Journal of Multivariate Analysis 143 (2016) 472–480

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Robust model-free feature screening via quantile correlation

Xuejun Ma, Jingxiao Zhang *

Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, PR China

ARTICLE INFO

Article history:
Received 5 June 2015
Available online 30 October 2015

AMS 2010 subject classifications:
62G08
62G20
62H20

Keywords:
Quantile correlation
Ultrahigh-dimensionality
Sure screening
Robustness

ABSTRACT

In this paper, we propose a new sure independence screening procedure based on quantile correlation (QC-SIS). The method not only is robust against outliers, but also can discover the nonlinear relationship between independent variables and dependent variable. We establish the sure screening property under certain technical conditions. Simulation studies are conducted to assess the performances of QC-SIS, sure independent screening (SIS), sure independent ranking and screening (SIRS), robust rank correlation screening (RRCS) and distance correlation-sure independent screening (DC-SIS). Results have shown the effectiveness and the flexibility of the proposed method. We also illustrate the QC-SIS through an empirical example.

1. Introduction

Ultrahigh-dimensional data have frequently appeared in a large variety of areas such as biomedical imaging, engineering, finance and so on. Due to the challenges of computational expediency, statistical accuracy, and algorithmic stability, the traditional methods with penalization do not perform well, such as LASSO [16], SCAD [2], MCP [17] and so on. To tackle the problems, Fan and Lv [3] proposed sure independent screening (SIS) based on Pearson correlation. However, SIS is not robust against outliers or influence points, and cannot discover the nonlinear relationship between dependent variable and independent variables.

Many approaches have been proposed to make up for the issues. In summary, three types of problems are of practical interests in sure independent screening. (i) Robust screening. Li et al. [10] proposed a robust rank correlation screening (RRCS) based on the Kendall's tau correlation. (ii) Model-Free screening. Hall and Miller [7] introduced a generalized Pearson correlation to select variables. Zhu et al. [18] developed a sure independent ranking and screening (SIRS) to select significant independent variables. Li et al. [11] proposed a sure independent screening based on distance correlation (DC-SIS). (iii) Nonparametric screening. Fan et al. [1,4] and Song et al. [15] considered ultrahigh-dimensional additive models and ultrahigh-dimensional varying coefficient models by ranking the magnitude of spline approximations of the nonparametric components. Liu et al. [12] developed a conditional correlation sure independence screening (CC-SIS) based on conditional correlation for varying coefficient models with ultrahigh dimensional independent variables. Besides, Fan et al. [5], Fan and Song [6] extended SIS from a linear model to a generalized linear model by marginal regression.

In the paper, we develop a new sure independence screening procedure based on quantile correlation (QC-SIS) capable of solving the first two problems of ultrahigh-dimensional screening. Compared with RRCS and SIRS, QC-SIS takes full advantage of data, since RRCS just uses the ranking of variables, and SIRS uses the ranking of dependent variables and independent variables. Compared with SIS, DC-SIS and nonparametric screening, QC-SIS is robust against influence points.

* Corresponding author.
E-mail address: zhjxiaoruc@163.com (J. Zhang).

http://dx.doi.org/10.1016/j.jmva.2015.10.010
0047-259X/

We systematically study the theoretical properties of the QC-SIS. The effectiveness and flexibility of the proposed methods are further illustrated by numerical studies and a real data application.

The paper is organized as follows. In Section 2, we introduce the proposed method of QC-SIS, and present the theoretical results. Simulation studies and an empirical example are presented in Section 3. The article concludes with a short discussion in Section 4. The technical proofs of the main results are presented in the Appendix.

2. Method

2.1. Some preliminaries

Quantile correlation (QC) was proposed by Li et al. [9]. For random variables Z and Y,

$$\mathrm{qcor}_\tau(Y,Z)=\frac{\mathrm{qcov}_\tau(Y,Z)}{\sqrt{\mathrm{var}\{\psi_\tau(Y-Q_{\tau,Y})\}\mathrm{var}(Z)}}=\frac{\mathrm{E}\{\psi_\tau(Y-Q_{\tau,Y})(Z-\mathrm{E}Z)\}}{\sqrt{(\tau-\tau^2)\mathrm{var}(Z)}}$$

where $\tau\in(0,1)$, $\mathrm{qcov}_\tau(Y,Z)=\mathrm{cov}\{I(Y-Q_{\tau,Y}>0),Z\}=\mathrm{E}\{\psi_\tau(Y-Q_{\tau,Y})(Z-\mathrm{E}Z)\}$, $Q_{\tau,Y}$ is the τth conditional quantile of Y, $\psi_\tau(u)=\tau-I(u<0)$. If Z is independent of Y, the $\mathrm{qcor}_\tau(Y,Z)=0$. On the other hand, If Z and Y are correlated, the $\mathrm{qcor}_\tau(Y,Z)\neq 0$.

Compared with Pearson correlation, QC is robust against outliers, heavy tailed distributions and influence points. Note that quantile correlation does not enjoy the symmetry property of classical correlation, but it is effective to measure the relationship between the dependent variable and the independent variable for regression models.

2.2. A new screening procedure

Let Y be the dependent variable, and $\boldsymbol{X}=(X_1,\ldots,X_p)^\top$ be the p-dimensional independent variables. $F(y|\boldsymbol{x})=P(y|\boldsymbol{X}=\boldsymbol{x})$ denotes the conditional distribution function of y given $\boldsymbol{x}$. Define two index sets: $\mathcal{A}=\{k,F(y|\boldsymbol{x})$ functionally depends on $X_k\}$, $\mathcal{I}=\{k,F(y|\boldsymbol{x})$ does not functionally depend on $X_k\}$. Further, we define accordingly: $\boldsymbol{x}_\mathcal{A}$, a $p_1\times 1$ vector, consisting of all active predictors X_k with $k\in\mathcal{A}$, and $\boldsymbol{x}_\mathcal{I}$, a $(p-p_1)\times 1$ vector, consisting of all inactive predictors X_k with $k\in\mathcal{I}$.

To facilitate presentation, we assume $\mathrm{E}(X_k)=0$ and $\mathrm{var}(X_k)=1$ for $k=1,\ldots,p$. Let $w=(w_1,\ldots,w_p)^\top$ be a p-vector each being

$$w_k=\mathrm{E}\{\mathrm{qcor}_\tau^2(Y,X_k)\},\quad k=1,\ldots,p. \tag{1}$$

From quantile correlation, we can get that $w_k=0$ if and only if X_k is independent of Y. QC-SIS is a robust model-free screening method because it is robust, and does not require a specific model structure.

Now, we provide a brief discussion on the relationship between QC-SIS and SIRS. There exists τ such that $Q_{\tau,Y}=y$, and $\mathrm{qcor}_\tau(Y,X_k)=-\mathrm{cov}\{I(Y<y),X_k\}/\sqrt{\tau-\tau^2}$, hence, $w_{QC-SIS,k}=\mathrm{E}\{\mathrm{qcor}_\tau^2(Y,X_k)\}=\mathrm{E}\big[\frac{1}{\tau-\tau^2}\mathrm{cov}^2\{I(Y<y),X_k\}\big]$. Combining $w_{SIRS,k}=\mathrm{E}\big[\mathrm{cov}^2\{I(Y<y),X_k\}\big]$, we can get that QC-SIS is weighted SIRS. Because QC-SIS not only utilize the information of Y and X_k, but also take advantage of quantile, QC-SIS is superior to SIRS, especially with regard to complex statistical models, such as Example 3 in simulation studies.

Based on Eq. (1), the sample estimate of w_k is defined as

$$\tilde{w}_k=\frac{1}{n}\sum_{s=1}^{n}\left\{\frac{1}{\sqrt{\tau_s-\tau_s^2}}\sum_{i=1}^{n}\frac{1}{n}\psi_{\tau_s}(Y_i-\widehat{Q}_{\tau_s,Y})X_{ik}\right\}^2 \tag{2}$$

where $0<\tau_1\leq\cdots\leq\tau_n<1$. In our implementation, we set $\tau_s=\frac{s}{n+1}$, $s=1,\ldots,n$.

2.3. Theoretical properties

To facilitate asymptotic property of our proposed method, we suppose τ_j satisfies the equality $Q_{\tau_j,Y}=Y_j, j=1,\ldots,n$, then Eq. (2) can be rewritten as

$$\tilde{w}_k=\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{\sqrt{\tau_j-\tau_j^2}}\sum_{i=1}^{n}\frac{1}{n}\psi_{\tau_j}(Y_i-Y_j)X_{ik}\right\}^2.$$

As shown in the proof of Theorem 2,

$$\hat{w}_k=\frac{n^3}{n(n-1)(n-2)}\tilde{w}_k$$