Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Q1

Correlation rank screening for ultrahigh-dimensional survival data

Q2 Jing Zhang, Yanyan Liu, Yuanshan Wu *

School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China

ARTICLE INFO

Article history:
Received 17 November 2015
Received in revised form 24 August 2016
Accepted 6 November 2016
Available online xxxx

Keywords:
Censored data
Correlation rank
Model-free screening
Sure independent screening
Survival data
Ultrahigh-dimensional data

ABSTRACT

With the recent explosion of ultrahigh-dimensional data, extensive work has been carried out for screening methods which can effectively reduce the dimensionality. However, censored survival data which often arise in clinical trials and genetic studies have been left greatly unexplored for ultrahigh-dimensional scenarios. A novel feature screening procedure is proposed for ultrahigh-dimensional survival data. Also established are the ranking consistency and the sure independent screening properties. Compared with the existing methods, the proposed screening procedure is invariant to the monotone transformation, known or unknown, of the response. Moreover, it can be readily applied to ultrahigh-dimensional complete data when the censoring rate is zero. Simulation studies demonstrate that the proposed procedure exhibits favorably in comparisons with the existing ones. As an illustration, the proposed method is applied to the mantle cell lymphoma study.

1. Introduction

With the rapid advance of technology, ultrahigh-dimensional data could be collected at a relatively low cost and have appeared in various fields such as genomics, imaging and economics. Because the dimensionality p_n increases very rapidly with sample size n, existing penalized variable selection methods, such as the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), the adaptive LASSO (Zou, 2006), the Dantzig selector (Candes and Tao, 2007) and the minimax concave penalty (MCP, Zhang, 2010) may not perform well (Fan et al., 2009).

To overcome ultrahigh dimensionality, Fan and Lv (2008) proposed a sure independence screening (SIS) method to reduce the dimension in the context of linear regression models, so that penalized variable selection methods are applicable. Such screening procedures have been extensively studied in various ultrahigh-dimensional contexts, such as generalized linear models (Fan and Song, 2010) and additive models (Fan et al., 2011). Furthermore, in order to avoid the specification of a particular model structure, Zhu et al. (2011) proposed a sure independent ranking and screening (SIRS) procedure for ultrahigh-dimensional data in the framework of the general multi-index models. Li et al. (2012b) proposed a model-free SIS procedure based on the distance correlation. Using the Kendall τ, Li et al. (2012a,b) proposed a robust screening procedure Q3 in the framework of the transformation models.

For censored ultrahigh-dimensional data, Fan et al. (2010) investigated the SIS method for the Cox proportional hazards model via ranking variables according to their respective univariate partial log-likelihoods. Zhao and Li (2012) proposed a

[*] Corresponding author.
E-mail address: shan@whu.edu.cn (Y. Wu).

http://dx.doi.org/10.1016/j.csda.2016.11.005
0167-9473/© 2016 Published by Elsevier B.V.

screening method based on the standardized marginal maximum partial likelihood estimators of the Cox model, and they also provided theoretical justification for the sure independent screening property. To relax the Cox model assumption, Gorst-Rasmussen and Scheike (2013) proposed a screening procedure for a general class of single-index hazard rate models. Based on Kendall's τ and via the inverse-probability-of-censoring weighting, Song et al. (2014) proposed a censored rank independence screening method which is shown to be robust against the potential outliers and to work for a general class of survival models. Wu and Yin (2015) developed a screening method which is designed to identify the covariates that contribute to the conditional quantile of the response. Recently, Zhou and Zhu (in press) proposed a censored version of the SIRS method by incorporating the weight of the inverse probability of censoring.

In a model-free fashion, we propose a novel correlation rank sure independent screening procedure (CR-SIS), which can naturally handle ultrahigh-dimensional survival data without any nonparametric approximation except for the Kaplan–Meier estimator. Compared with the existing procedures, our approach enjoys several distinctive advantages. Our procedure does not rely on any model assumption and works generally for nonlinear survival regression models. On the other hand, our approach is invariant under the monotone transformation of the response. These advantages greatly facilitate the implementation of the proposed method in real applications.

The rest of the article is organized as follows. In Section 2, we propose the CR-SIS procedure for both ultrahigh-dimensional complete and censored data. In Section 3, we establish the theoretical properties of the proposed procedure. Its finite-sample performances are evaluated in Section 4 via extensive simulation studies. In Section 5, we apply the proposed method to a recent study on mantle cell lymphoma. Section 6 concludes some remarks. All technical proofs are presented in the Appendix.

2. Screening procedures

Consider the conditional distribution function,

$$F(y|\mathbf{Z}) = P(Y \le y|\mathbf{Z}),$$

where Y denote the response variable and $\mathbf{Z} = (Z_1, \ldots, Z_{p_n})^{\mathrm{T}}$ the covariate vector. In an ultrahigh-dimensional setting, the dimensionality p_n, possibly depending on and greatly exceeding the sample size n, might increase with n at an exponential rate. To identify which covariates among the p_n ones contribute to the conditional distribution function of Y given $\mathbf{Z}$, we define the active covariate set as

$$\mathcal{A} = \{k : F(y|\mathbf{Z}) \text{ depends on } Z_k,\ k = 1, \ldots, p_n\}.$$

Without loss of generality, we assume throughout this article that $E(Z_k) = 0$ for $k = 1, \ldots, p_n$. Let $G(y) = P(Y \le y)$ denotes the unconditional distribution function of Y. Define $\mathbf{R}(Y) = E\{\mathbf{Z}G(Y)\}$, let $R_k(Y)$ be the kth element of $\mathbf{R}(Y)$, then $R_k(Y) = E\{Z_k G(Y)\} = \mathrm{cov}\{Z_k, G(Y)\}$, where Z_k denotes the kth element of $\mathbf{Z}$. Define

$$r_k = [R_k(Y)]^2,$$

where $k = 1, \ldots, p_n$, then r_k serves as the population version of our proposed marginal utility measure for the kth covariate.

Intuitively, the unconditional distribution function of Y, $G(y)$, compositing with Y, is expected to contain the whole information of Y. Consequently, r_k, which measures the correlation between $G(Y)$ and Z_k, could reflect the relationship between Y and Z_k. If Y and Z_k are independent, $G(Y)$ and Z_k should be independent; hence $r_k = 0$. On the other hand, it is reasonable to expect $r_k > 0$ if Y and Z_k are dependent. Under the framework of semiparametric regression, Zhu and Zhu (2009) proposed a distribution-weighted least squares estimator which can be deduced from the variant of $\mathrm{cov}\{Z_k, G(Y)\}$. Our proposed marginal utility r_k shares the spirit of their method. The SIRS method proposed by Zhu et al. (2011) adopted the dichotomous $I(Y < y)$ variable, while we use $G(y)$, which is continuous and thus expected to contain the whole information of Y. The correlation between $G(Y)$ and Z_k could be consequently appropriate to reflect the relationship between Y and Z_k. Furthermore, our method can naturally handle ultrahigh-dimensional survival data without any nonparametric approximation except for the routine Kaplan–Meier estimator. These remarkable properties motivate us to use r_k for feature screening in ultrahigh-dimensional data. We can see in the sequel that the proposed method indeed enjoys the ranking consistency property and also performs well in different scenarios.

Given a random sample $\{Y_i, \mathbf{Z}_i \equiv (Z_{i1}, \ldots, Z_{ip_n})^{\mathrm{T}}\}, i = 1, \ldots, n$, from the population $\{Y, \mathbf{Z} = (Z_1, \ldots, Z_{p_n})^{\mathrm{T}}\}$. It is desirable to derive an estimator of r_k based on the n independent and identical observations. For ease of presentation, we assume that the sample predictors are all centralized, that is, $n^{-1}\sum_{i=1}^{n} Z_{ik} = 0$ for $k = 1, \ldots, p_n$, where Z_{ik} is the kth element of $\mathbf{Z}_i$. Obviously, we can use the empirical distribution function, which is given by

$$\widehat{G}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y),$$

to estimate $G(y)$. Therefore, we propose an estimator for r_k, which takes the form of

$$\widehat{r}_k = \left\{\frac{1}{n}\sum_{i=1}^{n} Z_{ik}\widehat{G}_n(Y_i)\right\}^2.$$

Please cite this article in press as: Zhang, J., et al., Correlation rank screening for ultrahigh-dimensional survival data. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.11.005