ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Robust conditional nonparametric independence screening for ultrahigh-dimensional data

Shucong Zhang [a], Jing Pan [b,*], Yong Zhou [c]

[a] School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China
[b] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[c] School of Statistics, East China Normal University, Shanghai, China

ARTICLE INFO

Article history:
Received 27 November 2017
Received in revised form 24 July 2018
Accepted 5 August 2018
Available online xxxx

Keywords:
Feature screening
Semivarying coefficient models
Sure screening property
Ultrahigh-dimensional

ABSTRACT

This article novelly proposes a robust model-free screening procedure, which performs well for a variety of semivarying coefficient models. Under technical conditions, we show that it possesses the ranking consistency property and the sure screening property. Comprehensive simulation studies are conducted to demonstrate that it exhibits more competitive performance than existing screening methods.

1. Introduction

With major advances of data acquisition and storage technologies, ultrahigh dimensionality has become a typical data feature in statistical research fields, for example, gene expression microarray data, biomedical imaging data and so on. Following the sparsity principle, recent years have witnessed a large variety of well-developed variable selection methods including Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive Lasso (Zou, 2006) and Dantzig selector (Candes and Tao, 2007). Nevertheless, for ultrahigh-dimensional data, the dimensionality p diverges at an exponential rate with the sample size n. The aforementioned variable selection methods may not perform well in such scenario due to simultaneous challenges of computational expediency, statistical accuracy, and algorithm stability (Fan et al., 2009).

Since the sure independence screening (SIS) was innovatively introduced by Fan and Lv (2008), SIS has been further extended to many important model settings including generalized linear model (Fan and Song, 2010), additive model (Fan et al., 2011), varying coefficient model (Fan et al., 2014; Liu et al., 2014a), and general nonparametric model (Feng et al., 2018). Moreover, as SIS is essentially equivalent to a Pearson correlation learning method, various extensions of correlation measures have been proposed, such as robust rank correlation (Li et al., 2012a), distance correlation (Li et al., 2012b) and conditional distance correlation (Wen et al., 2018). See Liu et al. (2015) for a comprehensive overview of independence screening methods.

As we all know, incorporating exposure or confounding variables into feature screening procedures can improve statistical accuracy and interpretability. In the present literatures, NIS (Fan et al., 2014) and CC-SIS (Liu et al., 2014a) solve problems of variable screening very well under the varying coefficient model setting, but these two methods cannot characterize the conditional nonlinear relationship between the response and predictors given the exposure variable.

* Corresponding author.
E-mail address: panjing1233@163.com (J. Pan).

https://doi.org/10.1016/j.spl.2018.08.003
0167-7152/

Moreover, CDC-SIS (Wen et al., 2018) can be adjusted for the confounding variable, but it may lose efficacy when there exists extreme values in the response.

Based on the above consideration, our article aims to propose a new robust conditional nonparametric independence screening under a more general model framework. The newly proposed method is model-free, and it is based on the conditional correlation between each predictor and the indicator function of the response given some exposure variable. Without moment conditions on the response, we establish several desirable theoretical properties for it and design corresponding numerical studies.

The rest of paper is organized as follows. In Section 2, we consider a new ranking index for feature screening under a general model framework. Section 3 provides important theoretical properties of the proposed method. In Section 4, we carry out simulation studies to evaluate the finite sample performance of our method. Technical proofs are given in the Appendix A.

2. Robust nonparametric conditional independence screening

2.1. Model settings and A new index

Let Y be the response with support Ψ_Y, U be some exposure variable with support $\mathbb{U}$ and $\mathbf{X} = (X_1, \ldots, X_p)^\top$ be the p-dimensional predictor vector. We define $F(y|\mathbf{X}, U)$ to be the entire conditional distribution function of Y given $\mathbf{X}$ and U, and assume $F(y|\mathbf{X}, U)$ depends on $\mathbf{X}$ and U only through the combinations $\boldsymbol{\beta}^\top(U)\mathbf{X}$, that is, $F(y|\mathbf{X}, U) = F(y|\boldsymbol{\beta}^\top(U)\mathbf{X})$, where $\boldsymbol{\beta}(U) = \big(\beta_1(U), \ldots, \beta_p(U)\big)^\top$ vary smoothly with U. Further define the active set as

$$\mathcal{A} = \{k : F(y|\mathbf{X}, U) \text{ depends on } X_k \text{ for some } y \in \Psi_Y, \text{ given some } U \in \mathbb{U}\},$$

and the inactive set as $\mathcal{I} = \{1, \ldots, p\} \setminus \mathcal{A}$. Accordingly, we refer to $\mathbf{X}_\mathcal{A} = (X_k, k \in \mathcal{A})^\top$ as the active predictors and $\mathbf{X}_\mathcal{I}$ as the inactive ones. $\boldsymbol{\beta}_\mathcal{A}(U)$ and $\boldsymbol{\beta}_\mathcal{I}(U)$ can be similarly defined. Our primary goal is to select a submodel with a moderate scale which can almost contain all active predictors $\mathbf{X}_\mathcal{A}$.

To represent conveniently, we assume that $\mathbf{X}$ is conditionally centralized given U, namely $\widetilde{\mathbf{X}}(U) = \mathbf{X} - \mathrm{E}(\mathbf{X}|U)$. To measure the conditional dependence between Y and $\mathbf{X}$ given U, we define

$$\boldsymbol{\Gamma}(y, U) = \mathrm{E}\left[\widetilde{\mathbf{X}}(U)F(y|\mathbf{X}, U)|U\right].$$

By simple derivations, it is obtained that $\boldsymbol{\Gamma}(y, U) = \mathrm{cov}(\widetilde{\mathbf{X}}(U), I(Y \le y)|U)$. Following the idea of the marginal regression in Fan and Lv (2008), we consider the kth component of $\boldsymbol{\Gamma}(y, U)$, denoted by $\Gamma_k(y, U)$, $k = 1, \ldots, p$. To be specific, $\Gamma_k(y, U) = \mathrm{cov}(\widetilde{X}_k(U), I(Y \le y)|U)$, where $\widetilde{X}_k(U) = X_k - \mathrm{E}(X_k|U)$. Then we define $\omega_k(U) = \mathrm{E}\left[\Gamma_k^2(Y, U)|U\right]$, and a robust ranking index for feature screening can be defined as

$$\omega_k^* = \mathrm{E}\left[\frac{\omega_k(U)}{\mathrm{var}(X_k|U)}\right]. \tag{1}$$

It can be easily seen that this new index ω_k^* in (1) is closely related to that of Fan et al. (2014). The significant difference is that ω_k^* involves the conditional covariance between X_k and the indicator function $I(Y \le y)$ given U instead of $\mathrm{cov}(X_k, Y|U)$. Further, our metric is invariant under monotone transformations of the response, hence it captures the conditional nonlinear relationship between Y and X_k given U. All this motivates us to consider the new index ω_k^* to measure the explanatory importance of each X_k for Y.

2.2. A new nonparametric conditional independence screening procedure

In this part, we propose a new nonparametric conditional independence screening procedure based on the ranking index ω_k^*. Firstly, to obtain a sample estimate of ω_k^*, the local constant estimator for those involved conditional expectations is uniformly adopted. Given a random sample $\left\{(U_i, \mathbf{X}_i^\top, Y_i)^\top, i = 1, \ldots, n\right\}$ from the population $(U, \mathbf{X}^\top, Y)^\top$, the NW estimator of $\mathrm{E}\left[X_k I(Y \le y)|U = u\right]$ is given by

$$\widehat{\mathrm{E}}\left[X_k I(Y \le y)|U = u\right] = \sum_{i=1}^n \frac{K_h(U_i - u)X_{ik}I(Y_i \le y)}{\sum_{i=1}^n K_h(U_i - u)}, \tag{2}$$

where $K_h(t) = h^{-1}K(t/h)$, $K(t)$ is a kernel function, and h is a bandwidth. Similar to (2), we can derive the kernel regression estimates $\widehat{\mathrm{E}}\left[I(Y \le y)|U\right]$, $\widehat{\mathrm{E}}(X_k|U)$ and $\widehat{\mathrm{E}}(X_k^2|U)$. By plug-in method, the sample estimators of $\mathrm{var}(X_k|U)$ and $\Gamma_k(Y, U)$ can be expressed by $\widehat{\mathrm{var}}(X_k|U) = \widehat{\mathrm{E}}(X_k^2|U) - \left[\widehat{\mathrm{E}}(X_k|U)\right]^2$ and $\widehat{\Gamma}_k(\widetilde{Y}, U) = \widehat{\mathrm{E}}\left[X_k I(Y \le \widetilde{Y})|U\right] - \widehat{\mathrm{E}}(X_k|U)\widehat{\mathrm{E}}\left[I(Y \le \widetilde{Y})|U\right]$, where $\widetilde{Y}$ is an independent copy of Y. Then the sample estimate of ω_k^* is defined by

$$\widehat{\omega}_k^* = \frac{1}{n}\sum_{i=1}^n \frac{\widehat{\omega}_k(U_i)}{\widehat{\mathrm{var}}(X_k|U_i)} = \frac{1}{n}\sum_{i=1}^n \frac{\sum_{j=1}^n \widehat{\Gamma}_k^2(\widetilde{Y}_j, U)/n}{\widehat{\mathrm{var}}(X_k|U_i)}.$$

Please cite this article in press as: Zhang S., et al., Robust conditional nonparametric independence screening for ultrahigh-dimensional data. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.08.003.