Variable Selection and Feature Screening

Wanjun Liu and Runze Li

Abstract This chapter provides a selective review on feature screening methods for ultra-high dimensional data. The main idea of feature screening is reducing the ultra-high dimensionality of the feature space to a moderate size in a fast and efficient way and meanwhile retaining all the important features in the reduced feature space. This is referred to as the sure screening property. After feature screening, more sophisticated methods can be applied to reduced feature space for further analysis such as parameter estimation and statistical inference. This chapter only focuses on the feature screening stage. From the perspective of different types of data, we review feature screening methods for independent and identically distributed data, longitudinal data and survival data. From the perspective of modeling, we review various models including linear model, generalized linear model, additive model, varying-coefficient model, Cox model, etc. We also cover some model-free feature screening procedures.

1 Introduction

With the advent of modern technology for data collection, ultra-high dimensional datasets are widely encountered in machine learning, statistics, genomics, medicine, finance, marketing, etc. For example, in biomedical studies, huge numbers of magnetic resonance images (MRI) and functional MRI data are collected for each subject. Financial data is also of a high dimensional nature. Hundreds or thousands of financial instruments can be measured and tracked over time at very fine time

Wanjun Liu
Department of Statistics, the Pennsylvania State University, University Park, PA 16802, e-mail: wxl204@psu.edu

Runze Li
Department of Statistics and The Methodology Center, the Pennsylvania State University, University Park, PA 16802, e-mail: rzli@psu.edu

intervals for use in high frequency trading. This ultra-high dimensionality causes challenges in both computation and methodology. Scalability is the major challenge to ultra-high dimensional data analysis. Many traditional methods that perform well for low dimensional data do not scale to ultra-high dimensional data. Other issues such as high collinearity, spurious correlation, and noise accumulation (Fan and Lv 2008, 2010) brings in additional challenges. Therefore, variable selection and feature screening have been a fundamental problem in the analysis of ultra-high dimensional data. For example, the issue of spurious correlation is illustrated by a simple example in Fan and Lv (2008). Suppose we have a $n \times p$ dataset with sample size n and the p predictors independently follow the standard normal distribution. When $p \gg n$, the maximum absolute value of sample correlation coefficient among predictors can be very large. Figure 1 shows the distributions of the maximum absolute sample correlation with $n = 60$ and $p = 1000, 5000$. Though the predictors are generated independently, some of them can be highly correlated due to high-dimensionality.

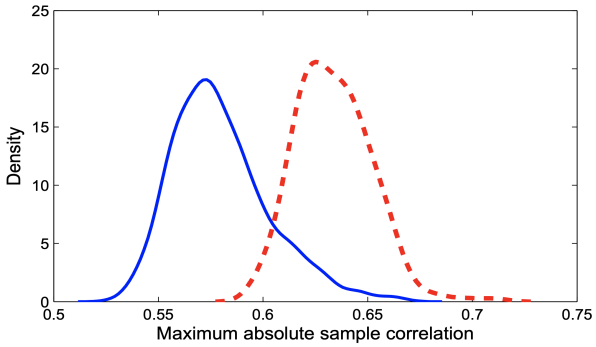

Fig. 1 Distributions of the maximum absolute sample correlation coefficient when $n = 60, p = 1000$ (solid curve) and $n = 60, p = 5000$ (dashed curve).

Over the past two decades, a large amount of variable selection approaches based on regularized M-estimation have been developed. These approaches include the Lasso (Tibshirani 1996), the SCAD (Fan and Li 2001), the Dantzig selector (Candes and Tao 2007), and the MCP (Zhang 2010), among others. However, these regularization methods may not perform well for ultra-high dimensional data due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability (Fan, Samworth and Wu 2009). To improve the statistical performance of regularization methods and reduce computational cost, a class of two stage approaches is proposed. In the first stage, we reduce the number of features from a very large scale to a moderate size in a computationally fast way. Then in the second stage, we further implement refined variable selection algorithms such as regularization methods to the features selected from the first stage. Ideally, we select all the important features and may allow a few unimportant features entering

our model in the first stage. The first stage is referred to as the feature screening stage. We will only focus on the feature screening stage in this chapter.

Suppose we have p features $X_1,\ldots,X_p$ in the feature space and denote the true index set of important variables by $\mathscr{M}_\star$. The definition of $\mathscr{M}_\star$ may vary across different models. For example, in a parametric model associated with true parameters $\beta^\star = (\beta_1^\star,\ldots,\beta_p^\star)^\top$, $\mathscr{M}_\star$ is typically defined to be

$$\mathscr{M}_\star = \{1 \le j \le p : \beta_j^\star \neq 0\}.$$

Our goal in the feature screening stage is to select a submodel $\widehat{\mathscr{M}} \subset \{1,\ldots,p\}$ with little computational cost such that $\mathscr{M}_\star \subset \widehat{\mathscr{M}}$ with high probability. This is referred to as the sure screening property.

Definition 1 (Sure Screening). Let $\mathscr{M}_\star$ be the true index set of important features and $\widehat{\mathscr{M}}$ be the index set of selected important variables by some feature screening procedure based on a sample of size n, then this feature screening procedure has the sure screening property if

$$\Pr(\mathscr{M}_\star \subset \widehat{\mathscr{M}}) \to 1 \text{ as } n \to \infty.$$

The sure screening property ensures that all the important features will be included in the selected submodel with probability approaching to 1 as the sample size goes to infinity. A trivial but less interesting choice of $\widehat{\mathscr{M}}$ is $\widehat{\mathscr{M}} = \{1,\ldots,p\}$, which always satisfies the definition of sure screening. Here we assume the number of true important features is much smaller than p. This kind of assumption is also known as sparsity assumption in the sense that most of the entries in the true parameter $\beta^\star$ are zero. Of interest is to find a $\widehat{\mathscr{M}}$ whose cardinality is much smaller than p (i.e., $|\widehat{\mathscr{M}}| \ll p$) and meanwhile the sure screening holds.

2 Marginal, Iterative and Joint Feature Screening

2.1 Marginal feature screening

The most popular feature screening method is the marginal feature screening, which ranks the importance of features based on marginal utility and thus is computationally attractive. More specifically, the marginal feature screening procedure assigns an index, say $\widehat{\omega}_j$, to the feature X_j for $j = 1,\ldots,p$. This index $\widehat{\omega}_j$ measures the dependence between the jth feature and the response variable. Then we can rank the importance of all features according to $\widehat{\omega}_j$ and include the features ranked on the top in the submodel. For example, in the setting of linear regression, the index $\widehat{\omega}_j$ is chosen to be the absolute value of marginal Pearson correlation between the jth feature and the response (Fan and Lv 2008). Features with larger absolute values of $\widehat{\omega}_j$ are more relevant to the response and thus are ranked on the top. As a result, we

include the top d_n features in the submodel,

$$\widehat{\mathscr{M}}_{d_n} = \{1 \leq j \leq p : \widehat{\omega}_j \text{ is among the top } d_n \text{ ones}\},$$

where d_n is some pre-specified threshold. Note that the marginal feature screening procedure only uses the information of jth feature and the response without looking at all other features and thus it can be carried out in a very efficient way. A large amount of literature have studied the sure screening property of various marginal feature screening methods, see Fan and Lv (2008), Fan, Samworth and Wu (2009), Fan, Feng and Song (2011), Li, Zhong and Zhu (2012), Fan, Ma and Dai (2014).

2.2 Iterative feature screening

As pointed out in Fan and Lv (2008), the marginal feature screening procedure may suffer from the following two issues:

1. Some unimportant features that are highly correlated with important features can have higher rankings than other important features that are relatively weakly related to the response.
2. An important feature that is marginally independent but jointly dependent on the response tends to have lower ranking.

The first issue says that the marginal feature screening has chance to include some unimportant features in the submodel. This is not a big issue for the purpose of feature screening. The second one is a bigger issue, which indicates that the marginal feature screening may fail to include all the important feature if it is marginally independent of the response. Absence of any important feature may lead to a biased estimation. To overcome the two aforementioned issues, one can apply an iterative feature screening procedure by iteratively carrying out the marginal screening procedure. This iterative procedure was first introduced by Fan and Lv (2008) and can be viewed as a natural extension of the marginal feature screening. At the kth iteration, we apply marginal feature screening to the features survived from the previous step and is typically followed by a regularization methods if a regression model is specified. Let $\widehat{\mathscr{M}_k}$ be the selected index set of important variables at the kth iteration and the final selected index set of important variables is given by $\widehat{\mathscr{M}} = \widehat{\mathscr{M}_1} \cup \widehat{\mathscr{M}_2} \cup \ldots$, the union of all selected index sets. For example, Fan and Lv (2008) uses the residuals computed from linear regression as the new response and iteratively applies marginal feature screening based on Pearson correlation. The iterative feature screening can significantly improve the simple marginal screening, but it can also be much more computationally expensive.

2.3 Joint feature screening

Another approach to improve the marginal screening is known as the joint screening screening (Xu and Chen 2014, Yang, Yu, Li and Buu 2016). Many regularization methods involves solving a optimization problem of the following form,

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \beta) \quad \text{subject to } \|\beta\|_0 \leq k, \tag{1}$$

where $\ell(\cdot,\cdot)$ is some loss function of negative log-likelihood function. It is quite challenging to solve the minimization problem in (1) especially in the ultra-high dimensional setting. The joint screening approach approximates the objective function by its Taylor's expansion and replaces the possibly singular Hessian matrix with some invertible matrix. After the approximation, one can solve such optimization problem iteratively in a fast manner. In many applications, one can obtain a closed form at each iteration for the joint screening approach.

2.4 Notations and organization

We introduce some notations used this chapter. Let $Y \in \mathbb{R}$ be the univariate response variable and $\mathbf{x} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ be the p-dimensional features. We observe a sample $\{(\mathbf{x}_i, Y_i)\}, i = 1, \ldots, n$ from the population $(\mathbf{x}, Y)$ with $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})^\top$. Let $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$ be the response vector and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ be the design matrix. We use $\mathbf{x}_{(j)}$ to denote the jth column of $\mathbf{X}$ and use $\mathbf{1}(\cdot)$ to denote the indicator function. For a vector $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, $\|\beta\|_q = (\sum_{j=1}^p |\beta_j|^q)^{1/q}$ denotes its ℓ_q norm for $0 \leq q \leq \infty$. In particular, $\|\beta\|_0 = \sum_{j=1}^p \mathbf{1}(|\beta_j| \neq 0)$ is the number of non-zero elements in β and $\|\beta\|_\infty = \max_{1 \leq j \leq p} |\beta_j|$. For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{p \times p}$, we use $\|\mathbf{M}\|_F$ and $\|\mathbf{M}\|_\infty$ to denote the Frobenius norm and supremum norm respectively. Let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be the smallest and largest eigenvalue of $\mathbf{M}$. Let $\mathscr{M}$ be a subset of $\{1, \ldots, p\}$ and $\beta_{\mathscr{M}}$, a sub-vector of β, consists of β_j for all $j \in \mathscr{M}$. We use $\mathscr{M}_\star$ to denote the true index set of important features and $\beta^\star = (\beta_1^\star, \ldots, \beta_p^\star)^\top$ denote the true parameter. We assume $|\mathscr{M}_\star| = s$ throughout this chapter, where $|\mathscr{M}_\star|$ denotes the cardinality of the set $\mathscr{M}_\star$.

In the rest of this chapter, we spend most of the efforts reviewing the marginal feature screening methods as the marginal feature screening is the most popular screening method. The iterative feature screening can be viewed as a natural extension of marginal feature screening. We will discuss the details on the iterative and joint screening methods in one or two particular examples.

The rest of this chapter is organized as follows. In section 3, we introduce the feature screening methods for independent and identically distributed data, which is the most common assumption in statistical modeling. Many different models have been developed for such data, including linear model, generalized linear model,

additive model, varying-coefficient model, etc. However, this assumption is usually violated in areas such as finance and economics. In section 4, we review the feature screening methods that are developed for longitudinal data, that is, data is collected over a period of time for each subject. In section 5, we review the feature screening methods for survival data, which is widely seen in reliability analysis in engineering, duration analysis in economics, and event history analysis in sociology, etc.

3 Independent and Identically Distributed Data

Independent and identically distributed (IID) data is the most common assumption in statistical literature and a large amount of feature screening methods have been developed for IID data. In this section, we review some of the widely used feature screening methods for such data. Throughout this section, we assume that $\{(\mathbf{x}_i, Y_i)\}, i = 1, \ldots, n$ is a random sample from the population $(\mathbf{x}, Y)$.

3.1 Linear model

Let us consider the linear regression model,

$$Y = \beta_0 + \mathbf{x}^\top \beta + \varepsilon, \tag{2}$$

where β_0 is the intercept, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional regression coefficient vector, and ε is the error term. In the ultra-high dimensional setting, the true regression coefficient vector $\beta^\star = (\beta_1^\star, \ldots, \beta_p^\star)^\top$ is assumed to be sparse, meaning that most of the coefficients $\beta_j^\star$ are 0. The true index set of the model is defined as

$$\mathscr{M}_\star = \{1 \le j \le p : \beta_j^\star \neq 0\}.$$

We call the features with indices in the set $\mathscr{M}_\star$ important features. Fan and Lv (2008) suggests ranking all features according to the marginal Pearson correlation coefficient between individual feature and the response and select the top features which have strong correlation with the response as important features. For a pre-specified value $\nu_n (0 < \nu_n < 1)$, the index set of selected features is given by

$$\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : |\widehat{\mathrm{corr}}(\mathbf{x}_{(j)}, \mathbf{y})| \text{ is among the top } \lfloor \nu_n n \rfloor \text{ largest ones}\},$$

where $\mathbf{x}_{(j)}$ is the jth column of $\mathbf{X}$, $\widehat{\mathrm{corr}}$ denotes the sample Pearson correlation, and $\lfloor \nu_n n \rfloor$ is the integer part of $\nu_n n$. This procedure achieves the goal of feature screening since it reduces the ultra-high dimensionality down to a relatively moderate scale $\lfloor \nu_n n \rfloor$. This procedure is referred to as the sure independence screening (SIS). Then appropriate regularization methods such as Lasso, SCAD and Dantzig selector can be further applied to the selected important features. The corresponding methods

are referred to as SIS-LASSO, SIS-SCAD and SIS-DS. This feature screening procedure is based on Pearson correlation and can be carried out in a extremely simple way at very low computational cost. In addition to the computational advantage, this SIS enjoys the sure screening property. Assume that the error is normally distributed and the following conditions hold,

(A1) $\min_{j\in\mathscr{M}_\star}\beta_j^\star \geq c_1 n^{-\kappa}$ and $\min_{j\in\mathscr{M}_\star}|\mathrm{cov}(\beta_j^{\star-1}Y,X_j)| \geq c_2$, for some $\kappa > 0$ and $c_1, c_2 > 0$.

(A2) There exists $\tau \geq 0$ and $c_3 > 0$ such that $\lambda_{\max}(\Sigma) \leq c_3 n^\tau$, where $\Sigma = \mathrm{cov}(\mathbf{x})$ is the covariance matrix of $\mathbf{x}$ and $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ.

(A3) $p > n$ and $\log p = O(n^\xi)$ for some $\xi \in (0, 1-2\kappa)$.

Fan and Lv (2008) showed that if $2\kappa+\tau < 1$, then with the choice of $\nu_n = cn^{-\theta}$ for some $0 < \theta < 1-2\kappa-\tau$ and $c > 0$, we have for some $C > 0$

$$\Pr(\mathscr{M}_\star \subset \widehat{\mathscr{M}_{\nu_n}}) \geq 1 - O(\exp\{-Cn^{1-2\kappa}/\log n\}). \tag{3}$$

Conditions (A1) requires certain order of minimal signal among the important features, condition (A2) rules out the case of strong collinearity and condition (A3) allows p grows exponentially with sample size n. Equation (3) shows that the SIS can reduce the exponentially growing dimension p down to a relatively small scale $d_n = \lfloor \nu_n n \rfloor = O(n^{1-\theta}) < n$, while include all important features in the submodel with high probability. The optimal choice of d_n relies on unknown parameters. It is common to assume $s/n \to 0$ where s is the number of important features. In practice, one can conservatively set $d_n = n-1$ or require $d_n/n \to 0$ with $d_n = n/\log n$. See more details in Fan and Lv (2008).

Marginal Pearson correlation is employed to rank the importance of features, SIS may suffer from the potential issues with marginal screening. On one hand, SIS may fail to select the important feature when it is jointly correlated but marginally uncorrelated with the response. On the other hand, the SIS tends to select unimportant features which are jointly uncorrelated but highly marginally correlated with the response. To address these issues, Fan and Lv (2008) also introduced an iterative SIS procedure (ISIS) by iteratively replacing the response with the residuals obtained from the linear regression using the selected features from the previous step. The ISIS works as follows. In the first iteration, we select a subset of k_1 features $\mathscr{A}_1 = \{X_{i_1}, \ldots, X_{i_{k_1}}\}$ using an SIS based model selection method such as SIS-LASSO or SIS-SCAD. Then we regress the response Y over the selected features $\mathscr{A}_1$ and obtain the residuals. We treat the residuals as the new responses and apply the same method to the remaining $k_2 = p - k_1$ features $\mathscr{A}_2 = \{X_{j_1}, \ldots, X_{j_{k_2}}\}$. We keep doing this until we get l disjoint subsets $\mathscr{A}_1, \ldots, \mathscr{A}_l$ such that $d = \sum_{i=1}^{l} |\mathscr{A}_i| < n$. We use the union $\mathscr{A} = \cup_{i=1}^{\ell} \mathscr{A}_i$ as the set of selected features. In practical implementation, we can choose, for example, the largest l such that $|\mathscr{A}| < n$. This iterative procedure makes those important features that are missed in the previous step possible to re-enter the selected model. In fact, after features in $\mathscr{A}_1$ entering into the model, those that are marginally weakly correlated with Y purely due to the presence of variables in $\mathscr{A}_1$ should now be correlated with the residuals.

3.2 Generalized linear model and beyond

A natural extension of SIS is applying the feature screening procedure to generalized linear models. Assume that the response Y is from an exponential family with the following canonical form

$$f_Y(y,\theta) = \exp\{y\theta - b(\theta) + c(y)\},$$

for some known functions $b(\cdot)$, $c(\cdot)$ and unknown parameter θ. Consider the following generalized linear model

$$E(Y|\mathbf{x}) = g^{-1}(\beta_0 + \mathbf{x}^\top \beta), \tag{4}$$

where $g(\cdot)$ is the link function, β_0 is a unknown scalar, and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional unknown vector. The linear regression model in (2) is just a special case of (4) by taking $g(\mu) = \mu$. Without loss of generality, we assume that all the features are standardized to have mean zero and standard deviation one. Fan and Song (2010) proposes a feature screening procedure for (4) by ranking the maximum marginal likelihood estimator (MMLE). For each $1 \le j \le p$, the MMLE $\widehat{\beta}_j^M$ is a 2-dimensional vector and defined as

$$\widehat{\beta}_j^M = (\widehat{\beta}_{j0}^M, \widehat{\beta}_{j1}^M)^\top = \arg\min_{\beta_{j0},\beta_{j1}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \beta_{j0} + \beta_{j1} X_{ij}), \tag{5}$$

where $\ell(y,\theta) = -y\theta + b(\theta) - c(y)$ is the negative log-likelihood function. The minimization problem in (5) can be rapidly computed and its implementation is robust since it only involves two parameters. Such a feature screening procedure ranks the importance of features according to their magnitude of marginal regression coefficients. The set of important features is defined as

$$\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : |\widehat{\beta}_{j1}^M| > \nu_n\},$$

where ν_n is some pre-specified threshold. As a result, we dramatically decrease the dimension from p to a moderate size by choosing a large ν_n and hence the computation is much more feasible after screening. Although the interpretations and implications of the marginal models are biased from the full model, it is suitable for the purpose of variable screening. In the linear regression setting, the MMLE ranking is equivalent to the marginal correlation ranking. However, the MMLE screening does not rely on the normality assumption and can be more easily to applied to other models. Under proper regularity conditions, Fan and Song (2010) established the sure screening property of the MMLE ranking. By taking $\nu_n = cn^{1-2\kappa}$ for some $0 < \kappa < 1/2$ and $c > 0$, we have

$$\Pr(\mathscr{M}_\star \subset \widehat{\mathscr{M}}_{\nu_n}) \to 1 \quad \text{as } n \to \infty.$$

See more about the details about the conditions in Fan and Song (2010). This MMLE procedure can handle the NP-dimensionality of order

$$\log p = o(n^{(1-2\kappa)\alpha/(\alpha+2)}),$$

where α is some positive parameter that characterizes the how fast the tail of distribution of features decay. For instance, $\alpha = 2$ corresponds to normal features and $\alpha = \infty$ corresponds to features that are bounded. When features are normal ($\alpha = 2$), the MMLE gives a weaker result than that of the SIS which permits $\log p = o(n^{1-2\kappa})$. However, MMLE allows non-normal features and other error distributions.

Fan, Samworth and Wu (2009) studied a very general pseudo-likelihood framework in which the aim is to find the parameter vector $\beta = (\beta_1, \ldots, \beta_p)^\top$ that is sparse and minimizes an objective function of the form

$$Q(\beta_0, \beta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta_0 + \beta^\top \mathbf{x}_i), \tag{6}$$

where the function $\ell(\cdot, \cdot)$ can be some loss function or negative log-likelihood function. This formulation in (6) includes a lot of important statistical models including

1. **Generalized linear models**: All generalized linear models, including logistic regression and Poisson log-linear models, fit very naturally into the framework.
2. **Classification**: Some common approaches to classification assume the response takes values in $\{-1, 1\}$ also fit the framework. For instance, support vector machine (Vapnik 2013) uses the hinge loss function $\ell(Y_i, \beta_0 + \mathbf{x}_i^\top \beta) = (1 - Y_i(\beta_0 + \mathbf{x}_i^\top \beta))_+$, while the boosting algorithm AdaBoost (Freund and Schapire 1997) uses $\ell(Y_i, \beta_0 + \mathbf{x}_i^\top \beta) = \exp\{-Y_i(\beta_0 + \mathbf{x}_i^\top \beta)\}$.
3. **Robust fitting**: Instead of the conventional least squares loss function, one may prefer a robust loss function such as the ℓ_1 loss $\ell(Y_i, \beta_0 + \mathbf{x}_i^\top \beta) = |Y_i - \beta_0 - \mathbf{x}_i^\top \beta|$ or the Huber loss (Huber 1964), which also fits into the framework.

Fan, Samworth and Wu (2009) suggests to rank the importance of features according to their marginal contributions to the magnitude of the likelihood function. This method can be viewed as a marginal likelihood ratio screening, as it builds on the increments of the log-likelihood. The marginal utility of the jth feature X_j is quantified by

$$L_j = \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^{n} \ell(Y_i, \beta_0 + X_{ij}\beta_j).$$

The idea is to compute the vector of marginal utilities $\mathbf{L} = (L_1, \ldots, L_p)^\top$ and rank the features according to the marginal utilities: the smaller L_j is, the more important X_j is. Note that in order to compute L_j , we only need to fit a model with two parameters, β_0 and β_j, so computing the vector $\mathbf{L}$ can be done very quickly and stably, even for an ultrahigh dimensional problem. The feature X_j is selected if the corresponding utility L_j is among the d_n smallest components of $\mathbf{L}$. Typically, we may take $d_n = \lfloor n/\log n \rfloor$. When d_n is large enough, it has high probability of select-

ing all of the important features. The marginal likelihood screening and the MMLE screening share a common computation procedure as both procedures solve p optimization problems over a two-dimensional parameter space. Fan and Song (2010) showed that these two procedures are actually equivalent in the sense that they both possess the sure screening property and that the number of selected variables of the two methods are of the same order of magnitude.

Fan, Samworth and Wu (2009) also proposes an iterative feature screening procedure, which consists of the following steps.

Step 1. Compute the vector of marginal utilities $\mathbf{L} = (L_1, \ldots, L_p)^\top$ and select the set $\widehat{\mathscr{A}_1} = \{1 \le j \le p : L_j \text{ is among the first } k_1 \text{ smallest ones}\}$. Then apply a penalized (pseudo)-likelihood, such as Lasso and SCAD, to select a subset $\widehat{\mathscr{M}}$.

Step 2. For each $j \in \{1, \ldots, p\}/\widehat{\mathscr{M}}$, compute

$$L_j^{(2)} = \min_{\beta_0, \beta_j, \beta_{\widehat{\mathscr{M}}}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{i,\widehat{\mathscr{M}}}^\top \beta_{\widehat{\mathscr{M}}} + X_{ij}\beta_j), \tag{7}$$

where $\mathbf{x}_{i,\widehat{\mathscr{M}}}$ denotes the sub-vector of $\mathbf{x}_i$ consisting of those elements in $\widehat{\mathscr{M}}$. Then select the set

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

Step 3. Use penalized likelihood to the features in set $\widehat{\mathscr{M}} \cup \widehat{\mathscr{A}_2}$,

$$\widehat{\beta}_2 = \underset{\beta_0, \beta_{\widehat{\mathscr{A}_2}}, \beta_{\widehat{\mathscr{M}}}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta_0 + \mathbf{x}_{i,\widehat{\mathscr{M}}}^\top \beta_{\widehat{\mathscr{M}}} + \mathbf{x}_{i,\widehat{\mathscr{A}_2}}^\top \beta_{\widehat{\mathscr{A}_2}}) + \sum_{j \in \widehat{\mathscr{M}} \cup \widehat{\mathscr{A}_2}} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is some penalty function such as Lasso or SCAD. The indices of $\widehat{\beta}_2$ that are non-zero yield a new estimated set $\widehat{\mathscr{M}}$.

Step 4. Repeat Step 2 and Step 3 and stop once $|\widehat{\mathscr{M}}| \ge d_n$.

Note that $L_j^{(2)}$ can be interpreted as the additional contribution of feature X_j given the presence of features in $\widehat{\mathscr{M}}$. The optimization problem in Step 2 is a low-dimensional problem which can be solved efficiently. An alternative approach in Step 2 is to substitute the fitted value $\widehat{\beta}_{\widehat{\mathscr{M}}_1}$ from the Step 1 into (7). Then the optimization in (7) only involves two parameters and is exactly an extension of Fan and Lv (2008). To see this, let $r_i = Y_i - \mathbf{x}_{i,\widehat{\mathscr{M}}}^\top \beta_{\widehat{\mathscr{M}}}$ denote the residual from the previous step and we choose the square loss function, then

$$\ell(Y_i, \beta_0 + \mathbf{x}_{i,\widehat{\mathscr{M}}}^\top \beta_{\widehat{\mathscr{M}}} + X_{ij}\beta_j) = (r_i - \beta_0 - \beta_j X_{ij})^2.$$

Without explicit definition of residuals, the idea of considering additional contribution to the response can be applied to a much more general framework.

3.3 Nonparametric regression models

Fan, Feng and Song (2011) proposes a nonparametric independence screening (NIS) for ultra-high dimensional additive model of the following form,

$$Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon, \tag{8}$$

where $m_j(X_j)$ is assumed to have mean zero for identifiability. The true index set of important features is defined as

$$\mathscr{M}_\star = \{1 \le j \le p : Em_j^2(X_j) > 0\}.$$

To identify the important features in (8), Fan, Feng and Song (2011) considers the following p marginal nonparametric regression problems

$$\min_{f_j \in L_2(P)} E(Y - f_j(X_j))^2, \tag{9}$$

where P denotes the joint distribution of $(\mathbf{x}, Y)$ and $L_2(P)$ is the family of square integrable functions under the measure P. The minimizer of (9) is $f_j = E(Y|X_j)$ and hence $Ef_j^2(X_j)$ can be used as marginal utility to measure the importance of feature X_j at population level. Given a random sample $\{(\mathbf{x}_i, Y_i)\}, i = 1, \ldots, n$, $f_j(x)$ can be estimated by a set of B-spline basis. Let $\mathbf{B}(x) = (B_1(x), \ldots, B_L(x))^\top$ be a B-spline basis and $\beta_j = (\beta_{j1}, \ldots, \beta_{jL})^\top$ be the corresponding coefficients for the B-spline basis associated with feature X_j. Consider the following least squares,

$$\widehat{\beta}_j = \arg\min_{\beta_j} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_j^\top \mathbf{B}(X_{ij}))^2.$$

Thus $f_j(x)$ can be estimated by $\widehat{f}_j(x) = \widehat{\beta}_j^\top \mathbf{B}(x)$. The index set of selected submodel is given by

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \le j \le p : \|\widehat{f}_j\|_n^2 \ge \nu_n\},$$

where $\|\widehat{f}_j\|_n^2 = n^{-1} \sum_{i=1}^n \widehat{f}_j(X_{ij})^2$ and ν_n is some pre-specified threshold. The NIS ranks the importance according to the marginal strength of the marginal nonparametric regression. Under the regularity conditions, Fan, Feng and Song (2011) shows that by taking $\nu_n = c_1 L n^{-2\kappa}$, we have

$$\Pr(\mathscr{M}_\star \subset \widehat{\mathscr{M}_{\nu_n}}) \ge 1 - sL[(8+2L)\exp(-c_2 n^{1-4\kappa} L^{-3}) + 6L\exp(-c_3 n L^{-3})],$$

where L is the number of B-spline basis, $s = |\mathscr{M}_\star|$ and c_2, c_3 are some positive constants. It follows that if

$$\log p = o(n^{1-4\kappa} L^{-3} + nL^{-3}), \tag{10}$$

then $\Pr(\mathscr{M}_\star \subset \widehat{\mathscr{M}_{\nu_n}}) \to 1$. It is worthwhile to point out that the number of spline bases L affects the order of dimensionality. Equation (10) shows that the smaller the number of basis functions, the higher the dimensionality that the NIS can handle. However, the number of basis functions cannot be too small since the approximation error would be too large if we only use a small number of basis functions. After the feature screening, a natural next step is to use penalized method for additive model such as penGAM proposed in Meier, Van de Geer and Bühlmann (2009) to further select important features. Similar to the iterative procedure in Fan, Samworth and Wu (2009), Fan, Feng and Song (2011) also introduces an iterative version of NIS, namely INIS-penGAM, by carrying out the NIS procedure and penGAM alternatively. We omit the details here.

Varying coefficient model is another important nonparametric statistical model that allows us to examine how the effects of features vary with some exposure variable. It is a natural extension of classical linear models with good interpretability and flexibility. Varying coefficient model arises frequently in economics, finance, epidemiology, medical science, ecology, among others. For an overview, see Fan and Zhang (2008). An example of varying coefficient model is the analysis of cross-country growth. Linear model is often used in the standard growth analysis. However, a particular country's growth rate will depend on its state of development and it would make much more sense if we treat the coefficients as functions of the state of development, which leads to a standard varying coefficient model (Fan and Zhang 2008). In this example, state of development is the exposure variable.

Consider the following varying coefficient model,

$$Y = \sum_{j=1}^{p} \beta_j(U) X_j + \varepsilon, \tag{11}$$

where U is some observable univariate exposure variable and the coefficient $\beta_j(\cdot)$ is a smooth function of variable U. In the form of (11), the features X_j enter the model linearly. Such nonparametric formulation allows nonlinear interactions between the exposure variable and the features. The true index set of important features is defined as

$$\mathscr{M}_\star = \{1 \le j \le p : E(\beta_j^2(U)) > 0\},$$

with model size $s = |\mathscr{M}_\star|$. Fan, Ma and Dai (2014) considered a nonparametric screening procedure by ranking a measure of the marginal nonparametric contribution of each feature given the exposure variable. For each feature $X_j, j = 1, \ldots, p$, consider the following marginal regression

$$\min_{a_j, b_j} E[(Y - a_j - b_j X_j)^2 | U]. \tag{12}$$

Let $a_j(U)$ and $b_j(U)$ be the solution to (12) and we have

$$b_j(U) = \frac{\mathrm{Cov}[X_j, Y|U]}{\mathrm{Var}[X_j|U]} \text{ and } a_j(U) = E(Y|U) - b_j(U)E(X_j|U).$$

The marginal contribution of X_j for the response can be characterized by

$$\omega_j = \|a_j(U) + b_j(U)X_j)\|^2 - \|a_0(U)\|^2, \tag{13}$$

where $a_0(U) = E[Y|U]$ and $\|f\|^2 = Ef^2$. By some algebra, it can be seen that

$$\omega_j = E\left[\frac{(\mathrm{Cov}[X_j, Y|U])^2}{\mathrm{Var}[X_j|U]}\right].$$

This marginal utility ω_j is closely related to the conditional correlation between X_j and Y since $\omega_j = 0$ if and only if $\mathrm{Cov}[X_j, Y|U] = 0$. On the other hand, if we assume $\mathrm{Var}[X_j|U] = 1$, then the marginal utility ω_j is the same as the measure of marginal functional coefficient $\|b_j(U)\|^2$.

Suppose we have a random sample $\{(\mathbf{x}_i, Y_i, U_i)\}, i = 1, \ldots, n$. Similar to the setting of additive model, we can estimate $a_j(U)$, $b_j(U)$ and $a_0(U)$ using B-spline technique. Let $\mathbf{B}(U) = (B_1(U), \ldots, B_L(U))^\top$ be a B-spline basis and the coefficients of B-splines can be estimated by the following marginal regression problems

$$(\widehat{\eta}_j, \widehat{\theta}_j) = \min_{\eta_j, \theta_j} n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{B}(U_i)^\top \eta_j - \mathbf{B}(U_i)^\top \theta_j X_{ij})^2,$$

$$\widehat{\eta}_0 = \min_{\eta_0} n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{B}(U_i)^\top \eta_0)^2,$$

where $\eta_0 = (\eta_{0_1}, \ldots, \eta_{0_L})^\top$, $\eta_j = (\eta_{j_1}, \ldots, \eta_{j_L})^\top$, and $\theta_j = (\theta_{j_1}, \ldots, \theta_{j_L})^\top$ are the B-spline coefficients for $a_0(U)$, $a_j(U)$ and $b_j(U)$, respectively. As a result, $\widehat{a}_j(U), \widehat{b}_j(U)$ and $\widehat{a}_0(U)$ can be estimated by

$$\widehat{a}_j(U) = \mathbf{B}(U)^\top \widehat{\eta}_j, \widehat{b}_j(U) = \mathbf{B}(U)^\top \widehat{\theta}_j, \text{ and } \widehat{a}_0(U) = \mathbf{B}(U)^\top \widehat{\eta}_0.$$

The sample marginal utility for screening is

$$\widehat{\omega}_j = \|\widehat{a}_j(U) + \widehat{b}_j(U)\|_n^2 - \|\widehat{a}_0(U)\|_n^2,$$

where $\|f(U)\|_n^2 = n^{-1} \sum_{i=1}^n f(U_i)^2$. The submodel is selected by

$$\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : \widehat{\omega}_j \ge \nu_n\}.$$

Under regularity conditions, Fan, Ma and Dai (2014) established the sure screening property for their proposed screening procedure if the dimensionality satisfies $\log p = o(n^{1-4\kappa} L^{-3})$ for some $0 < \kappa < 1/4$, which is of the same order for the additive model setting. An iterative nonparametric independence screening procedure is also introduced in Fan, Ma and Dai (2014), which repeatedly applies the feature screening procedure followed by a moderate-scale penalized method such as group-SCAD (Wang, Li and Huang 2008).

Instead of using the marginal contribution in (13) to rank the importance of features, Liu, Li and Wu (2014) proposed a screening procedure based on conditional

correlation for varying-coefficient model. Given U, the conditional correlation between X_j and Y is defined as the conditional Pearson correlation

$$\rho(X_j, Y|U) = \frac{\text{cov}(X_j, Y|U)}{\sqrt{\text{cov}(X_j, X_j|U)\text{cov}(Y, Y|U)}}.$$

Then $E[\rho^2(X_j, Y|U)]$ can be used as a marginal utility to evaluate the importance of X_j at population level. It can be estimated by the kernel regression (Liu, Li and Wu 2014). The features with high conditional correlations will be included in the selected submodel. This procedure can be viewed as a natural extension of the SIS by conditioning on the exposure variable U.

3.4 Model-free feature screening

In previous sections, we have discussed model-based feature screening procedures for ultra-high dimensional data, which requires us to specify the underlying true model structure. However, it is quite challenging to correctly specify the model structure on the regression function in high-dimensional modeling. Misspecification of the data generation mechanism could lead to large bias. In practice, one may do not know what model to use unless the dimensionality of feature space is reduced to a moderate size. To achieve greater realism, model-free feature screening is necessary for high-dimensional modeling. In this section, we review several model-free feature screening procedures.

Recall that under the parametric modeling, the true index set of important features $\mathscr{M}_\star$ is defined as the indices of nonzero elements in $\beta^\star$. Since no assumption is made on the specification of the model, there is no such true parameter $\beta^\star$ and thus we need to redefine the true index set of important features $\mathscr{M}_\star$. Let Y be the response variable and $\mathbf{x} = (X_1, \ldots, X_p)^\top$ be the p-dimensional covariate vector. Define the index set of important features as

$$\mathscr{M}_\star = \{1 \le j \le p : F(y|\mathbf{x}) \text{ functionally depends on } X_j \text{ for any } y \in \Psi_y\},$$

where $F(y|\mathbf{x}) = \Pr(Y < y|\mathbf{x})$ is the conditional distribution function of Y given $\mathbf{x}$ and Ψ_y is the support of Y. This indicates that conditional on $\mathbf{x}_{\mathscr{M}_\star}$, Y is statistically independent of $\mathbf{x}_{\mathscr{M}_\star^c}$, where $\mathbf{x}_{\mathscr{M}_\star}$ is a s-dimensional sub-vector of $\mathbf{x}$ consisting of all X_j with $j \in \mathscr{M}_\star$ and $\mathscr{M}_\star^c$ is the complement of $\mathscr{M}_\star$.

Zhu, Li, Li and Zhu (2011) considered a general model framework under which $F(y|\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{B}^\top \mathbf{x}_{\mathscr{M}_\star}$, where $\mathbf{B}$ is a $s \times K$ unknown parameter matrix. In other words, we assume $F(y|\mathbf{x}) = F(y|\mathbf{B}^\top \mathbf{x}_{\mathscr{M}_\star})$. Note that $\mathbf{B}$ may not be identifiable. What is identifiable is the space spanned by the columns of $\mathbf{B}$. However, the identifiability of $\mathbf{B}$ is of no concern here because our primary goal is to identify important features rather than estimating $\mathbf{B}$ itself. This general framework covers a wide range of existing models including the linear regression model, generalized

linear models, the partially linear model (Hardle, Liang and Gao 2012), the single-index model (Hardle, Hall and Ichimura 1993), and the partially linear single-index model (Carroll, Fan, Gijbels and Wand 1997), etc. It also includes the transformation regression model with a general transformation $h(Y)$.

Zhu, Li, Li and Zhu (2011) proposes a unified screening procedure for this general framework. Without loss of generality, assume $E(X_j) = 0$ and $Var(X_j) = 1$. Define $\Omega(y) = E[\mathbf{x}F(y|\mathbf{x})]$. It then follows by the law of iterated expectations that $\Omega(y) = E[\mathbf{x}E(\mathbf{1}(Y < y|\mathbf{x}))|\mathbf{x}] = \text{cov}(\mathbf{x}, \mathbf{1}(Y < y))$. Let $\Omega_j(y)$ be the jth element of $\Omega(y)$ and define

$$\omega_j(y) = E(\Omega_j^2(y)),\ j = 1, \ldots, p.$$

Under certain conditions, Zhu, Li, Li and Zhu (2011) showed that

$$\max_{j \in \mathscr{M}_\star^c} \omega_j < \min_{j \in \mathscr{M}_\star} \omega_j \quad \text{uniformly for } p,$$

and $\omega_j = 0$ if and only if $cov(\mathbf{B}^\top \mathbf{x}_{\mathscr{M}_\star}, X_j) = 0$. These results reveal that the quantity ω_j is in fact a measure of the correlation between the marginal covariate X_j and the linear combination $\mathbf{B}^\top \mathbf{x}_{\mathscr{M}_\star}$ and hence can be used as a marginal utility. Here are some insights. If X_j and Y are independent, so are X_j and $\mathbf{1}(Y < y)$. Consequently, $\Omega_j(y) = 0$ for all $y \in \Psi_y$ and $\omega_j = 0$. On the other hand, if X_j and Y are dependent, then there exists some $y \in \Psi_y$ such that $\Omega_j(y) \neq 0$, and hence ω_j must be positive. In practice, one can employ the sample estimate of ω_j to rank the features. Given a random sample $\{(\mathbf{x}_i, Y_i)\}, i = 1, \ldots, n$, and assume the features are standardized in the sense that $n^{-1}\sum_{i=1}^n X_{ij} = 0$ and $n^{-1}\sum_{i=1}^n X_{ij}^2 = 1$ for all j. A natural estimator for ω_j is

$$\tilde{\omega}_j = \frac{1}{n}\sum_{k=1}^n \left\{ \frac{1}{n}\sum_{i=1}^n X_{ij}\mathbf{1}(Y_i < Y_k) \right\}^2.$$

An equivalent expression of $\tilde{\omega}_j$ is $\widehat{\omega}_j = n^2/(n-1)(n-2)\tilde{\omega}_j$, which is the corresponding U-statistic of $\tilde{\omega}_j$. We use $\widehat{\omega}_j$ as the marginal utility to select important features and the selected submodel is given by

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \leq j \leq p : \widehat{\omega}_j > \nu_n\}.$$

This procedure is referred to as sure independent ranking screening (SIRS). Zhu, Li, Li and Zhu (2011) established the consistency in ranking (CIR) property of the SIRS, which is a stronger result than the sure screening property. It states that if $p = o(\exp(an))$ for some fixed $a > 0$, then there exists some constant $s_\delta \in (0, 4/\delta)$ where $\delta = \min_{j \in \mathscr{M}_\star} \omega_j - \max_{j \in \mathscr{M}_\star^c} \omega_j$ such that

$$\Pr(\max_{j \in \mathscr{M}_\star^c} \widehat{\omega}_j < \min_{j \in \mathscr{M}_\star} \widehat{\omega}_j) \geq 1 - 4p\exp\{n\log(1 - \delta s_\delta/4)/3\}. \tag{14}$$

Since $p = o(\exp(an))$, the right hand side of (14) approaches to 1 with an exponential rate as $n \to \infty$. Therefore, SIRS ranks all important features above unimportant features with high probability. Provided that an ideal threshold is available,

this property would lead to consistency in selection, that is, a proper choice of the threshold can perfectly separate the important and unimportant features. In practice, one can choose the threshold with the help of extra artificial auxiliary variables. The idea of introducing auxiliary variables for thresholding was first proposed by Luo, Stefanski and Boos (2006) to tune the entry significance level in forward selection, and then extended by Wu, Boos and Stefanski (2007) to control the false selection rate of forward regression in linear model. Zhu, Li, Li and Zhu (2011) extended this idea to choose the threshold for feature screening as follows. We generate d auxiliary variables $\mathbf{z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$ such that $\mathbf{z}$ is independent of $\mathbf{x}$ and Y and we regard $(p+d)$-dimensional vector $(\mathbf{x}^\top, \mathbf{z}^\top)^\top$ as the new features. The normality of $\mathbf{z}$ here is not critical here. We know that $\min_{j\in\mathscr{M}_\star} \omega_j > \max_{l=1,\ldots,d} \omega_{p+l}$ since we know $\mathbf{z}$ is truly unimportant features. Given a random sample, we know $\min_{k\in\mathscr{M}_\star} \widehat{\omega}_k > \max_{l=1,\ldots,d} \widehat{\omega}_{p+l}$ holds with high probability according to the consistence in ranking property. Let $C_d = \max_{l=1,\ldots,d} \widehat{\omega}_{p+l}$, the set of selected features is given by

$$\widehat{\mathscr{M}_{C_d}} = \{1 \le k \le p : \widehat{\omega}_k > C_d\}.$$

Li, Zhong and Zhu (2012) proposed a model-free feature screening procedure based on the distance correlation. This procedure does not impose any model assumption on $F(y|\mathbf{x})$. Let $\mathbf{u} \in \mathbb{R}^{d_u}$ and $\mathbf{v} \in \mathbb{R}^{d_v}$ be two random vectors. The distance correlation measures the distance between the joint characteristic function of $(\mathbf{u}, \mathbf{v})$ and the product of marginal characteristic functions of $\mathbf{u}$ and $\mathbf{v}$ (Székely, Rizzo and Bakirov 2007). To be precise, let $\phi_{\mathbf{u}}(\mathbf{t})$ and $\phi_{\mathbf{v}}(\mathbf{s})$ be the characteristic functions of $\mathbf{u}$ and $\mathbf{v}$ respectively, and $\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s})$ be the joint characteristic function of $(\mathbf{u},\mathbf{v})$. The squared distance covariance is defined as

$$\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) = \int_{\mathbb{R}^{d_u+d_v}} |\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s}) - \phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s})|^2 w(\mathbf{t},\mathbf{s}) d\mathbf{t} d\mathbf{s},$$

where $w(\mathbf{t},\mathbf{s})$ is some weight function. With a proper choice of the weight function, the squared distance covariance can be expressed in the following closed form,

$$\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) = S_1 + S_2 - 2S_3,$$

where $S_j, j = 1,2,3$ are defined as

$$\begin{aligned}
S_1 &= E\{\|\mathbf{u}-\tilde{\mathbf{u}}\|_{d_u}\|\mathbf{v}-\tilde{\mathbf{v}}\|_{d_v}\},\\
S_2 &= E\{\|\mathbf{u}-\tilde{\mathbf{u}}\|_{d_u}\}E\{\|\mathbf{v}-\tilde{\mathbf{v}}\|_{d_v}\},\\
S_3 &= E\{E(\|\mathbf{u}-\tilde{\mathbf{u}}\|_{d_u}|\mathbf{u})E(\|\mathbf{v}-\tilde{\mathbf{v}}\|_{d_v}|\mathbf{v})\},
\end{aligned}$$

where $(\tilde{\mathbf{u}},\tilde{\mathbf{v}})$ is an independent copy $(\mathbf{u},\mathbf{v})$ and $\|\mathbf{a}\|_d$ stands for the Euclidean norm of $\mathbf{a} \in \mathbb{R}^d$. The distance correlation (DC) between $\mathbf{u}$ and $\mathbf{v}$ is defined as

$$\mathrm{dcorr}(\mathbf{u},\mathbf{v}) = \frac{\mathrm{dcov}(\mathbf{u},\mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u},\mathbf{u})\mathrm{dcov}(\mathbf{v},\mathbf{v})}}.$$

The distance correlation has many appealing properties. The first property is that distance correlation is closely related to the Pearson correlation. If U and V are two univariate normal random variables, the distance correlation $\mathrm{dcorr}(U,V)$ is a strictly increasing function of $|\rho|$, where ρ is the Pearson correlation between U and V. This property implies that the DC-based marginal feature screening procedure is equivalent to the SIS in Fan and Lv (2008) for linear regression if features and errors are normally distributed. The second property is that $\mathrm{dcorr}(\mathbf{u},\mathbf{v}) = 0$ if and only if $\mathbf{u}$ and $\mathbf{v}$ are independent (Székely, Rizzo and Bakirov 2007). Note that two univariate random variables U and V are independent if and only if U and $T(V)$, a strictly monotone transformation of V, are independent. This implies that a DC-based feature screening procedure can be more effective than the Pearson correlation based procedure since DC can capture both linear and nonlinear relationship between U and V. In addition, DC is well-defined for multivariate random vectors, thus DC-based screening procedure can be directly used for grouped predictors and multivariate response. These remarkable properties make distance correlation a good candidate for feature screening.

Given a random sample $\{(\mathbf{u}_i,\mathbf{v}_i)\}, i = 1,\ldots,n$ from $(\mathbf{u},\mathbf{v})$, the squared distance covariance between $\mathbf{u}$ and $\mathbf{v}$ is estimated by $\widehat{\mathrm{dcov}}^2(\mathbf{u},\mathbf{v}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3$, where

$$\widehat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{u}_i - \mathbf{u}_j\|_{d_u}\|\mathbf{v}_i - \mathbf{v}_j\|_{d_v},$$

$$\widehat{S}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{u}_i - \mathbf{u}_j\|_{d_u}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{v}_i - \mathbf{v}_j\|_{d_v},$$

$$\widehat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \|\mathbf{u}_i - \mathbf{u}_k\|_{d_u}\|\mathbf{v}_j - \mathbf{v}_k\|_{d_v}.$$

Similarly, we can define the sample distance covariances $\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{u})$ and $\widehat{\mathrm{dcov}}(\mathbf{v},\mathbf{v})$. Accordingly, the sample distance correlation between $\mathbf{u}$ and $\mathbf{v}$ is defined by

$$\widehat{\mathrm{dcorr}}(\mathbf{u},\mathbf{v}) = \frac{\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{v})}{\sqrt{\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{u})\widehat{\mathrm{dcov}}(\mathbf{v},\mathbf{v})}}.$$

Let $\mathbf{y} = (Y_1,\ldots,Y_q)^\top$ be the response vector with support Ψ_y, and $\mathbf{x} = (X_1,\ldots,X_p)^\top$ be the covariate vector. Here we allow the response to be univariate or multivariate and assume q is a fixed number. For each $j = 1,\ldots,p$, we can calculate the sample distance correlation $\widehat{\mathrm{dcorr}}(X_j,\mathbf{y})$. Based on the fact that $\mathrm{dcorr}(X_j,\mathbf{y}) = 0$ if and only if X_j and $\mathbf{y}$ are independent, $\widehat{\mathrm{dcorr}}(X_j,\mathbf{y})$ can be used as a marginal utility to rank the importance of X_j. Therefore, the set of important variables is defined as

$$\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : \widehat{\mathrm{dcorr}}(X_j,\mathbf{y}) > \nu_n\},$$

for some pre-specified threshold ν_n. This model-free feature screening procedure is known as DC-SIS. Under certain moment assumptions and with the choice of $\nu_n = cn^{-\kappa}$ for some constants c and κ, Li, Zhong and Zhu (2012) showed that DC-SIS enjoys the sure screening property. This DC-SIS allows for arbitrary regression relationship of Y onto $\mathbf{x}$, regardless of whether it is linear or nonlinear. It also permits univariate and multivariate responses, regardless of whether it is continuous, discrete, or categorical. Note that the SIRS in Zhu, Li, Li and Zhu (2011) requires that $F(y|\mathbf{x})$ depends on $\mathbf{x}$ through a linear combination $\mathbf{B}^\top \mathbf{x}_{\mathscr{M}_\star}$. Comparing with SIRS, this DC-SIS is completely model-free and it does not require any model assumption on the relationship between features and the response. Another advantage of DC-SIS is that it can be directly utilized for screening grouped variables and multivariate responses while SIRS can only handle univariate response. An iterative version of DC-SIS was proposed in Zhong and Zhu (2015) to address the issues of marginal feature screening.

3.5 Feature screening for categorical data

Plenty of feature screening methods have been proposed for models where both the features and the response are continuous. In practice, we are also interested in the situation where features and/or response are categorical data. Fan and Fan (2008) proposed a marginal t-test screening for the linear discriminant analysis and showed that it has the sure screening property. Fan and Song (2010) proposed a maximum marginal likelihood screening for generalized linear models and rank variables according to the magnitudes of coefficient, which can be applied directly to the logistic regression.

Mai and Zou (2012) introduced a nonparametric screening method based on Kolmogorov-Smirnov distance for binary classification. It does not require any modeling assumption and thus is robust and has wide applicability. Let Y be the label and takes value in $\{-1,1\}$ and let $F_{+j}(x)$ and $F_{-j}(x)$ denote the conditional CDF of X_j given $Y = 1,-1$, respectively. Define

$$K_j = \sup_{-\infty<x<\infty} |F_{+j}(x) - F_{-j}(x)|.$$

The sample version of K_j is defined as

$$K_{nj} = \sup_{-\infty<x<\infty} |\widehat{F}_{+j}(x) - \widehat{F}_{-j}(x)|,$$

where $\widehat{F}_{+j}(x)$ and $\widehat{F}_{-j}(x)$ are the empirical CDF of X_j given $Y = 1,-1$ respectively. This screening procedure is called Kolmogorov filter due to the fact that K_{nj} is actually the Kolmogorov-Smirnov test statistic for testing the equivalence of two distributions. By definition, K_{nj} is invariant under any strictly monotone univariate transformations applied to individual feature. Mai and Zou (2012) recommended

using the Kolmogorov filter to select the submodel

$$\widehat{\mathscr{M}_{d_n}} = \{1 \le j \le p : K_{nj} \text{ is among the first } d_n \text{ largest ones}\}.$$

Mai and Zou (2015) extended the idea of Kolmogorov filter to a wide variety of applications including multi-class classification, Poisson regression and so on by slicing the response. The resulting procedure is a nonparametric model-free feature screening procedure that works with discrete, categorical or continuous features.

Cui, Li and Zhong (2015) developed an effective model-free and robust feature screening procedure for ultra-high dimensional discriminant analysis with a possibly diverging number of classes. Without specifying a regression model, define the true index set of important features by

$$\mathscr{M}_\star = \{1 \le j \le p : F(y|\mathbf{x}) \text{ functionally depends on } X_j\}.$$

Let Y be a categorical response with K categories $\{y_1, \ldots, y_K\}$, and assume X is a continuous univariate feature. Let $F(x|Y) = \Pr(X \le x|Y)$ be the conditional distribution function of X given Y. Denote by $F(x) = \Pr(X \le x)$ the unconditional distribution function of X and $F_k(x) = \Pr(X \le x|Y = y_k)$ the conditional distribution function of X given $Y = y_k$. If $F_k(x) = F(x)$ for all x and $k = 1, \ldots, K$, then X and Y are independent. Based on this observation, Cui, Li and Zhong (2015) proposed the following index

$$\mathrm{MV}(X|Y) = E[Var(F(X|Y))]$$

to measure the dependence between X and Y. Let $p_k = \Pr(Y = y_k) > 0$, then $\mathrm{MV}(X|Y)$ can be written as

$$\mathrm{MV}(X|Y) = \sum_{k=1}^{K} p_k \int (F_k(x) - F(x))^2 dF(x). \tag{15}$$

Equation (15) implies that $\mathrm{MV}(X|Y)$ can be represented as the weighted average of Cramer-von Mises distances between the conditional distribution of X given $Y = y_k$ and the unconditional distribution function of X. Cui, Li and Zhong (2015) showed that $\mathrm{MV}(X|Y) = 0$ if and only if X and Y are statistically independent. Another appealing property of $\mathrm{MV}(X|Y)$ is that it characterizes both linear and nonlinear relationships, making it a good marginal utility for ultra-high dimensional discriminant analysis.

Let $\{(X_i, Y_i)\}, i = 1, \ldots, n$ be a random sample from the population (X, Y). Define $\widehat{p}_k = n^{-1} \sum_{i=1}^n \mathbf{1}(Y_i = y_k)$, $\widehat{F}(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \le x)$, and $\widehat{F}_k(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \le x, Y_i = y_k)/\widehat{p}_k$. Based on the Cramer-von Mises representation (15), $\mathrm{MV}(X|Y)$ can be estimated by its sample counterpart

$$\widehat{\mathrm{MV}}(X|Y) = n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \widehat{p}_k (\widehat{F}_k(X_i) - \widehat{F}(X_i))^2.$$

For each of the features $X_j, j = 1\ldots,p$, we can compute the sample version of the index $\widehat{\mathrm{MV}}(X_j|Y)$ between X_j and Y. We select the submodel by

$$\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : \widehat{\mathrm{MV}}(X_j|Y) > \nu_n\},$$

for some pre-specified threshold ν_n. This MV-based screening procedure is referred to as MV-SIS. The sure screening property holds for MV-SIS under very mild moment conditions of features and it does not require the regression function of Y onto $\mathbf{x}$ to be linear. It is worth noting that MV-SIS is insensitive to heavy-tailed distributions of features and potential outliers due to the robustness of conditional distribution function. Furthermore, the sure screening property holds even when number of classes diverges.

In reality, one may also encounter the situation in which both the features and the response are categorical. Huang, Li and Wang (2014) proposed a chi-square based feature screening procedure for such situation. The ides is to construct a chi-square test statistic for each pair of feature and response. Let $Y_i \in \{1,\ldots,K\}$ be the class label of response, and $\mathbf{x}_i = (X_{i1},\ldots,X_{ip})^\top$ be the associated categorical features. For simplicity, assume each X_{ij} is binary though the method and theory can be readily applied to multi-class categorical features. Define $\Pr(Y_i = k) = \pi_{yk}, \Pr(X_{ij} = k) = \pi_{jk}$, and $\Pr(Y_i = k_1, X_{ij} = k_2) = \pi_{yj,k_1k_2}$. Those quantities can be estimated by their sample counterparts $\widehat{\pi}_{yk} = n^{-1}\sum_{i=1}^n \mathbf{1}(Y_i = k)$, $\widehat{\pi}_{jk} = n^{-1}\sum_{i=1}^n \mathbf{1}(X_{ij} = k)$ and $\widehat{\pi}_{yj,k_1k_2} = n^{-1}\sum_{i=1}^n \mathbf{1}(Y_i = k_1)\mathbf{1}(X_{ij} = k_2)$, respectively. Subsequently, for each feature, a chi-square type statistic can be constructed as

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

which is a natural estimator of

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

Obviously, features with larger values of $\widehat{\Delta}_j$ are more relevant to the response. As a result, the submodel is selected by

$$\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : \widehat{\Delta}_j > \nu_n\},$$

where $\nu_n > 0$ is some pre-specified threshold. Note that $n\widehat{\Delta}_j$ has an asymptotic distribution χ^2_{K-1}, where χ^2_{K-1} is the chi-squared distribution with degrees of freedom $K-1$. Then $\widehat{\mathscr{M}}_{\nu_n}$ can be defined in terms of p-value. Let $\widehat{p}_j = \Pr(\chi^2_{K-1} > n\widehat{\Delta}_j)$ and $\widehat{\mathscr{M}}_{\nu_n}$ can be equivalently expressed as $\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : \widehat{p}_j < p_{\nu_n}\}$ for some $0 < p_{\nu_n} < 1$. When the number of categories of features are different from each other, then features involving more categories are more likely to have larger Δ_j values, regardless of whether the feature is important or not. Based on this observation,

it is more appropriate to use the p-values $\widehat{p}_j$ as the marginal utility to select the important features instead of using $\widehat{\Delta}_j$. Assume the jth feature has R_j categories, the p-value $\widehat{p}_j$ can be obtained from the Pearson chi-squared test of independence with degrees of freedom $(K-1)(R_j-1)$.

Huang, Li and Wang (2014) suggested using the following maximum ratio criterion to determine how many features should be included in the submodel. Let $\{k_1,\ldots,k_p\}$ be a permutation of $\{1,\ldots,p\}$ such that $\widehat{\Delta}_{k1} \geq \widehat{\Delta}_{k2} \geq \cdots \geq \widehat{\Delta}_{kp}$. Recall that the true model size is $|\mathscr{M}_\star| = s$. As long as $j+1 \leq s$, we should have $\widehat{\Delta}_{k_j}/\widehat{\Delta}_{k_{j+1}} \to c_{j,j+1}$ in probability for some $c_{j,j+1} > 0$. On the other hand, if $j > s$, we should have both $\widehat{\Delta}_{k_j}$ and $\widehat{\Delta}_{k_{j+1}}$ converge towards to 0 in probability. If their convergence rates are of the same order, we should have $\widehat{\Delta}_{k_j}/\widehat{\Delta}_{k_{j+1}} = O(1)$. If $j = s$, we expect $\widehat{\Delta}_{k_j} \to c_j > 0$ while $\widehat{\Delta}_{k_{j+1}} \to 0$ in probability. This makes the ratio $\widehat{\Delta}_{k_j}/\widehat{\Delta}_{k_{j+1}} \to \infty$. They suggest selecting the top $\widehat{d}$ features as submodel where

$$\widehat{d} = \underset{0\leq j\leq p-1}{\arg\max}\, \widehat{\Delta}_{k_j}/\widehat{\Delta}_{k_{j+1}}$$

and $\widehat{\Delta}_0$ is defined to be 1. That is, we include the $\widehat{d}$ features with the largest $\widehat{\Delta}_j$ in the submodel.

4 Time-dependent Data

4.1 Longitudinal data

Instead of observing independent and identically distributed data, one may observe longitudinal data, that is, the features may change over time. More precisely, longitudinal data, also known as panel data, is a collection of repeated observations of the same subjects over a period of time. Longitudinal data differs from cross-sectional data in that it follows the same subjects over a period of time, while cross-sectional data are collected from different subjects at each time point. Longitudinal data is often seen in economy, finance studies, clinical psychology, etc. For example, longitudinal data is often seen in event studies, which tries to analyze what factors drive abnormal stock returns over time, or how stock prices react to merger and earnings announcements.

Time-varying coefficient model is widely used for modeling longitudinal data. Consider the following time-varying coefficient model,

$$y(t) = \mathbf{x}(t)^\top \beta(t) + \varepsilon(t),\ t \in T, \tag{16}$$

where $\mathbf{x}(t) = (X_1(t),\ldots,X_p(t))^\top$ are the p-dimensional covariates, $\beta(t) = (\beta_1(t),\ldots,\beta_p(t))^\top$ are the time-varying coefficients, $\varepsilon(t)$ is a mean zero stochastic process, and T is the time interval in which the measurements are taken. In model (16), t need not

to be calendar time, for instance, we can set t to be the age of a subject. In general, it is assumed that T is a closed and bounded interval in $\mathbb{R}$. The goal is to identify the set of true important variables, which is defined as

$$\mathscr{M}_\star = \{1 \le j \le p : \|\beta_j(t)\|_2 \neq 0\},$$

where $\|\beta(t)\|_2 = \frac{1}{|T|}\int_T \beta^2(t)dt$ and $|T|$ is the length of T.

Suppose there is a random sample of n independent subjects $\{\mathbf{x}_i(t), Y_i(t)\}, i = 1,\ldots,n$ from model (16). Let t_{ik} and m_i be the time of the kth measurement and the number of repeated measurement for the ith subject. $Y(t_{ik})$ and $\mathbf{x}_i(t_{ik}) = (X_{i1}(t_{ik}),\ldots,X_{ip}(t_{ik}))^\top$ are the ith subject's observed response and covariates at time t_{ik}. Based on the longitudinal observations, the model can be written as

$$Y_i(t_{ik}) = \mathbf{x}_i(t_{ik})^\top \beta(t_{ik}) + \varepsilon_i(t_{ik}),$$

where $\beta(t_{ik}) = (\beta_1(t_{ik}),\ldots,\beta_p(t_{ik}))^\top$ is the coefficient at time t_{ik}. Song, Yi and Zou (2014) considered a marginal time-varying coefficient model for each $j = 1,\ldots,p$,

$$Y_i(t_{ik}) = \beta_j(t_{ik})X_{ij}(t_{ik}) + \varepsilon_i(t_{ik}).$$

Let $\mathbf{B}(t) = (B_1(t),\ldots,B_L(t))^\top$ be a B-spline basis on the time interval T, where L is the dimension of the basis. For the ease of presentation, we use the same B-spline basis for all $\beta_j(t)$. Under smoothness conditions, each $\beta_j(t)$ can be approximated by the linear combination of B-spline basis functions. For each j, consider marginal weighted least square estimation based on B-spline basis

$$\widehat{\gamma}_j = \arg\min_{\gamma_{jl}} \sum_{i=1}^{n} w_i \sum_{k=1}^{m_i} \left(Y_i(t_{ik}) - \sum_{l=1}^{L} X_{ij}(t_{ik})B_l(t_{ik})\gamma_{jl} \right)^2,$$

where $\gamma_j = (\gamma_{j1},\ldots,\gamma_{jL})^\top$ is the unknown parameter and $\widehat{\gamma}_j = (\widehat{\gamma}_{j1},\ldots,\widehat{\gamma}_{jL})^\top$ is its estimate. Choices of w_i can be 1 or $1/m_i$, that is equal weights to observations or equal weights to subjects. See Song, Yi and Zou (2014) for more details on how to obtain $\widehat{\gamma}_j$. The B-spline estimator of $\beta_k(t)$ is given by $\widehat{\beta}_j(t) = \widehat{\gamma}_j^\top \mathbf{B}(t)$. The selected set of features is given by

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \le j \le p : \|\widehat{\beta}_j(t)\|_2 \ge \nu_n\},$$

where ν_n is a pre-specified threshold. To evaluate $\|\widehat{\beta}_j(t)\|_2$, one can take N equally spaced time points $t_1 < \cdots < t_N$ in T, and compute $\|\widehat{\beta}_{Nj}(t)\|_2 = N^{-1}\sum_{i=1}^{N}\widehat{\beta}_j^2(t_i)$. As long as N is large enough, $\|\widehat{\beta}_{Nj}(t)\|_2$ should be close enough to $\|\widehat{\beta}_j(t)\|_2$. This varying-coefficient independence screening is referred to as VIS and enjoys the sure screening property (Song, Yi and Zou 2014). An iterative VIS (IVIS) was also introduced in Song, Yi and Zou (2014), which utilizes the additional contribution of unselected features by conditioning on the selected features that survived the previous step.

Cheng, Honda, Li and Peng (2014) proposed a similar nonparametric independence screening method for the time varying-coefficient model. In their setting, they allow some of the important features simply have constant effects, i.e.,

$$Y_i(t_{ik}) = \sum_{j=1}^{q} X_{ij}(t_{ik})\beta_j + \sum_{j=q+1}^{p} X_{ij}(t_{ik})\beta_j(t_{ik}) + \varepsilon_i(t_{ik}).$$

The first q coefficients $\beta_j, j = 1,\ldots,q$ do not change over time. Cheng, Honda, Li and Peng (2014) points out that it is very important to identify the nonzero constant coefficients because treating a constant coefficient as time varying will yield a convergence rate that is slower than $\sqrt{n}$.

Both Song, Yi and Zou (2014) and Cheng, Honda, Li and Peng (2014) ignore the covariance structure of $\varepsilon(t)$ and carry out the feature screening on a working independence structure. Chu, Li and Reimherr (2016) extended the VIS by incorporating within-subject correlation and dynamic error structure. They also allow baseline variables in their model, which are believed to have impact on the response based on empirical evidence or relevant theories and are not subject to be screened. These baseline features are called Z-variables and the longitudinal features to be screened are called X-variables. Consider the following model,

$$Y_i(t_{ik}) = \sum_{j=1}^{q} \beta_j(t_{ik})Z_{ij}(t_{ik}) + \sum_{j=1}^{p} \beta_j(t_{ik})X_{ij}(t_{ik}) + \varepsilon_i(t_{ik}), \tag{17}$$

where Z-variables are the known important variables by prior knowledge and X-variables are ultra-high dimensional features. It is assumed that $\varepsilon_i(t)$ have variances that vary across time, are independent across i (between subjects) and correlated across t (within the same subject). Incorporating the error structure into the model estimation is expected to increase screening accuracy. Chu, Li and Reimherr (2016) proposed a working model without any X-variables to estimate the covariance structure,

$$Y_i(t_{ik}) = \beta_0^w + \sum_{l=1}^{q} \beta_l^w(t_{ik})Z_{il}(t_{ik}) + \varepsilon_i^w(t_{ik}). \tag{18}$$

Although model (18) is mis-specified, valuable information about the covariance structure can still be gained. Standard ordinary least squares and regression spline technique can be applied to (18) (Huang, Wu and Zhou 2004), and we can obtain the corresponding residuals $r_i(t_{ik})$. Let $V(t_{ik})$ be a working variance function for $\varepsilon(t_{ik})$ and it can be approximated by $V(t_{ik}) \approx \sum_{l=1}^{L} \alpha_l B_l(t_{ik})$, where $B_1(t),\ldots,B_L(t)$ is a B-spline basis. The coefficients $\alpha_l, l = 1,\ldots,L$ can be estimated by minimizing the following least squares

$$\widehat{\alpha} = (\widehat{\alpha}_1,\ldots,\widehat{\alpha}_L)^\top = \min_{\alpha_1,\ldots,\alpha_L} \sum_{i=1}^{n}\sum_{k=1}^{m_i} \left(r_i^2(t_{ik}) - \sum_{l=1}^{L} \alpha_l B_l(t_{ik}) \right)^2.$$

Then define $\widehat{V}(t_{ik}) = \sum_{l=1}^{L} \widehat{\alpha}_l B_l(t_{ik})$. Denote by $\mathbf{R}_i$ the $m_i \times m_i$ working correlation matrix for the ith subject. A parametric model can be used to estimate the working correlation matrix. These models include autoregressive (AR) structure, stationary or non-stationary M-dependent correlation structure, parametric families such as the Matern. Now assume we obtain the working correlation matrix $\mathbf{R}_i$ based on some parametric model, then the weight matrix for ith subject is given by

$$\mathbf{W}_i = \frac{1}{m_i} \widehat{\mathbf{V}}_i^{-1/2} \mathbf{R}_i^{-1} \widehat{\mathbf{V}}_i^{-1/2},$$

where $\widehat{\mathbf{V}}_i$ is the $m_i \times m_i$ diagonal matrix consisting of the time-varying variance

$$\widehat{\mathbf{V}}_i = \begin{pmatrix} \widehat{V}(t_{i1}) & 0 & \dots & 0 \\ 0 & \widehat{V}(t_{i2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widehat{V}(t_{im_i}) \end{pmatrix}.$$

For each j, define a marginal time-varying model with the jth X-variable,

$$Y_i(t_{ik}) = \sum_{l=1}^{q} \beta_{lj} Z_{il}(t_{ik}) + \beta_j(t_{ik}) X_{ij}(t_{ik}) + \varepsilon_i(t_{ik}). \tag{19}$$

Using the B-spline technique and the weight matrix $\mathbf{W}_i$, one can obtain the weighted least squares estimate for model (19), and thus the fitted value $\widehat{Y}^{(j)}(t_{ik})$, see Chu, Li and Reimherr (2016) for a detailed description. Then the weighted mean squared errors are given by

$$\widehat{u}_j = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \widehat{\mathbf{y}}_i^{(j)})^\top \mathbf{W}_i (\mathbf{y}_i - \widehat{\mathbf{y}}_i^{(j)}),$$

where $\mathbf{y}_i = (Y(t_{i1}), \dots, Y(t_{im_i})^\top$ and $\widehat{\mathbf{y}}_i^{(j)} = (\widehat{Y}^{(j)}(t_{i1}), \dots, \widehat{Y}^{(j)}(t_{im_i}))^\top$. Note that a small value of $\widehat{u}_j$ indicates a strong marginal association between the jth feature and the response. Thus, the selected set of important variables is given by

$$\widehat{\mathscr{M}}_{\nu_n} = \{1 \le j \le p : \widehat{u}_j \le \nu_n\}.$$

This procedure has sure screening property, meaning that with probability tending to 1, all important variables will be included in the submodel defined by $\widehat{\mathscr{M}}_{\nu_n}$ provided certain conditions are satisfied. See the supplementary material of Chu, Li and Reimherr (2016).

Different form the B-spline techniques, Xu, Zhu and Li (2014) proposed a generalized estimating equation (GEE) based sure screening procedure for longitudinal data. Without risk of confusion, we slightly abuse the notations here. Let $\mathbf{y}_i = (Y_{i1}, \dots, Y_{im_i})^\top$ be the response vector for the ith subject, and $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{im_i})^\top$ be the corresponding $m_i \times p$ matrix of features. Suppose the conditional mean of Y_{ik}

given $\mathbf{x}_{ik}$ is

$$\mu_{ik}(\beta) = E(Y_{ik}|\mathbf{x}_{ik}) = g^{-1}(\mathbf{x}_{ik}^\top \beta),$$

where $g(\cdot)$ is a known link function, and β is a p-dimensional unknown parameter vector. Let $\mathbf{A}_i(\beta)$ be an $m_i \times m_i$ diagonal matrix with kth diagonal element $\sigma_{ik}^2(\beta) = \text{Var}(Y_{ik}|\mathbf{x}_{ik})$, and $\mathbf{R}_i$ be an $m_i \times m_i$ working correlation matrix. The GEE estimator of β is defined to be the solution to

$$\mathbf{G}(\beta) = n^{-1} \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{A}_i^{1/2}(\beta) \mathbf{R}_i^{-1} \mathbf{A}_i^{1/2}(\beta)(\mathbf{y}_i - \mu_i(\beta)) = \mathbf{0}, \tag{20}$$

where $\mu_i(\beta) = (\mu_{i1}(\beta), \ldots, \mu_{im_i}(\beta))^\top$. Let $\mathbf{g}(\beta) = (g_1(\beta), \ldots, g_p(\beta))^\top = E(\mathbf{G}(\beta))$. Then $g_j(\mathbf{0})$ can be used as a measure of the dependence between the response and the jth feature. Let $\widehat{\mathbf{R}}_i$ be an estimate of $\mathbf{R}_i$. Then $\widehat{\mathbf{G}}(\mathbf{0})$ is defined as

$$\widehat{\mathbf{G}}(\mathbf{0}) = n^{-1} \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{A}_i^{1/2}(\mathbf{0}) \widehat{\mathbf{R}}_i^{-1} \mathbf{A}_i^{-1/2}(\mathbf{0})(\mathbf{y}_i - \mu_i(\mathbf{0})).$$

Hence, we would select the set of important features using

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \le j \le p : |\widehat{G}_j(\mathbf{0})| > \nu_n\},$$

where $\widehat{G}_j(\mathbf{0})$ is the jth component of $\widehat{\mathbf{G}}(\mathbf{0})$ and ν_n is a pre-specified threshold. If we consider the linear regression model $Y_i = \mathbf{x}_i^\top \beta + \varepsilon_i$, the GEE function in (20) reduces to

$$\mathbf{G}(\mathbf{0}) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i (Y_i - \mathbf{x}_i^\top \beta).$$

Therefore, for any given ν_n, the GEE based screening (GEES) selects the submodel using

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \le k \le p : n^{-1} |\mathbf{x}_{(j)}^\top \mathbf{y}| > \nu_n\},$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{x}_{(j)}$ is the jth column of the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, which coincides with the original SIS proposed in Fan and Lv (2008). One desiring property of GEES is that even the working correlation matrix structure of $\widehat{\mathbf{R}}$ is mis-specified, all the important features will be retained by the GEES with probability approaching to 1.

4.2 Time-series data

The analysis of time-series data is common in economics and finance. For example, the market model in finance relates the return of an individual stock to the return of a market index or another individual stock. Another example is the term structure of interest rates in which the time evolution of the relationship between interest

rates with different maturities is investigated. In this section, we briefly review some feature screening methods in time series. The SIS (Fan and Lv 2008) was originally proposed for linear regression and assume the random errors follow normal distribution. Yousuf (2018) analyzes the theoretical properties of SIS for high dimensional linear models with dependent and/or heavy tailed covariates and errors. They also introduced a generalized least squares screening (GLSS) procedure which utilizes the serial correlation present in the data. With proper assumptions on the moment, the strength of dependence in the error and covariate processes, Yousuf (2018) established the sure screening properties for both screening procedures. GLSS is shown to outperform SIS in many cases since GLSS utilizes the serial correlation when estimating the marginal effects.

Yousuf (2018)'s work is limited to the linear model and ignore some unique qualities of time series data. The dependence structure of longitudinal data is too restrictive to cover the type of dependence present in most time series. Yousuf and Feng (2018) studied a more general time series setting. Let $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$ be the response time series, and let $\mathbf{x}_{t-1} = (X_{t-1,1}, \ldots, X_{t-1,m})^\top$ denote the m predictor series at time $t-1$. Given that the lags of these predictor series are possible covariates, let $\mathbf{z}_{t-1} = (\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-h}) = (Z_{t-1,1}, \ldots, Z_{t-1,p})$ denote the p-dimensional vector of covariates, where $p = mh$. The set of important covariates is defined as

$$\mathscr{M}_\star = \{1 \le j \le p : F(y_t | Y_{t-1}, \ldots, Y_{t-h}, \mathbf{z}_{t-1}) \text{ functionally depends on } Z_{t-1,j}\},$$

where $F(y_t|\cdot)$ is the conditional distribution function of Y_t. The value h represents the maximum lag order for the response and predictor series. The value of h can be pre-specified by the user, or can be determined by some data driven method. Yousuf and Feng (2018) proposed a model-free feature screening method based on the partial distance correlation (PDC). More specifically, the PDC between $\mathbf{u}$ and $\mathbf{v}$, controlling for $\mathbf{z}$, is defined as

$$\text{pdcor}(\mathbf{u}, \mathbf{v}; \mathbf{z}) = \frac{\text{dcor}^2(\mathbf{u}, \mathbf{v}) - \text{dcor}^2(\mathbf{u}, \mathbf{z})\text{dcor}^2(\mathbf{v}, \mathbf{z})}{\sqrt{1 - \text{dcor}^4(\mathbf{u}, \mathbf{z})}\sqrt{1 - \text{dcor}^4(\mathbf{v}, \mathbf{z})}}, \tag{21}$$

if $\text{dcor}(\mathbf{u}, \mathbf{z}), \text{dcor}(\mathbf{v}, \mathbf{z}) \neq 1$, otherwise $\text{pdcor}(\mathbf{u}, \mathbf{v}; \mathbf{z}) = 0$. For more details and interpretation of PDC, see Székely and Rizzo (2014). $\text{pdcor}(\mathbf{u}, \mathbf{v}; \mathbf{z})$ can be estimated by its sample counterpart $\widehat{\text{pdcor}}(\mathbf{u}, \mathbf{v}; \mathbf{z})$ which replaces dcor by $\widehat{\text{dcor}}$ in (21).

The corresponding feature screening procedure PDC-SIS was introduced in Yousuf and Feng (2018). They first define the conditioning vector for the lth lag of predictor series k as $\mathscr{S}_{k,l} = (Y_t, \ldots, Y_{t-h}, X_{t-1,k}, \ldots, X_{t-l+1,k})$ with $1 \le l \le h$. Besides that a certain number of lags of Y_t are included in the model, the conditioning vector also includes all lower order lags for each lagged covariate of interest. By including the lower order lags in the conditioning vector, PDC-SIS tries to shrink towards submodels with lower order lags. For convenience, let $\mathbf{C} = \{\mathscr{S}_{1,1}, \ldots, \mathscr{S}_{m,1}, \mathscr{S}_{1,2}, \ldots, \mathscr{S}_{m,h}\}$ denote the set of conditioning vectors where $C_{k+(l-1)*m} = \mathscr{S}_{k,l}$ is the conditioning vector for covariate $Z_{t-1,(l-1)*m+k}$. The selected submodel is

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \le j \le p : |\widehat{\text{pdcor}}(Y_t, Z_{t-1}; C_j)| \ge \nu_n\}. \tag{22}$$

The PDC-SIS attempts to utilize the time series structure by conditioning on previous lags of the covariates. Yousuf and Feng (2018) also proposed a different version, namely PDC-SIS+, to improve the performance of PDC-SIS. Instead of only conditioning on the previous lags of one covariate, PDC-SIS+ also conditions on additional information available from previous lags of other covariates as well. To attempt this, PDC-SIS+ identifies strong conditional signals at each lag level and add them to the conditioning vector for all higher order lag levels. By utilizing this conditioning scheme we can pick up on hidden significant variables in more distant lags, and also shrink toward models with lower order lags by controlling for false positives resulting from high autocorrelation, and cross-correlation.

5 Survival Data

5.1 Cox model

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems. This topic is referred to as reliability theory or reliability analysis in engineering, duration analysis or duration modeling in economics, and event history analysis in sociology. It is inevitable to analyze survival data in many scientific studies since the primary outcomes or responses are subject to be censored. The Cox model (Cox 1972) is the most commonly used regression model for survival data. Let T be the survival time and $\mathbf{x}$ be the p-dimensional covariate vector. Consider the following Cox proportional hazard model

$$h(t|\mathbf{x}) = h_0(t)\exp\{\mathbf{x}^\top \beta\}, \tag{23}$$

where $h_0(t)$ is the unknown baseline hazard functions. In survival analysis, survival time T is typically censored by the censoring time C. Denote the observed time by $Z = \min\{T, C\}$ and the event indicator by $\delta = \mathbf{1}(T \le C)$. For simplicity we assume that T and C are conditionally independent given $\mathbf{x}$ and the censoring mechanism is non-informative. The observed data is an independently and identically distributed random sample $\{(\mathbf{x}_i, z_i, \delta_i)\}, i = 1, \ldots, n$. Let $t_1^0 < \cdots < t_N^0$ be the ordered distinct observed failure times and (k) index its associate covariates $\mathbf{x}_{(k)}$. $\mathscr{R}(t)$ denotes the risk set right before the time t: $\mathscr{R}(t) = \{i : z_i \ge t\}$. Under Equation (23), the likelihood function is

$$L(\beta) = \prod_{k=1}^{N} h_0(z_{(k)})\exp(\mathbf{x}_{(k)}^\top \beta) \prod_{i=1}^{n} \exp\{H_0(z_i)\exp\{\mathbf{x}_i^\top \beta\}\},$$

where $H_0(t) = \int_0^t h_0(s)ds$ is the corresponding cumulative baseline hazard function. Consider the 'least informative' nonparametric modeling for H_0 with the form $H_0(t) = \sum_{k=1}^N h_k \mathbf{1}(t_k^0 \le t)$, then $H_0(z_i) = \sum_{k=1}^N h_k \mathbf{1}(i \in \mathscr{R}(t_k^0))$. Consequently the log-likelihood becomes

$$\ell(\beta) = \sum_{k=1}^{N} \{\log(h_k) + \mathbf{x}_{(k)}^\top \beta\} - \sum_{i=1}^{n} \left\{ \sum_{k=1}^{N} h_k \mathbf{1}(i \in \mathscr{R}(t_k^0)) \exp(\mathbf{x}_i^\top \beta) \right\}. \tag{24}$$

Given β, the maximizer of (24) is given by $\widehat{h}_k = 1/\sum_{i \in \mathscr{R}(t_j^0)} \exp\{\mathbf{x}_i^\top \beta\}$. Plugging in the maximizer, the log-likelihood function can be written as

$$\ell(\beta) = \left(\sum_{i=i}^{n} \delta_i \mathbf{x}_i^\top \beta - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{k \in \mathscr{R}(t_i)} \exp\{\mathbf{x}_k^\top \beta\} \right\} \right), \tag{25}$$

which is also known as the partial likelihood (Cox 1972).

5.2 Feature screening for Cox model

A marginal feature screening procedure is developed in Fan, Feng and Wu (2010). The marginal utility $\widehat{u}_j$ of the feature X_j is defined as the maximum of the partial likelihood only with respect to X_j,

$$\widehat{u}_j = \max_{\beta_j} \left(\sum_{i=i}^{n} \delta_i X_{ij} \beta_j - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{k \in \mathscr{R}(t_i)} \exp\{X_{kj} \beta_j\} \right\} \right). \tag{26}$$

Here X_{ij} is the jth element of $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})^\top$. Intuitively, a larger marginal utility indicates that the associated feature contains more information about the survival outcome. One can rank all features according to the marginal utilities from the largest to the smallest and define the selected submodel

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \le j \le p : \widehat{u}_j > \nu_n\}.$$

Zhao and Li (2012) proposed to fit a marginal Cox model for each feature, namely the hazard function has the form $h_0(t)\exp\{\beta_j X_j\}$ for feature X_j. Let $N_i(t) = \mathbf{1}(z_i \le t, \delta_i = 1)$ be independent counting process for each subject i and $Y_i(t) = \mathbf{1}(z_i \ge t)$ be the at-risk processes. For $k = 0, 1, \ldots,$ define

$$S_j^{(k)}(t) = n^{-1} \sum_{i=1}^{n} X_{ij}^k Y_i(t) h(t|\mathbf{x}_i),$$

$$S_j^{(k)}(\beta, t) = n^{-1} \sum_{i=1}^{n} X_{ij}^k Y_i(t) \exp\{\beta X_{ij}\}.$$

Then the maximum marginal partial likelihood estimator $\widehat{\beta}_j$ is defined as the solution to the following estimating equation

$$U_j(\beta) = \sum_{i=1}^{n} \int_0^C \left\{ X_{ij} - \frac{S_j^{(1)}(\beta,t)}{S_j^{(0)}(\beta,t)} \right\} dN_i(t) = 0. \tag{27}$$

Define the information to be $I_j(\beta) = -\partial U_j(\beta)/\partial\beta$. The submodel of selected important feature is given by

$$\widehat{\mathscr{M}_{\nu_n}} = \{1 \le j \le p : I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j| \ge \nu_n\}.$$

Zhao and Li (2012) also proposed a practical way to choose the threshold $\widehat{\nu}_n$ such that the proposed method has control on the false positive rate, which is the proportion of unimportant features incorrectly selected, i.e., $|\widehat{\mathscr{M}_{\nu_n}} \cap \mathscr{M}_\star^c|/|\mathscr{M}_\star^c|$. The expected false positive rate can be written as

$$E\left(\frac{|\widehat{\mathscr{M}_{\nu_n}} \cap \mathscr{M}_\star^c|}{|\mathscr{M}_\star^c|}\right) = \frac{1}{p-s} \sum_{j \in \mathscr{M}_\star^c} \Pr(I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j| \ge \nu_n).$$

Zhao and Li (2012) showed that $I_j(\widehat{\beta}_j)^{1/2}\widehat{\beta}_j$ has an asymptotically standard normal distribution. Therefore, the expected false positive rate is $2(1-\Phi(\nu_n))$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal. The false positive rate decreases to 0 as p increases with n. In practice, we can first fix the number of false positives f that we are willing to tolerate, which corresponds to a false positive rate of $f/(p-s)$. Because s is unknown, we can be conservative by letting $\nu_n = \Phi^{-1}(1-f/p)$, so that the expected false positive is less than f. The choice of ν_n is also related to a false discovery rate (FDR). By definition, the FDR is $|\mathscr{M}_\star^c \cap \widehat{\mathscr{M}_{\nu_n}}|/|\widehat{\mathscr{M}_{\nu_n}}|$, which is the false positive rate multiplying by $|\mathscr{M}_\star^c|/|\widehat{\mathscr{M}_\nu}|$. Since $|\mathscr{M}_\star^c|/|\widehat{\mathscr{M}_\nu}| \le p/|\widehat{\mathscr{M}_\nu}|$, in order to control the false positive rate at level $q = f/p$, we can control the FDR at level $f/|\widehat{\mathscr{M}_{\nu_n}}|$. This proposed method is called the principled Cox sure independence screening procedure (PSIS) and we summarize the PSIS as follows.

Step 1. Fit a marginal Cox model for each feature and obtain the parameter estimate $\widehat{\beta}_j$ and variance estimate $I_j(\widehat{\beta}_j)^{-1}$.

Step 2. Fix the false positive rate $q = f/p$ and set $\nu_n = \Phi(1-q/2)$.

Step 3. Select the feature X_j if $I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j| \ge \nu_n$.

Zhao and Li (2012) showed that this PSIS enjoys the sure screening property and is able to control the false positive rate. Under regularity conditions (see Appendix in Zhao and Li (2012)), if we choose $\nu_n = \Phi^{-1}(1-q/2)$, and $\log p = O(n^{1/2-\kappa})$ for some $\kappa < 1/2$, then there exists constants $c_1, c_2 > 0$ such that

$$\Pr(\mathscr{M} \subset \widehat{\mathscr{M}_{\nu_n}}) \ge 1 - s\exp(-c_1 n^{1-2\kappa})$$

and

$$E\left(\frac{|\widehat{\mathscr{M}_{\nu_n}} \cap \mathscr{M}_\star^c|}{|\mathscr{M}_\star^c|}\right) \leq q + c_2 n^{-1/2}.$$

Distinguished from marginal screening procedure in Fan, Feng and Wu (2010) and Zhao and Li (2012), Yang, Yu, Li and Buu (2016) proposed a joint screening procedure based on the joint likelihood for the Cox's model. They considered the constrained partial likelihood

$$\widehat{\beta}_m = \arg\max_\beta \ell(\beta) \text{ subject to } \|\beta\|_0 \leq m, \tag{28}$$

where $\ell(\beta)$ is defined in (25) and m is some pre-specified integer and is assumed to be greater than the number of nonzero elements in the true parameter $\beta^\star$. The constraint $\|\beta^\star\|_0 \leq m$ guarantees that the solution $\widehat{\beta}_m$ is sparse. However, it is almost impossible to solve the constrained problem (28) in the high-dimensional setting directly. Alternatively, one can approximate the likelihood function by its Taylor expansion. Let γ be in the neighborhood of β, then

$$\ell(\gamma) \approx \ell(\beta) + (\gamma-\beta)^\top \ell'(\beta) + \frac{1}{2}(\gamma-\beta)^\top \ell''(\beta)(\gamma-\beta), \tag{29}$$

where $\ell'(\beta)$ and $\ell''(\beta)$ are the first and second gradient of $\ell(\beta)$, respectively. When $p > n$, the Hessian matrix $\ell''(\beta)$ is not invertible. To deal with the singularity of $\ell''(\beta)$ and save computational costs, Yang, Yu, Li and Buu (2016) further approximated $\ell(\gamma)$ only including the diagonal elements in $\ell''(\beta)$,

$$g(\gamma|\beta) = \ell(\beta) + (\gamma-\beta)^\top \ell'(\beta) - \frac{u}{2}(\gamma-\beta)^\top \mathbf{W}(\gamma-\beta), \tag{30}$$

where u is a scaling constant to be specified and $\mathbf{W}$ is a diagonal matrix with $\mathbf{W} = -\text{diag}\{\ell''(\beta)\}$. Then the original problem can be approximated by

$$\max_\gamma g(\gamma|\beta) \text{ subject to } \|\gamma\|_0 \leq m. \tag{31}$$

Since $\mathbf{W}$ is a diagonal matrix, there is a closed form solution to (31) and thus the computational cost is low. In fact, the maximizer of $g(\gamma|\beta)$ without the constraint is

$$\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)^\top = \beta + u^{-1}\mathbf{W}^{-1}\ell'(\beta).$$

Denote the order statistics of $\tilde{\gamma}_j$ by $|\tilde{\gamma}_{(1)}| \geq |\tilde{\gamma}_{(2)}| \geq \cdots \geq |\tilde{\gamma}_{(p)}|$. The solution to (31) is given by $\widehat{\gamma} = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p)^\top$ with $\widehat{\gamma}_j = \tilde{\gamma}_j \mathbf{1}\{|\tilde{\gamma}_j| > |\tilde{\gamma}_{(m+1)}|\} := H(\tilde{\gamma}_j; m)$. We summarize the joint feature screening as follows.

Step 1. Initialize $\beta^{(0)} = \mathbf{0}$.
Step 2. Set $t = 0, 1, 2, \ldots$ and iteratively conduct Step 2a and Step 2b until the algorithm converges.

Step 2a. Compute $\tilde{\gamma}^{(t)}$ and $\tilde{\beta}^{(t)}$ where $\tilde{\gamma}^{(t)} = \beta^{(t)} + u_t^{-1}\mathbf{W}^{-1}(\beta^{(t)})\ell'(\beta^{(t)})$, and $\tilde{\beta}^{(t)} = (H(\tilde{\gamma}_1^{(t)};m),\ldots,H(\tilde{\gamma}_p^{(t)};m))^\top$. Set $\mathscr{M}_t = \{j : \tilde{\beta}_j^{(t)} \neq 0\}$.

Step 2b. Update β by $\beta^{(t+1)} = (\beta_1^{(t+1)},\ldots,\beta_p^{(t+1)})^\top$ as follows. If $j \notin \mathscr{M}_t$, set $\beta_j^{(t+1)} = 0$; otherwise, set $\{\beta_j^{(t+1)} : j \in \mathscr{M}_t\}$ be the maximum partial likelihood estimate of the submodel $\mathscr{M}_t$.

This procedure is referred to as sure joint screening (SJS) procedure. Yang, Yu, Li and Buu (2016) showed the sure screening property of the SJS under proper regularity conditions. This SJS is expected to perform better than the marginal screening procedure when there are features that are marginally independent of the survival time, but not jointly independent of the survival time. In practical implementation, Yang, Yu, Li and Buu (2016) suggested setting $m = \lfloor n/\log n \rfloor$ in practice based on their numerical studies.

Acknowledgements:

This work was supported by a NSF grant DMS 1820702 and NIDA, NIH grant P50 DA039838. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF, NIH or NIDA.

References

Candes, E. and Tao, T. (2007), 'The dantzig selector: Statistical estimation when p is much larger than n', *The Annals of Statistics* **35**(6), 2313–2351.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997), 'Generalized partially linear single-index models', *Journal of the American Statistical Association* **92**(438), 477–489.

Cheng, M.-Y., Honda, T., Li, J. and Peng, H. (2014), 'Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data', *The Annals of Statistics* **42**(5), 1819–1849.

Chu, W., Li, R. and Reimherr, M. (2016), 'Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data', *The Annals of Applied Statistics* **10**(2), 596.

Cox, D. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **34**(2), 87–22.

Cui, H., Li, R. and Zhong, W. (2015), 'Model-free feature screening for ultrahigh dimensional discriminant analysis', *Journal of the American Statistical Association* **110**(510), 630–641.

Fan, J. and Fan, Y. (2008), 'High dimensional classification using features annealed independence rules', *The Annals of Statistics* **36**(6), 2605.

Fan, J., Feng, Y. and Song, R. (2011), 'Nonparametric independence screening in sparse ultra-high-dimensional additive models', *Journal of the American Statistical Association* **106**(494), 544–557.

Fan, J., Feng, Y. and Wu, Y. (2010), High-dimensional variable selection for cox's proportional hazards model, *in* 'Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown', Institute of Mathematical Statistics, pp. 70–86.

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J. and Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Fan, J. and Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**(1), 101.

Fan, J., Ma, Y. and Dai, W. (2014), 'Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models', *Journal of the American Statistical Association* **109**(507), 1270–1284.

Fan, J., Samworth, R. and Wu, Y. (2009), 'Ultrahigh dimensional feature selection: Beyond the linear model', *The Journal of Machine Learning Research* **10**, 2013–2038.

Fan, J. and Song, R. (2010), 'Sure independence screening in generalized linear models with np-dimensionality', *The Annals of Statistics* **38**(6), 3567–3604.

Fan, J. and Zhang, W. (2008), 'Statistical methods with varying coefficient models', *Statistics and its Interface* **1**(1), 179.

Freund, Y. and Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences* **55**(1), 119–139.

Hardle, W., Hall, P. and Ichimura, H. (1993), 'Optimal smoothing in single-index models', *The Annals of Statistics* **21**(1), 157–178.

Hardle, W., Liang, H. and Gao, J. (2012), *Partially linear models*, Springer Science & Business Media.

Huang, D., Li, R. and Wang, H. (2014), 'Feature screening for ultrahigh dimensional categorical data with applications', *Journal of Business & Economic Statistics* **32**(2), 237–244.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004), 'Polynomial spline estimation and inference for varying coefficient models with longitudinal data', *Statistica Sinica* **14**, 763–788.

Huber, P. J. (1964), 'Robust estimation of a location parameter', *The Annals of Mathematical Statistics* **35**(1), 73–101.

Li, R., Zhong, W. and Zhu, L. (2012), 'Feature screening via distance correlation learning', *Journal of the American Statistical Association* **107**(499), 1129–1139.

Liu, J., Li, R. and Wu, R. (2014), 'Feature selection for varying coefficient models with ultrahigh-dimensional covariates', *Journal of the American Statistical Association* **109**(505), 266–274.

Luo, X., Stefanski, L. A. and Boos, D. D. (2006), 'Tuning variable selection procedures by adding noise', *Technometrics* **48**(2), 165–175.

Mai, Q. and Zou, H. (2012), 'The Kolmogorov filter for variable screening in high-dimensional binary classification', *Biometrika* **100**(1), 229–234.

Mai, Q. and Zou, H. (2015), 'The fused Kolmogorov filter: a nonparametric model-free screening method', *The Annals of Statistics* **43**(4), 1471–1497.

Meier, L., Van de Geer, S. and Bühlmann, P. (2009), 'High-dimensional additive modeling', *The Annals of Statistics* **37**(6B), 3779–3821.

Song, R., Yi, F. and Zou, H. (2014), 'On varying-coefficient independence screening for high-dimensional varying-coefficient models', *Statistica Sinica* **24**(4), 1735.

Székely, G. J. and Rizzo, M. L. (2014), 'Partial distance correlation with methods for dissimilarities', *The Annals of Statistics* **42**(6), 2382–2412.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics* **35**(6), 2769–2794.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288.

Vapnik, V. (2013), *The Nature of Statistical Learning Theory*, Springer science & business media.

Wang, L., Li, H. and Huang, J. Z. (2008), 'Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements', *Journal of the American Statistical Association* **103**(484), 1556–1569.

Wu, Y., Boos, D. D. and Stefanski, L. A. (2007), 'Controlling variable selection by the addition of pseudovariables', *Journal of the American Statistical Association* **102**(477), 235–243.

Xu, C. and Chen, J. (2014), 'The sparse mle for ultrahigh-dimensional feature screening', *Journal of the American Statistical Association* **109**(507), 1257–1269.

Xu, P., Zhu, L. and Li, Y. (2014), 'Ultrahigh dimensional time course feature selection', *Biometrics* **70**(2), 356–365.

Yang, G., Yu, Y., Li, R. and Buu, A. (2016), 'Feature screening in ultrahigh dimensional Cox's model', *Statistica Sinica* **26**, 881.

Yousuf, K. (2018), 'Variable screening for high dimensional time series', *Electronic Journal of Statistics* **12**(1), 667–702.

Yousuf, K. and Feng, Y. (2018), 'Partial distance correlation screening for high dimensional time series', *arXiv preprint arXiv:1802.09116* .

Zhang, C.-H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics* **38**(2), 894–942.

Zhao, S. D. and Li, Y. (2012), 'Principled sure independence screening for Cox models with ultra-high-dimensional covariates', *Journal of Multivariate Analysis* **105**(1), 397–411.

Zhong, W. and Zhu, L. (2015), 'An iterative approach to distance correlation-based sure independence screening', *Journal of Statistical Computation and Simulation* **85**(11), 2331–2345.

Zhu, L., Li, L., Li, R. and Zhu, L. (2011), 'Model-free feature screening for ultrahigh-dimensional data', *Journal of the American Statistical Association* **106**(496), 1464–1475.