Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Ultrahigh dimensional feature screening via projection

Xingxiang Li, Guosheng Cheng *, Liming Wang, Peng Lai, Fengli Song

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China

ARTICLE INFO

Article history:
Received 18 May 2016
Received in revised form 5 March 2017
Accepted 11 April 2017
Available online xxxx

Keywords:
Feature screening
Projection theory
Sure screening property
Ranking consistency property

ABSTRACT

This work is concerned with feature screening for linear model with multivariate responses and ultrahigh dimensional covariates. Instead of utilizing the correlation between every response and covariate, the linear space spanned by the multivariate responses is considered in this paper. Based on the projection theory, each covariate is projected on the linear space spanned by the multivariate responses, and a new screening procedure called projection screening (PS) is proposed. The sure screening and ranking consistency properties are established under some regular conditions. To solve some difficulties in marginally feature screening for linear model and enhance the screening performance of the proposed procedure, an iterative projection screening (IPS) procedure is constructed. The finite sample properties of the proposed procedure are assessed by Monte Carlo simulation studies and a real-life data example is analysed.

1. Introduction

Recent rapid advances of scientific techniques have led to data explosion in many fields, where ultrahigh dimensionality of predictors becomes a significant character; examples can be seen in genomics, imaging and finance, to name but a few. When the number of predictors p is much larger than the sample size n, many statistical methods cannot be applied. To make the underdetermined statistical inference possible for ultrahigh dimensional problems, sparsity assumption that only a small set of important variables contributes to the response was proposed. This leads that feature screening and variable selection play important roles in ultrahigh dimensional problems.

There are numerous statistical literatures related to variable selection for various models, such as the Lasso (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the adaptive Lasso (Zou, 2006), the Dantzig selector (Candes and Tao, 2007), the MCP (Zhang, 2010) and so on. However, the methods introduced above may not perform well when $\log(p) = O(n^a)$, for some $a \in (0, 1/2)$ for ultrahigh dimensional data because of the simultaneous challenges of computational cost, statistical accuracy and algorithmic stability (Fan et al., 2009).

These challenges call for new statistical techniques for ultrahigh dimensional data. Marginal feature screening becomes indispensable and has caused much attention since that Fan and Lv (2008) proposed the feature screening approach for the linear model called sure independence screening (SIS) and demonstrated that SIS could theoretically filter out many irrelevant variables and keep all relevant variables. Fan et al. (2009) proposed a nonparametric independence screening (NIS) procedure for additive models based on B-splines method. Fan and Song (2010) developed a maximum marginal likelihood screening procedure for generalized linear models. Zhu et al. (2011) proposed a sure independent ranking and screening (SIRS) procedure to screen significant predictors in multi-index models under some linearity assumptions, and the screening procedure satisfied the ranking consistency property. Li et al. (2012a) proposed a robust rank correlation screening

* Corresponding author. Fax: +86 25 58731160.
E-mail address: chenggs@nuist.edu.cn (G. Cheng).

http://dx.doi.org/10.1016/j.csda.2017.04.006
0167-9473/

procedure based on the Kendall τ rank correlation for semiparametric models such as transformation regression models and single-index models. Mai and Zou (2012) proposed a variable screening method for binary classification with ultrahigh dimensional predictors based on the Kolmogorov–Smirnov test. Liu et al. (2014) developed a marginal sure screening procedure for varying coefficient models based conditional Pearson correlation. Lai et al. (2017) proposed a feature screening technique for the ultrahigh dimensional data with responses missing at random.

All the aforementioned screening procedures only handle univariate responses. However, feature screening for multivariate responses or grouped predictors is often of great interest in some scientific fields, such as pathway analyses. Li et al. (2012b) developed a model-free feature screening procedure for ultrahigh dimensional data based on distance correlation (DC), which could be used for multiple responses and grouped predictors. DC is a measurement to evaluate the dependence relationship between two random vectors, but its computational cost may be expensive when sample size n or dimension of random vector is large. Investigating the relationship between the response and predictor in linear model is an extremely important and widely studied statistical problem, from the point of view of both practical applications and theory. In this paper, we consider the ultrahigh dimensional linear model with multivariate responses and aim to propose a new feature screening procedure. In order to measure the relationship between the predictor and multivariate responses simultaneously, we attempt to construct a new screening index using the projection of each predictor onto the space spanned by multivariate responses. To make each predictor on the same criterion for measurement, we standardize each predictor with mean 0 and variance 1 and use the norm-squared of projection as a screening index. Then we show that new index shares the sure screening property under certain conditions. Similarly to the iterative SIS and SIRS procedures (Fan and Lv, 2008; Zhu et al., 2011), we also propose an iterative algorithm of our new screening method. This is due to the fact that irrelevant variables which are highly correlated with the relevant variables can have a high priority for being selected in marginal screening procedure and a relevant variable can be marginally uncorrelated but jointly correlated with the response. The iterative procedure is used to resolve this issue effectively. Computationally, the proposed screening procedure is very simple and fast to implement.

The rest of this paper is organized as follows. In Section 2, we describe the methodological details of the PS and further study its theoretical property. In Section 3, the finite sample performance is studied by Monte Carlo simulations and a real data analysis. Section 4 concludes the paper. All proofs are given in the Appendix.

2. Projection Screening (PS)

2.1. Method

Consider linear model with multivariate responses

$$\mathbf{y} = \mathbf{B}^\top \mathbf{X} + \boldsymbol{\varepsilon},$$

where response vector $\mathbf{y} = \{Y_1, \ldots, Y_q\}^\top$, $\mathbf{B} = (\beta_{jk})^{p\times q}, j = 1, \ldots, p, k = 1, \ldots, q$, is a matrix of coefficients, $\mathbf{X} = \{X_1, \ldots, X_p\}^\top$ is p-dimensional covariate vector, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_q)^\top$ is the random error vector with mean $\mathbf{0}^{(q\times 1)}$, and each ε_k has finite variance $\sigma_k^2, k = 1, \ldots, q$. Without loss of generality, we assume that $E(X_j) = 0$ and $Var(X_j) = 1$ for $j = 1, \ldots, p$, and this leads that $E(Y_j) = 0$.

This paper aims to investigate the correlation between $\mathbf{y}$ and $X_j, j = 1, \ldots, p$. Instead of considering the correlation between response $Y_k, k = 1, \ldots, q$ and covariate individually, our paper is concerned with correlation between the linear space $\mathcal{H}$ spanned by the multivariate responses and covariate. $\mathcal{H} = \{\mathbf{a}^\top \mathbf{y} : \text{ for any } \mathbf{a} \in \mathbb{R}^q\}$. Assume that there is no collinearity among Y_k's, then the dimension of $\mathcal{H}$ is identically equal to q. More strictly, we assume $\mathbf{B}^\top\mathbf{B}$ is non-singular. Based projection theory, we can project each covariate on the linear space $\mathcal{H}$ and use the projection to construct a screening index. The projection of X_j onto the linear space $\mathcal{H}$ is given by

$$X_j^* = E(X_j\mathbf{y}^\top)(E\mathbf{y}\mathbf{y}^\top)^{-1}\mathbf{y},$$

which leads us to utilize the norm-squared of this projection as a marginal utility screening index

$$\omega_j = E(X_j\mathbf{y}^\top)(E\mathbf{y}\mathbf{y}^\top)^{-1}E(\mathbf{y}X_j). \tag{2.1}$$

Intuitively, if X_j and all Y_k's, $k = 1, \ldots, q$ are linearly independent, then $X_j \perp \mathcal{H}$ and $\omega_j = 0$. On the other hand, if X_j and some Y_k's are linearly correlated, there exists $\mathbf{a}_j \neq \mathbf{0}^{(q\times 1)}$ such that $X_j^* = \mathbf{a}_j^\top\mathbf{y}$, hence ω_j must be positive. This remarkable property allows us to utilize ω_j to conduct a feature screening procedure for linear model with multivariate responses.

We denote the sample design matrix of $\mathbf{X}$ as $\mathbf{U} = (U_1, \ldots, U_p)$, where $U_j = (X_{1j}, \ldots, X_{nj})^\top, j = 1, \ldots, p$. Let $\mathbf{V} = (V_1, \ldots, V_q)$ denote the sample matrix of $\mathbf{y}$, where $V_k = (Y_{1k}, \ldots, Y_{nk})^\top, k = 1, \ldots, q$. We assume that the estimator of $(E\mathbf{y}\mathbf{y}^\top)^{-1}$ exists when $q \leq n$. We next derive the sample estimate of ω_j to rank all the predictors,

$$\hat{\omega}_j = \left(\frac{1}{n}U_j^\top\mathbf{V}\right)\left(\frac{1}{n}\mathbf{V}^\top\mathbf{V}\right)^{-1}\left(\frac{1}{n}\mathbf{V}^\top U_j\right). \tag{2.2}$$

Please cite this article in press as: Li, X., et al., Ultrahigh dimensional feature screening via projection. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.04.006