SCIENCE CHINA
Mathematics

Invited Articles
• REVIEWS •

October 2015 Vol. 58 No. 10: 2033–2054
doi: 10.1007/s11425-015-5062-9

A selective overview of feature screening for ultrahigh-dimensional data

LIU JingYuan[1,2,3], ZHONG Wei[2,1,3] & LI RunZe[4,*]

[1]*Department of Statistics, School of Economics, Xiamen University, Xiamen* 361005*, China;*
[2]*Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen* 361005*, China;*
[3]*Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen* 361005*, China;*
[4]*Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, PA* 16802-2111*, USA*
Email: jingyuan@xmu.edu.cn, wzhong@xmu.edu.cn, rzli@psu.edu

Received March 24, 2015; accepted July 23, 2015; published online August 20, 2015

Abstract High-dimensional data have frequently been collected in many scientific areas including genome-wide association study, biomedical imaging, tomography, tumor classifications, and finance. Analysis of high-dimensional data poses many challenges for statisticians. Feature selection and variable selection are fundamental for high-dimensional data analysis. The sparsity principle, which assumes that only a small number of predictors contribute to the response, is frequently adopted and deemed useful in the analysis of high-dimensional data. Following this general principle, a large number of variable selection approaches via penalized least squares or likelihood have been developed in the recent literature to estimate a sparse model and select significant variables simultaneously. While the penalized variable selection methods have been successfully applied in many high-dimensional analyses, modern applications in areas such as genomics and proteomics push the dimensionality of data to an even larger scale, where the dimension of data may grow exponentially with the sample size. This has been called ultrahigh-dimensional data in the literature. This work aims to present a selective overview of feature screening procedures for ultrahigh-dimensional data. We focus on insights into how to construct marginal utilities for feature screening on specific models and motivation for the need of model-free feature screening procedures.

Keywords correlation learning, distance correlation, sure independence screening, sure joint screening, sure screening property, ultrahigh-dimensional data

MSC(2010) 62H12, 62H20

Citation: Liu J Y, Zhong W, Li R Z. A selective overview of feature screening for ultrahigh-dimensional data. Sci China Math, 2015, 58: 2033–2054, doi: 10.1007/s11425-015-5062-9

1 Introduction

High-dimensional data have frequently been collected in a large variety of areas such as genomics, biomedical imaging, tomography, tumor classifications, and finance. Analysis of high-dimensional data poses many challenges for statisticians. Donoho [10] convincingly demonstrated the need for developing new statistical methodologies and theories for high-dimensional data analysis. Fan and Li [20] presented a comprehensive overview of statistical challenges with high-dimensionality in various statistical problems. Fan et al. [18] described various challenges in analysis of big data. Analysis of high-dimensional data calls for new statistical methodologies and theories. Feature selection and variable selection are fundamental for high-dimensional data analysis.

*Corresponding author

© Science China Press and Springer-Verlag Berlin Heidelberg 2015

The sparsity principle, which assumes that only a small number of predictors contribute to the response, is frequently adopted and deemed useful in the analysis of high-dimensional data. Following this general principle, a large number of variable selection approaches have been developed in the recent literature to estimate a sparse model and select significant variables simultaneously. Fan and Lv [22] provided a review of variable selection methods for high-dimensional data. The nonnegative garrote [4,54], the least absolute shrinkage and selection operator (LASSO) [48] and the smoothly clipped absolute deviation (SCAD) [19] are the most popular approaches for selecting significant variables and estimating regression coefficients simultaneously. While the aforementioned variable selection methods have been successfully applied in many high-dimensional analyses, modern applications in areas such as genomics and proteomics push the dimensionality of data to an even larger scale, where the dimension of data may grow exponentially with the sample size. This has been called ultrahigh-dimensional data in the literature. Such ultrahigh-dimensional data present simultaneous challenges of computational expediency, statistical accuracy and algorithm stability [25]. It is difficult to directly apply the aforementioned variable selection methods to those ultrahigh-dimensional statistical learning problems, due to computational complexity inherent in those methods.

To address those challenges, Donoho [11, 12] showed that the individual equivalence of the minimal L_1-norm solution and the minimal L_0-norm solution. Candes and Tao [5] further extended [11, 12] idea and proposed the Dantzig selector for a linear regression model when the number of predictors is much greater than the sample size. Independence learning has been proposed to select significant genes between treatment and control groups for macroarray data by using a two-sample test in [13, 14, 24, 46]. Fan and Li [21] emphasized the importance of feature screening in ultrahigh-dimensional data analysis, and proposed sure independence screening in the context of linear regression models. Since the seminar work of Fan and Li [21], feature screening for ultrahigh-dimensional data received a lot of attentions in the literature. Many authors have developed various sure independence screening procedures. This article aims to provide a selective overview on this topic. Most feature screening procedures can be classified into two categories: Model-based and model-free feature screening procedures.

In Sections 2–4, we concentrate on model-based feature screening procedures. Specifically, we provide motivations and insights into feature screening procedures for linear models in Section 2, and give a brief introduction of independent screening procedure based on Pearson's correlation, generalized Pearson and rank correlation for linear models and transformation linear models. In Section 3, we first introduce feature screening procedure for generalized linear models. We further discuss feature screening for a general parametric framework which include generalized linear models, robust regression, quantile regression and linear classification. Section 4 is devoted to an overview for feature screening methods for nonparametric regression models including additive models and varying coefficient models, the two most popular nonparametric models in the literature of statistics.

In practice, we typically have data with a huge number of candidate variables, but we have little information that the actual model is linear or follows any other specific parametric, nonparametric or semiparametric form. Thus, it is of great interest to develop model-free feature screening procedures for ultrahigh-dimensional data. By model-free, it means that one does not need to impose a specific model structure on regression functions to carry out a screening procedure. One way to achieve the model-free goal is to develop feature screening procedures for a general class of models which include most commonly-used parametric, nonparametric and semiparametric models as special cases thereof. Another way is to construct feature screening procedure via tests of independence, which naturally do not require to specify a model for the regression functions. We provide a selective overview of model-free feature screening procedures for ultrahigh-dimensional data in Section 5.

Before we pursue further, let us introduce notation. Throughout this paper, $\{\boldsymbol{x}_i, Y_i\}$, $i = 1, \ldots, n$ is assumed to be an independent and identically distributed random sample from a population $\{\boldsymbol{x}, Y\}$, where $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is the p-dimensional predictor variables, and Y is the response variable. Denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, the $n \times p$ design matrix. Furthermore, denote by X_{ij} the i-th observation of the j-th variable. Thus, $\boldsymbol{x}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$. Let ε be a general random error and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ with ε_i being an $n \times 1$ vector of random errors. Let $\mathcal{M}_*$ stand for the true

model with the size $s = |\mathcal{M}_*|$, and $\widehat{\mathcal{M}}$ is the selected model with the size $d = |\widehat{\mathcal{M}}|$. The definitions of $\mathcal{M}_*$ and $\widehat{\mathcal{M}}$ may be different for different models and contexts.

2 Linear models and transformation linear models

This section is devoted to a review of screening procedures for linear model and its variants. Let us start with Pearson correlation and linear regression models.

2.1 Pearson Correlation and linear models

Consider a linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a $p \times 1$ vector of parameters, When the dimension p is greater than the sample size n, the least squares estimator of $\boldsymbol{\beta}$ is not well defined due to the singularity of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. A useful technique in the classical linear regression to deal with singularity of the design matrix $\boldsymbol{X}$ is the ridge regression, defined by

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda I_p)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y},$$

where λ is a ridge parameter. It is observed that if $\lambda \to 0$, then $\hat{\boldsymbol{\beta}}_\lambda$ tends to the least squares estimator, if it is well-defined; and if $\lambda \to \infty$, then $\lambda\hat{\boldsymbol{\beta}}_\lambda$ tends to $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$. This implies that $\hat{\boldsymbol{\beta}}_\lambda \propto \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$. In practice, all covariates and the response are marginally standardized so that their means and variances equals 0 and 1, respectively. Then $\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ becomes the vector consists of the sample version of Pearson correlations between the response and individual covariate. This is the motivation of using Pearson correlation as a marginal utility for feature screening. Specifically, denote

$$\omega_j = \frac{1}{n}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{Y}, \quad \text{for} \quad j = 1, 2, \ldots, p. \tag{2.2}$$

Here, it is assumed that both $\boldsymbol{X}_j$ and $\boldsymbol{Y}$ are marginally standardized. Thus, ω_j indeed is the sample correlation between the j-th predictor and the response variable.

Fan and Lv [21] suggested ranking all predictors according to $|\omega_j|$ and select the top predictors which are relatively strongly correlated with the response. To be specific, for any given $\gamma \in (0, 1)$, the $[\gamma n]$ top ranked predictors are selected to obtain the submodel

$$\widehat{\mathcal{M}}_\gamma = \{1 \leqslant j \leqslant p : |\omega_j| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \tag{2.3}$$

where $[\gamma n]$ denotes the integer part of γn. It reduces the ultrahigh-dimensionality down to a relatively moderate scale $[\gamma n]$, i.e., the size of $\widehat{\mathcal{M}}_\gamma$, and then the well-established penalized variable selection methods is applied to the submodel $\widehat{\mathcal{M}}_\gamma$. This screening procedure defined in (2.3) is referred to as sure independence screening (SIS) in the literature.

Before we pursue further, let us examine a simple, but very useful case in which Y is coded as 1 for case (disease) and -1 for control (normal). Then ω_j is proportional to $\bar{x}_{j+} - \bar{x}_{j-}$, where the $\bar{x}_{j+}$ and $\bar{x}_{j-}$ are the sample mean of samples with $Y = 1$ and -1, respectively. Thus, ranking ω_j is about the same as ranking the t-value of two-sample t-test under the common variance assumption. Fan and Fan [15] proposed the t-test statistic for two-sample mean problem as a marginal utility for feature screening and establish its theoretical properties. Methods in multiple tests include false discovery rate method [1, 2] are to derive a critical value or proper thresholding for all t-values. Feature screening based on ω_j is distinguished from the multiple test methods in that the screening method aims to rank the importance of predictors rather than directly judge whether an individual variable is significant. Thus, further fine-tuning analysis based on the selected variables from the screening step is necessary. Note that it is

not involved any numerical optimization to evaluate ω_j or $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$. Thus, feature screening based on the Pearson correlation can be carried out in a simple way at low computational burden. In addition to the computational advantage, Fan and Lv [21] showed that this feature screening procedure also enjoys nice theoretical property.

For (2.1), the true model is defined as

$$\mathcal{M}_* = \{1 \leqslant j \leqslant p : \beta_j \neq 0\}$$

with the size $s = |\mathcal{M}_*|$. Under the sparsity assumption, the size s should be smaller than n. The success of the feature screening procedure depends on whether the selected subset $\widehat{\mathcal{M}}_\gamma$ is able to contain the set $\mathcal{M}_*$ of all the important variables. Fan and Lv [21] proved that that $\mathcal{M}_* \subset \widehat{\mathcal{M}}_\gamma$ with an overwhelming probability under the following technical conditions:

(a1) The covariate vector follows an elliptical contoured distribution [27] and the concentration property (see [21] for details about the concentration property).

(a2) The variance of the response variable is finite. For some $\kappa \geqslant 0$ and $c_1, c_2 > 0$, $\min_{j\in\mathcal{M}_*}|\beta_j| \geqslant c_1 n^{-\kappa}$ and $\min_{j\in\mathcal{M}_*}|\mathrm{cov}(\beta_j^{-1}Y, X_j)| \geqslant c_2$.

(a3) For some $\tau \geqslant 0$ and $c_3 > 0$, the largest eigenvalue of Σ satisfies $\lambda_{\max}(\mathrm{cov}(\boldsymbol{x})) \leqslant c_3 n^{\tau}$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

(a4) There exists a $\xi \in (0, 1-2\kappa)$ such that $\log p = O(n^{\xi})$, and $p > n$.

Under the above conditions, if $2\kappa + \tau < 1$ then there exists some $0 < \theta < 1 - 2\kappa - \tau$ such that when $\gamma \sim cn^{-\theta}$ with $c > 0$. Under Conditions (a1)–(a4), assume that $2\kappa + \tau < 1$ and the true model size $s \leqslant [\gamma n]$, Fan and Lv [21] showed that for some $C > 0$,

$$P(\mathcal{M}_* \subset \widehat{\mathcal{M}}_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log n)) \to 1, \tag{2.4}$$

as $n \to \infty$. The property in (2.4) is referred to as sure screening property. It ensures that under certain conditions, with probability tending to one, the submodel selected by SIS would not miss any truly important predictor, and hence the false negative rate can be controled. The sure screening property is essential for implementation of all screening procedures in practice, since any post-screening variable selection method (e.g., penalized regression) is based on the screened submodels. It is worth pointing out that Conditions (a1)–(a4) certainly are not the weakest conditions to establish this property. They are only used to facilitate the technical proofs from a theoretical point of view. Although these conditions are sometimes difficult to check in practice, the numerical studies in [21] demonstrated that the SIS can efficiently shrink the ultrahigh dimension p down to a relatively large scale $O(n^{1-\theta})$ for some $\theta > 0$ and still can contain all important predictors into the submodel $\widehat{\mathcal{M}}_\gamma$ with probability approaching one as n tends to ∞.

However, the screening procedure may fail when some key conditions are not valid. For example, when a predictor X_j for some j is jointly correlated but marginally uncorrelated with the response, then $\mathrm{cov}(Y, X_j) = 0$. As a result, Condition (a2) is not valid. In such situations, the SIS may fail to select this important variable. On the other hand, the SIS tends to select the unimportant predictor which is jointly uncorrelated but highly marginally correlated with the response. To refine the screening performance, Fan and Lv [21] 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. Subsection 3.2 below will introduce iterative feature screening procedures under a general framework, to better control the false negative rate in the finite sample case than the one-step screening procedures. Another way to cope with the situation with some individual covariates being jointly correlated but marginally uncorrelated with the response is the forward regression. Wang [51] carefully studied the property of forward regression with ultrahigh-dimensional predictors. To achieve the sure screening property, the forward regression requires that there exist two positive constants $0 < \tau_1 < \tau_2 < \infty$ such that $\tau_1 < \lambda_{\min}(\mathrm{cov}(\boldsymbol{x})) \leqslant \lambda_{\max}(\mathrm{cov}(\boldsymbol{x})) < \tau_2$. Wang [51] further proposed using the extended BIC [7] to determine the size of the active predictor set.

2.2 Generalized correlation, rank correlation and transformation linear models

The SIS procedure proposed in [21] performs well for the linear regression model with ultrahigh-dimensional predictors. It is well known that the Pearson correlation is used to measure linear dependence. In the presence of nonlinearity, one may try to make a transformation such as the Box-Cox transformation on the covariates. This motivates people to consider the Pearson correlation between a transformed covariate and the response as the marginal utility.

To capture both linearity and nonlinearity, Hall and Miller [28] defined the generalized correlation between the j-th predictor X_j and Y to be,

$$\rho_g(X_j, Y) = \sup_{h\in\mathcal{H}} \frac{\mathrm{cov}\{h(X_j), Y\}}{\sqrt{\mathrm{var}\{h(X_j)\}\mathrm{var}(Y)}}, \quad \text{for each} \ \ j = 1, \ldots, p, \tag{2.5}$$

where $\mathcal{H}$ is a class of functions including all linear functions. For example, it is a class of polynomial functions up to a given degree. Remark that if $\mathcal{H}$ is restricted to be a class of all linear functions, $\rho_g(X_j, Y)$ is the absolute value of Pearson correlation between X_j and Y. Therefore, $\rho_g(X, Y)$ is considered as a generalization of the conventional Pearson correlation. Suppose that $\{(X_{ij}, Y_i), i = 1, 2, \ldots, n\}$ is a random sample from the population (X_j, Y). The generalized correlation $\rho_g(X_j, Y)$ can be estimated by

$$\widehat{\rho}_g(X_j, Y) = \sup_{h\in\mathcal{H}} \frac{\sum_{i=1}^n \{h(X_{ij}) - \bar{h}_j\}(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n \{h^2(X_{ij}) - \bar{h}_j^2\} \sum_{i=1}^n (Y_i - \overline{Y})^2}}, \tag{2.6}$$

where

$$\bar{h}_j = n^{-1}\sum_{i=1}^n h(X_{ij}) \quad \text{and} \quad \overline{Y} = n^{-1}\sum_{i=1}^n Y_i.$$

In practice, $\mathcal{H}$ can be defined as a set of cubic splines [28]

The above generalized correlation is able to characterize both linear and nonlinear relationships between two random variables. Thus, Hall and Miller [28] proposed using the generalized correlation $\rho_g(X_j, Y)$ as a marginal screening utility and ranking all predictors based on the magnitude of estimated generalized correlation $\widehat{\rho}_g(X_j, Y)$. Therefore, given a suitable cutoff, one can select predictors with higher rankings and thus reduce the ultrahigh-dimensionality to a relatively low scale. In addition, Hall and Miller [28] introduced a bootstrap method to determine a cutoff. Let $r(j)$ be the ranking of the j-th predictor X_j, i.e., X_j has the $r(j)$ largest empirical generalized correlation of all. Let $r^*(j)$ be the ranking of the j-th predictor X_j using the bootstrapped sample. Then, a nominal $(1-\alpha)$-level two-sided prediction interval of the ranking, $[\widehat{r}_-(j), \widehat{r}_+(j)]$, is computed based on the bootstrapped $r^*(j)$'s. Hall and Miller [28] recommended a criterion to regard the predictor X_j as influential if $\widehat{r}_+(j) < \frac{1}{2}p$ or some smaller fraction of p, such as $\frac{1}{4}p$. Therefore, the proposed generalized correlation ranking reduces the ultrahigh p down to the size of the selected model

$$\widehat{\mathcal{M}}_k = \{j : \widehat{r}_+(j) < kp\},$$

where $0 < k < 1/2$ is a constant multiplier to control the size of the selected model $\widehat{\mathcal{M}}_k$. Although the generalized correlation ranking can detect both linear and nonlinear features in the ultrahigh-dimensional problems, how to choose an optimal transformation $h(\cdot)$ remains an open issue and the associated sure screening property is needed to justify.

Alternative to make transformations on predictors, one may make a transformation on the response and define a correlation between the transformed response and a covariate. A general transformation regression model is defined to be

$$H(Y_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \tag{2.7}$$

Li et al. [35] proposed the rank correlation as a measure of the importance of each predictor by imposing

an assumption on strict monotonicity on $H(\cdot)$ in (2.7). Instead of the sample Pearson correlation defined in Subsection 2.1, Li et al. [35] proposed the marginal rank correlation

$$\omega_j = \frac{1}{n(n-1)} \sum_{i \neq \ell}^{n} I(X_{ij} < X_{\ell j}) I(Y_i < Y_\ell) - \frac{1}{4}, \tag{2.8}$$

to measure the importance of the j-th predictor X_j. Note that the marginal rank correlation equals a quarter of the Kendall τ correlation between the response and the j-th predictor. According to the magnitudes of all ω_j's, the feature screening procedure based on the rank correlation selects a submodel

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leqslant j \leqslant p : |\omega_j| > \gamma_n\},$$

where γ_n is the predefined threshold value.

Because the Pearson correlation is not robust against heavy-tailed distributions, outliers or influence points, the rank correlation can be considered as an alternative way to robustly measure the relationship between two random variables. Rather than using the Pearson correlation of SIS [21], Li et al. [35] referred this rank correlation based feature screening procedure to as a robust rank correlation screening (RRCS) procedure to deal with ultrahigh-dimensional data. From the definition of the marginal rank correlation, it is robust against heavy-tailed distributions and invariant under monotonic transformation, which implies that there is no need to estimate the transformation $H(\cdot)$. This may save a lot of computational cost and is another advantage of RRCS over the feature screening procedure based on generalized correlation.

For (2.7) with $H(\cdot)$ being an unspecified strictly increasing function, Li et al. [35] defined the true model as

$$\mathcal{M}_* = \{1 \leqslant j \leqslant p : \beta_j \neq 0\}$$

with the size $s = |\mathcal{M}_*|$. Suppose that $\min_{j \in \mathcal{M}_*} E|X_j|$ is a positive constant free of p. Under some technical conditions, Li et al. [35] proved that the RRCS enjoys the sure screening property, i.e.,

$$P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}_{\gamma_n}) \geqslant 1 - 2s \exp(-c_2 n^{1-2\kappa}) \to 1, \quad \text{as} \ \ n \to \infty \tag{2.9}$$

hold for some constant c_2 provided that $\gamma_n = c_3 n^{-\kappa}$ for some constant c_3, $p = O(\exp(n^\delta))$ for some $\delta \in (0, 1)$ satisfying $\delta + 2\kappa < 1$ for any $\kappa \in (0, 0.5)$. Li et al. [35] further demonstrated that the RRCS procedure is robust to outliers and influence points in the observations.

3 Generalized linear models and beyond

Consider a simple linear regression model

$$Y = \beta_{j0} + X_j \beta_{j1} + \varepsilon_j^*, \tag{3.1}$$

where ε_j^* is a random error with $E(\varepsilon_j^* \,|\, X_j) = 0$. The discussion in Subsection 2.1 actually implies that we may use the magnitude of the least squares estimate $\hat{\beta}_{j1}$ or the residual sum of squares RSS_j to rank importance of predictors. Specifically, let $\boldsymbol{Y}$ and $\boldsymbol{X}_j$ both be marginal standardized so that their sample means equals 0 and sample variance equal 1. Thus, $\hat{\beta}_{j1} = n^{-1} \boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{Y}$, which equals ω_j defined in (2.2), is used to measure the importance of X_j. Furthermore, note that $\mathrm{RSS}_j \overset{\mathrm{def}}{=} \|\boldsymbol{Y} - \hat{\beta}_{j0} - \boldsymbol{X}_j \hat{\beta}_{j1}\|^2 = \|\boldsymbol{Y}\|^2 - n\|\hat{\beta}_{j1}\|^2$, as $\boldsymbol{X}_j$ is marginally standardized (i.e., $\|\boldsymbol{X}_j\|^2 = n$), and the least squares estimate $\hat{\beta}_{j0} = \bar{Y} = 0$. Thus, a greater magnitude of $\hat{\beta}_{j1}$, or equivalently a smaller RSS_j, results in a higher rank of the j-th predictor X_j. The same rationale can be used to develop the marginal utilities for a much broader class of models, with a different estimate of β_{j1} or a more general definition of loss function than RSS, such as the negative likelihood function for the generalized linear models. In fact, the negative likelihood function is a monotonic function of RSS in the Gaussian linear models.

3.1 Generalized linear models

Assume that given the predictor vector $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$, the conditional distribution of Y belongs to an exponential family, whose density function has the canonical form

$$f_{Y|\boldsymbol{x}}(y \,|\, \boldsymbol{x}) = \exp\{y\theta(\boldsymbol{x}) - b(\theta(\boldsymbol{x})) + c(y)\}, \tag{3.2}$$

for some known functions $b(\cdot)$ and $c(\cdot)$. Further assume $E(Y \,|\, \boldsymbol{x}) = b'\{\theta(\boldsymbol{x})\} = g^{-1}(\beta_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$. Denote the negative log-likelihood of i-th observation $\{\boldsymbol{x}_i, Y_i\}$ to be $\ell(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, Y_i)$. Note that the minimizer of the negative log-likelihood $\sum_{i=1}^n \ell(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, Y_i)$ is not well defined when $p > n$.

Parallel to the least squares estimate for (3.1), assume that each predictor is standardized to have mean zero and standard deviation one, and define the maximum marginal likelihood estimator (MMLE) $\widehat{\boldsymbol{\beta}}_j^M$ for the j-th predictor X_j as

$$\widehat{\boldsymbol{\beta}}_j^M = (\widehat{\beta}_{j0}^M, \widehat{\beta}_{j1}^M) = \arg\min_{\beta_{j0}, \beta_{j1}} \sum_{i=1}^n \ell(Y_i, \beta_{j0} + \beta_{j1} X_{ij}). \tag{3.3}$$

Similar to the marginal least squares estimate for (3.1), it is reasonable to consider the magnitude of $\widehat{\beta}_{j1}^M$ as a marginal utility to rank the importance of X_j and select a submodel for a given prespecified threshold γ_n,

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leqslant j \leqslant p : |\widehat{\beta}_{j1}^M| \geqslant \gamma_n\}. \tag{3.4}$$

This was proposed for feature screening in generalized linear model by Fan and Song [26]. To establish the theoretical properties of MMLE, Fan and Song [26] defined the population version of the marginal likelihood maximizer as

$$\boldsymbol{\beta}_j^M = (\beta_{j0}^M, \beta_{j1}^M) = \arg\min_{\beta_{j0}, \beta_{j1}} E\ell(Y, \beta_{j0} + \beta_{j1} X_j).$$

They first showed that the marginal regression parameters $\beta_{j1}^M = 0$ if and only if $\mathrm{cov}(Y, X_j) = 0$ for $j = 1, \ldots, p$. Thus, when the important variables are correlated with the response, $\beta_{j1}^M \neq 0$. Define the true model as $\mathcal{M}_* = \{1 \leqslant j \leqslant p : \beta_j \neq 0\}$ with the size $s = |\mathcal{M}_*|$. Fan and Song [26] showed that under some conditions if $|\mathrm{cov}(Y, X_j)| \geqslant c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$ and some $c_1 > 0$, then

$$\min_{j \in \mathcal{M}_*} |\beta_{j1}^M| \geqslant c_2 n^{-\kappa},$$

for some $c_2, \kappa > 0$. Thus, the marginal signals β_{j1}^M's are stronger than the stochastic noise provided that X_j's are marginally correlated with Y. Under some technical assumptions, Fan and Song [26] proved that the MMLEs are uniformly convergent to the population values and established the sure screening property of the MMLE screening procedure. That is, if $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$, where $k_n = b'(K_n B + B) + m_0 K_n^\alpha / s_0$, B is the upper bound of the true value of β_j^M, K_n is the supremum norm of $\boldsymbol{x}$, then for any $c_3 > 0$, there exists some $c_4 > 0$ such that

$$P\Big(\max_{1 \leqslant j \leqslant p} |\widehat{\beta}_{j1}^M - \beta_{j1}^M| \geqslant c_3 n^{-\kappa}\Big) \leqslant p\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^\alpha)\}.$$

In addition, assume that $|\mathrm{cov}(Y, X_j)| \geqslant c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$ and take $\gamma_n = c_5 n^{-\kappa}$ with $c_5 \leqslant c_2/2$, the following inequality holds,

$$P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}_{\gamma_n}) \geqslant 1 - s\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^\alpha)\} \to 1, \tag{3.5}$$

as $n \to \infty$. It further implies that the MMLE can handle the ultrahigh-dimensionality as high as

$\log p = o(n^{1-2\kappa})$ for the logistic model with bounded predictors, and $\log p = o(n^{(1-2\kappa)/4})$ for the linear model without the joint normality assumption.

Fan and Song [26] further discussed the size of the selected model $\widehat{\mathcal{M}}_{\gamma_n}$ in the asymptotic sense. Under some regularity conditions, they showed that with probability approaching one, $|\widehat{\mathcal{M}}_{\gamma_n}| = O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}$, where the constant κ determines how large the threshold γ_n is, and $\lambda_{\max}(\Sigma)$ is the maximum eigenvalue of the covariance matrix Σ of predictors $\boldsymbol{x}$, which controls how correlated the predictors are. If $\lambda_{\max}(\Sigma) = O(n^{\tau})$, the size of $\widehat{\mathcal{M}}_{\gamma_n}$ has the order $O(n^{2\kappa+\tau})$.

3.2 Feature screening under a general parametric framework

As discussed in the beginning of this section, one may use the RSS of marginal simple linear regression as a marginal utility for feature screening. The RSS criterion can be extended to a general statistical framework.

Suppose that we are interested in exploring the relationship between $\boldsymbol{x}$ and Y. A general statistical framework is to minimize an objective function

$$Q(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} L(Y_i, \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i), \tag{3.6}$$

where $L(\cdot,\cdot)$ is a loss function, and $\boldsymbol{\beta}$ is a regression coefficient and is assumed to be sparse. By taking different loss functions, many commonly-used statistical frameworks can be unified under (3.6). Let us provide a few examples under this model framework.

1. Linear models. Let $L(Y_i, \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) = (Y_i - \beta_0 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)^2$. This leads to the least squares method for a linear regression model.

2. Generalized linear models. Let $L(Y_i, \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$ be the negative logarithm of (quasi)-likelihood function of Y_i given $\boldsymbol{x}_i$, $\ell(Y_i, \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$ defined in the Subsection 3.1. This leads to the likelihood method for generalized linear model, which includes normal linear regression models, logistic regression model and Poisson log-linear models as special cases.

3. Quantile linear regression. In the presence of heteroscedastic errors, people may consider quantile regression instead of the least squares estimate. For a given quantile level $\alpha \in (0,1)$, define $\rho_\alpha(u) = u\{\alpha - I(u<0)\}$, the quantile loss function, where $I(A)$ is the indicator function of a set A. Taking $L(Y_i, \beta_0+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) = \rho_\alpha(Y_i-\beta_0-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$ leads to the quantile regression. In particular, $\alpha = 1/2$ corresponds to the median regression or the least absolute deviation method.

4. Robust linear regression. In the presence of outliers or heavy-tailed errors, robust linear regression has been proved to be more suitable than the least squares method. The commonly-used loss function includes the L_1-loss: $\rho_1(u) = |u|$, the Huber loss function: $\rho_2(u) = (1/2)|u|^2 I(|u| \leqslant \delta) + \delta\{|u| - (1/2)\delta\}I(|u| > \delta)$ (see [34]). Setting $L(Y_i, \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) = \rho_k(Y_i - \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$, $k = 1$, or 2 leads a robust linear regression.

5. Classification. In machine learning, the response variable typically is set to be the input class label such as $Y_1 = \{-1, +1\}$ as in the classification method in the statistical literature. The hinge loss function is defined to be $h(u) = \{(1-u) + |1-u|\}/2$ (see [50]). Taking $L(Y_i, \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) = h(Y_i - \beta_0 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$ in (3.6) leads a classification rule.

Similar to RSS_j defined in the beginning of this section, a natural marginal utility of the j-th predictor is

$$L_j = \min_{\beta_{j0},\beta_{j1}} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \beta_{j0} + X_{ij}\beta_{j1}).$$

According to the definition, the smaller the value L_j, the more important the corresponding j-th predictor. This was first proposed in [25], in which numerical studies clearly demonstrates the potential of this marginal utility.

As mentioned in the end of Section 2, the marginal feature screening methodology may fail if the

predictor is marginally uncorrelated but jointly related with the response, or jointly uncorrelated with the response but has higher marginal correlation than some important features. Fan et al. [25] proposed an iterative feature screening procedure under the general statistical framework (3.6). The proposed iterative procedure consists of the following steps.

S1. Compute the vector of marginal utilities $(L_1, \ldots, L_p)$ and select the set $\widehat{\mathcal{A}}_1 = \{1 \leqslant j \leqslant p : L_j \text{ is among the first } k_1 \text{ smallest of all}\}$. Then apply a penalized method to the model with index set $\widehat{\mathcal{A}}_1$, such as the LASSO [48] and SCAD [19], to select a subset $\widehat{\mathcal{M}}$.

S2. Compute, for each $j \in \{1, \ldots, p\}/\widehat{\mathcal{M}}$,

$$L_j^{(2)} = \min_{\beta_0, \boldsymbol{\beta}_{\widehat{\mathcal{M}}}, \beta_j} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}}^{\mathrm{T}} \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + X_{ij}\beta_j),$$

where $\boldsymbol{x}_{i,\widehat{\mathcal{M}}}$ denotes the sub-vector of $\boldsymbol{x}_i$ consisting of those elements in $\widehat{\mathcal{M}}$. $L_j^{(2)}$ can be considered as the additional contribution of the j-th predictor given the existence of predictors in $\widehat{\mathcal{M}}$. Then, select the set

$$\widehat{\mathcal{A}}_2 = \{j \in \{1, \ldots, p\}/\widehat{\mathcal{M}} : L_j^{(2)} \text{ is among the first } k_2 \text{ smallest of all}\}.$$

S3. Employ a penalized (pseudo-)likelihood method such as the LASSO and SCAD on the combined set $\widehat{\mathcal{M}} \cup \widehat{\mathcal{A}}_2$ to select an updated selected set $\widehat{\mathcal{M}}$, i.e., use the penalized likelihood to obtain

$$\widehat{\boldsymbol{\beta}}_2 = \mathop{\arg\min}_{\beta_0, \boldsymbol{\beta}_{\widehat{\mathcal{M}}}, \boldsymbol{\beta}_{\widehat{\mathcal{A}}_2}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \boldsymbol{x}_{i,\widehat{\mathcal{M}}}^{\mathrm{T}} \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + \boldsymbol{x}_{i,\widehat{\mathcal{A}}_2}^{\mathrm{T}} \boldsymbol{\beta}_{\widehat{\mathcal{A}}_2}) + \sum_{j \in \widehat{\mathcal{M}} \cup \widehat{\mathcal{A}}_2} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function such as LASSO or SCAD. Thus the indices set of $\widehat{\boldsymbol{\beta}}_2$ that are none-zero yield a new updated set $\widehat{\mathcal{M}}$.

S4. Repeat S2 and S3 until $|\widehat{\mathcal{M}}| \leqslant d$, where d is the prescribed number and $d \leqslant n$. The indices set $\widehat{\mathcal{M}}$ is the final selected submodel.

This iterative procedure extends the Iterative SIS [21], without explicit definition of residuals, to a general statistical framework. It also allows the procedure to delete predictors from the previously selected set.

In addition, to reduce the false selection rates, Fan et al. [25] further introduced a variation of their iterative feature screening procedure. One can partition the sample into two halves at random and apply the previous iterative procedure separately to each half to obtain the two submodels, denoted by $\widehat{\mathcal{M}}^{(1)}$ and $\widehat{\mathcal{M}}^{(2)}$. By the sure screening property, both sets may satisfy $P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}^{(h)}) \to 1$, as $n \to \infty$, for $h = 1, 2$, where $\mathcal{M}_*$ is the true model set. Then, the intersection $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}^{(1)} \cap \widehat{\mathcal{M}}^{(2)}$ can be considered as the final estimated set of $\mathcal{M}_*$. This estimate also satisfies that $P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}) \to 1$. This variant can effectively reduce the false selection rates in the feature screening stage.

3.3 Sure joint screening procedure

The iterative feature screening procedure significantly improves the simple marginal screening in that it can select weaker but still significant predictors and delete inactive predictors which are spuriously marginally correlated with the response. However, Xu and Chen [52] claimed that the gain of the iterative procedure is apparently built on higher computational cost and increased complexity. To this end, they proposed the sparsity restricted MLE (SRMLE) method for the generalized linear models and demonstrated the SRMLE retains the virtues of the iterative procedure in a conceptually simpler and computationally cheaper manner.

Suppose that a random sample was collected from a ultrahigh-dimensional generalized linear model

and its the log-likelihood function is

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})Y_i - b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}. \tag{3.7}$$

Xu and Chen [52] defined the SRMLE of $\boldsymbol{\beta}$ to be

$$\widehat{\boldsymbol{\beta}}_{[k]} = \arg\max_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}), \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leqslant k, \tag{3.8}$$

where $\|\boldsymbol{\beta}\|_0$ denotes the number of nonzero entries of $\boldsymbol{\beta}$. Let $\widehat{\mathcal{M}} = \{1 \leqslant j \leqslant p : \widehat{\boldsymbol{\beta}}_{[k]j} \neq 0\}$ correspond to nonzero entries of $\widehat{\boldsymbol{\beta}}_{[k]}$. For a given number k, the SMLE method can be considered as a joint-likelihood-supported feature screening procedure. Of course, the maximization problem in (3.8) cannot be solved when p is large. To tackle the computation issue, Xu and Chen [52] proposed the following approximation. For a given $\boldsymbol{\beta}$, $\ell_n(\boldsymbol{\gamma})$ is approximated by

$$h_n(\boldsymbol{\gamma}; \boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^{\mathrm{T}} S_n(\boldsymbol{\beta}) - (u/2)\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2, \tag{3.9}$$

for some scaling parameter $u > 0$, where $\|\cdot\|_2$ denotes the L_2 norm and $S_n(\boldsymbol{\beta}) = \ell_n'(\boldsymbol{\beta})$ is the score function. The first two terms in (3.9) come from the first-order Taylor's expansion of $\ell_n(\boldsymbol{\beta})$, while the third term is viewed as a quadratic penalty to avoid $\boldsymbol{\gamma}$ too far away from $\boldsymbol{\beta}$. The advantage of this approximation is that there is a closed form solution for $\boldsymbol{\beta}$ at each step. Specifically, given the t-th sparse solution $\boldsymbol{\beta}^{(t)}$, we can update $\boldsymbol{\beta}^{(t)}$ by $\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\gamma}} h_n(\boldsymbol{\gamma}; \boldsymbol{\beta}^{(t)})$, subject to $\|\boldsymbol{\gamma}\|_0 \leqslant k$, whose solution has an explicit form $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{H}(\boldsymbol{g}; k)$, where

$$\boldsymbol{g} = (g_1, \ldots, g_p) = \boldsymbol{\beta}^{(t)} + u^{-1}\boldsymbol{x}^{\mathrm{T}}\{Y - b'(\boldsymbol{x}\boldsymbol{\beta}^{(t)})\}, \quad \boldsymbol{H}(\boldsymbol{g}; k) = [H(g_1; r_k), \ldots, H(g_p; r_k)]^{\mathrm{T}}$$

with r_k being the k-th largest value of $\{|g_1|, \ldots, |g_p|\}$ and $H(g_j; r_k) = \gamma_j I(|g_j| > r_k)$, which is a hard-thresholding rule. Thus, Xu and Chen [52] referred this algorithm to as iterative hard thresholding (IHT) algorithm. The authors further showed that the IHT algorithm has the ascent property, i.e., under some conditions, $\ell_n(\boldsymbol{\beta}^{(t+1)}) \geqslant \ell_n(\boldsymbol{\beta}^{(t)})$. This property features the SRMLE a promising method for feature screening. Therefore, one can start with an initial $\boldsymbol{\beta}^{(0)}$ and carry out the above iteration until $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2$ falls below some tolerance level. Xu and Chen [52] further demonstrated that the SRMLE procedure enjoys the sure screening property, i.e., under some technical conditions, $P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}) \to 1$, as $n \to \infty$.

4 Nonparametric regression models

In practice, parametric models such as linear and generalized linear models may lead to model misspecification. Nonparametric models become very useful in the absence of priori information about the model structure for the regression and may be used to enhance the model flexibility, especially for the ultrahigh-dimensional data with much challenge to check model assumptions. Therefore, the feature screening techniques for the nonparametric models naturally drew great attention from researchers.

4.1 Additive models

Fan et al. [16] proposed a nonparametric independence screening (NIS) for the ultrahigh-dimensional additive model

$$Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon, \tag{4.1}$$

where $\{m_j(X_j)\}_{j=1}^p$ have mean 0 for identifiability. An intuitive population level marginal screening utility is $E(f_j^2(X_j))$, where $f_j(X_j) = E(Y \mid X_j)$ is the projection of Y onto X_j. With the sample $\{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$, $f_j(x)$ can be estimated via a normalized B-spline basis $\boldsymbol{B}_j(x) = \{B_{j1}(x), \ldots, B_{jd_n}(x)\}^{\mathrm{T}}$:

$$\hat{f}_{nj}(x) = \hat{\boldsymbol{\beta}}_j^{\mathrm{T}} \boldsymbol{B}_j(x), \quad 1 \leqslant j \leqslant p, \tag{4.2}$$

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

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

Thus the screened model index set is

$$\widehat{\mathcal{M}}_\nu = \{1 \leqslant j \leqslant p : \|\hat{f}_{nj}\|_n^2 \geqslant \nu_n\}, \tag{4.3}$$

for some predefined threshold value ν_n, with $\|\hat{f}_{nj}\|_n^2 = n^{-1} \sum_{i=1}^n \hat{f}_{nj}(X_{ij})^2$. In practice, ν_n can be determined by the random permutation idea [55], i.e., ν_n is taken as the qth quantile of $\|\hat{f}_{nj}^*\|_n^2$, where $0 \leqslant q \leqslant 1$, and $\hat{f}_{nj}^*$ is estimated in the same fashion as above, but based on the random permuted decouple $\{(\boldsymbol{x}_{\pi(i)}, Y_i), i = 1, \ldots, n\}$, and $\{\pi(1), \ldots, \pi(n)\}$ is a random permutation of the index $\{1, \ldots, n\}$.

Fan et al. [16] advocated the sure screening property of NIS of including the true model $\mathcal{M}_* = \{j : Em_j(X_j)^2 > 0\}$ based on a set of regularity conditions: (b1) The r-th derivative of f_j is Lipschitz of order α for some $r > 0$, $\alpha \in (0, 1]$ and $s = r + \alpha > 0.5$; (b2) the marginal density function of X_j is bounded away from 0 and infinity; (b3) the signal of the active components do not vanish, i.e., $\min_{j \in \mathcal{M}_*} E\{f_j^2(X_j)\} \geqslant c_1 d_n n^{-2\kappa}$, $0 < \kappa < s/(2s+1)$ and $c_1 > 0$; (b4) the sup norm $\|m\|_\infty$ is bounded, the number of spline basis d_n satisfies $d_n^{2s-1} \leqslant c_2 n^{-2\kappa}$ for some $c_2 > 0$; and the i.i.d. random error ε_i satisfies the sub-exponential tail probability: For any $B_1 > 0$, $E\{\exp(B_2|\varepsilon_i|) \mid \boldsymbol{x}_i\} < B_2$ for some $B_2 > 0$. Under Conditions (b1)–(b4),

$$P(\mathcal{M}_* \subset \widehat{\mathcal{M}}_\nu) \to 1, \tag{4.4}$$

for $p = \exp\{n^{1-4\kappa} d_n^{-3} + n d_n^{-3}\}$. In addition, if $\operatorname{var}(Y) = O(1)$, then the size of the selected model $|\widehat{\mathcal{M}}_\nu|$ is bounded by the polynomial order of n and the false selection rate is under control.

Fan et al. [16] further refined the NIS by two strategies. The first strategy consists of the iterative NIS along with the post-screening penalized method for additive models (ISIS-penGAM) — first conduct the regular NIS with data-driven threshold; then apply penGAM [43] to further select components based on the screened model; reapply the NIS with the response Y replaced by the residual from the regression using selected components; repeat the process until convergence. The second strategy is the greedy INIS — in each NIS step, a predefined small model size p_0, often taken to be 1 in practice, is used instead of the data-driven cutoff ν_n. The false positive rate can be better controlled by this method, especially when the covariates are highly correlated or conditionally correlated.

4.2 Varying coefficient models

Varying coefficient models are another class of popular nonparametric regression models in the literature and defined by

$$Y = \sum_{j=1}^{p} \beta_j(u) X_j + \varepsilon, \tag{4.5}$$

where $\beta_j(u)$'s are coefficient functions. An intercept can be included by setting $X_1 \equiv 1$. Fan et al. [23] extended the NIS for the additive model [16] to this varying coefficient model. From a different perspective, Fan et al. [39] proposed a conditional correlation screening procedure based on the kernel regression

approach.

Fan et al. [23] started from the marginal regression problem for each X_j, $j = 1, \ldots, p$, i.e., finding $a_j(u)$ and $b_j(u)$ to minimize $E\{(Y - a_j(u) - b_j(u)X_j)^2 \,|\, u\}$. Thus, the marginal contribution of X_j for predicting Y can be described as

$$u_j = E\{a_j(u) + b_j(u)X_j\}^2 - E\{a_0(u)^2\},$$

where $a_0(u) = E(Y \,|\, u)$. With the sample $\{(u_i, \boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$, where $\boldsymbol{x}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, $a_j(u), b_j(u)$ and $a_0(u)$ can be estimated based on the normalized B-spline basis $\boldsymbol{B}(u) = \{B_1(u), \ldots, B_{d_n}(u)\}^{\mathrm{T}}$, i.e.,

$$\hat{a}_j(u) = \boldsymbol{B}(u)^{\mathrm{T}}\hat{\boldsymbol{\eta}}_j, \quad \hat{b}_j(u) = \boldsymbol{B}(u)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_j, \quad \hat{a}_0(u) = \boldsymbol{B}(u)^{\mathrm{T}}\hat{\boldsymbol{\eta}}_0,$$

where $(\hat{\boldsymbol{\eta}}_j^{\mathrm{T}}, \hat{\boldsymbol{\theta}}_j^{\mathrm{T}})^{\mathrm{T}} = (Q_{nj}^{\mathrm{T}}Q_{nj})^{-1}Q_{nj}^{\mathrm{T}}\boldsymbol{Y}$, $\hat{\boldsymbol{\eta}}_0 = (\mathcal{B}_n^{\mathrm{T}}\mathcal{B}_n)^{-1}\mathcal{B}_n^{\mathrm{T}}\boldsymbol{Y}$, and

$$Q_{nj} = \begin{pmatrix} \boldsymbol{B}(u_1)^{\mathrm{T}}, & X_{1j}\boldsymbol{B}(u_1)^{\mathrm{T}} \\ \vdots & \vdots \\ \boldsymbol{B}(u_n)^{\mathrm{T}}, & X_{1j}\boldsymbol{B}(u_n)^{\mathrm{T}} \end{pmatrix}_{n\times 2d_n}, \quad \mathcal{B}_n = \begin{pmatrix} \boldsymbol{B}(u_1)^{\mathrm{T}} \\ \vdots \\ \boldsymbol{B}(u_n)^{\mathrm{T}} \end{pmatrix}_{n\times d_n}, \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1}.$$

Therefore, the sample marginal utility for screening is

$$\hat{\mu}_{nj} = \|\hat{a}_j(u) + \hat{b}_j(u)X_j\|_n^2 - \|\hat{a}_0(u)\|_n^2, \tag{4.6}$$

where $\|\cdot\|_n^2$ is the sample average, defined in the same fashion as in the last section. The selected model is then $\widehat{\mathcal{M}}_\tau = \{1 \leqslant j \leqslant p : \hat{u}_{nj} \geqslant \tau_n\}$ for the pre-specified data-driven threshold τ_n.

Fan et al. [23] proved the sure screening property $P(\mathcal{M}_* \subset \widehat{\mathcal{M}}_\tau) \to 1$ for $p = o(\exp\{n^{1-4\kappa}d_n^{-3}\})$, $\kappa > 0$, of the proposed method under similar regularity conditions with [16]. An additional bounded constraint is imposed on u, and both X_j and $E(Y \,|\, \boldsymbol{x}, u)$, as well as the random error ε, should satisfy the sub-exponential tail probability. Furthermore, the iterative screening and greedy iterative screening can be conducted similarly with [16].

From a different point of view, Liu et al. [39] proposed another sure independent screening procedure for the ultrahigh-dimensional varying coefficient model (4.5) based on the conditional correlation learning (CC-SIS). Since (4.5) is indeed a linear model when conditioning on u, the conditional correlation between each predictor X_j and Y given u is defined similarly as the Pearson correlation:

$$\rho(X_j, Y \,|\, u) = \frac{\mathrm{cov}(X_j, Y \,|\, u)}{\sqrt{\mathrm{cov}(X_j, X_j \,|\, u)\mathrm{cov}(Y, Y \,|\, u)}},$$

where $\mathrm{cov}(X_j, Y \,|\, u) = E(X_jY \,|\, u) - E(X_j \,|\, u)E(Y \,|\, u)$, and the population level marginal utility to evaluate the importance of X_j is $\rho_{j0}^* = E\{\rho^2(X_j, Y \,|\, u)\}$. To estimate ρ_{j0}^*, or equivalently, the conditional means $E(Y \,|\, u)$, $E(Y^2 \,|\, u)$, $E(X_j \,|\, u)$, $E(X_j^2 \,|\, u)$ and $E(X_jY \,|\, u)$, with sample $\{(u_i, \boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, the kernel regression is adopted, for example,

$$\widehat{E}(Y \,|\, u) = \sum_{i=1}^n \frac{K_h(u_i - u)Y_i}{\sum_{i=1}^n K_h(u_i - u)},$$

where $K_h(t) = h^{-1}K(t/h)$ is a rescaled kernel function. Therefore, all the plug-in estimates for the conditional means and conditional correlations $\widehat{\rho}^2(X_j, Y \,|\, u_i)$'s are obtained, and thus also that of ρ_{j0}^*:

$$\widehat{\rho}_j^* = \frac{1}{n}\sum_{i=1}^n \widehat{\rho}^2(X_j, Y \,|\, u_i). \tag{4.7}$$

And the screened model is defined as $\widehat{\mathcal{M}} = \{j : 1 \leqslant j \leqslant p,\ \widehat{\rho}_j^*$ ranks among the first $d\}$, where the size $d = [n^{4/5}/\log(n^{4/5})]$ follows the hard threshold by Fan et al. [21] but modified for the nonparametric regression, where $[a]$ is the integer part of a.

The authors proved the sure screening property of CC-SIS under some regularity conditions: (c1) the density function of u has continuous second-order derivative; (c2) the kernel function $K(\cdot)$ is bounded uniformly over its finite support; (c3) X_jY, Y^2, X_j^2 satisfy the sub-exponential tail probability uniformly; (c4) all the conditional means, along with their first- and second-order derivatives are finite and the conditional variances are bounded away from 0.

Furthermore, the authors showed that the population level signal ρ_{j0}^* of the active and inactive predictors can be well separated under the linearity condition [57] and when $\min_{j\in\mathcal{M}_*}\rho_{i0}^*$ does not vanish, where $\mathcal{M}_* = \{1 \leqslant j \leqslant p :\ \beta_j(u) \neq 0$ for some $u\}$ is the true model. With these additional conditions, the authors also systematically studied the ranking consistency [57] of CC-SIS:

$$\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^*\} > 0 \quad \text{in probability,} \tag{4.8}$$

i.e., with an overwhelming probability, the truly important predictors have larger $\widehat{\rho}_j^*$ than the unimportant ones.

4.3 Heterogeneous nonparametric models

As is known, nonparametric quantile regression is useful to analyze the heterogeneous data, by separately studying different conditional quantiles of the response given the predictors. As to the ultrahigh-dimensional data, He et al. [32] proposed an adaptive nonparametric screening approach based on this quantile regression methodology.

At any quantile $\alpha \in (0, 1)$, the true sparse model is defined as

$$\mathcal{M}_\alpha = \{1 \leqslant j \leqslant p :\ Q_\alpha(Y \,|\, \boldsymbol{x}) \text{ functionally depends on } X_j\},$$

where $Q_\alpha(Y \,|\, \boldsymbol{x})$ is the αth conditional quantile of Y given $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$. Thus, their population screening criteria can be defined as

$$q_{\alpha j} = \{Q_\alpha(Y \,|\, X_j) - Q_\alpha(Y)\}^2, \tag{4.9}$$

with $Q_\alpha(Y \,|\, X_j)$, the α-th conditional quantile of Y given X_j, and $Q_\alpha(Y)$, the unconditioned quantile of Y. Thus Y and X_j are independent if and only if $q_{\alpha j} = 0$ for every $\alpha \in (0, 1)$. Notice that $Q_\alpha(Y \,|\, X_j)$ $= \mathrm{argmin}_f E[\rho_\alpha(Y - f(X_j))]$, where $\rho_\alpha(z) = z[\alpha - I(z < 0)]$ is the quantile loss function.

To estimate $q_{\alpha j}$ based on the sample $\{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$, one may consider the B-spline approximation based on a set of basis functions $\boldsymbol{B}(x) = \{B_1(x), \ldots, B_{d_n}(x)\}^{\mathrm{T}}$, i.e., one considers $\widehat{Q}_\alpha(Y \,|\, x)$ $= \boldsymbol{B}(x)^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}_j$, where $\hat{\boldsymbol{\Gamma}} = \mathrm{argmin}_{\boldsymbol{\Gamma}} \sum_{i=1}^n \rho_\alpha(Y_i - \boldsymbol{B}(X_{ij})^{\mathrm{T}}\boldsymbol{\Gamma})$. Furthermore, $Q_\alpha(Y)$ in (4.9) might be estimated by $F_{Y,n}^{-1}(\alpha)$, the αth sample quantile function based on $Y_1, \ldots, Y_n$. Therefore, $q_{\alpha j}$ is estimated by

$$\hat{q}_{\alpha j} = \frac{1}{n}\sum_{i=1}^n \{\boldsymbol{B}(X_{ij})^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}_j - F_{Y,n}^{-1}(\alpha)\}^2, \tag{4.10}$$

and the selected submodel $\widehat{\mathcal{M}}_\gamma = \{1 \leqslant j \leqslant p : \hat{q}_{\alpha j} \geqslant \gamma_n\}$ for some $\gamma_n > 0$.

To guarantee the sure screening property and control the false selection rate, the r-th derivative of $Q_\alpha(Y \,|\, X_j)$ is required to satisfy the Lipschitz condition of order c, where $r + c > 0.5$; the active predictors need to have strong enough marginal signals; the conditional density function of Y given $\boldsymbol{x}$ is locally bounded away from 0 and infinity — this relaxes the sub-exponential tail probability condition in literature; however, global upper and lower bounds are needed for the marginal density functions of X_j's; and additional restriction on d_n, the number of basis functions is applied.

5 Model-free feature screening

Feature screening procedures reviewed in Sections 2–4 were developed for a class of specific models. In high-dimensional modelling, it may be very challenging to specify the model structure on the regression function without priori information. Zhu et al. [57] advocated model-free feature screening procedures for ultrahigh-dimensional data. This section is devoted to review of several feature screening procedures without imposing a specific model structure on regression function.

5.1 Sure independent ranking screening procedure

Let Y be the response variable with support Ψ_y. Here Y can be both univariate and multivariate. Let $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a covariate vector. Zhu et al. [57] developed the notion of active predictors and inactive predictors without specifying a regression model. We consider the conditional distribution function of Y given $\boldsymbol{x}$, denoted by $F(y \,|\, \boldsymbol{x}) = P(Y < y \,|\, \boldsymbol{x})$. Define the true model

$$\mathcal{M}_* = \{k : F(y \,|\, \boldsymbol{x}) \text{ functionally depends on } X_k \text{ for some } y \in \Psi_y\},$$

if $k \in \mathcal{M}_*$, X_k is referred to as an active predictor, otherwise it is referred to as an inactive predictor. Again $|\mathcal{M}_*| = s$. Let $\boldsymbol{x}_{\mathcal{M}_*}$, an $s \times 1$ vector, consist of all X_k with $k \in \mathcal{M}_*$. Similarly, let $\boldsymbol{x}_{\mathcal{M}_*^c}$, a $(p-s) \times 1$ vector, consist of all inactive predictors X_k with $k \in \mathcal{M}_*^c$.

Zhu et al. [57] considered a general model framework under which a unified screening approach was developed. Specifically, they considered that $F(y \,|\, \boldsymbol{x})$ depends on $\boldsymbol{x}$ only through $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*}$ for some $p_1 \times K$ constant matrix $\boldsymbol{B}$. In other words, assume that

$$F(y \,|\, \boldsymbol{x}) = F_0(y \,|\, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*}), \tag{5.1}$$

where $F_0(\cdot \,|\, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*})$ is an unknown distribution function for a given $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*}$.

Many existing models with either continuous or discrete response are special examples of (5.1). For example, many existing regression models can be written in the following form:

$$h(Y) = f_1(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*}) + \boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*} + f_2(\boldsymbol{\beta}_3^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*})\varepsilon, \tag{5.2}$$

for a continuous response and

$$g\{E(Y \,|\, \boldsymbol{x})\} = f_1(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*}) + \boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{x}_{\mathcal{M}_*}, \tag{5.3}$$

for a binary or count response by extending the framework of generalized linear model to semiparametric regression, where $h(\cdot)$ is a monotone function, $g(\cdot)$ is a link function, $f_2(\cdot)$ is a nonnegative function, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are unknown coefficients, and it is assumed that ε is independent of $\boldsymbol{x}$. Here, $h(\cdot), g(\cdot), f_1(\cdot)$ and $f_2(\cdot)$ may be either known or unknown. Clearly, (5.2) is a special case of (5.1) if $\boldsymbol{B}$ is chosen to be a basis of the column space spanned by $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ for (5.2) and by $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for (5.3). Meanwhile, it is seen that (5.2) with $h(Y) = Y$ includes the following special cases: The linear regression model, the partially linear model [31] and the single-index model [30], and (5.3) is the generalized partially linear single-index model [6]. (5.2) also includes the transformation regression model for a general transformation $h(Y)$. As a consequence, a feature screening approach developed under (5.1) offers a unified approach that works for a wide range of existing models.

As before, assume that $E(X_k) = 0$ and $\mathrm{Var}(X_k) = 1$ for $k = 1, \ldots, p$. By the law of iterated expectations, it follows that $E[\boldsymbol{x}E\{\mathbf{1}(Y < y) \,|\, \boldsymbol{x}\}] = \mathrm{cov}\{\boldsymbol{x}, \mathbf{1}(Y < y)\}$. To avoid specifying a structure for regression function, Zhu et al. [57] considered the correlation between X_j and $\mathbf{1}(Y < y)$ to measure the importance of the j-th predictor instead of the correlation between X_j and Y. Specifically, define

$\boldsymbol{M}(y) = E\{\boldsymbol{x}F(y\,|\,\boldsymbol{x})\}$, and let $\Omega_k(y)$ be the k-th element of $\boldsymbol{M}(y)$. Define

$$\omega_k = E\{\Omega_k^2(Y)\}, \quad k = 1, \ldots, p. \tag{5.4}$$

Zhu et al. [57] advocated ω_k as the marginal utility measure for predictor ranking. Intuitively, if X_k and Y are independent, $\Omega_k(y) = 0$ for any $y \in \Psi_y$ and $\omega_k = 0$. On the other hand, if X_k and Y are related, then there exists some $y \in \Psi_y$ such that $\Omega_k(y) \neq 0$, and hence ω_k must be positive. This is the motivation to employ the sample estimate of ω_k to rank all the predictors.

Given a random sample $\{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$ from $\{\boldsymbol{x}, Y\}$. For ease of presentation, we assume that the sample predictors are all marginally standardized. For any given y, the sample moment estimator of $\boldsymbol{M}(y)$ is

$$\widehat{\boldsymbol{M}}(y) = n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i \mathbf{1}(Y_i < y).$$

Consequently, a natural estimator for ω_k is

$$\hat{\omega}_k = \frac{1}{n}\sum_{j=1}^{n} \widehat{\Omega}_k^2(Y_j) = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n} X_{ik}\mathbf{1}(Y_i < Y_j)\right\}^2, \quad k = 1, \ldots, p,$$

where $\widehat{\Omega}_k(y)$ denotes the k-th element of $\widehat{\boldsymbol{M}}(y)$, and X_{ik} denotes the k-th element of $\boldsymbol{x}_i$. [57] proposed ranking all the candidate predictors $X_k, k = 1, \ldots, , p$, according to $\widehat{\omega}_k$ from the largest to smallest, and then selecting the top ones as the active predictors.

Motivated by the following fact, Zhu et al. [57] established the consistency in ranking property for their procedure. If $K = 1$ and $\boldsymbol{x} \sim N_p(\mathbf{0}, \sigma^2 I_p)$ with unknown σ^2. It follows by a direct calculation that

$$\boldsymbol{M}(y) = E\{\boldsymbol{x}F_0(y\,|\,\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})\} = c(y)\boldsymbol{\beta},$$

where $c(y) = \|\boldsymbol{\beta}\|^{-1}\int_{-\infty}^{\infty} vF_0(y\,|\,v\|\boldsymbol{\beta}\|)\phi(v; 0, \sigma^2)\,dv$ with $\phi(v; 0, \sigma^2)$ being the density function of $N(0, \sigma^2)$ at v. Thus, if $E\{c^2(Y)\} > 0$, then

$$\max_{k\in\mathcal{M}_*^c} \omega_k < \min_{k\in\mathcal{M}_*} \omega_k, \tag{5.5}$$

and $\omega_k = 0$ if and only if $k \in \mathcal{M}_*^c$. This implies that the quantity ω_k may be used for feature screening in this setting. Under much more general conditions than the normality assumption, Zhu et al. [57] showed that (5.5) holds. The authors further established a concentration inequiality for $\hat{\omega}_k$ and, $\max_{k\in\mathcal{M}_*^c} \hat{\omega}_k < \min_{k\in\mathcal{M}_*} \hat{\omega}_k$, which is referred to as the property of consistency in ranking (CIR). Due to this property, Zhu et al. [57] referred their procedure to as sure independent ranking screening (SIRS) procedure. They studied several issues related to implementation of the SIRS method. Since the SIRS possesses the CIR property, the authors further developed a soft cutoff value for $\hat{\omega}_j$'s to determine which indices should be included in $\widehat{\mathcal{M}_*}$ by extend the idea of introducing auxiliary variables for thresholding proposed by Luo et al. [40]. Zhu et al. [57] 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. Lin et al. [38] proposed an improved version of the SIRS procedure for a setting in which the relationship between the response and an individual predictor is symmetric.

5.2 Feature screening via distance correlation

The SIRS procedure was proposed for multi-index models that include many commonly-used models in order to be a model-free screening procedure. Another strategy to achieve model-free is to employ the measure of independence to efficiently detect linearity and nonlinearity between predictors and the response variable and construct feature screening procedures for ultrahigh-dimensional data. Li et al. [37] proposed a SIS procedure based on the distance correlation [47]. Unlike the Pearson correlation coefficient,

rank correlation coefficient and generalized correlation coefficient, which all are defined for two random variables, the distance covariance are defined from two random vectors which are allowed to have different dimensions.

The distance correlation between two general random vector $\boldsymbol{U} \in R^{q_1}$ and $\boldsymbol{V} \in R^{q_2}$ is defined by

$$\mathrm{dcov}^2(\boldsymbol{U},\boldsymbol{V}) = \int_{R^{q_1+q_2}} \|\phi_{\boldsymbol{U},\boldsymbol{V}}(\boldsymbol{t},\boldsymbol{s}) - \phi_{\boldsymbol{U}}(\boldsymbol{t})\phi_{\boldsymbol{V}}(\boldsymbol{s})\|^2 w(\boldsymbol{t},\boldsymbol{s})d\boldsymbol{t}d\boldsymbol{s}, \tag{5.6}$$

where $\phi_{\boldsymbol{U}}(\boldsymbol{t})$ and $\phi_{\boldsymbol{V}}(\boldsymbol{s})$ are the marginal characteristic functions of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, $\phi_{\boldsymbol{U},\boldsymbol{V}}(\boldsymbol{t},\boldsymbol{s})$ is the joint characteristic function of $\boldsymbol{U}$ and $\boldsymbol{V}$, and $w(\boldsymbol{t},\boldsymbol{s}) = \{c_{q_1}c_{q_2}\|\boldsymbol{t}\|_{q_1}^{1+q_1}\|\boldsymbol{s}\|_{q_2}^{1+q_2}\}^{-1}$ with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$. Here and hereafter, $\|\boldsymbol{a}\| = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}$ stands for the Euclidean norm of $\boldsymbol{a}$ if $\boldsymbol{a}$ is a real vector, and $\|\phi\|^2 = \phi\bar{\phi}$ for a complex-valued function ϕ with $\bar{\phi}$ being the conjugate of ϕ. Székely et al. [47] proved that $\mathrm{dcov}^2(\boldsymbol{U},\boldsymbol{V}) = S_1+S_2-2S_3$, where $S_1 = E(\|\boldsymbol{U}-\widetilde{\boldsymbol{U}}\|\|\boldsymbol{V}-\widetilde{\boldsymbol{V}}\|)$, $S_2 = E(\|\boldsymbol{U}-\widetilde{\boldsymbol{V}}\|)E(\|\boldsymbol{U}-\widetilde{\boldsymbol{V}}\|)$ and $S_3 = E\{E(\|\boldsymbol{U}-\widetilde{\boldsymbol{U}}\|\,|\,\boldsymbol{U})E(\|\boldsymbol{V}-\widetilde{\boldsymbol{V}}\|\,|\,\boldsymbol{V})\}$, and $(\widetilde{\boldsymbol{U}},\widetilde{\boldsymbol{V}})$ is an independent copy of $(\boldsymbol{U},\boldsymbol{V})$. Thus, the distance covariance between $\boldsymbol{U}$ and $\boldsymbol{V}$ can be estimated by plugging-in its sample counterpart. Specifically, the distance covariance between $\boldsymbol{U}$ and $\boldsymbol{V}$ is estimated by $\widehat{\mathrm{dcov}}^2(\boldsymbol{U},\boldsymbol{V}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3$ with

$$\begin{aligned}
\widehat{S}_1 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\boldsymbol{U}_i-\boldsymbol{U}_j\|\|\boldsymbol{V}_i-\boldsymbol{V}_j\|,\\
\widehat{S}_2 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\boldsymbol{U}_i-\boldsymbol{U}_j\|\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\boldsymbol{V}_i-\boldsymbol{V}_j\|, \quad \text{and}\\
\widehat{S}_3 &= \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n \|\boldsymbol{U}_i-\boldsymbol{U}_l\|\|\boldsymbol{V}_j-\boldsymbol{V}_l\|
\end{aligned}$$

based on an iid random sample $\{(\boldsymbol{U}_i,\boldsymbol{V}_i), i = 1,\ldots,n\}$ from the population $(\boldsymbol{U},\boldsymbol{V})$.

Accordingly, the distance correlation between $\boldsymbol{U}$ and $\boldsymbol{V}$ is defined by

$$\mathrm{dcorr}(\boldsymbol{U},\boldsymbol{V}) = \mathrm{dcov}(\boldsymbol{U},\boldsymbol{V})/\sqrt{\mathrm{dcov}(\boldsymbol{U},\boldsymbol{V})\mathrm{dcov}(\boldsymbol{V},\boldsymbol{V})},$$

and estimated by plugging in the corresponding estimate of distance covariances. From the definition of distance covariance, it can be seen that if $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent, then $\mathrm{dcov}^2(\boldsymbol{U},\boldsymbol{V}) = 0$. Székely et al. [47] theoretically proved that $\mathrm{dcorr}(\boldsymbol{U},\boldsymbol{V}) = 0$ if and only if $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent. Furthermore, $\mathrm{dcorr}(U,V)$ is strictly increasing in the absolute value of the Pearson correlation between two univariate normal random variables U and V.

Let $\boldsymbol{y} = (Y_1,\ldots,Y_q)^{\mathrm{T}}$ be the response vector and $\boldsymbol{x} = (X_1,\ldots,X_p)^{\mathrm{T}}$ be the predictor vector. Motivated by the appealing properties of distance correlation, Li et al. [37] proposed a SIS procedure to rank the importance of the k-th predictor X_k for the response using its distance correlation $\widehat{\omega}_j = \widehat{\mathrm{dcorr}}^2(X_j,\boldsymbol{y})$. Then, a set of important predictors with large $\widehat{\omega}_j$ is selected and denoted by $\widehat{\mathcal{M}} = \{j : \widehat{\omega}_j \geqslant cn^{-\kappa}, \text{ for } 1 \leqslant j \leqslant p\}$.

Suppose that both $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfy the sub-exponential tail probability uniformly in p, i.e.,

$$\sup_p \max_{1\leqslant j\leqslant p} E\{\exp(s\|X_j\|_1^2)\} < \infty,$$

and $E\{\exp(s\|\boldsymbol{y}\|_q^2)\} < \infty$, for $s > 0$. Li et al. [37] proved that $\widehat{\omega}_j$ is consistent to ω_j uniformly in p,

$$P\Big(\max_{1\leqslant j\leqslant p}|\widehat{\omega}_j - \omega_j| \geqslant cn^{-\kappa}\Big) \leqslant O(p[\exp\{-c_1n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2n^{\gamma})]), \tag{5.7}$$

for any $0 < \gamma < 1/2-\kappa$, some constants $c_1 > 0$ and $c_2 > 0$. If the minimum distance correlation of active predictors is further assumed to satisfy $\min_{j\in\mathcal{M}_*}\omega_j \geqslant 2cn^{-\kappa}$, for some constants $c > 0$ and $0 \leqslant \kappa < 1/2$,

as $n \to \infty$,

$$P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}) \geqslant 1 - O(s[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]) \to 1, \tag{5.8}$$

where s is the size of $\mathcal{M}_*$, the indices set of the active predictors is defined by

$$\mathcal{M}_* = \{1 \leqslant j \leqslant p : F(\boldsymbol{y} \,|\, \boldsymbol{x}) \text{ functionally depends on } X_j \text{ for some } \boldsymbol{y} \in \Psi_y\},$$

without specifying a regression model. The result in (5.7) shows that the feature screening procedure enjoys the sure screening property without assuming any regression function of $\boldsymbol{y}$ on $\boldsymbol{x}$, and therefore it has been referred to as distance correlation based SIS (DC-SIS for short). Clearly, the DC-SIS provides a unified alternative to existing model-based sure screening procedures. If (U, V) follows a bivariate normal distribution, then $\mathrm{dcorr}(U, V)$ is a strictly monotone increasing function of the Pearson correlation between U and V. Thus, the DC-SIS is asymptotically equivalent to the SIS proposed in [21] for normal linear models. The DC-SIS is model-free since it does not require to impose a specific model structure on the relationship between the response and the predictors. From the definition of the distance correlation, the DC-SIS can be directly employed for screening grouped variables, and it can be directly utilized for ultrahigh-dimensional data with multivariate responses. Feature screening for multivariate responses and/or grouped predictors is of great interest in pathway analysis [8].

As discussed in Subsection 3.2, there exist some predictors that are marginally uncorrelated with or independent of, but jointly are not independent of the response. Similar to marginal screening procedures introduced in Sections 2 and 3, the SIRS and DC-SIS both will fail to select such predictors. Thus, an iterative screening procedure may be considered. Since both SIRS and DC-SIS do not rely on the regression function (i.e., the conditional mean of the response given the predictor). Thus, the iterative procedure introduced in Subsection 3.2 cannot be applied for SIRS and the DC-SIS because it may be difficult to obtain the residuals. Motivated by the idea of partial residuals, one may apply the SIRS and DC-SIS to the residual of the predictors on the selected predictor and the response. This strategy was used to construct an iterative procedure for the SIRS by [57]. Zhong and Zhu [56] applied this strategy to the DC-SIS and developed a general iterative procedure for the DC-SIS, which consists of three following steps:

T1. Apply the DC-SIS for $\boldsymbol{y}$ and $\boldsymbol{x}$ and select p_1 predictors, which are denoted by $\boldsymbol{x}_{\mathcal{M}_1} = \{X_1^{(1)}, \ldots, X_{p_1}^{(1)}\}$, where $p_1 < d$, where d is user-specified model size (for example, $d = [n/\log n]$). Let $\widehat{\mathcal{M}} = \mathcal{M}_1$.

T2. Denote by $\boldsymbol{X}_1$ and $\boldsymbol{X}_1^c$ the corresponding design matrix of variables in $\widehat{\mathcal{M}}$ and $\widehat{\mathcal{M}}^c$, respectively. Define $\boldsymbol{X}_{\mathrm{new}} = \{\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\}\boldsymbol{X}_1^c$. Apply the DC-SIS for $\boldsymbol{y}$ and $\boldsymbol{x}_{\mathrm{new}}$ and select p_2 predictors $\boldsymbol{x}_{\mathcal{M}_2} = \{X_1^{(2)}, \ldots, X_{p_2}^{(2)}\}$. Update $\widehat{\mathcal{M}} = \widehat{\mathcal{M}} \cup \mathcal{M}_2$.

T3. Repeat Step T2 until the total number of selected predictors reaches d. The final model we select is $\widehat{\mathcal{M}}$.

5.3 Feature screening for high-dimensional categorial data

Analysis of high-dimensional categorical data are common in scientific researches. For example, it is of interest to identify which genes are associated with a certain types of tumors. The types of tumors are categorical and may be binary if researchers are concerned with only case and control. This is a typical example of categorical responses. Genetic markers such as SNPs are categorical covariates. 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. Even when the covariance matrix is known, Bickel and Levina [3] showed that the Fisher discriminant analysis performs poorly in a minimax sense due to the diverging spectra. Fan and Fan [15] demonstrated further that almost all linear discriminants can perform as poorly as the random guessing. To this end, it is important to choose a subset of important features for high-dimensional classification and discriminant analysis. In this section, we review some recent work on feature screening for ultrahigh-dimensional categorical data.

Let Y be a categorical response with K classes $\{y_1, y_2, \ldots, y_K\}$. If an individual covariate X_j is associated with the response Y, then $\mu_{jk} = E(X_j \,|\, Y = y_k)$ are likely different from the population mean $\mu_j = E(X_j)$. Thus, it is intuitive to use the test statistic for multi-sample mean problem as a marginal utility for feature screening. Based on this idea, Tibshirani et al. [49] proposed a nearest shrunken centroid approach to cancer class prediction from gene expression data. Fan and Fan [15] 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 the important features with probability 1 under some technical conditions.

Although the variable screening based on two-sample t-statistic [15] 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 [41] proposed a new feature screening method for binary classification based on the Kolmogorov-Smirnov statistic. For ease of notation, relabel $Y = +1, -1$ be the class label. Let $F_{+j}(x)$ and $F_{-j}(x)$ denote the conditional cumulative distribution function of X_j given $Y = +1, -1$, respectively. Thus, if X_j and Y are independent, then $F_{+j}(x) \equiv F_{-j}(x)$. Thus, nonparametric test of independence may be used to construct a marginal utility for feature screening. Mai and Zou [41] proposed the following Kolmogorov-Smirnov statistic to be a marginal utility for feature screening,

$$\omega_j = \sup_{x \in \mathbb{R}} |F_{+j}(x) - F_{-j}(x)|, \tag{5.9}$$

which can be estimated by $\omega_{nj} = \sup_{x \in \mathbb{R}} |\widehat{F}_{+j}(x) - \widehat{F}_{-j}(x)|$, where $\widehat{F}_{+j}(x)$ and $\widehat{F}_{-j}(x)$ are the corresponding empirical conditional cumulative distribution functions. Mai and Zou [41] named this feature screening method as the Kolmogorov filter which sets the selected subset as

$$\widehat{\mathcal{M}} = \{1 \leqslant j \leqslant p : K_{nj} \text{ ranks among the first } d_n \text{ largest of all}\}. \tag{5.10}$$

Under the conditions that

$$\min_{j \in \mathcal{M}_*} \{K_j\} - \max_{j \in \mathcal{M}_*^c} \{K_j\} > (\log p/n)^{1/2}$$

and $|\mathcal{M}_*| < d_n$, Mai and Zou [41] showed that the sure screening property of the Kolmogorov filter holds with probability approaching 1, i.e., $P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}) \to 1$ as $n \to \infty$.

The Kolmogorov filter proposed by Mai and Zou [41] based on the Kolmogorov-Smirnov statistic is model-free and robust to heavy-tailed distributions of predictors and the presence of potential outliers. However, it is limited to the binary classification. To achieve broader applications, Cui et al. [9] proposed a new sure independence screening using mean variance index for ultrahigh-dimensional discriminant analysis. It not only retains the advantages of the Kolmogorov filter, but also allows the categorical response having a diverging number of classes in the order of $O(n^\kappa)$ with some $\kappa \geqslant 0$.

To emphasize the number of classes being allowed to diverge as the sample size n grows, we use K_n for K. Denote by $F_j(x) = \mathbb{P}(X_j \leqslant x)$ the unconditional distribution function of the j-th feature X_j and $F_{jk}(x) = \mathbb{P}(X_j \leqslant x \,|\, Y = y_k)$ the conditional distribution function of X_j given $Y = y_k$. If X_j and Y are statistically independent, then $F_{jk}(x) = F_j(x)$ for any k and x. This motivates one to consider the following marginal utility

$$MV(X_j \,|\, Y) = E_{X_j}[\mathrm{Var}_Y(F(X_j \,|\, Y))]. \tag{5.11}$$

This index is named as the mean variance index of X_j given Y. Cui et al. [9] showed that

$$MV(X_j \,|\, Y) = \sum_{k=1}^{K_n} p_k \int [F_{jk}(x) - F_j(x)]^2 dF_j(x),$$

which a weighted average of Cramér-von Mises distances between the conditional distribution function

of X_j given $Y = y_k$ and the unconditional distribution function of X_j. They further showed that $MV(X_j \mid Y) = 0$ if and only if X_j and Y are statistically independent. Thus, the proposal of [9] essentially is to employ a test of independence statistic as a marginal utility of feature screening for ultrahigh-dimensional categorical data.

Let $\{(X_{ij}, Y_i) : 1 \leqslant i \leqslant n\}$ be a random sample of size n from the population (X_j, Y). Then, $MV(X_j \mid Y)$ can be estimated by its sample counterpart

$$\widehat{MV}(X_j \mid Y) = \frac{1}{n} \sum_{k=1}^{K_n} \sum_{i=1}^{n} \hat{p}_k [\hat{F}_{jk}(X_{ij}) - \hat{F}_j(X_{ij})]^2, \tag{5.12}$$

where $\hat{p}_k = n^{-1} \sum_{i=1}^{n} I\{Y_i = y_k\}$, $\hat{F}_{jk}(x) = n^{-1} \sum_{i=1}^{n} I\{X_{ij} \leqslant x, Y_i = y_k\}/\hat{p}_k$, $\hat{F}_j(x) = n^{-1} \sum_{i=1}^{n} I\{X_{ij} \leqslant x\}$ and $I\{\cdot\}$ denotes the indicator function.

Without specifying any classification model, Cui et al. [9] defined the important feature subset by

$$\mathcal{M}_* = \{j : F(y \mid \boldsymbol{x}) \ \text{ functionally depends on } \ X_j \ \text{ for some } \ y = y_k\}.$$

$\widehat{MV}(X_j \mid Y)$ is used to rank the importance of all features and the subset $\widehat{\mathcal{M}} = \{j : \widehat{MV}(X_j \mid Y) \geqslant cn^{-\tau}, \text{ for } 1 \leqslant j \leqslant p\}$ is selected. This procedure is referred to as the MV-based sure independence screening (MV-SIS). Assume that (1) For some positive constants c_1 and c_2,

$$c_1/K_n \leqslant \min_{1 \leqslant k \leqslant K_n} p_k \leqslant \max_{1 \leqslant k \leqslant K_n} p_k \leqslant c_2/K_n,$$

and $K_n = O(n^{\kappa})$ for $\kappa \geqslant 0$; (2) For some positive constants $c > 0$ and $0 \leqslant \tau < 1/2$,

$$\min_{j \in \mathcal{M}_*} MV(X_j \mid Y) \geqslant 2cn^{-\tau}.$$

Under the above conditions, Cui et al. [9] proved that

$$P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}) \geqslant 1 - O(s \exp\{-bn^{1-(2\tau+\kappa)} + (1+\kappa)\log n\}), \tag{5.13}$$

where b is a positive constant and $s = |\mathcal{M}_*|$. As $n \to \infty$, $P(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}) \to 1$ for nonpolynomial (NP)-dimensionality problem $\log p = O(n^{\alpha})$, where $\alpha < 1 - 2\tau - \kappa$ with $0 \leqslant \tau < 1/2$ and $0 \leqslant \kappa < 1 - 2\tau$. Thus, the sure screening property of the MV-SIS holds.

It is worth pointing out that the MV-SIS is directly applicable for the setting in which the response is continuous, but the feature variables are categorical (e.g., the covariates are SNP). To implement the MV-SIS, one can simply use $MV(Y \mid X_j)$ as the marginal utility for feature screening. When both response and feature variables are categorical, it is not difficult to use a test of independence statistic as marginal utility for feature screening. Huang et al. [33] employed the Pearson χ^2-test statistic for independence as a marginal utility for feature screening. They further established the sure screening procedure of their screening procedure under mild conditions.

6 Concluding remarks and future research

In this paper, we have provided a selective overview on feature screening for ultrahigh-dimensional data. We focus on the settings in which the observations are independent and identically distribution random samples from a populations. We briefly described a variety of feature screening procedures for linear models, generalized linear models, nonparametric regression models and several model-free feature screening procedures. Below we outline several topics for future research.

Feature screening has been intensively studied during the last decade. Thus, it is frequent that several screening procedures are available for a specific models. For example, the SIS [21], RRCS, SIRS and DC-

SIS can be used for feature screening under linear regression models. Under a set of certain conditions, they all enjoy sure screening property. A natural question raised here is how to integrate the results from these different screening procedures together to reduce false positive rate and improve the accuracy of the results. This is certainly an interesting research topic to investigate. To our best knowledge, there is no existing work to address such problems.

It is of great interest to investigate gene×gene interactions in many medical studies, and it is of primary interest to identify gene×environment interaction in social behavioral science researches. Thus, it is important to construct effective feature screening procedures for interactions. This is conceptual simple, but is computational challenging because the dimensionality becomes much higher than the number of collected variables. Furthermore, the variables representing the interactions likely are highly correlated. This may break down the imposed conditions to derive the sure screening property. Li et al. [36] proposed a fast two-stage algorithm to screening interactions in linear models under the weak heredity assumption that if an interaction is significant, then at least one of the main effects are significant. Hao and Zhang [29] developed a forward-selection based procedure to identify interaction effects in a greedy forward fashion under the strong heredity condition. More research are welcome to contribute this research topic.

Longitudinal studies have been conducted in many research fields. Many longitudinal studies collect thousands of variables such as a huge number of genetic markers, while the number of subjects typically is in the order of hundreds. Thus, feature screening is a necessary step before conducting confirmatory statistical analysis. Xu et al. [53] proposed a feature screening procedure for longitudinal data based on marginal generalized estimating equation methods and Song et al. [45] extended the nonparametric independence screening procedure for time-varying coefficient model with longitudinal data. For longitudinal data analysis, it is of importance to incorporate the within-subject correlation and heteroscedasticity to improve existing procedure by reducing false negative rate. Thus, more research to construct effective feature screening procedure is of crucial importance.

Time-to-event responses are frequently collected in many fields. For example, the survival time in a cancer research, the time to relapse in a smoking cession study, and the time to default in a loan of housing or a card are typically time to event response. It becomes more and more common to collect a huge number of covariates such as millions of genetic markers in a cancer research along with some time-to-event responses or phenotypes. Thus, feature screening for survival data certainly is of great interest. Fan et al. [17] proposed a SIS procedure for the Cox model based on marginal partial likelihood estimate, and this procedure was further studied by Zhao and Li [55]. It is clear that more researches on feature screening for time-to-event data are needed to develop in future.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 11401497 and 11301435), the Fundamental Research Funds for the Central Universities (Grant No. T2013221043), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, the Fundamental Research Funds for the Central Universities (Grant No. 20720140034), National Institute on Drug Abuse, National Institutes of Health (Grant Nos. P50 DA036107 and P50 DA039838), and National Science Foundation (Grant No. DMS1512422). The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institute on Drug Abuse, National Institutes of Health, National Science Foundation or National Natural Science Foundation of China. Runze Li sincerely thanks the editor-in-chief, Professor Ya-Xiang Yuan, for his kind invitation to write this article. The authors thank reviewers for their constructive comments, which led to significant improvement of this work. First two authors contributed equally to this work.

References

1 Benjamini Y, Hochberg T. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc Ser B, 1995, 57: 289–300

2 Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist, 2004, 29: 1165–1188

3 Bickel P J, Levina E. Some theory for Fishers linear discriminant function, "Naive Bayes," and some alternatives when

there are many more variables than observations. Bernoulli, 2004, 10: 989–1010
4 Breiman L. Better subset regression using the nonnegative garrote. Technometrics, 1995, 37: 373–384
5 Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). Ann Statist, 2007, 35: 2313–2404
6 Carroll R J, Fan J, Gijbels I, et al. Generalized partially linear single-index models. J Amer Statist Assoc, 1997, 92: 477–489
7 Chen J, Chen Z. Extended Bayesian information criterion for model selection with large model spaces. Biometrika, 2008, 85: 759–771
8 Chen L S, Paul D, Prentice R L, et al. A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. J Amer Statist Assoc, 2011, 106: 1345–1360
9 Cui H, Li R, Zhong W. Model-free feature screening for ultrahigh-dimensional discriminant analysis. J Amer Statist Assoc, 2015, 110: 630–641
10 Donoho D L. High-dimensional data: The curse and blessings of dimensionality. Los Angeles: Amer Math Soc Conference Math Challenges of 21st Century, 2000
11 Donoho D L. Neighborly polytopes and sparse solution of underdetermined linear equations. Technical Report, Department of Statistics. Stanford: Stanford University, 2004
12 Donoho D L. For most large undetermined systems of linear equations the minimal ℓ_1-norm solution is the sparsest solution. Commun Pure Appl Math, 2006, 59: 797–829
13 Dudoit S, Shaffer J P, Boldrick J C. Multiple hypothesis testing in microarray experiments. Stat Sci, 2003, 18: 71–103
14 Efron B. Correlation and large-scale simultaneous significance testing. J Amer Statist Assoc, 2007, 102: 93–103
15 Fan J, Fan Y. High-dimensional classification using features annealed independence rules. Ann Statist, 2008, 36: 2605–2637
16 Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high dimensional additive models. J Amer Statist Assoc, 2011, 106: 544–557
17 Fan J, Feng Y, Wu Y. Ultrahigh dimensional variable selection for Cox's proportional hazards model. IMS Collection, 2010, 6: 70–86
18 Fan J, Han F, Liu H. Challenges of Big Data analysis. Nat Sci Rev, 2014, 1: 293–314
19 Fan J, Li R. Variable selection via nonconcave penalized likelihood and it oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360
20 Fan J, Li R. Statistical challenges with high dimensionality: Feature selection in knowledge discoverry. In: Sanz-Sole M, Soria J, Varona J L, et al., eds. Proceedings on the International Congress of Mathematicians, vol. III. Freiburg: European Math Soc, 2006, 595–622
21 Fan J, Lv J. Sure independence screening for ultrahigh-dimensional feature space (with discussion). J Roy Stat Soc Ser B, 2008, 70: 849–911
22 Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Stat Sinica, 2010, 20: 101–148
23 Fan J, Ma Y, Dai W. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. J Amer Statist Assoc, 2014, 109: 1270–1284
24 Fan J, Ren Y. Statistical analysis of DNA microarray data. Clin Cancer Res, 2006, 12: 4469–4473
25 Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: Beyond the linear model. J Mach Learn Res, 2009, 10: 1829–1853
26 Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Statist, 2010, 38: 3567–3604
27 Fang K, Kotz S, Ng K. Symmetric Multivariate and Related Distributions. London: Chapman & Hall, 1990
28 Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. J Comput Graph Stat, 2009, 18: 533–550
29 Hao N, Zhang H H. Interaction Screening for Ultrahigh-Dimensional Data. J Amer Statist Assoc, 2014, 109: 1285–1301
30 Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. Ann Statist, 1993, 21: 157–178
31 Härdle W, Liang H, Gao J T. Partially Linear Models. Germany: Springer Phisica-Verlag, 2000
32 He X, Wang L, Hong H G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann Statist, 2013, 41: 342–369
33 Huang D, Li R, Wang H. Feature screening for ultrahigh-dimensional categorical data with applications. J Bus Econ Stat, 2014, 32: 237–244
34 Huber P J. Robust estimation of a location parameter. Ann Math Stat, 1964, 35: 73–101
35 Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Statist, 2012, 40: 1846–1877
36 Li J, Zhong W, Li R, et al. A fast algorithm for detecting gene-gene interactions in genome-wide association studies. Ann Appl Stat, 2014, 8: 2292–2318
37 Li R, Zhong W, Zhu L P. Feature screening via distance correlation learning. J Amer Statist Assoc, 2012, 107: 1129–1139

38 Lin L, Sun J, Zhu L X. Nonparametric feature screening. Comput Stat Data Anal, 2013, 67: 162–174
39 Liu J, Li R, Wu R. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. J Amer Statist Assoc, 2014, 109: 266–274
40 Luo X, Stefanski L A, Boos D D. Tuning variable selection procedure by adding noise. Technometrics, 2006, 48: 165–175
41 Mai Q, Zou H. The Kolmogorov filter for variable screening in high-dimensional binary classification. Biometrika, 2013, 100: 229–234
42 Mai Q, Zou H, Yuan M. A direct approach to sparse discriminant analysis in ultra-high dimensions. Biometrika, 2012, 99: 29–42
43 Meier I, Geer V, Bühlmann P. High-dimensional additive modeling. J Roy Stat Soc Ser B, 2009, 71: 1009–1030
44 Pan R, Wang H, Li R. On the ultrahigh-dimensional linear discriminant analysis problem with a diverging number of classes. J Amer Statist Assoc, 2015, in press
45 Song R, Yi F, Zou H. On varying-coefficient independence screening for high-dimensional varying-coefficient models. Stat Sinica, 2014, 24: 1735–1752
46 Storey J D, Tibshirani R. Statistical significance for genome-wide studies. Proc Natl Acad Sci USA, 2003, 100: 9440–9445
47 Székely G J, izzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. Ann Statist, 2007, 35: 2769–2794
48 Tibshirani R. Regression shrinkage and selection via lasso. J Roy Stat Soc Ser B, 1996, 58: 267–288
49 Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA, 2002, 99: 6567–6572
50 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995
51 Wang H. Forward regression for ultra-high dimensional variable screening. J Amer Statist Assoc, 2009, 104: 1512–1524
52 Xu C, Chen J. The sparse MLE for ultrahigh-dimensional feature screening. J Amer Statist Assoc, 2014, 109: 1257–1265
53 Xu P, Zhu L X, Li Y. Ultrahigh dimensional time course feature selection. Biometrics, 2014, 70: 356–365
54 Yuan M, Lin Y. On the nonnegative garrote estimator. J Roy Stat Soc Ser B, 2007, 69: 143–161
55 Zhao D S, Li Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. J Multivariate Anal, 2012, 105: 397–411
56 Zhong W, Zhu L P. An iterative approach to distance correlation based sure independence screening. J Stat Comput Sim, 2014, doi:10.108000949655.2014.928820
57 Zhu L P, Li L, Li R, et al. Model-free feature screening for ultrahigh-dimensional data. J Amer Statist Assoc, 2011, 106: 1464–1475