Statistica Sinica **33** (2023), 1809-1830
doi:https://doi.org/10.5705/ss.202020.0174

A NEW MODEL-FREE FEATURE SCREENING PROCEDURE FOR ULTRAHIGH-DIMENSIONAL INTERVAL-CENSORED FAILURE TIME DATA

Jing Zhang, Mingyue Du, Yanyan Liu and Jianguo Sun

Zhongnan University of Economics and Law, The Hong Kong Polytechnic University, Wuhan University and University of Missouri

Abstract: Screening important features based on ultrahigh-dimensional data has become an important task in statistical analysis. As such, several screening procedures have been proposed for various types of studies or data, including complete data and right-censored failure time data. In this study, we consider ultrahigh-dimensional interval-censored failure time data. Such data occur frequently in medical follow-up studies, among others, and include right-censored data as a special case, but for which few works exist. For the problem, a distance correlation-based sure independent screening procedure is proposed. The new approach is model-free and does not require estimating survival functions, unlike most existing nonparametric screening procedures for failure time data. We establish the sure screening property and the ranking consistency of the proposed method, and conduct an extensive simulation study, which suggests that the proposed procedure works well for practical situations. Finally, we apply the proposed method to a set of real data on Alzheimer's disease, which motivated this study.

Key words and phrases: Distance correlation, interval-censored data, model-free screening, sure screening property, ultrahigh-dimensional data.

1. Introduction

Screening important features based on ultrahigh-dimensional data has become an important task in statistical analysis, and various screening procedures have been proposed. For example, an early work was that of Fan and Lv (2008), who proposed a sure independence screening (SIS) procedure under the framework of a linear regression model. Many authors have since extended the SIS procedure to different models, including the generalized linear model (Fan and Song (2010)), additive model (Fan, Feng and Song (2011)), and multi-index model (Zhu et al. (2011)), and Li, Zhong and Zhu (2012) provided a distance correlation-based SIS procedure. We discuss the same problem, but unlike these prior works,

Corresponding author: Yanyan Liu, School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China. E-mail: liuyy@whu.edu.cn.

we consider a type of incomplete data, namely, interval-censored (IC) failure time data (Sun (2006)). By IC data, we mean that the failure time of interest is known, or is observed to belong to an interval, instead of being observed exactly. Such data commonly occur in fields with periodic-ups, especially medical studies, such as clinical studies. Furthermore, IC data include right-censored data as a special case, and their analysis is much more difficult than that of the latter, owing to their more complicated structures.

An example of IC data that motivated this study is that arising from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal follow-up study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). In the study, one variable of interest is the AD conversion time, defined as the time from the baseline visit date to the AD conversion. Because the participants are examined intermittently or periodically, the exact time of the AD conversion is not available, and instead only IC data are available on the variable. In other words, if it happens, the occurrence of the AD conversion is only known to be within an interval. For the covariates or factors of interest, the study consists of information on a large number of single nucleotide polymorphisms (SNPs) for each participant, with the aim of detecting SNPs that have significant effects on the risk of developing AD. For example, questions of interest include which of the SNPs are individually most associated with the AD conversion, and how one can predict the AD conversion by using the SNPs and other information.

Many authors have considered generalizations of the SIS procedure to the screening of important features based on right-censored failure time data. In general, these procedures can be classified into two types, namely, model-based methods (Tibshirani (2009); Fan, Feng and Wu (2010); Zhao and Li (2012); Gorst-Rasmussen and Scheike (2013)) and model-free methods (Song et al. (2014); Wu and Yin (2015); Zhang, Liu and Wu (2017); Zhou and Zhu (2017); Liu, Zhang and Zhao (2018); Zhang et al. (2018); Lin, Liu and Hao (2018); Zhang, Liu and Cui (2021)). However, few SIS methods have been developed for IC failure time data. Note that one simple generalization is to consider the procedures developed for right-censored data that involve the nonparametric estimation of survival functions, and then to simply replace the estimators with Turnbull's estimator for the IC data (Turnbull (1976)). One such method, the only existing screening method for IC failure time data, is that of Hu et al. (2020), who generalized the screening method for right-censored data proposed by Zhang et al. (2018) to case-II IC data. A drawback of such an approach is the significant computational burden, because there is no closed form for Turnbull's estimator.

Several regularized methods have been proposed in the literature for variable or covariate selection based on high-dimensional IC data. For example, Wu and Cook (2015) proposed such a procedure for the proportional hazards model, with the baseline hazard function being a piecewise constant function, and gave an EM algorithm for maximizing the penalized log-likelihood function. Scolas et al. (2016) extended the adaptive Lasso procedure under a flexible parametric mixture cure model structure, and Zhao et al. (2020) proposed a broken adaptive ridge regression procedure that combines the strengths of quadratic regularization and the adaptive weighted bridge shrinkage. However, these methods either do not apply or face the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability if the dimension or the number of covariates p is ultrahigh, in the sense that $p = \exp(n^{\alpha})$, with n denoting the sample size and $\alpha > 0$ (Fan, Samworth and Wu (2009)).

In the following, we propose a SIS procedure for ultrahigh-dimensional IC failure time data by employing the distance correlation (DC) between a redefined response and each predictor as the dependence measure. As pointed out by Székely, Rizzo and Bakirov (2007), the DC of two random vectors is equal to zero if and only if they are independent, and therefore can be used as a sensitive dependence measure. One major advantage of the proposed approach is that it does not involve a nonparametric estimation of the survival function or any complicated numerical optimization, and thus is easy to implement and converges quickly. Furthermore, it is model free, and thus robust to model misspecification. In addition, it applies to general and complex IC data and, thus, also to the mixture of left-, interval-, and right-censored data. We also establish the large-sample properties of the proposed method, including the sure screening property and the rank consistency.

The remainder of the article is organized as follows. First we introduce some notation and assumptions in Section 2, and also provide some background on the DC. The proposed model-free screening procedure is presented in Section 3, and in Section 4 we establish the theoretical properties of the proposed approach. Section 5 presents results obtained from simulation studies conducted to evaluate the finite-sample performance of the method, which indicate that it works well for practical situations. To determine the selection threshold value for the proposed screening procedure, we propose and investigate the performance of a data-driven log-ratio criterion in Section 6, and Section 7 applies the method to the AD example discussed above. Section 8 concludes the paper.

2. Notation, Assumptions, and DC

Consider a failure time study that consists of n independent subjects, and let T denote the failure time of interest. For each T or subject, suppose there exist two monitoring or observation times, denoted by U and V with $U < V$, and T is observed only to be in one of three situations: T is between U and V or interval-censored, T is larger than V or right-censored, or T is less than U or left-censored. That is, only IC data are available on T, often referred to as case-II IC data (Sun (2006)). Define the indicator variables $\Delta_1 = I(T < U)$, $\Delta_2 = I(U \leq T < V)$, and $\Delta_3 = 1 - \Delta_1 - \Delta_2$, and let T_i, U_i, V_i, Δ_{1i}, Δ_{2i}, and Δ_{3i} denote T, U, V, Δ_1, Δ_2, and Δ_3 respectively, defined above and associated with subject i $(i = 1, \ldots, n)$. Then, the observed data on the T_i can be summarized as $\{ U_i, V_i, \Delta_{1i}, \Delta_{2i}, \Delta_{3i} : i = 1, 2, \ldots, n \}$.

For each study subject, suppose there exists a p-dimensional vector of covariates denoted by $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^{\mathrm{T}}$, and p can be ultrahigh. Let $S(t|\boldsymbol{Z}) = P(T > t|\boldsymbol{Z})$ denote the survival function for a subject with the covariate $\boldsymbol{Z}$, and by following Song et al. (2014) and others, define

$$\mathcal{A} = \{k : S(t|\boldsymbol{Z}) \text{ functionally depends on } Z_k \text{ for } t \geq 0, k = 1, \ldots, p\},$$

the index set of the active covariates or the covariates that have some effects on T. To understand the definition above, let $\boldsymbol{Z}_{\mathcal{A}}$ denote the sub-vector of $\boldsymbol{Z}$ containing all of the active covariates or the components in $\mathcal{A}$. Then, $\mathcal{A}$ means that $S(t|\boldsymbol{Z})$ depends on $\boldsymbol{Z}$ only through $\boldsymbol{Z}_{\mathcal{A}}$ or $S(t|\boldsymbol{Z}) = S(t|\boldsymbol{Z}_{\mathcal{A}})$, for any t. Suppose that the goal is to determine or estimate $\mathcal{A}$ using a screening procedure. Because our proposed screening utility is based on the DC, we first introduce this concept. Let $\boldsymbol{u}$ and $\boldsymbol{v}$ denote two random vectors, and $\phi_{\boldsymbol{u},\boldsymbol{v}}(t, s)$, $\phi_{\boldsymbol{u}}(t)$, and $\phi_{\boldsymbol{v}}(s)$ be the characteristic functions of $(\boldsymbol{u}, \boldsymbol{v})$, $\boldsymbol{u}$, and $\boldsymbol{v}$, respectively. The DC between $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as the square root of

$$\mathrm{dcorr}^2(\boldsymbol{u}, \boldsymbol{v}) = \frac{\mathrm{dcov}^2(\boldsymbol{u}, \boldsymbol{v})}{\sqrt{\mathrm{dcov}^2(\boldsymbol{u}, \boldsymbol{u})\, \mathrm{dcov}^2(\boldsymbol{v}, \boldsymbol{v})}},$$

where $\mathrm{dcov}^2(\boldsymbol{u}, \boldsymbol{v})$ denotes the distance covariance between $\boldsymbol{u}$ and $\boldsymbol{v}$, and is given by

$$\mathrm{dcov}^2(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{c_{d_u} c_{d_v}} \int_{R^{d_u + d_v}} \frac{\|\phi_{\boldsymbol{u},\boldsymbol{v}}(t, s) - \phi_{\boldsymbol{u}}(t)\phi_{\boldsymbol{v}}(s)\|^2}{\|t\|_{d_u}^{1+d_u} \|s\|_{d_v}^{1+d_v}} \, \mathrm{d}t \, \mathrm{d}s \,.$$

In the above,

$$c_{d_u} = \pi^{(1+d_u)/2} \Big/ \Gamma\left(\frac{1+d_u}{2}\right), \quad c_{d_v} = \pi^{(1+d_v)/2} \Big/ \Gamma\left(\frac{1+d_v}{2}\right),$$

where $\Gamma(\cdot)$ denotes the gamma function, and d_u and d_v denote the dimensions of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Székely, Rizzo and Bakirov (2007) proved that $\mathrm{dcorr}^2(\boldsymbol{u}, \boldsymbol{v}) = 0$ if and only if $\boldsymbol{u}$ and $\boldsymbol{v}$ are independent. This suggests that the DC can be used as a sensitive measure of dependence, and Li, Zhong and Zhu (2012) developed a DC-based screening procedure for the complete data situation. In the following, we focus on case-II interval-censored data.

3. DC Screening Procedure

As discussed above, a natural screening utility based on the DC is $\mathrm{dcorr}^2(Z_k, T)$, the square of the DC between Z_k and T. However, $\mathrm{dcorr}^2(Z_k, T)$ cannot be estimated directly using the incomplete IC data. Consider more general data that includes interval-censored, right-censored, and left-censored observations. Note that both right-censored and left-censored observations can be viewed as special cases of interval-censored observations. Define two new variables L and H, representing the length and endpoint, respectively, of the time interval within which the event time lies, as

$$\begin{aligned} L &= \Delta_1 U + \Delta_2 (V - U) + \Delta_3 (\eta - V)\,, \\ H &= \Delta_1 \cdot U + \Delta_2 \cdot V + \Delta_3 \cdot V\,, \end{aligned}$$

where η is a large constant. In practice, any large number can be used for η, such as $\eta = 10^6$, which we use in our numerical studies. Note that $H = V$ if T is either between U and V (interval-censored) or larger than V (right-censored), and $H = U$ if T is less than U (left-censored). In other words, H represents either the left or the right endpoint of the observed interval. According to Székely, Rizzo and Bakirov (2007), the following four conditions are equivalent:

(1) $\mathrm{dcorr}^2(Z_k, (L, H)) = 0$;

(2) (L, H) and Z_k are independent;

(3) $(a \cdot L + b, c \cdot H + d)$ and Z_k are independent for any constants $a, c \neq 0$ and constants b, d;

(4) $\mathrm{dcorr}^2(Z_k, (a \cdot L + b, c \cdot H + d)) = 0$ for any constants $a, c \neq 0$ and constants b, d.

Note that if $\mathrm{dcorr}^2(Z_k,(L,H))>0$, the value of $\mathrm{dcorr}^2(Z_k,(a\cdot L+b,c\cdot H+d))$ depends on a and c when $a\neq c$. To eliminate the influence of a and c, we first standardize (L,H) marginally, as follows:

$$L^*=\frac{L-\mu_1}{\sigma_1}\,,\ H^*=\frac{H-\mu_2}{\sigma_2}\,,$$

where $\mu_1=\mathrm{E}(L)$, $\mu_2=\mathrm{E}(H)$, $\sigma_1=\mathrm{sd}(L)$, and $\sigma_2=\mathrm{sd}(H)$, and $\mathrm{E}(\cdot)$ and $\mathrm{sd}(\cdot)$ represent the expectation and standard deviation, respectively, of the corresponding variable. Define $\boldsymbol{Y}=(L^*,H^*)$ and

$$\omega_k=\mathrm{dcorr}^2(Z_k,\boldsymbol{Y})=\frac{\mathrm{dcov}^2(Z_k,\boldsymbol{Y})}{\sqrt{\mathrm{dcov}^2(Z_k,Z_k)\mathrm{dcov}^2(\boldsymbol{Y},\boldsymbol{Y})}}\,, \tag{3.1}$$

which will serve as the population quantity of the proposed marginal utility measure for ranking the dependence between the covariate Z_k and the failure time T.

A key feature of $\omega_k=\mathrm{dcorr}^2(Z_k,\boldsymbol{Y})$ is that, unlike $\mathrm{dcorr}^2(Z_k,(a\cdot L+b,c\cdot H+d))$, $\mathrm{dcorr}^2(Z_k,(a\cdot L^*+b,c\cdot H^*+d))$ does not depend on a and c. An additional discussion on their comparison is given below. Based on remark 3 in Székely, Rizzo and Bakirov (2007), $\mathrm{dcov}^2(Z_k,\boldsymbol{Y})$ can be partitioned as

$$\mathrm{dcov}^2(Z_k,\boldsymbol{Y})=S_{k1}+S_{k2}-2S_{k3}\,.$$

In the above, $S_{k1}=\mathrm{E}(\|Z_k-\widetilde{Z}_k\|_1\|\boldsymbol{Y}-\widetilde{\boldsymbol{Y}}\|_2)$, $S_{k2}=\mathrm{E}(\|Z_k-\widetilde{Z}_k\|_1)\mathrm{E}(\|\boldsymbol{Y}-\widetilde{\boldsymbol{Y}}\|_2)$, and $S_{k3}=\mathrm{E}\{\mathrm{E}(\|Z_k-\widetilde{Z}_k\|_1|Z_k)\mathrm{E}(\|\boldsymbol{Y}-\widetilde{\boldsymbol{Y}}\|_2|\boldsymbol{Y})\}$, with $(\widetilde{Z}_k,\widetilde{\boldsymbol{Y}})$ denoting an independent copy of $(Z_k,\boldsymbol{Y})$, and $\|\cdot\|_1$ and $\|\cdot\|_2$ denoting the Euclidean norm.

Let (L_i,H_i), for $i=1,\ldots,n$, be a random sample from (L,H). Let $(\widehat{\mu}_1,\widehat{\sigma}_1)$ and $(\widehat{\mu}_2,\widehat{\sigma}_2)$ denote the sample mean, and sample standard error of L_i and H_i, respectively. This yields the standardized data $\{\,\widehat{\boldsymbol{Y}}_i=(\widehat{L}_i^*,\widehat{H}_i^*):\ i=1,\ldots,n\,\}$, where

$$\widehat{L}_i^*=\frac{L_i-\widehat{\mu}_1}{\widehat{\sigma}_1}\,,\ \widehat{H}_i^*=\frac{H_i-\widehat{\mu}_2}{\widehat{\sigma}_2}.$$

Then, the empirical estimator of $\mathrm{dcov}^2(Z_k,\boldsymbol{Y})$ is given by

$$\widehat{\mathrm{dcov}}^2(Z_k,\boldsymbol{Y})=\widehat{S}_{k1}+\widehat{S}_{k2}-2\widehat{S}_{k3},$$

where

$$\widehat{S}_{k1}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|Z_{ki}-Z_{kj}|\|\widehat{\boldsymbol{Y}}_i-\widehat{\boldsymbol{Y}}_j\|_2,$$

$$\widehat{S}_{k2} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|Z_{ki} - Z_{kj}| \cdot \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\widehat{\boldsymbol{Y}}_i - \widehat{\boldsymbol{Y}}_j\|_2,$$
$$\widehat{S}_{k3} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|Z_{ki} - Z_{kl}| \cdot \|\widehat{\boldsymbol{Y}}_j - \widehat{\boldsymbol{Y}}_l\|_2.$$

Similarly, we can obtain the empirical estimators of $\widehat{\mathrm{dcov}}^2(Z_k, Z_k)$ and $\widehat{\mathrm{dcov}}^2(\boldsymbol{Y}, \boldsymbol{Y})$ and, accordingly, an empirical estimator of ω_k,

$$\widehat{\omega}_k = \frac{\widehat{\mathrm{dcov}}^2(Z_k, \boldsymbol{Y})}{\sqrt{\widehat{\mathrm{dcov}}^2(Z_k, Z_k)\widehat{\mathrm{dcov}}^2(\boldsymbol{Y}, \boldsymbol{Y})}}. \tag{3.2}$$

Based on the discussion above and the property of DC (Székely, Rizzo and Bakirov (2007)), the estimator $\widehat{\omega}_k$ is expected to fluctuate around zero if Z_k is an inactive covariate, and to be away from zero otherwise. In other words, we can select those candidate covariates with top values of $\widehat{\omega}_k$ as active covariates. This motivates the estimate of $\mathcal{A}$ or the screening procedure given by

$$\widehat{\mathcal{A}} = \{k : \widehat{\omega}_k \geq cn^{-\kappa},\ k = 1, \ldots, p\},$$

for some prespecified threshold constants c and κ, as discussed below.

In practice, it is difficult to obtain the constants c and κ or the cutoff threshold $cn^{-\kappa}$, which is used to separate the active and inactive sets. Following the thresholding rule of Fan and Lv (2008), we instead propose ranking $\{\widehat{\omega}_k,\ k = 1, \ldots, p\}$ from the largest to the smallest, and selecting the covariates based on the top ones or estimating $\mathcal{A}$ by

$$\widetilde{\mathcal{A}} = \{j : \widehat{\omega}_j \text{ is amongst the first } d_0 \text{ largest of all } \widehat{\omega}_k,\ k = 1, \ldots, p\}.$$

In the above, d_0 is a predetermined positive integer and suggested to be $d_0 = \lceil n/\log n \rceil$ by Fan and Lv (2008). This choice of d_0 has been widely adopted in the literature on screening procedures (Zhu et al. (2011); Li, Zhong and Zhu (2012); Song et al. (2014)). More discussion on this is given below, along with a new maximum log-ratio criterion.

4. Asymptotic Properties

In this section, we establish the sure screening and ranking consistency properties of the model-free screening procedure proposed in the previous sections. For this, we need the following conditions.

C1. There exists a positive constant s_0 such that for all $0 < s \leq 2s_0$, we have that

$$\sup_p \max_{1\leq k\leq p} E\{\exp(s\|G\|_1^2)\} < \infty,$$

for $G = Z_k$, and $\tilde{V}$ with $\tilde{V} = \max\{U, V\}$, V, or U for interval-censored, right-censored, or left-censored observations, respectively.

C2. The minimum distance correlation of active predictors satisfies $\min_{k\in\mathcal{A}} \omega_k \geq 2cn^{-\kappa}$, for some constants $c > 0$ and $\kappa \in [0, 1/2)$.

Condition C1 is a common assumption required by most existing screening procedures. Condition C2 requires that the values of the marginal utilities between each active variable and the response are not too small. This is a typical assumption in the literature on feature screening, and similar to condition 3 of Fan and Lv (2008), condition 2 of Li, Zhong and Zhu (2012), and conditions 2 and 5 of Wu and Cook (2015), among others. We first present the sure screening property of the proposed screening method, and then the ranking consistency property.

Theorem 1. *Under condition* C1*, for any* $0 < \gamma < 1/2 - \kappa$*, there exist positive constants* $c_1 > 0$ *and* $c_2 > 0$ *such that*

$$P\left(\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\right) \leq O\left(p\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp(-c_2 n^{\gamma})\right]\right).$$

In addition, under conditions C1 *and* C2*, we have that*

$$P\left(\mathcal{A} \subseteq \widehat{\mathcal{A}}\right) \geq 1 - O\left(a\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp(-c_2 n^{\gamma})\right]\right),$$

where $a = |\mathcal{A}|$ *is the cardinality of* $\mathcal{A}$*.*

Theorem 2. *Assume* $\omega_k = 0$ *holds for* $k \notin \mathcal{A}$*. Then, under conditions* C1 *and* C2*, there exist positive constants* $c_1 > 0$ *and* $c_2 > 0$ *such that*

$$P\left(\max_{k\notin\mathcal{A}} \widehat{\omega}_k < \min_{k\in\mathcal{A}} \widehat{\omega}_k\right) \geq 1 - O\left(p\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp(-c_2 n^{\gamma})\right]\right).$$

The proofs of the above results are provided in the Supplementary Material. Note that although the idea used here is similar to that discussed in Li, Zhong and Zhu (2012), the proof here is much more complicated. One main reason is that, unlike Li, Zhong and Zhu (2012), we have to estimate the redefined response $\boldsymbol{Y} = (L^*, H^*)$ in addition to estimating the distance correlation $\mathrm{dcorr}^2(Z_k, \boldsymbol{Y})$. More specifically, we have to obtain the exponential tail probability bound for $P(|\mathrm{dcov}^2(Z_k, \boldsymbol{Y}) - \widehat{\mathrm{dcov}}^2(Z_k, \boldsymbol{Y})| \geq \epsilon)$ in order to prove Theorem 1. For this, one

needs to compute $P(|\mathrm{dcov}^2(Z_k, \boldsymbol{Y}) - \widetilde{\mathrm{dcov}}^2(Z_k, \boldsymbol{Y})| \geq \epsilon)$ and $P(|\widehat{\widetilde{\mathrm{dcov}}}^2(Z_k, \boldsymbol{Y}) - \widetilde{\mathrm{dcov}}^2(Z_k, \boldsymbol{Y})| \geq \epsilon)$. The calculation of the former can be obtained in the same way as in Li, Zhong and Zhu (2012), but determining the latter is new. More details can be found in the Supplementary Material.

Theorem 1 guarantees that the proposed DC-based SIS will retain the active set $\mathcal{A}$, and Theorem 2 shows the reasonability or validity of the proposed dependence measure. Together, these results state that if $p = \exp\{n^\alpha\}$ with $0 \leq \alpha < (1 - 2\kappa)/3$, we have that

$$\lim_{n\to\infty} P\Big(\mathcal{A} \subseteq \widehat{\mathcal{A}}\Big) = 1, \quad \lim_{n\to\infty} P\left(\max_{k\notin\mathcal{A}} \widehat{\omega}_k < \min_{k\in\mathcal{A}} \widehat{\omega}_k\right) = 1.$$

This proves that the DC values of all active predictors can be expected to be larger than those of all inactive variables asymptotically, and all active covariates can be selected with probability approaching one as $n \to \infty$. In other words, it is reasonable to choose covariates or predictors for which the ω_k are among the d_0 largest ones.

5. Simulation Studies

To assess the finite-sample performance of the DC-based SIS procedure proposed in the previous sections, we conducted a sequence of simulation studies under different model settings and distributional assumptions of the predictor variables. Note that instead of the proposed approach, a naive alternative method is to directly set $Y = (U, V, \Delta_1, \Delta_2, \Delta_3)$, and then to build the marginal utility ω_k as above. For comparison, we also considered this alternative procedure, referred to as DC-SIS1; the proposed method is referred to as DC-SIS. To measure the performance, following Li, Zhong and Zhu (2012), we considered the following quantities:

(i) $\mathcal{S}$: the minimum model size required to include all active variables. In the tables below, we report the average value of $\mathcal{S}$ and a robust estimate of its standard deviation, denoted as MMS and RSD, respectively, where RSD is defined as IQR/1.34, with IQR denoting the interquartile range of $\mathcal{S}$ over all replications (Huang and Zhu (2016)).

(ii) $\mathcal{P}_e$: the selection proportion that each active variable is selected into the model, with the model size $d_0 = \lceil n/\log n \rceil$, where $\lceil x \rceil$ denotes the integer part of x.

(iii) $\mathcal{P}_a$: the selection proportion that all active variables are selected into the

model, with the model size d_0, as above.

It is apparent that an effective screening procedure is expected to yield $\mathcal{S}$ close to the true minimum model size and both $\mathcal{P}_e$ and $\mathcal{P}_a$ close to one.

To generate the observed data, for subject i, we first randomly generated a sequence of observation times from the uniform distribution $\mathrm{U}(0, \tau)$, denoted by $t_{1i} < t_{2i} < \cdots < t_{mi}$. If $T_i < t_{1i}$, we set the censoring type for subject i as left-censored, and if $T_i > t_{mi}$, we set the censoring type for subject i as right-censored. Otherwise, we set the censoring type as interval-censored, and U_i and V_i as the largest observation time smaller than T_i and the smallest observation time greater than T_i, respectively. The constant τ, which controls the percentages of left-censored, interval-censored, and right-censored observations, was chosen to yield $20\%, 60\%$, and 20%, respectively. The failure time of interest is assumed to follow the Cox proportional hazards model, the nonlinear model, or the general transformation model, as described below. The results given below are based on $p = 2000$ or 4000, $n = 100$ or 200, and $\eta = 10^6$, with 500 replications.

Setup 1. In this case, the failure time of interest T_i was generated from the Cox proportional hazards model given by

$$\lambda(t|\boldsymbol{Z}_i) = \lambda_0(t) \exp(\boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_0)\,,$$

where $\lambda_0(t) = (t - 0.5)^2$ and $\boldsymbol{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = (0.8^{|i-j|})$, for $i, j = 1, \ldots, p$. Here, we set $\boldsymbol{\beta}_0 = (1, 0.8, 0.6, 0.4, 0.2, 0, \ldots, 0)^{\mathrm{T}}$, that is, there are five active covariates, and the index set of active covariates is $\mathcal{A} = \{1, 2, 3, 4, 5\}$. Table 1 presents the simulation results of MMS, RSD, $\mathcal{P}_e$, and $\mathcal{P}_a$ for these two procedures. The results show that the proposed procedure DC-SIS performs well for all cases, and is clearly better than DC-SIS1, especially for small sample sizes.

Setup 2. Here, we generated the failure time of interest from the nonlinear survival model with interactions given by

$$T = (2 + \sin Z_1)^2 + 0.5(1 + Z_5)^{-3} + 3(Z_{10}^2 + Z_{10}) + 0.5 Z_1 Z_{10} + \epsilon\,,$$

where $\epsilon \sim N(0, 1)$. Note that we have three active covariates in Z_1, Z_5, and Z_{10}, giving the index set of the active covariates $\mathcal{A} = \{1, 5, 10\}$. For the covariates, we considered three situations:

a) $\boldsymbol{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = (0.8^{|i-j|})$, for $i, j = 1, \ldots, p$;

b) The first component Z_1 was generated from a Bernoulli distribution with success probability 0.5, and the remaining $(p - 1)$ components $(Z_2, \ldots, Z_p)$

Table 1. The simulation results on $\mathcal{S}$ (the minimum model size needed to include all active covariates), $\mathcal{P}_e$ (the selection proportions for each active covariate), and $\mathcal{P}_a$ (the selection proportion for all active covariates) for setup 1

					$\mathcal{P}_e$					
p	n	Method	MMS	RSD	Z_1	Z_2	Z_3	Z_4	Z_5	$\mathcal{P}_a$
2,000	100	DC-SIS1	8.69	0.75	1.000	1.000	1.000	0.998	0.960	0.960
		DC-SIS	5.13	0.00	1.000	1.000	1.000	1.000	0.998	0.998
	200	DC-SIS1	5.07	0.00	1.000	1.000	1.000	1.000	1.000	1.000
		DC-SIS	5.00	0.00	1.000	1.000	1.000	1.000	1.000	1.000
4,000	100	DC-SIS1	16.99	0.75	1.000	1.000	0.998	0.992	0.914	0.914
		DC-SIS	5.18	0.00	1.000	1.000	1.000	1.000	0.998	0.998
	200	DC-SIS1	5.04	0.00	1.000	1.000	1.000	1.000	1.000	1.000
		DC-SIS	5.00	0.00	1.000	1.000	1.000	1.000	1.000	1.000

MMS: the average value of $\mathcal{S}$ among 500 replications; RSD: defined as IQR/1.34, where IQR denotes the interquartile range of $\mathcal{S}$ over 500 replications; DC-SIS1: another DC-based SIS procedure, where we directly set $Y = (U, V, \Delta_1, \Delta_2, \Delta_3)$; DC-SIS: the proposed method.

were assumed to follow a multivariate normal distribution $N_{p-1}(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma} = (0.8^{|i-j|})$, for $i, j = 1, \ldots, (p-1)$.

c) All covariates $Z_k (k = 1, \ldots, p)$ were generated independently from a Bernoulli distribution with success probability 0.5.

Table 2 gives the simulation results, containing the same quantities as in Table 1. The results again indicate that the proposed screening procedure gives reasonable performance for the complicated nonlinear model with interactions. In particular, the procedure appears to perform better for the continuous covariate than it does for the categorical covariate, but the difference between the types of covariates decreases when the sample size increases. In other words, the proposed method seems to give good performance for both continuous and categorical covariates when the sample size is large. Furthermore, as seen in Table 2, the proposed procedure DC-SIS significantly outperforms the naive DC-based SIS procedure DC-SIS1 for the complicated nonlinear model, especially for the cases with categorical covariates.

To evaluate the performance of the proposed procedure under different observation schedules, we considered another setup. Specifically, we first randomly generated a sequence of observation times from the uniform distribution $\mathrm{U}(0, \tau)$, denoted by $t_1 < t_2 < \cdots < t_m$. For each subject i and at each observation time t_j, a random variable was then independently generated from a Bernoulli distribution with success probability 0.6. If the value is one, we assume that subject

Table 2. The simulation results on $\mathcal{S}$ (the minimum model size needed to include all active covariates), $\mathcal{P}_e$ (the selection proportions for each active covariate), and $\mathcal{P}_a$ (the selection proportion for all active covariates) for setup 2

						$\mathcal{P}_e$			
p	n	Case	Method	MMS	RSD	Z_1	Z_5	Z_{10}	$\mathcal{P}_a$
2,000	100	a)	DC-SIS1	35.33	11.19	0.734	0.986	0.998	0.726
			DC-SIS	6.19	2.99	1.000	1.000	1.000	1.000
		b)	DC-SIS1	853.55	1,242.35	0.084	0.954	0.996	0.076
			DC-SIS	30.90	8.21	0.788	0.998	1.000	0.786
		c)	DC-SIS1	1,491.33	434.70	0.010	0.004	0.012	0.000
			DC-SIS	41.35	8.21	1.000	0.812	1.000	0.812
	200	a)	DC-SIS1	10.09	2.24	0.996	1.000	1.000	0.996
			DC-SIS	5.32	2.24	1.000	1.000	1.000	1.000
		b)	DC-SIS1	488.65	483.77	0.240	1.000	1.000	0.240
			DC-SIS	10.07	2.24	0.996	1.000	1.000	0.996
		c)	DC-SIS1	1,520.49	429.29	0.022	0.010	0.022	0.000
			DC-SIS	3.34	0.00	1.000	1.000	1.000	1.000
4,000	100	a)	DC-SIS1	47.58	14.18	0.716	0.980	0.988	0.698
			DC-SIS	6.36	2.99	0.998	1.000	1.000	0.998
		b)	DC-SIS1	1,757.48	2,524.81	0.052	0.922	0.998	0.052
			DC-SIS	9.77	22.39	0.660	0.996	1.000	0.658
		c)	DC-SIS1	3,103.47	786.94	0.006	0.006	0.004	0.000
			DC-SIS	83.49	27.05	0.988	0.690	1.000	0.682
	200	a)	DC-SIS1	10.51	2.24	0.988	1.000	1.000	0.988
			DC-SIS	5.43	2.24	1.000	1.000	1.000	1.000
		b)	DC-SIS1	979.00	1,020.15	0.186	1.000	1.000	0.186
			DC-SIS	10.05	1.68	0.998	1.000	1.000	0.998
		c)	DC-SIS1	3,049.69	728.54	0.008	0.004	0.006	0.000
			DC-SIS	7.21	0.00	1.000	0.982	1.000	0.982

Case a): $\boldsymbol{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$; Case b): Z_1 is generated from a Bernoulli distribution with success probability 0.5; Case c): all covariates Z_k $(k = 1, \ldots, p)$ are generated from a Bernoulli distribution with success probability 0.5; MMS: the average value of $\mathcal{S}$ among 500 replications; RSD: defined as IQR/1.34, where IQR denotes the interquartile range of $\mathcal{S}$ over 500 replications; DC-SIS1: the naive DC-based SIS procedure that set $Y = (U, V, \Delta_1, \Delta_2, \Delta_3)$; DC-SIS: the proposed method.

i is observed at t_j; otherwise, the subject is not observed. In addition to the percentages of left-, right-, and interval-censored observations being 20%, 20%, and 60%, respectively, we also considered cases with percentages 20%, 60%, and 20%, and 60%, 20%, and 20%, denoted as cases I, II, and III, respectively. Table 3 gives the results obtained with $m = 20$ and under case a). These results suggest that the proposed method again performs well for this situation. Furthermore, as expected, the results indicate that the proposed method is robust with respect to the percentages of left-, right-, and interval-censored observations.

Table 3. The simulation results on $\mathcal{S}$ (the minimum model size needed to include all active covariates), $\mathcal{P}_e$ (the selection proportions for each active covariate), and $\mathcal{P}_a$ (the selection proportion for all active covariates) for setup 2

						$\mathcal{P}_e$			
p	n	Ratio	Method	MMS	RSD	Z_1	Z_5	Z_{10}	$\mathcal{P}_a$
2000	100	I	DC-SIS1	1,490.87	425.75	0.008	0.012	0.016	0.000
			DC-SIS	6.53	2.24	1.000	1.000	0.998	0.998
		II	DC-SIS1	1,529.61	359.33	0.008	0.006	0.006	0.000
			DC-SIS	8.67	2.24	1.000	1.000	0.978	0.978
		III	DC-SIS1	1,508.07	424.07	0.008	0.010	0.012	0.000
			DC-SIS	11.29	3.73	0.964	0.998	0.996	0.958
	200	I	DC-SIS1	1,509.70	435.07	0.018	0.010	0.012	0.000
			DC-SIS	5.22	1.49	1.000	1.000	1.000	1.000
		II	DC-SIS1	1,516.77	413.43	0.014	0.028	0.022	0.000
			DC-SIS	6.98	1.49	1.000	1.000	1.000	1.000
		III	DC-SIS1	1,495.15	427.61	0.024	0.026	0.016	0.002
			DC-SIS	8.24	2.99	1.000	1.000	1.000	1.000

I: left-censored, right-censored, interval-censored rates of $20\%, 20\%, 60\%$; II: left-censored, right-censored, interval-censored rates of $20\%, 60\%, 20\%$; III: left-censored, right-censored, interval-censored rates of $60\%, 20\%, 20\%$; MMS: the average value of $\mathcal{S}$ among 500 replications; RSD: defined as IQR/1.34, where IQR denotes the interquartile range of $\mathcal{S}$ over 500 replications; DC-SIS1: the naive DC-based SIS procedure that set $Y = (U, V, \Delta_1, \Delta_2, \Delta_3)$; DC-SIS: the proposed method.

Setup 3. In this setup, we generated the failure time of interest T_i from the transformation model

$$H(T) = -\boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{Z} + \epsilon,$$

where we took $H(t) = \log\{0.5(e^{2t} - 1)\}$, $\boldsymbol{\beta}_0 = (1, 0.7, \mathbf{0}_6, 0.8, 1.0, \mathbf{0}_{p-10})^{\mathrm{T}}$, and $\boldsymbol{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = (0.5^{|i-j|})$, for $i, j = 1, \ldots, p$. Here, we have four active covariates in Z_1, Z_2, Z_9, and Z_{10}, giving the index set of active covariates $\mathcal{A} = \{1, 2, 9, 10\}$. We considered three choices for the distribution of ϵ, namely, the standard normal distribution, the standard logistic distribution, and the type-I extreme value distribution. The obtained simulation results on the same quantities as above are presented in Table 4, which again suggest that the proposed procedure DC-SIS gives satisfactory results and seems to be robust with respect to the error distribution. In addition, the proposed procedure DC-SIS again outperforms the DC-SIS1 procedure.

To further assess the performance of the proposed screening procedure in terms of $\mathcal{S}$, Figure 1 gives box plots of $\mathcal{S}$ obtained based on 500 replications under the three setups discussed above for the left-censored, right-censored, and interval-censored rates being 20%, 20%, and 60%, respectively. Again, they indi-

Table 4. The simulation results on $\mathcal{S}$ (the minimum model size needed to include all active covariates), $\mathcal{P}_e$ (the selection proportions for each active covariate), and $\mathcal{P}_a$ (the selection proportion for all active covariates) for setup 3

						$\mathcal{P}_e$				
p	n	F_ϵ	Method	MMS	RSD	Z_1	Z_2	Z_9	Z_{10}	$\mathcal{P}_a$
2,000	100	Norm	DC-SIS1	44.79	17.91	0.934	0.936	0.826	0.918	0.668
			DC-SIS	5.45	0.75	0.998	0.998	0.992	0.996	0.984
		Logistic	DC-SIS1	120.69	83.21	0.850	0.802	0.664	0.752	0.388
			DC-SIS	16.97	5.97	0.984	0.980	0.914	0.974	0.858
		Extreme	DC-SIS1	41.28	19.59	0.946	0.938	0.838	0.900	0.688
			DC-SIS	7.42	0.75	1.000	0.994	0.976	0.994	0.966
	200	Norm	DC-SIS1	4.86	0.00	1.000	1.000	0.996	0.998	0.994
			DC-SIS	4.01	0.00	1.000	1.000	0.998	1.000	0.998
		Logistic	DC-SIS1	17.21	1.49	0.992	0.988	0.962	0.992	0.940
			DC-SIS	4.26	0.00	1.000	1.000	0.998	1.000	0.998
		Extreme	DC-SIS1	5.45	0.00	1.000	1.000	0.994	1.000	0.994
			DC-SIS	4.02	0.00	1.000	1.000	1.000	1.000	1.000
4,000	100	Norm	DC-SIS1	83.56	41.04	0.910	0.890	0.794	0.870	0.568
			DC-SIS	7.24	0.75	0.994	0.998	0.976	1.000	0.968
		Logistic	DC-SIS1	227.90	142.91	0.766	0.706	0.612	0.704	0.266
			DC-SIS	23.19	8.21	0.970	0.960	0.906	0.946	0.810
		Extreme	DC-SIS1	92.87	39.18	0.890	0.870	0.820	0.872	0.586
			DC-SIS	8.94	0.75	0.996	0.992	0.972	0.992	0.952
	200	Norm	DC-SIS1	5.99	0.00	0.998	0.998	0.998	1.000	0.994
			DC-SIS	4.03	0.00	1.000	1.000	1.000	1.000	1.000
		Logistic	DC-SIS1	18.52	3.73	0.996	0.982	0.938	0.988	0.904
			DC-SIS	4.15	0.00	1.000	1.000	1.000	1.000	1.000
		Extreme	DC-SIS1	5.29	0.00	1.000	0.998	0.998	0.998	0.994
			DC-SIS	4.02	0.00	1.000	1.000	1.000	1.000	1.000

F_ϵ: the distribution of ϵ; MMS: the average value of $\mathcal{S}$ among 500 replications; RSD: defined as IQR/1.34, where IQR denotes the interquartile range of $\mathcal{S}$ over 500 replications; DC-SIS1: the naive DC-based SIS procedure that set $Y = (U, V, \Delta_1, \Delta_2, \Delta_3)$; DC-SIS: the proposed method.

cate that the proposed procedure gives excellent performance and can screen out the inactive covariates, as expected.

6. A Data-Driven Log-Ratio Criterion

As discussed above, to apply the proposed screening procedure and other existing screening procedures, one needs to choose a selection threshold value. Here, a common choice is given by Fan and Lv (2008), who suggested a hard cutoff value $d_0 = \lceil n/\log n \rceil$. In general, using this choice relies on the sparsity

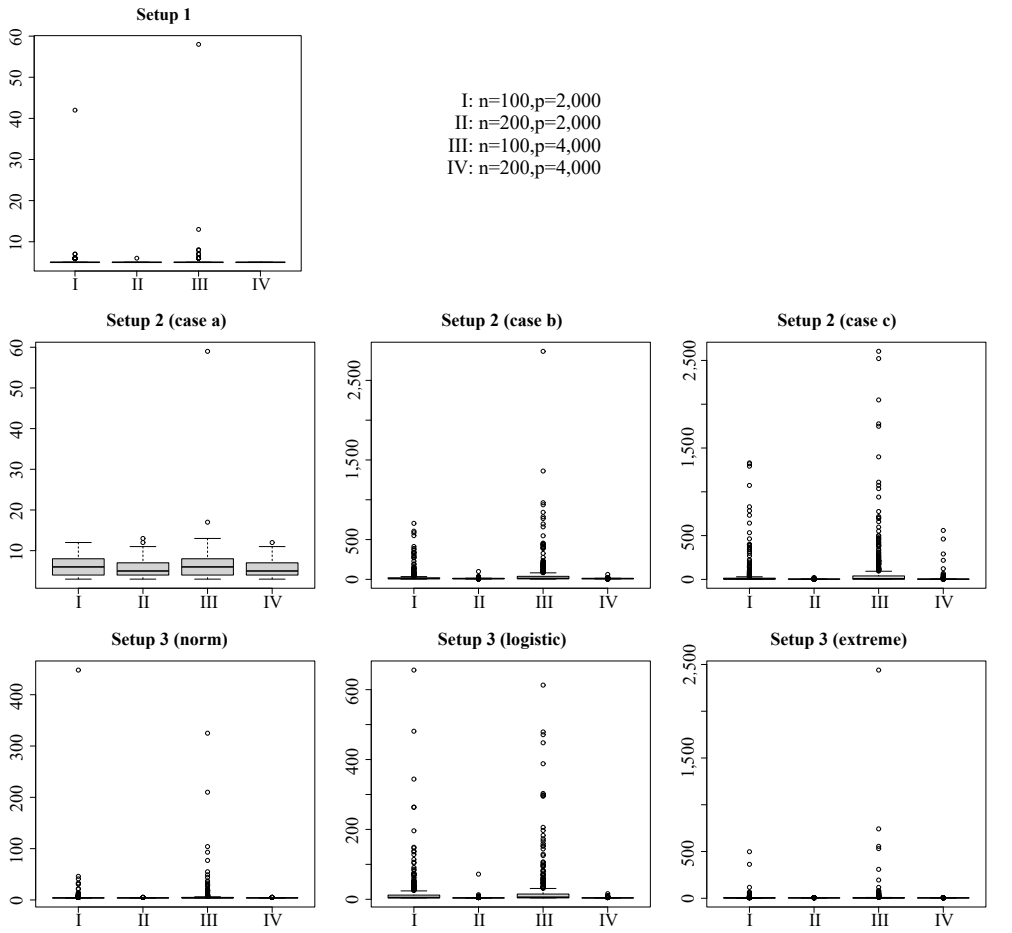

Figure 1. Boxplots of $\mathcal{S}$ for setups 1 (top), 2 (middle), and 3 (bottom).

assumption, meaning that only a small number of variables or covariates are truly associated with the response variable of interest. This is often true in many situations, including gene selection and biomedical imaging data analysis. On the other hand, it is apparent that the sparsity assumption may not hold, and a data-driven criterion for the choice would be useful. Based on Theorem 2, we propose the following log-ratio criterion.

Let $\widehat{\omega}_k$ be defined as above and $\widehat{\omega}_{(k)}$ denote the kth largest of all $\{\widehat{\omega}_k : k = 1, \ldots, p\}$. Consider the sequence $\{\log(\widehat{\omega}_{(k)}/\widehat{\omega}_{(k+1)})\}$. In general, one can expect that the sequence will sometimes increase and sometimes decrease when k corresponds to active covariates. Furthermore, it tends to give the largest value when k approaches the boundary between the active and inactive covariates,

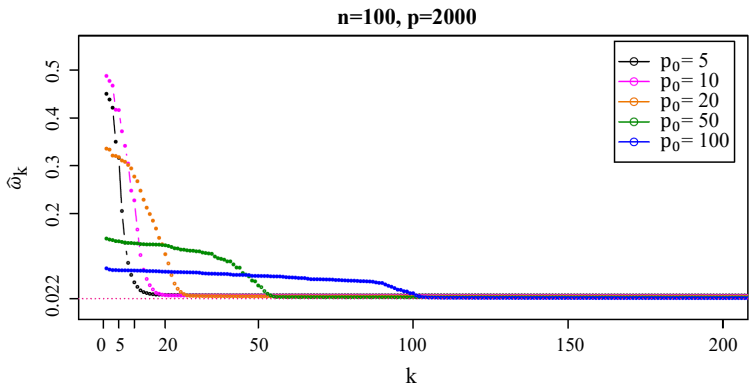

Figure 2. The plot of $\widehat{\omega}_k$ with the true model size $p_0 = 5, 10, 20, 50, 100$.

and then decreases among the inactive covariates. This motivates the following log-ratio criterion for choosing the selection threshold value

$$d_0^* = \underset{1 \le j \le (p-1)}{\operatorname{argmax}} \log \frac{\widehat{\omega}_{(k)}}{\widehat{\omega}_{(k+1)}}. \tag{6.1}$$

To investigate the behavior of the sequence defined above, and to assess the performance of the log-ratio criterion, we repeated the above simulation study under setup 1 with $n = 100$, $p = 2000$, and the true model size being $p_0 = 5, 10, 20, 50,$ or 100. Figure 2 plots the first 200 $\widehat{\omega}_{(k)}$ obtained from $\{\widehat{\omega}_k : k = 1, \ldots, p\}$, with each $\widehat{\omega}_k$ being the average value over 500 replications. Figure 3 shows a plot of the corresponding $\log(\widehat{\omega}_{(k)}/\widehat{\omega}_{(k+1)})$. Note that in both figures, the solid points denote the active covariates. One can see from Figure 2 that all of the active covariates are located on the left side of the inactive covariates, and there seems to exist a change point from which the value of ω goes smoothly. Figure 3 indicates that the ratio $\log(\widehat{\omega}_{(k)}/\widehat{\omega}_{(k+1)})$ indeed behaves as described above, and confirms the effectiveness of the proposed log-ratio criterion.

7. An Application

In this section, we apply the model-free DC-based screening procedure proposed in the previous sections to data from the ADNI. The participants in the study were examined intermittently for various factors, including demographic and clinical factors and SNPs, and are classified into three groups based on their cognitive condition: cognitively normal, mild cognitive impairment, and AD. Among others, one variable of interest is the time (in years) from the baseline visit date to the AD conversion. Owing to the design of the study, only interval-

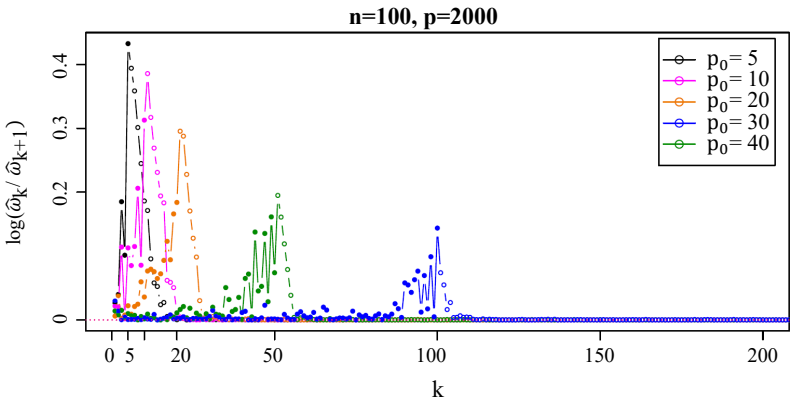

Figure 3. The plot of $\log(\widehat{\omega}_{(k)}/\widehat{\omega}_{(k+1)})$ with the true model size $p_0 = 5, 10, 20, 50, 100$.

censored data are available on the AD conversion time. In the following, we are interested in identifying the SNPs that have significant effects on the AD conversion time.

In the analysis below, following Li et al. (2017), we focus on the 280 participants in the mild cognitive impairment group, for whom complete information is available about the four demographic and clinical factors identified as significantly associated with the AD conversion by Li et al. (2017). They are the participants' Alzheimer's Disease Assessment Scale Score of 13 items (ADAS13), the Rey auditory verbal learning test score of immediate recall (RAVLT.i), the functional assessment questionnaire score (FAQ), and the MRI volume of middle temporal gyrus (MidTemp). In addition to these four covariates, we consider 162,194 SNPs, coded as 0, 1, or 2. Note that although there are far more SNPs in the data, most of them are constants for the subjects considered here, and thus removed. Note that most predictors are categorical and based on the simulation results given in Section 5, the proposed method may not perform well for the categorical covariates when the sample is small. On the other hand, as seen in Section 5, the proposed method is expected to perform well for both categorical and continuous covariates for the given sample size here.

Figure 4 presents the estimated dependence measures $\widehat{\omega}_k$ given by the model-free screening procedure proposed in the previous sections for the 162,194 SNPs, plus the four demographic and clinical factors, which correspond to $k = 1, \ldots, 4$. It is apparent that, as expected, the four demographic and clinical factors have higher correlations with the AD conversion time than the individual SNPs do. To determine the number of active SNPs, we should choose the 45 SNPs with the highest $\widehat{\omega}_k$ by using the selection rule $[n/\log(n)]$. To see this, Figure 5 gives the

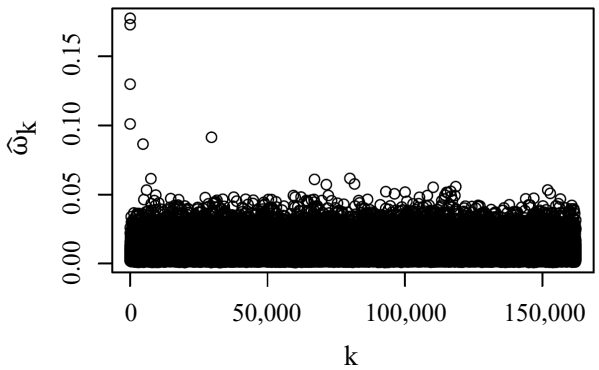

Figure 4. The $\widehat{\omega}_k$ values for all covariates for the AD example.

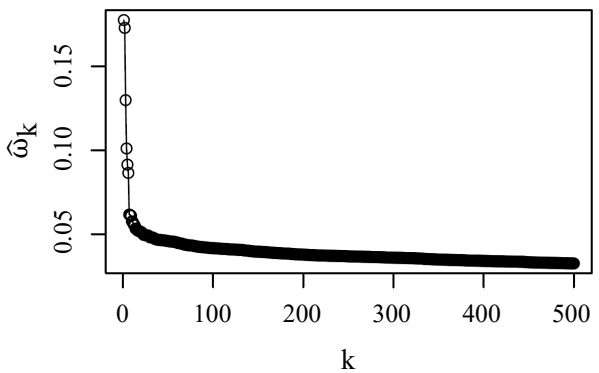

Figure 5. The 500 largest $\widehat{\omega}_k$ values for the AD example.

largest 500 $\widehat{\omega}_k$ from the largest to the smallest, and Figure 6 displays the ratio sequence $\log(\widehat{\omega}_{(k)}/\widehat{\omega}_{(k+1)})$ for the first 50 ratios. The former suggests that there may be at most 100 active covariates for the AD conversion time. Furthermore, by using the proposed log-ratio criterion, the latter indicates that there may be around 10 or fewer than 20 active covariates.

8. Conclusion

We have considered the variable selection or identifying of important or relevant variables for ultrahigh-dimensional IC failure time data, and proposed a model-free screening procedure for this problem. To develop the proposed approach, a marginal utility ω_k was derived based on the DC between a redefined response or failure time variable and each covariate. One major advantage of the proposed screening method is that it does not involve a nonparametric estimation

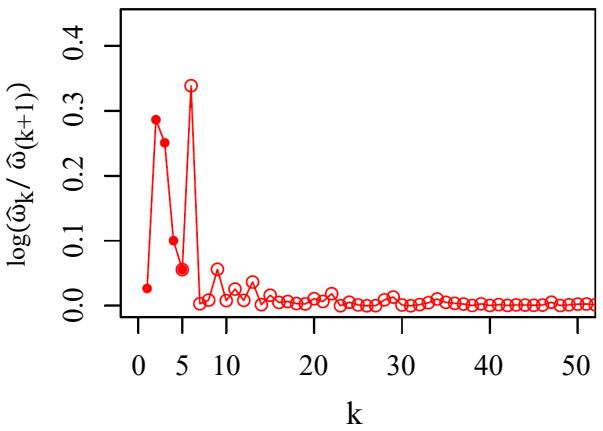

Figure 6. The scatter points of $\log(\widehat{\omega}_{(k)}/\widehat{\omega}_{(k+1)})$ for the AD example.

of a survival function or any complicated numerical optimization. Thus, it can be easily implemented and the computation is fast. Furthermore, the procedure was shown to have the sure screening and ranking consistency properties, and a data-driven log-ratio criterion was presented to determine the selection threshold value. Numerical results indicated that the proposed methodology works well for practical situations.

Noted that, as is the case of SIS procedures, the proposed DC-based screening procedure is a marginal approach developed based on marginal utilities. One drawback of these marginal methods is that they may not perform well for situations in which covariates are jointly important, but marginally unimportant. For this situation, one can generalize the proposed method to an iterative approach that can take the correlation among covariates into account, or develop a new screening method by directly incorporating the correlation information among the covariates. Another common issue related to the feature screening of ultrahigh-dimensional data that is not discussed much in the literature is the analysis of the data after the feature screening or variable selection. Although many methods have been developed for low-dimensional or high-dimensional data with respect to estimation or simultaneous variable selection and estimation, there are few studies on the joint analysis of the two steps or stages.

Supplementary Material

The online Supplementary Material includes detailed proofs of Theorems 1 and 2.

Acknowledgments

The authors wish to thank the co-editor, associate editor and two reviewers for their helpful comments and suggestions. Dr. Zhang's research was partially supported by the National Natural Science Foundation of China (NSFC 11901581) and the Natural Science Foundation of Hubei Province (2021CFB502), Dr. Liu's research was partially supported by the National Natural Science Foundation of China (NSFC 11971362), and Dr. Sun's research was partially supported by Washington University Institute of Clinical and Translational Sciences grant (# CTSA 131).

References

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544–557.

Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications-A Festschrift for Lawrence D. Brown* (Edited by J. O. Berger, T. T. Cai and I. M. Johnstone), 70–86. Institute of Mathematical Statistics, Beachwood.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10**, 2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–3604.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75**, 217–245.

Hu, Q., Zhu, L., Liu, Y., Sun, J., Srivastava, D. K. and Robison, L. L. (2020). Nonparametric screening and feature selection for ultrahigh-dimensional case II interval-censored failure time data. *Biom. J.* **62**, 1909–1925.

Huang, Q. and Zhu, Y. (2016). Model-free sure screening via maximum correlation. *J. Multivariate Anal.* **148**, 89–106.

Li, K., Chan, W., Doody, R. S., Quinn, J. and Luo, S. (2017). Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *J. Alzheimer's Dis.* **58**, 361–371.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129–1139.

Lin, Y., Liu, X. and Hao, M. (2018). Model-free feature screening for high-dimensional survival data. *Sci. China Math.* **61**, 1617–1636.

Liu, Y., Zhang, J. and Zhao, X. (2018). A new nonparametric screening method for ultrahigh-dimensional survival data. *Comput. Statist. Data Anal.* **119**, 74–85.

Scolas, S., El Ghouch, A., Legrand, C. and Oulhaj, A. (2016). Variable selection in a flexible parametric mixture cure model with interval-censored data. *Stat. Med.* **35**, 1210–1225.

Song, R., Lu, W., Ma, S. and Jeng, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101**, 799–814.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer, New York.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

Tibshirani, R. (2009). Univariate shrinkage in the Cox model for high dimensional data. *Stat. Appl. Genet. Mol. Biol.* **8**, 1–18.

Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped, censored and truncated data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **38**, 290–295.

Wu, Y. and Cook, R. (2015). Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics* **71**, 782–791.

Wu, Y. and Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102**, 65–76.

Zhao, H., Wu, Q., Li, G. and Sun, J. (2020). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *J. Amer. Statist. Assoc.* **115**, 204–216.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivariate Anal.* **105**, 397–411.

Zhang, J., Liu, Y. and Cui, H. (2021). Model-free feature screening via distance correlation for ultrahigh dimensional survival data. *Statist. Papers* **62**, 2711–2738.

Zhang, J., Liu, Y. and Wu, Y. (2017). Correlation rank screening for ultrahigh-dimensional survival data. *Comput. Statist. Data Anal.* **108**, 121–132.

Zhang, J., Yin, G., Liu, Y. and Wu, Y. (2018). Censored cumulative residual independent screening for ultrahigh-dimensional survival data. *Lifetime Data Anal.* **24**, 273–292.

Zhou, T. and Zhu, L. (2017). Model-free feature screening for ultrahigh dimensional censored regression. *Stat. Comput.* **27**, 947–961.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106**, 1464–1475.

Jing Zhang

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, China.

E-mail: jing66@zuel.edu.cn

Mingyue Du

Department of Mathematics and Statistics, The Hong Kong Polytechnic University, Hong Kong, China.

E-mail: dummoon@163.com

Yanyan Liu

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

E-mail: liuyy@whu.edu.cn

Jianguo Sun
Department of Statistics, University of Missouri, Columbia, Missouri 65211, USA.
E-mail: sunj@missouri.edu

(Received May 2020; accepted September 2021)