all the components of ω 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.

2.6.2 Numerical Studies

We used the linear model (1.1) with standard Gaussian predictors and the noise ε is generated from two different distributions: the standard normal distribution and the standard t distribution with one degree of freedom. We chose $n = 200, p = 1000$. The sizes s of the true models, i.e., the numbers of nonzero coefficients, were chosen to be 8 and the coefficients of the nonzero components of the p-vectors β were chosen to be 5. We consider three designs for the covariance matrix of X as follows: (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.

We compared the performances of four methods: SIS, screening with ANOVA test, screening with Spearman correlation and screening with Polyserial correlation. For SIS, we used the original continuous y value. For ANOVA test, Spearman correlation

and Polyserial correlation: when $y_i < Q1$, y_i is labeled as 1; when $Q1 \leq y_i < Q2$, y_i is labeled as 2; when $Q2 \leq y_i < Q3$, y_i is labeled as 3; when $y_i > Q3$, y_i is labeled as 4, where $Q1$, $Q2$ and $Q3$ are the first, second and third quantile of Y.

We used the median number of correctly selected predictors and proportion of times that the screened predictor set contained the true model to evaluate the performances of the procedures. Table 2.9 summarized the simulation results and we can draw the following conclusions:

1. With standard normal noise, screening with Polyserial correlation performs almost as good as SIS.

2. with t distribution noise, screening with Polyserial correlation outperforms the other three methods.

3. Generally speaking, the performance of Polyserial correlation is the best.

ρ	ε	SIS	Spearman	ANOVA	Polyserial
0	$\mathbf{N}(0,1)$	8	8	8	8
		0.988	0.948	0.962	0.982
0	$t(1)$	7	8	8	8
		0.386	0.888	0.928	0.966
0.5	$\mathbf{N}(0,1)$	8	8	8	8
		0.976	0.928	0.946	0.966
0.5	$t(1)$	6	8	8	8
		0.278	0.882	0.890	0.952
0.8	$\mathbf{N}(0,1)$	8	8	8	8
		0.668	0.586	0.616	0.646
0.8	$t(1)$	6	8	8	8
		0.244	0.524	0.556	0.578

Table 2.9: Results of simulation in Section 2.6.2: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model for SIS, Spearman, ANOVA and Polyserial

Chapter 3

Robust Feature Screening for Mixed Type of Data

3.1 Motivating Examples

*Example*1: Arrhythmia Data Set

This Arrhythmia data set was contributed by Dr. H. Altay Guvenir to the UC-Irvine Machine Learning Respository. The data set can be downloaded from *https* : *//archive.ics.uci.edu/ml/datasets/Arrhythmia*. There are 452 patient records and 279 attributes, 206 of which are continuous variables and the rest are nominal. The aim is to distinguish normal from abnormal heartbeat behavior based on ECG (Electrocardiogram) data. The main challenges in processing this data set are the limited number of samples compared to the number of attributes and attribute values belonging to both continuous and categorical types.

*Example*2: Asthma Data Set

The association between SNPs at $ORMDL3$ gene and the risk of childhood asthma was studied by Miriam F. Moffatt et al.[48]. The data set can be downloaded from the Gene Expression Omnibus (GEO) database at the website of the National Center for Biotechnology Information (NCBI) with accession number $GSE8052$. The data set consists of 268 cases and 136 controls with both SNP genotype and gene expression data available. The original genome-wide study reported that the SNPs on chromosome $17q21$ where $ORMDL3$ is located, were strongly associated with childhood asthma. The authors also found that these SNPs were highly correlated with gene expression of $ORMDL3$, which is also associated with asthma. This motivated us to assess the overall genetic effect of $ORMDL3$ on the occurrence of childhood asthma, by jointly analyzing SNP and gene expression data.

The studies for single type of data in *Chapter*2 prepared us to find a robust procedure for mixed type of 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.

3.2 Method

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an n-vector response, $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_p})^T$ be an $n \times p$ design matrix. For each pair of $\mathbf{Y}$ and $\mathbf{X}_j$, we want to perform a test and the p-value of the test indicates the significance of the marginal relationship between the response and

the predictor.

For continuous response $\mathbf{Y}$: When the predictor $\mathbf{X_j}$ is continuous, we can perform a B-spline fit. Consider the model

$$\mathbf{Y} = f_j(\mathbf{X}_j) + \boldsymbol{\epsilon}.$$

$f_j(x)$ can be estimated via a B-spline basis $\mathbf{B}_j(x) = \{\mathbf{B}_{j1}(x), \ldots, \mathbf{B}_{jd}(x)\}^T$:

$$\hat{f}_j(x) = \hat{\beta}_j^T \mathbf{B}_j(x),$$

where $\hat{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})^T$ is obtained through the least squares regression:

$$\hat{\beta}_j = \underset{\beta_j \in \mathbb{R}^d}{argmin} \sum_{i=1}^{n} (\mathbf{Y}_i - \beta_j^T \mathbf{B}_j(\mathbf{X}_{ij})).$$

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 different groups, then we can 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 statistics would be:

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

This test statistics follows an $\mathbf{F}$ distribution with degrees of freedom $K-1$ and $n-K$.

And we can get a p-value from the ANOVA test. If we choose 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 approximates a χ^2 distribution with $K-1$ degrees of freedom and we can get a p-value from the K-W test.

For discrete response $\mathbf{Y}$: When the predictor $\mathbf{X_j}$ is continuous, we can treat different values of the response as different groups, then we can 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 statistics would be:

$$\mathbf{F} = \frac{\sum_{i=1}^{K} n_i (\overline{X}_{ji\cdot} - \overline{X}_{j\cdot\cdot})^2/(K-1)}{\sum_{i=1}^{K} \sum_{l=1}^{n_i} (X_{jil} - \overline{X}_{ji\cdot})^2/(n-K)}.$$

This test statistics follows an $\mathbf{F}$ distribution with degrees of freedom $K-1$ and $n-K$. And we can get a p-value from the ANOVA test. If we choose Kruskal-Wallis test, we need to rank the predictor, 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 test statistic approximates a χ^2 distribution with $K-1$ degrees of freedom and we can get a p-value from the K-W test. When the predictor $\mathbf{X_j}$ is discrete, we can perform a Chi-square test. Suppose $\mathbf{Y}_i \in \{\mathbf{1}, \ldots, \mathbf{K_1}\}$ and $\mathbf{X}_{ij} \in \{\mathbf{1}, \ldots, \mathbf{K_2}\}$. 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 $\hat{\pi}_{yj,k_1k_2} = n^{-1}\sum \mathbf{I}(Y_i = k_1)\mathbf{I}(X_{ij} = k_2)$. Our Chi-square test statistics 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.

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 decreasing order and select a submodel

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

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.

3.3 Simulation Studies

*Example*1: We considered the linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Half of the predictors are generated from standard Gaussian distribution, the other half are Binary predictors. The noise $\boldsymbol{\epsilon}$ is generated from two different distributions: the standard normal distribution and the t distribution with three degree of freedom. We consider two designs for the covariance matrix of X as follows: (1) $\boldsymbol{\Sigma}_1 = \mathbf{I}_{p\times p}$; (2) $\boldsymbol{\Sigma}_3 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.8$. We chose $(n, p) = (400, 1000)$, $s = 8$, $d = [n/\log n]$ and the coefficients of the nonzero components of the p-vectors β to be 5. For each model we simulated 500 data sets.

*Example*1.1:

Same as *Example*1 except that $\boldsymbol{y} = X^2\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

*Example*1.2:

Same as *Example*1 except that $\boldsymbol{y} = sin(X)\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

*Example*2: Let $g_1(x) = x$, $g_2(x) = x^2$ and $g_3(x) = sin(x)$. $\boldsymbol{y} = 5g_1(\mathbf{X}_1) + 5g_2(\mathbf{X}_2) + 5g_3(\mathbf{X}_3) + 5g_1(\mathbf{X}_4) + 5g_2(\mathbf{X}_5) + 5g_3(\mathbf{X}_6)$, where $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are continuous predictors and $\mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6$ are binary predictors. The other settings are the same with *Example*1.

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. Table 3.1 summarized the simulation results and we can draw the following conclusions:

1. Both tests perform better with standard normal noise and independent predictors according to higher proportions of predictors containing the true model selected.

2. Generally speaking, both methods perform well and the performances of the two methods are comparable.

Model	0 & $\mathbf{N}(0,1)$	0 & $\mathbf{t}(3)$	0.8 & $\mathbf{N}(0,1)$	0.8 & $\mathbf{t}(3)$
*Example*1	8	8	8	8
ANOVA & NIS	0.984	0.976	0.900	0.894
*Example*1	8	8	8	8
K-W &*NIS*	0.982	0.962	0.880	0.876
*Example*1.1	8	8	8	8
ANOVA & NIS	0.748	0.738	0.636	0.620
*Example*1.1	8	8	8	8
K-W &*NIS*	0.882	0.888	0.824	0.820
*Example*1.2	8	8	8	8
ANOVA & NIS	1.000	0.998	0.996	0.992
*Example*1.2	8	8	8	8
K-W &*NIS*	0.998	0.998	0.990	0.990
*Example*2	6	6	6	6
ANOVA & NIS	0.986	0.978	1.000	1.000
*Example*2	6	6	6	6
K-W &*NIS*	0.996	0.996	1.000	1.000

Table 3.1: Results of simulation in Section 3.3: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model

*Example*3: Same as *Example*1 except that $(n, p) = (100, 200)$ and $s = 6$.

*Example*3.1:

Same as *Example*3 except that $\boldsymbol{y} = X^2\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

*Example*3.2:

Same as *Example*3 except that $\boldsymbol{y} = sin(X)\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

*Example*4: Same as *Example*2 except that $(n, p) = (100, 200)$.

We used the median number of correctly selected predictors and proportion of times that the screened predictor set contained the true model to evaluate the performances of the procedures. Table 3.2 summarized the simulation results and we can draw the following conclusions:

1. With the decrease of the sample size, the performance of both methods become worse.

2. Generally speaking, the performances of K-W test is a little better than ANOVA test.

*Example*5: To make the simulation mimic the motivating arrhythmia data set, we choose $(n, p, s) = (450, 250, 6)$ and Y is distributed, conditional on $\mathbf{X} = \mathbf{x}$, as $Bin(1, p(\mathbf{x}))$, with $log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$. The other settings are the same as *Example*1.

*Example*5.1:

Same as *Example*5 except that $log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = \mathbf{x}^2\boldsymbol{\beta} + \varepsilon$.

*Example*5.2:

Same as *Example*5 except that $log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = sin(\mathbf{x})\boldsymbol{\beta} + \varepsilon$.

*Example*6: Same as *Example*2 except that $(n, p) = (450, 250)$ and Y is distributed, conditional on $\mathbf{X} = \mathbf{x}$, as $Bin(1, p(\mathbf{x}))$, with $log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}) = 5g_1(\mathbf{X}_1) + 5g_2(\mathbf{X}_2) + 5g_3(\mathbf{X}_3) + 5g_1(\mathbf{X}_4) + 5g_2(\mathbf{X}_5) + 5g_3(\mathbf{X}_6)$.

We used the median number of correctly selected predictors and proportion of

times that the screened predictor set contained the true model to evaluate the performances of the procedures. Table 3.3 summarized the simulation results and we can draw the following conclusions:

1. Both tests perform better with standard normal noise and independent predictors according to higher proportions of predictors containing the true model selected.

2. Generally speaking, both methods perform well and the performances of the two methods are comparable.

Model	0 & $\mathbf{N}(0,1)$	0 & $\mathbf{t}(3)$	0.8 & $\mathbf{N}(0,1)$	0.8 & $\mathbf{t}(3)$
$Example3$	6	6	6	6
$ANOVA\,\&\,NIS$	0.946	0.938	0.876	0.876
$Example3$	6	6	6	6
K-W $\&NIS$	0.938	0.932	0.876	0.872
$Example3.1$	6	6	6	6
$ANOVA\,\&\,NIS$	0.846	0.826	0.756	0.740
$Example3.1$	6	6	6	6
K-W $\&NIS$	0.822	0.778	0.718	0.684
$Example3.2$	6	6	6	6
$ANOVA\,\&\,NIS$	0.936	0.948	0.902	0.872
$Example3.2$	6	6	6	6
K-W $\&NIS$	0.952	0.966	0.920	0.900
$Example4$	6	6	6	6
$ANOVA\,\&\,NIS$	0.724	0.696	0.644	0.642
$Example4$	6	6	6	6
K-W $\&NIS$	0.774	0.726	0.674	0.674

Table 3.2: Results of simulation in Section 3.3: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model

Model	0 & $\mathbf{N}(0,1)$	0 & $\mathbf{t}(3)$	0.8 & $\mathbf{N}(0,1)$	0.8 & $\mathbf{t}(3)$
$Example5$	6	6	6	6
$ANOVA\,\&\,NIS$	0.740	0.720	0.530	0.528
$Example5$	6	6	5	5
K-W $\&NIS$	0.718	0.706	0.484	0.482
$Example5.1$	5	5	5	5
$ANOVA\,\&\,NIS$	0.416	0.406	0.318	0.306
$Example5.1$	6	6	6	5
K-W $\&NIS$	0.606	0.588	0.506	0.486
$Example5.2$	6	6	6	6
$ANOVA\,\&\,NIS$	0.934	0.914	0.792	0.754
$Example5.2$	6	6	6	6
K-W $\&NIS$	0.892	0.886	0.748	0.730
$Example6$	6	6	5	5
$ANOVA\,\&\,NIS$	0.586	0.580	0.494	0.404
$Example6$	6	6	6	5
K-W $\&NIS$	0.692	0.648	0.522	0.468

Table 3.3: Results of simulation in Section 3.3: Median numbers (top numbers) of correctly selected variables and proportions (bottom numbers) of times that the screened predictor set contained the true model

3.4 Real Data Analysis

3.4.1 Arrhythmia Data Set

In this section, we applied our screening procedures to the Arrhythmia data set. There are 452 rows, each representing the medical record of the different patient. There are 279 attributes, such as age, sex, height, weight and patients' ECG related data. The data set is labeled with 16 different classes. Class 1 corresponds to the normal ECG with no arrhythmia and class 16 refers to unlabeled patient. Class 2 to 15 correspond to different types of arrhythmia. The data set is heavily biased towards the no arrhythmia case with 245 patients belonging to class 1. The original

data contains columns with both missing values and single valued columns having the same value for all the patient records. These columns were deleted from the data set. The resulting data set contained 452 instances and 257 features.

Because the data set is heavily biased towards the no arrhythmia case, we first considered labeling the patients into two categories: no arrhythmia and all the other cases. Then we applied our screening procedure to the data set. To measure the classification accuracy, we used 10-fold cross validation. For continuous features, we used ANOVA test. For categorical features, we used Chi-square test. The features are selected based on the p-value of the selected tests. The number of features selected is $d_t = [n_t/\log n_t]$, where n_t is sample size of the training set. Then we applied the generalized linear model with SCAD penalty to the reduced feature space and get estimates for the test set. The classification accuracy can be calculated using the estimates and the true values of the test set. We repeated the whole procedure 100 times.

From the study by Gupta et al.[26], we know that the performance of Random Forest is quite well compared with other classification methods. We compared the performance of our method with random forest on the same data set, the results are summarized in Table 3.4. From the table, we can see that, with a much smaller model size and less computation time, the mean classification accuracy of our method is comparable to the Random Forest.

We also applied our screening procedure to the whole data set. After screening, the reduced feature space contains 73 features. Then we applied 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 $DI167$, $DI169$, $DII199$, $DII211$, $DII277$. Then we applied the random forest method to the whole data set. For comparison purpose, we listed the top 12 important features selected by model accuracy and Gini index below. Mean decrease in model accuracy: $DII224$, $DII91$, $DII277$, $DII93$, $DII228$, $DII234$, $DII199$, $DII103$, $DII179$, $DII76$, QRS duration, and $DII250$. Mean decrease in Gini index: $DII224$, $DII277$, QRS duration, $DII199$, $DII197$, $DII91$, $DII179$, $DII93$, $DII228$, $DI167$, $DII177$, and $DI169$. Mitra and Samanta[46] also studied the Arrhythmia data set. In their study, they got 18 features as reduced feature set as follows: Sex, QRS duration, $DII49$, $DII76$, $DII91$, $DII103$, $DII112$, $DI163$, $DI167$, $DI169$, $DII173$, $DII199$, $DII207$, $DII211$, $DII261$, $DII267$, $DII271$, and $DII277$.

Table 3.5 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.

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

Table 3.4: Results of the Arrhythmia data set:Mean values of the model size, classification accuracy and time (in seconds)

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

Table 3.5: Features selected by at least two methods

3.4.2 Asthma Data Set

In this section, we applied our screening procedures to the Asthma data set. The data set consists of 268 cases and 136 controls , indicating whether the child has the Asthma or not. There are 54675 continuous features, which are gene expression calculated by RMA Express software, and 160 discrete features, such as family ID, sex, country and SNP type. The original data contains rows with missing values and columns with single value. These rows and columns were deleted from the data set. The resulting data set contained 251 instances and 54802 features.

We applied our screening procedure to the data set. To measure the classification accuracy, we used 10-fold cross validation. For continuous features, we used ANOVA test. For categorical features, we used Chi-square test. The features are selected based on the p-value of the selected tests. The number of features selected is $d_t = [n_t/\log n_t]$, where n_t is sample size of the training set. Then we applied the generalized linear

model with SCAD penalty to the reduced feature space and get estimates for the test set. We can calculate the classification accuracy using the estimates and the true values of the test set. We repeated the whole procedure 100 times. The mean and median classification accuracy are 74.67% and 74.78%,respectively.

We also applied our screening procedure to the whole data set. After screening, the reduced feature space contains 45 features. Then we applied the generalized linear model with SCAD penalty to the reduced feature space. We got 17 features in the final model: 1559587_at, 1560842_a_at, 201017_at, 208359_s_at, 208534_s_at, 212486_s_at, 215649_s_at, 227561_at, 231592_at, 232688_at, 233946_at, 236278_at, 237083_at, 238573_at, 239992_at, 241630_at, and 243320_at. And their corresponding GenBank Accession Numbers are: $AL831859$, $BC042736$, $BG149698$, NM_004981, NM_006989, $N20923$, $AF217536$, $W73819$, $AV646335$, $AU144829$, $AL512690$, $AV705309$, $H46176$, $H19488$, $BF063430$, $AA742279$, and $H09564$.

Moffatt et al.[47] reported 10 SNPs on chromosome 17q21 that were strongly associated with childhood asthma as follows: $rs9303277$, $rs11557467$, $rs8067378$, $rs2290400$, $rs7216389$, $rs4795405$, $rs8079416$, $rs4795408$, $rs3894194$, and $rs3859192$. If we only use the SNPs data, after screening, the reduced feature space contains 45 features. And all the 10 SNPs listed above were contained in the reduced feature space. Then we applied the generalized linear model with SCAD penalty to the reduced feature space. We got 7 features in the final model: sex, $rs1106769$, $rs4795369$, $rs907092$, $rs9303277$, $rs11557467$, and $rs7211770$. We can see that $rs9303277$ and $rs11557467$ are the top two SNPs in their list.

From the study by Huang et al. [32], we know that the SNPs are highly correlated with gene expression, which is also associated with childhood asthma. And we can consider gene expression as the mediator between the SNPs and the disease. This might be the reason we only select gene expression features in our final model. For each of the 10 SNPs and each of the gene expression features we selected, we performed a one-way ANOVA test to see whether there is an association between the SNP and gene expression. In table 3.6, for each of the 10 SNPs, we listed three gene expression features with the smallest p-values. From the table, we can see that, for each SNP, we have at least one gene expression that is highly associated with it.

In our data set, we have 122 SNP features and 54675 gene expression features. The number of gene expression features is much larger than the number of SNP features. This might be another reason we only select gene expression features in our final model. We want both SNP features and gene expression features in the model, so we applied our screening procedure to the SNP features and gene expression features separately. After screening, we got 45 SNP features and 45 gene expression features. Then we applied the generalized linear model with SCAD penalty to the reduced feature space containing those 45 SNP features and 45 gene expression features. There are 18 features in the final model: sex, $rs1106769$, $rs9303277$, $rs7211770$, 1559258_a_at, 1560842_a_at, 208359_s_at, 208534_s_at, 212486_s_at, 215649_s_at, 227561_at, 232688_at, 235168_at, 236278_at, 236615_at, 238573_at, 239992_at, and 243320_at. The three SNP features selected are also selected when we only use the SNPs data. And $rs9303277$ is reported by Moffatt et al.[47]. For the gene expression features,

11 out of 14 features are also selected when we apply our screening procedure to the whole data set.

rs9303277	243320_at	1559587_at	233946_at
	0.0027	0.0101	0.0291
rs11557467	243320_at	212486_s_at	215649_s_at
	0.0026	0.0105	0.0141
rs8067378	243320_at	1559587_at	215649_s_at
	0.0026	0.0083	0.0174
rs2290400	215649_s_at	1559587_at	233946_at
	0.0140	0.0164	0.0218
rs7216389	215649_s_at	212486_s_at	233946_at
	0.0145	0.0173	0.0187
rs4795405	208534_s_at	243320_at	215649_s_at
	0.0013	0.0534	0.0592
rs8079416	208534_s_at	241630_at	227561_at
	0.0002	0.0358	0.0802
rs4795408	208534_s_at	241630_at	227561_at
	0.0007	0.0471	0.0940
rs3894194	208534_s_at	241630_at	212486_s_at
	0.0007	0.0635	0.1122
rs3859192	241630_at	208534_s_at	243320_at
	0.0030	0.0882	0.1316

Table 3.6: ANOVA test of association between SNP and gene expression

Bibliography

[1] H. Akaike. *Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, vol. 1, 267-281.* Akademinai Kiado, Budapest, 1973.

[2] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

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

[4] J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008.

[5] J. Chen and Z. Chen. Extended bayesian information criterion for model selection with large model space. *Biometrika*, 95:759–771, 2008.

[6] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[7] H. Cui, R. Li, and W. Zhong. Model-free feature screening for ultra-high dimensional discriminant analysis. *Journal of the American Statistical Association*, 110:630–641, 2015.

[8] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *The Annals of Statistics*, 32:407–499, 2004.

[10] J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36:2605–2637, 2008.

[11] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, 106:544–557, 2011.

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

[13] J. Fan and R. Li. Statistical challenges with high-dimensionality: feature selection in knowledge discovery. *Proceedings of International Congress of Mathematicians*, 3:595–622, 2006.

[14] J. Fan and R. Li. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

[15] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 70:849–911, 2008.

[16] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

[17] J. Fan and J. Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions On Information Theory*, 57:5467–5484, 2011.

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

[19] J. Fan, R. Samworth, and Y. Wu. Ultra-high dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

[20] J. Fan and R. Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38:3567–3604, 2010.

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

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

[23] W. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[24] J. Geoman, S. van de Geer, F. de Kort, and H. van Houwelingen. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, 20:93–99, 2004.

[25] C. Gu and D. Xiang. Cross-validating non-gaussian data : Generalized approximate cross-validation revisited. *Journal of Computational and Graphical Statistics*, 10:581–591, 2001.

[26] V. Gupta, S. Srinivasan, and S. Kudli. Prediction and classification of cardiac arrhythmia.

[27] P. Hall and H. Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18:533–550, 2009.

[28] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical LearningData Mining, Inference and Prediction, 2nd edn.* Springer, New York, 2009.

[29] M. Holden, S. Deng, L. Wojnowski, and B. Kulle. Gsea-snp: Applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24:27842785, 2008.

[30] D. Huang, R. Li, and H. Wang. Feature screening for ultrahigh-dimensional categorical data with applications. *Journal of Business & Economic Statistics*, 32:237–244, 2014.

[31] J. Huang, S. Ma, and C. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

[32] Y. Huang, T. VanderWeele, and X. Lin. Joint analysis of snp and gene expression data in genetic association studies of complex diseases. *The Annals of Applied Statistics*, 8:352–376, 2014.

[33] S. Kim and D. Volsky. Page: Parametric analysis of gene set enrichment. *Bioinformatics*, 6:144, 2005.

[34] Y. Kim and C. Gu. Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B*, 66:337–356, 2004.

[35] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16:375–390, 2006.

[36] W. Kruskal and W. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47:583–621, 1952.

[37] G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, 40:1846–1877, 2012.

[38] R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107:1129–1139, 2012.

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

[40] Y. Liu and Y. Wu. Variable selection via a combination of the l_0 and l_1 penalties. *Journal of Computational and Graphical Statistics*, 16:782–798, 2007.

[41] X. Ma and J. Zhang. Robust model-free feature screening via quantile correlation. *Journal of Multivariate Analysis*, 143:472–480, 2016.

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

[43] C. Mallows. Some comments on cp. *Technometrics*, 15:661–675, 1973.

[44] U. Mansmann and R. Meister. Testing differential gene expression in functional groups. goeman's global test versus an ancova approach. *Methods of Information in Medicine*, 44:449–453, 2005.

[45] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70:53–71, 2008.

[46] M. Mitra and R. Samanta. Cardiac arrhythmia classification using neural networks with selected features. *Procedia Technology*, 10:76–84, 2013.

[47] M. Moffatt, M. Kabesch, L. Liang, A. Dixon, D. Strachan, S. Heath, and et al. Genetic variants regulating ormdl3 expression contribute to the risk of childhood asthma. *Nature*, 448:470–473, 2007.

[48] M. Moffatt, M. Kabesch, L. Liang, and A. Dixon et al. Genetic variants regulating ormdl3 expression contribute to the risk of childhood asthma. *Nature*, 448:470–473, 2007.

[49] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[50] R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12:758–765, 1984.

[51] U. Olsson, F. Drasgow, and N. Dorans. The polyserial correlation coefficient. *Psychometrika*, 47:337–347, 1982.

[52] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000.

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

[54] C. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13:6897–705, 1985.

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

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

[57] H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104:1512–1524, 2009.

[58] H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B*, 71:671–683, 2009.

[59] H. Wang, R. Li, and C. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94:553–558, 2007.

[60] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81:1278–1283, 2007.

[61] L. Wang, G. Chen, and H. Li. Group scad regression analysis for microarray time course gene expression. *Bioinformatics*, 23:1486–1494, 2007.

[62] T. Wang and L. Zhu. Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102:1141–1151, 2011.

[63] S. Wu, X. Shen, and C. Geyer. Adaptive regularization using the entire solution surface. *Biometrika*, 96:513–527, 2009.

[64] T. Wu and K. Lange. Coordinate descent procedures for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244, 2008.

[65] Q. Xiong, N. ancona, E. Hauser, S. Mukherjee, and T. Furey. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Research*, 22:386397, 2012.

[66] C. Xu and J. Chen. The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*, 109:1257–1265, 2014.

[67] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006.

[68] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942, 2010.

[69] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics*, 37:3468–3497, 2009.

[70] H. Zhong, X. Yang, L. Kaplan, C.Molony, and E. Schadt. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *The American Journal of Human Genetics*, 86:581–591, 2010.

[71] L. Zhu, L. Li, R. Li, and L. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106:1464–1475, 2011.

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

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

[74] H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35:2173–2192, 2007.

[75] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533, 2008.