ARTICLE IN PRESS

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

Fused mean–variance filter for feature screening

Niansheng Tang *, Xiaodong Yan, Jinhan Xie, Xianwen Ding, Zhiqiang Wang

Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming 650091, PR China

ARTICLE INFO

Article history:
Received 5 February 2017
Received in revised form 14 October 2017
Accepted 20 October 2017
Available online xxxx

Keywords:
Fused mean–variance filter
Ranking consistency property
Slicing method
Sure screening property
Ultrahigh dimensional data

ABSTRACT

A new model-free screening approach called as the slicing fused mean–variance filter is proposed for ultrahigh dimensional data analysis. The new method has the following merits: (i) its implementation does not require specifying a regression form of predictors and response variables; (ii) it can deal with various types of covariates and response variables including continuous, discrete and categorical variables; (iii) it works well even when the covariates/random errors are heavy-tailed, or the predictors are strongly correlated, or there are outliers; (iv) it is unsensitive to the slicing scheme. Under some regularity conditions, the sure screening and ranking consistency properties are established for the proposed procedure without assuming any moment conditions on the predictors. Simulation studies are conducted to investigate the finite sample performance of the proposed procedure. A real data example is illustrated to the proposed procedure.

1. Introduction

Ultrahigh dimensional data are often encountered in genomics, bioinformatics, proteomics and economics. In ultrahigh dimensional data analysis, it is commonly assumed that the number of explanatory variables may grow exponentially with sample size but only a small number of explanatory variables contribute to response variable. To this end, various model-based feature screening approaches have been proposed to simultaneously estimate a sparse model and select the significant predictors for ultrahigh dimensional data. For example, Fan and Lv (2008) proposed a sure independent screening (SIS) procedure and an iterated sure independence screening (ISIS) procedure in linear regression models with Gaussian covariates and responses by ranking the marginal Pearson correlations; Fan and Song (2010) extended the SIS procedure to generalized linear models, and presented a more general version of the independent learning by ranking the maximum marginal likelihoods or the maximum marginal likelihood estimates; Fan et al. (2011) developed a nonparametric independence screening (NIS) method by ranking the importance of predictors via the magnitude of nonparametric components in sparse ultrahigh dimensional additive models; Chang et al. (2013) proposed a screening method for linear regression models and generalized linear models based on the marginal empirical likelihood ratio. The aforementioned screening methods only work well when the posited working models are correctly specified (Zhu et al., 2011), but they perform poorly in the presence of model misspecification.

To address the aforementioned issue for ultrahigh dimensional data analysis, some model-free feature screening procedures have been developed in recent years. For example, Zhu et al. (2011) proposed a sure independent ranking and screening (SIRS) procedure to screen the significant explanatory variables under a unified model framework, which includes a lot of widely used parameter and nonparametric models; Li et al. (2012a, b) proposed a robust rank correlation screening (RCS) procedure based on the Kendall τ correlation coefficient between response and explanatory variable;

* Correspondence to: Department of Statistics, Yunnan University, Kunming 650091, PR China
E-mail address: nstang@ynu.edu.cn (N. Tang).

https://doi.org/10.1016/j.csda.2017.10.008
0167-9473/

Li et al. (2012a, b) developed a SIS procedure using the distance correlation between two random variables to replace the Pearson correlation in marginal correlation screening; He et al. (2013) presented a quantile-adaptive-based nonlinear independence screening procedure (QAS); Mai and Zou (2013) proposed a sure feature screening procedure based on the Kolmogorov distance for binary classification problems, but the Kolmogorov filter screening is unavailable when response variable takes more than two values. Recently, Cui et al. (2015) developed a new marginal feature screening procedure for ultrahigh dimensional discriminant analysis problem based on the empirical conditional distribution function (MVS), which is easily implemented without involving the numerical optimization and is robust to model misspecification, heavy-tailed distributions of explanatory variables and outliers, but they only studied the scenario that response variable is categorical and explanatory variables are continuous. To address the shortcomings of Mai and Zou (2013) and Cui et al. (2015), Mai and Zou (2015) proposed a nonparametric model-free screening procedure based on the fused Kolmogorov filter (FKF) together with the slicing technique. The FKF screening procedure works well for many types of covariates and response variable including continuous, discrete and categorical variables, and is invariant under univariate monotone transformation of variables. But the FKF screening procedure is computationally heavy in that calculating the Kolmogorov–Smirnov statistic involves the numerical optimization problem.

In this article, our main purpose is to develop an effective and computationally feasible feature screening procedure for ultrahigh dimensional data analysis. The proposed screening procedure can be available for various types of covariates and response variable including discrete, categorical and continuous variables, and is robust to model misspecification, outliers and heavy-tailed distributions of explanatory variables, and is also model-free without specifying a regression model of explanatory variables and response variable and is easily implemented without involving the numerical optimization problem. To this end, we propose a marginal slicing feature screening procedure, which is referred to as the slicing fused mean–variance (FMV) screening, based on the empirical conditional distribution function of explanatory variable given response variable. We study its asymptotic properties and show the sure screening and ranking consistency properties under some regularity conditions.

The rest of this article is organized as follows. The slicing FMV screening method is introduced in Section 2. Section 3 studies its theoretical properties under some regularity conditions. Simulation studies are conducted to investigate the finite sample performance of the proposed method in Section 4. In Section 5, a real data example is used to illustrate the proposed screening procedure. Technical details are presented in the Appendix.

2. Method

2.1. Slicing fused mean–variance screening method

Let Y be the categorical response with R classes $\{y_1, \ldots, y_R\}$, and X be the continuous explanatory variable with the support $\mathbb{R}_X$. Define $F(x) = \Pr(X \le x)$ as the unconditional distribution function of X, and $F_r(x) = \Pr(X \le x|Y = y_r)$ as the conditional distribution function of X given $Y = y_r$. An explanatory variable X is independent of response variable Y if and only if $F_r(x) = F(x)$ for any $x \in \mathbb{R}_X$ and $r = 1, \ldots, R$. Due to the aforementioned fact, Cui et al. (2015) considered the following index

$$\mathrm{MV}(X|Y) = E_X[\mathrm{var}_Y\{F(X|Y)\}]$$

for measuring the dependence between X and Y, and showed that $\mathrm{MV}(X|Y) = \sum_{r=1}^{R} p_r \int \{F_r(x) - F(x)\}^2 dF(x)$ and $\mathrm{MV}(X|Y) = 0$ if and only if X and Y are statistically independent, where $F(x|Y) = \Pr(X \le x|Y)$ and $p_r = \Pr(Y = y_r) > 0$ for $r = 1, \ldots, R$. Given the observed data set $\{(X_i, Y_i) : i = 1, \ldots, n\}$, an empirical estimator of $\mathrm{MV}(X|Y)$ is given by

$$\widehat{\mathrm{MV}}(X|Y) = \frac{1}{n}\sum_{r=1}^{R}\sum_{j=1}^{n} \hat{p}_r\{\hat{F}_r(X_j) - \hat{F}(X_j)\}^2,$$

where $\hat{p}_r = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = y_r)$ in which $I(\cdot)$ is an indicative function, $\hat{F}(x) = \frac{1}{n} I(X_i \le x)$, and $\hat{F}_r(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x, Y_i = y_r)/\hat{p}_r$.

Motivated by Cui et al. (2015), our main purpose is to extend their developed method for a categorical response to a continuous response variable or a general categorical response variable Y taken countable values (e.g., Poisson random variable) with the support $\mathbb{R}_Y$. To this end, we define the following index

$$\begin{aligned}\mathrm{MV}_j &= E_{X_j}[\mathrm{var}_Y\{F(X_j|Y)\}] \\ &= \iint \{F_j(x|Y = y) - F_j(x)\}^2 \mathrm{d}F_j(x)\mathrm{d}F_Y(y)\end{aligned} \tag{2.1}$$

for measuring the dependence between X_j and Y, where $F_Y(y) = \Pr(Y \le y)$, $F_j(x) = \Pr(X_j \le x)$, and $F_j(x|Y = y)$ represents the conditional distribution function of X_j given Y evaluated at $Y = y$. It is easily shown that $\mathrm{MV}_j = 0$ if and only if X_j is independent of Y, which implies that we can use MV_j to identify the significant explanatory variables in ultrahigh dimensional data analysis.

It is rather difficult to compute MV_j when $F_j(x)$ or $F_Y(y)$ are unknown. Following the widely adopted idea, we use its empirical version to estimate MV_j. Thus, when Y is a categorical response having a growing number of classes in the order