Journal of Statistical Planning and Inference 232 (2024) 106153

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Feature screening via concordance indices for left-truncated and right-censored survival data

Li-Pang Chen

Department of Statistics, National Chengchi University, Taipei, Taiwan, ROC

ARTICLE INFO

Keywords:
Biased sampling
Incomplete data
Marginal correlation
Sure screening property
Ultrahigh-dimensionality

ABSTRACT

Ultrahigh-dimensional data analysis has been a popular topic in decades. In the framework of ultrahigh-dimensional setting, feature screening methods are key techniques to retain informative covariates and screen out non-informative ones when the dimension of covariates is extremely larger than the sample size. In the presence of incomplete data caused by censoring, several valid methods have also been developed to deal with ultrahigh-dimensional covariates for time-to-event data. However, little approach is available to handle feature screening for survival data subject to biased sample, which is usually induced by left-truncation. In this paper, we extend the C-index estimation proposed by Hartman et al. (2023) to develop a valid feature screening procedure to deal with left-truncated and right-censored survival data subject to ultrahigh-dimensional covariates. The sure screening property is also rigorously established to justify the proposed method. Numerical results also verify the validity of the proposed procedure.

1. Introduction

Left-truncated and right-censored (LTRC) data has been an attractive topic in lifetime data analysis. The key challenges include biased sampling caused by prevalent cohort studies and incomplete response induced by right-censoring. In the presence of covariates in datasets, a large body of methods have been developed under various regression models. To name a few, Huang and Qin (2013) proposed the estimation equation approach for the additive hazards model. Chen and Yi (2021) proposed the augmented pseudo likelihood method for the Cox model. Chen (2019a) considered LTRC data with cure models. Chen (2019b) discussed the additive hazards model under the LTRC data.

In the modern statistical analysis, high-dimensionality is a ubiquitous feature in datasets. The crucial impact of high-dimensional data is the involvement of irrelevant covariates. To address it, regularization methods are widely used to do variable selection. Under LTRC data, several methods have been discussed when the dimension of covariates is smaller than the sample size. For example, Chen and Yi (2020) proposed the focus information criterion for the Cox model. Chen (2020) considered the penalized likelihood function under the additive hazards model. McGough et al. (2021) adopted the regularization methods under the Cox model. Recently, Chen and Qiu (2023) proposed the boosting method to do variable selection under length-biased sampling with the truncation time following the uniform distribution. However, when the dimension of covariates is extremely larger than the sample size, known as *ultrahigh-dimensionality*, existing methods are no longer valid.

To address variable selection for ultrahigh-dimensional data, *feature screening* is perhaps a widely used strategy. The key idea of feature screening is to take the correlation between the response and the covariate as a signal, and use it to retain truly informative covariates. Since the seminal work of Fan and Lv (2008), a large number of research papers have emerged for handling feature

E-mail address: lchen723@nccu.edu.tw.

https://doi.org/10.1016/j.jspi.2024.106153
Received 3 November 2022; Received in revised form 3 October 2023; Accepted 1 February 2024
Available online 10 February 2024

screening when a dataset is complete, such as the distance correlation (e.g., Li et al., 2012), the score function approach (e.g., Zhao and Li, 2014), the concordance index (C-index; e.g., Ma et al., 2017), and the rank method (e.g., Chen, 2023). When the response is time-to-event and is incomplete, several methods have been developed under different censoring mechanisms, including right-censoring (e.g., Chen, 2021; Chen and Yi, 2022) and interval-censoring (e.g., Hu et al., 2020). In contrast, in the presence of left-truncation with unspecified distribution of the truncation time. rare methods have been available to deal with this challenge. It is expected that existing methods fail to handle feature screening when datasets suffer from left-truncation.

To fill out this research gap, in this paper we explore the feature screening for LTRC data. Our strategy is motivated by the C-index proposed by Hartman et al. (2023). Following the formulation in Hartman et al. (2023), we take their C-index as a function of the parameter to derive an estimating function, and then transform it as a signal, which enables us to do feature screening. To justify the validity of the feature screening procedure, we also establish the sure screening property with the rigorous proof. Finally, we conduct numerical studies, including simulation and real data analysis, to assess the performance of the proposed method.

The remainder is organized as follows. In Section 2, we introduce data structure and basic idea of the C-index method. In Section 3, we present our main result of feature screening. In addition, we discuss the sure screening property based on the proposed method, and the proof is placed in Section 4. Simulation studies and real data analysis are available in Sections 5 and 6, respectively. Finally, we conclude the article with discussions in Section 7.

2. Preliminary

2.1. Data structure

For an individual in the target disease population, let ξ be the calendar time of the recruitment and let u and v denote the calendar time of the initiating event and the failure event, respectively, which satisfy $u < \xi < v$. Let $\widetilde{T} = v - u$ be the lifetime and $\widetilde{A} = \xi - u$ be the truncation time. Let τ_{A} be a constant such that $P(\widetilde{A} \leq \tau_{\mathrm{A}}) > 0$. In the data collection, one can observe the lifetime if $\widetilde{T} \geq \widetilde{A}$; otherwise the lifetime cannot be recorded, which is called left-truncation. Let (A, T) denote $\left(\widetilde{A}, \widetilde{T}\right)\Big|\widetilde{T} \geq \widetilde{A}$ to indicate that such an individual is eligible for the recruitment so that measuring (A, T) is possible. Let $\mathbf{X} \triangleq \left(X_{(1)}, \ldots, X_{(p)}\right)^{\top}$ be a p-dimensional vector of covariates, where $X_{(k)}$ represents the kth component. Following existing frameworks (e.g., Fan and Lv, 2008; Ma et al., 2017; Chen, 2023), without loss of generality, we consider $E(X_{(k)}) = 0$ and $\operatorname{var}(X_{(k)}) = 1$ for $k = 1, \ldots, p$.

In addition to left-truncation, recruited individuals may suffer from right-censoring. Let C be the residual censoring time for a recruited subject, which is recorded from the recruitment point. Let τ_{C} denote a constant such that $P(C > \tau_{\mathrm{C}}) > 0$. Let $Y = \min\{T, A + C\}$ be the observed survival time and let $\Delta = \mathbb{I}(T \leq A + C)$ be the indicator of a failure event with $\mathbb{I}(\cdot)$ being an indicator function. Here we impose some standard assumptions for LTRC data:

(A1) $(\widetilde{T}, \mathbf{X}) \perp\!\!\!\perp \widetilde{A}|\mathbf{X}$, where $\perp\!\!\!\perp$ indicates the independence;
(A2) $T \perp\!\!\!\perp C|\mathbf{X}$.

Condition (A1) says that, given the covariates $\mathbf{X}$, the population failure time is independent of the truncation time. In addition, the covariates $\mathbf{X}$ are non-informative to the truncation time. This assumption comes from the literature of left-truncation (e.g., Chen and Yi, 2021; Huang and Qin, 2013). Condition (A2) is a standard assumption in survival analysis, which shows that the recruited censoring time is non-informative and is independent of the observed failure time. Suppose we have an observed sample of n subjects $\left(Y_i, \Delta_i, A_i, \mathbf{X}_i\right)$ that have the same distribution of $(Y, \Delta, A, \mathbf{X})$ for $i = 1, \ldots, n$.

In this study, we primarily focus on ultrahigh-dimensional data, where the dimension p is dependent on the sample size n and might be diverging. That is, following the scenario in Fan and Lv (2008), the relationship between p and n can be characterized as $p = \exp\{O(n^r)\}$ for some $r > 0$. In the ultrahigh-dimensional setting, most covariates might be irrelevant and rare covariates are informative to the response. Our goal is to identify the active set

$$\mathcal{I} = \left\{k : X_{(k)} \text{ is an informative covariate for } \widetilde{T}\right\}$$

that contains all relevant covariates for the response $\widetilde{T}$ with size $|\mathcal{I}| < n$, reflecting that the number of truly informative covariates is small. If $\widetilde{T}$ is completely observed, then one can directly adopt existing methods (e.g., Fan and Lv, 2008; Li et al., 2012; Ma et al., 2017; Chen, 2023) to do feature screening and estimate $\mathcal{I}$. In the presence of LTRC, however, we have only Y instead of $\widetilde{T}$. Directly implementing Y to feature screening may falsely exclude important covariates because the selected covariates are correlated to Y instead of $\widetilde{T}$.

2.2. Overview of C-index for LTRC

Under LTRC data, Hartman et al. (2023) proposed the following C-index estimator that is a function of a p-dimensional vector of parameters $\boldsymbol{\beta}$:

$$C(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\widehat{G}(Y_i)\right\}^{-2}\mathbb{I}(\mathbf{X}_i^{\top}\boldsymbol{\beta} > \mathbf{X}_j^{\top}\boldsymbol{\beta}, \tau_{\mathrm{A}} < Y_i < Y_j, Y_i < \tau_{\mathrm{C}}, \delta_i = 1, A_j \leq Y_i)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\widehat{G}(Y_i)\right\}^{-2}\mathbb{I}(\tau_{\mathrm{A}} < Y_i < Y_j, Y_i < \tau_{\mathrm{C}}, \delta_i = 1, A_j \leq Y_i)}, \tag{1}$$

where $\widehat{G}(y) = \int_0^t \widehat{F}(y-a)d\widehat{H}(a)$ with $\widehat{F}(y)$ and $\widehat{H}(a)$ being estimators of survivor functions of the censoring time $F(y)$ and the truncation time $H(a)$, respectively. According to the discussion in Hartman et al. (2023), it can be shown that the limiting value of (1) is equal to the target concordance probability

$$P(\mathbf{X}_i^\top \boldsymbol{\beta} > \mathbf{X}_j^\top \boldsymbol{\beta} | \tau_\mathrm{A} < \widetilde{T}_i < \widetilde{T}_j, \widetilde{T}_i < \tau_\mathrm{C}) \tag{2}$$

that is free of the truncation distribution and the censoring distribution.

As discussed in Section 3.2, two estimates $\widehat{F}(\cdot)$ and $\widehat{H}(\cdot)$ are required to be consistent estimators of $F(\cdot)$ and $H(\cdot)$ so that the theoretical result can be established. In the following discussion, we implement the nonparametric maximum likelihood estimator (NPMLE) to estimate $H(\cdot)$:

$$\widehat{H}(a) = \left(\sum_{i=1}^n \frac{1}{\widehat{S}(A_i)}\right)^{-1} \left(\sum_{i=1}^n \frac{\mathbb{I}(A_i \leq a)}{\widehat{S}(A_i)}\right), \tag{3}$$

where $\widehat{S}(y)$ is the Kaplan–Meier estimator of the survivor function of $\widetilde{T}$. As shown in Wang (1991), the NPMLE (3) is a consistent estimator of $H(a)$. In addition, for the estimation of $F(y)$, we adopt the Kaplan–Meier estimator by pooling the study subjects with their differences in covariates ignored, which is shown to be consistent (e.g., Wang, 1987).

3. Main results

3.1. Feature screening via C-index

Motivated by Ma et al. (2017), to seek a feature $\mathbf{X}_i^\top \boldsymbol{\beta}$ that predicts the response $\widetilde{T}_i$ under the LTRC structure, the concordance probability (2) enables us to achieve this goal. Following the discussion in Section 2.2, it suffices to consider the C-index (1). However, the indicator function $\mathbb{I}(\mathbf{X}_i^\top \boldsymbol{\beta} > \mathbf{X}_j^\top \boldsymbol{\beta})$ in (1) is discrete, which may cause computational and theoretical challenges.

To address the concern of the indicator function, we follow the similar discussion in Ma et al. (2017) and smoothly approximate the indicator function $\mathbb{I}(\mathbf{X}_i^\top \boldsymbol{\beta} > \mathbf{X}_j^\top \boldsymbol{\beta})$ by $\Phi\left\{(\mathbf{X}_i^\top \boldsymbol{\beta} - \mathbf{X}_j^\top \boldsymbol{\beta})/h\right\}$, where $\Phi(\cdot)$ is the distribution function of the standard normal distribution and h is the bandwidth. Then (1) can be re-written as

$$\begin{aligned} C(\boldsymbol{\beta}) &= \frac{\sum_{i=1}^n \sum_{j=1}^n \left\{\widehat{G}(Y_i)\right\}^{-2} \mathbb{I}(\tau_\mathrm{A} < Y_i < Y_j, Y_i < \tau_\mathrm{C}, \delta_i = 1, A_j \leq Y_i) \Phi\left\{(\mathbf{X}_i^\top \boldsymbol{\beta} - \mathbf{X}_j^\top \boldsymbol{\beta})/h\right\}}{\sum_{i=1}^n \sum_{j=1}^n \left\{\widehat{G}(Y_i)\right\}^{-2} \mathbb{I}(\tau_\mathrm{A} < Y_i < Y_j, Y_i < \tau_\mathrm{C}, \delta_i = 1, A_j \leq Y_i)} \\ &\triangleq \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \widehat{\omega}_{ij} \Phi\left(\frac{\mathbf{X}_i^\top \boldsymbol{\beta} - \mathbf{X}_j^\top \boldsymbol{\beta}}{h}\right) \end{aligned} \tag{4}$$

with

$$\widehat{\omega}_{ij} \triangleq \frac{\left\{\widehat{G}(Y_i)\right\}^{-2} \mathbb{I}(\tau_\mathrm{A} < Y_i < Y_j, Y_i < \tau_\mathrm{C}, \delta_i = 1, A_j \leq Y_i)}{\frac{1}{n(n-1)} \sum_{i'=1}^n \sum_{j'=1}^n \left\{\widehat{G}(Y_{i'})\right\}^{-2} \mathbb{I}(\tau_\mathrm{A} < Y_{i'} < Y_{j'}, Y_{i'} < \tau_\mathrm{C}, \delta_{i'} = 1, A_{j'} \leq Y_{i'})},$$

which can be regarded as weights of $\Phi\left(\frac{\mathbf{X}_i^\top \boldsymbol{\beta} - \mathbf{X}_j^\top \boldsymbol{\beta}}{h}\right)$.

Moreover, taking the derivative of (4) with respect to $\boldsymbol{\beta}$ yields the estimating function

$$\begin{aligned} \mathbf{g}(\boldsymbol{\beta}) &\triangleq \frac{\partial}{\partial \boldsymbol{\beta}} C(\boldsymbol{\beta}) \\ &= \frac{1}{n(n-1)} \sum_{i,j=1}^n \widehat{\omega}_{ij} \phi\left(\frac{\mathbf{X}_i^\top \boldsymbol{\beta} - \mathbf{X}_j^\top \boldsymbol{\beta}}{h}\right) \frac{\mathbf{X}_i - \mathbf{X}_j}{h}, \end{aligned} \tag{5}$$

where $\phi(\cdot)$ is the standard normal density function. Let $g_k(\boldsymbol{\beta})$ denote the kth component of (5). Moreover, by the spirit of the score test screening (e.g., Zhao and Li, 2014), we compute $hg_k(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \mathbf{0}_p$, which reflects the numerator of the score statistic for a hypothesis $\beta_k = 0$ and can be taken as a sensible screening statistic. That is, we define

$$\begin{aligned} \widehat{\rho}_k &\triangleq hg_k(\mathbf{0}_p) \\ &= \frac{1}{n(n-1)} \sum_{i,j=1}^n \widehat{\omega}_{ij} \left(X_{i,(k)} - X_{j,(k)}\right) \end{aligned} \tag{6}$$

and adopt it to screen the covariates. Given a thresholding value γ_n, the estimated active set is given by

$$\widehat{\mathcal{I}} = \left\{k : \widehat{\rho}_k > \gamma_n \text{ for } k = 1, \ldots, p\right\}, \tag{7}$$

and the size of $\widehat{\mathcal{I}}$ is less than n. Now, and hereafter, we call the feature screening procedure (7) *CI-LTRC*.

3.2. Sure screening property

In this section, we discuss theoretical results of the proposed CI-LTRC method. We first impose the following conditions that are used to derive the desired result:

(C1) $\widehat{G}(y)$ is a consistent estimator of $G_0(y)$, where $G_0(y)$ is the true function of $G(y) = \int_0^t F(y-a)dH(a)$.
(C2) The covariate $X_{i,(k)}$ with $k = 1, \ldots, p$ is uniformly bounded. That is, there exists a constant $\kappa > 0$ such that $\sup_{i=1,\ldots,n} \left|X_{i,(k)}\right| < \kappa$.
(C3) Assume that $\min_{1\le k\le |\mathcal{I}|} \widehat{\rho}_k \ge Q_1 n^{-2/(2+\alpha)}$ for some positive constants Q_1 and α.

Condition (C1) says that the estimator $\widehat{G}(y)$ converges in probability to $G_0(y)$ for all $y > 0$ when $n \to \infty$. Condition (C2) reflects that the covariates are bounded above. Condition (C3) indicates that marginal signal (6) of active covariates cannot be small (e.g., Fan and Lv 2008, Condition 3).

Based on three conditions (C1)-(C3), we present the probabilistic bound of $\widehat{\rho}_k$ in (6) and the sure screening property of $\widehat{\mathcal{I}}$ in (7). Detailed descriptions are summarized in the following theorem:

Theorem 3.1. *Suppose that regularity conditions (C1), (C2) and (C3) hold.*

(a) the probabilistic bound: There exist positive constants Q_1, c, K_1, and α, such that

$$\begin{aligned} &P\Big(\max_{1\le k\le p}\left|\widehat{\rho}_k-\rho_k\right| \ge Q_1 n^{-2/(2+\alpha)}\Big) \\ &\le p\left[c\exp\left\{-\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2}\right\}+2\exp\left\{-\frac{d_n Q_1^2 n^{2\alpha/(2+\alpha)}}{2n\left(64neK_1^2+4K_1Q_1n^{-2/(2+\alpha)}\right)}\right\}\right], \end{aligned} \tag{8}$$

where

$$\rho_k = \frac{E\left[\left\{G_0(Y_i)\right\}^{-2}\mathbb{I}(\tau_A < Y_i < Y_j, Y_i < \tau_C, \delta_i = 1, A_j \le Y_i)(X_{i,(k)} - X_{j,(k)})\right]}{E\left[\left\{G_0(Y_{i'})\right\}^{-2}\mathbb{I}(\tau_A < Y_{i'} < Y_{j'}, Y_{i'} < \tau_C, \delta_{i'} = 1, A_{j'} \le Y_{i'})\right]}$$

and $d_n = \left[\frac{n}{2}\right]$ is the greatest integer and is less than $n/2$.

(b) the sure screening property: By (a) and Condition (C3), we have that

$$\begin{aligned} P\left(\widehat{\mathcal{I}} \supseteq \mathcal{I}\right) \ge 1 - q&\left[c\exp\left\{-\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2}\right\}\right. \\ &\left.+2\exp\left\{-\frac{d_n Q_1^2 n^{2\alpha/(2+\alpha)}}{2n\left(64neK_1^2+4K_1Q_1n^{-2/(2+\alpha)}\right)}\right\}\right], \end{aligned}$$

where q is the cardinality of $\mathcal{I}$.

Similar to the result in existing work (e.g., Li et al., 2012; Ma et al., 2017; Chen, 2021, 2023), (8) indicates that the absolute difference $\left|\widehat{\rho}_k - \rho_k\right|$ is bounded above by $Q_1 n^{-2/(2+\alpha)}$ with a large probability, and verifies that the proposed method is able to handle the nonpolynomial (NP) dimensionality. In addition, the result (a) also implies the result (b), which says that the CI-LTRC method is able to retain all the truly important covariates with an overwhelming probability under the LTRC data structure.

4. Proof of Theorem 3.1

In this section, we provide the theoretical derivations for Theorem 3.1. We first prove the result (a), and then use it to derive part (b). The required technical lemmas are placed in Appendix.

4.1. Proof of (a):

We first express $\widehat{\rho}_k - \rho_k$ as

$$\begin{aligned} \widehat{\rho}_k - \rho_k = \frac{1}{n(n-1)}\sum_{i,j=1}^{n}&\Bigg(\widehat{w}_{ij}\left(X_{i,(k)} - X_{j,(k)}\right) \\ &-\frac{E\left[\left\{G_0(Y_i)\right\}^{-2}\mathbb{I}(\tau_{\mathrm{A}} < Y_i < Y_j, Y_i < \tau_{\mathrm{C}}, \delta_i = 1, A_j \le Y_i)(X_{i,(k)} - X_{j,(k)})\right]}{E\left[\left\{G_0(Y_{i'})\right\}^{-2}\mathbb{I}(\tau_{\mathrm{A}} < Y_{i'} < Y_{j'}, Y_{i'} < \tau_{\mathrm{C}}, \delta_{i'} = 1, A_{j'} \le Y_{i'})\right]}\Bigg). \end{aligned} \tag{9}$$

By Condition (C1), we have that $\sup_y \left|\widehat{G}(y) - G_0(y)\right| \to 0$, and by the consistency of U-statistics (e.g., van der Vaart, 1998, Chapter 12), we have that, as n is sufficiently large,

$$\frac{1}{n(n-1)}\sum_{i'=1}^{n}\sum_{j'=1}^{n}\left\{\widehat{G}(Y_{i'})\right\}^{-2}\mathbb{I}(\tau_{\mathrm{A}} < Y_{i'} < Y_{j'}, Y_{i'} < \tau_{\mathrm{C}}, \delta_{i'} = 1, A_{j'} \le Y_{i'})$$

$$\begin{aligned}&\xrightarrow{p} E\left[\left\{G_0(Y_{i'})\right\}^{-2}\mathbb{I}(\tau_{\mathrm{A}} < Y_{i'} < Y_{j'}, Y_{i'} < \tau_{\mathrm{C}}, \delta_{i'} = 1, A_{j'} \le Y_{i'})\right]\\ &\triangleq D.\end{aligned}$$

Then (9) gives that

$$\begin{aligned}\widehat{\rho}_k - \rho_k = &\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{1}{D}\Bigg(\left\{\widehat{G}(Y_i)\right\}^{-2}\mathbb{I}(\mathcal{S}_{ij})\left(X_{i,(k)} - X_{j,(k)}\right)\\ &- E\left[\left\{G_0(Y_i)\right\}^{-2}\mathbb{I}(\mathcal{S}_{ij})(X_{i,(k)} - X_{j,(k)})\right]\Bigg)\end{aligned} \tag{10}$$

with $\mathcal{S}_{ij} \triangleq \left\{\tau_{\mathrm{A}} < Y_i < Y_j, Y_i < \tau_{\mathrm{C}}, \delta_i = 1, A_j \le Y_i\right\}$. By adding and subtracting additional terms, we further express (10) as

$$\widehat{\rho}_k - \rho_k \triangleq A_1 + A_2, \tag{11}$$

where

$$\begin{aligned}A_1 = &\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{1}{D}\Bigg(\left[\left\{\widehat{G}(T_i)\right\}^{-2} - \left\{G_0(T_i)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})\Big\{\left(X_{i,(k)} - X_{j,(k)}\right)\\ &- E\left(X_{i,(k)} - X_{j,(k)}\right)\Big\} + \left[\left\{\widehat{G}(T_i)\right\}^{-2} - \left\{G_0(T_i)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})E\left(X_{i,(k)} - X_{j,(k)}\right)\Bigg)\end{aligned}$$

and

$$\begin{aligned}A_2 = &\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{1}{D}\Bigg(\left\{G_0(T_i)\right\}^{-2}\mathbb{I}(\mathcal{S}_{ij})\left(X_{i,(k)} - X_{j,(k)}\right)\\ &- E\left[\left\{G_0(T_i)\right\}^{-2}\mathbb{I}(\mathcal{S}_{ij})\left(X_{i,(k)} - X_{j,(k)}\right)\right]\Bigg).\end{aligned}$$

In the remaining derivation, we examine A_1 and A_2 separately.

Examine A_1:

For A_1, we define $V_{i,(k)} \triangleq (X_{i,(k)}, T_i)$ and obtain that

$$\begin{aligned}A_1 = &\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{1}{D}\Bigg(\left[\left\{\widehat{G}(T_i)\right\}^{-2} - \left\{G(T_i)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})\Big\{\left(X_{i,(k)} - X_{j,(k)}\right)\\ &- E\left(X_{i,(k)} - X_{j,(k)}\right)\Big\} + \left[\left\{\widehat{G}(T_i)\right\}^{-2} - \left\{G(T_i)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})E\left(X_{i,(k)} - X_{j,(k)}\right)\Bigg)\\ \triangleq &\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\widehat{\varphi}(V_{i,(k)}, V_{j,(k)})\\ = &\,\mathbb{P}\widehat{\varphi},\end{aligned}$$

where the symbol $\mathbb{P}$ represents the empirical measure defined in Appendix.

Let $\mathcal{F}$ denote the class of the "working" functions of $G(\cdot)$ that aim to estimate the true survivor function. Define a class of functions

$$\begin{aligned}\mathcal{H} = &\Bigg\{\varphi(V_{i,(k)}, V_{j,(k)}) \triangleq \left[\{G(y)\}^{-2} - \left\{G_0(y)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})\left\{\left(X_{i,(k)} - X_{j,(k)}\right) - E\left(X_{i,(k)} - X_{j,(k)}\right)\right\}\\ &+ \left[\{G(y)\}^{-2} - \left\{G_0(y)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})E\left(X_{i,(k)} - X_{j,(k)}\right) : G(y) \in \mathcal{F}\Bigg\},\end{aligned}$$

where function $\varphi(\cdot,\cdot)$ in $\mathcal{H}$ differs from $\widehat{\varphi}(\cdot,\cdot)$ in that the former function involves $G(y)$ whereas the latter function contains $\widehat{G}(y)$.

Now we want to apply Lemma A.1 to yield the desired result. To this end, we verify the required two conditions of Lemma A.1. Since the survivor functions are monotone and the support of $V_{i,(k)}$ is bounded, by Corollary 2.7.2 of van der Vaart and Wellner (1996), the bracketing number $N_{[\ ]}(\epsilon, \mathcal{H}, \|\cdot\|_{q,\mathcal{P}})$ is bounded above. It remains to verify that $\sup_{\varphi\in\mathcal{H}}\|\varphi\|_\infty \le 1$.

Noting that for $\varphi \in \mathcal{H}$, we have that

$$\begin{aligned}\|\varphi\|_\infty \le &\left\|\left[\left\{G(T_i)\right\}^{-2} - \left\{G_0(T_i)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})\left\{\left(X_{i,(k)} - X_{j,(k)}\right) - E\left(X_{i,(k)} - X_{j,(k)}\right)\right\}\right\|_\infty\\ &+ \left\|\left[\left\{G(T_i)\right\}^{-2} - \left\{G_0(T_i)\right\}^{-2}\right]\mathbb{I}(\mathcal{S}_{ij})E\left(X_{i,(k)} - X_{j,(k)}\right)\right\|_\infty\\ \le &\left\|\left\{G(T_i)\right\}^{-2} - \left\{G_0(T_i)\right\}^{-2}\right\|_\infty\left\|\mathbb{I}(\mathcal{S}_{ij})\right\|_\infty\left\|\left(X_{i,(k)} - X_{j,(k)}\right) - E\left(X_{i,(k)} - X_{j,(k)}\right)\right\|_\infty\end{aligned}$$

$$+\left\| \left\{G(T_i)\right\}^{-2} - \left\{G_0(T_i)\right\}^{-2} \right\|_\infty \left\| \mathbb{I}(S_{ij}) \right\|_\infty \left\| E\left(X_{i,(k)} - X_{j,(k)}\right) \right\|_\infty, \tag{12}$$

where the second step is due to the triangle inequality of the infinity norm. We particularly examine the right-hand side of (12) by considering those $G(y) \in \mathcal{F}$ satisfying Condition (C1). Then we have that

$$\begin{aligned} \|\varphi\|_\infty &\leq o_p\left(\frac{1}{\sqrt{n}}\right) \times 1 \times \left\{ \left\| X_{i,(k)} - X_{j,(k)} \right\|_\infty + 2E\left(\left| X_{i,(k)} - X_{j,(k)} \right| \right) \right\} \\ &\leq o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{13}$$

When n is large enough, taking the supremum on (13) gives $\sup_{\varphi \in \mathcal{H}} \|\varphi\|_\infty \leq 1$. Therefore, by Lemma A.1, we have that

$$P\left(\sup_{\varphi \in \mathcal{H}, \|\varphi\|_{2,\mathcal{P}} \leq n^{-1/(2+\alpha)}} |\mathbb{P}\varphi - \mathcal{P}\varphi| \geq Q_1 n^{-2/(2+\alpha)} \right) \leq c \exp\left\{ -\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2} \right\} \tag{14}$$

and

$$P\left(\sup_{\varphi \in \mathcal{H}, \|\varphi\|_{2,\mathcal{P}} > n^{-1/(2+\alpha)}} \frac{|\mathbb{P}\varphi - \mathcal{P}\varphi|}{\|\varphi\|_{2,\mathcal{P}}^{1-\alpha/2}} \geq Q_1 n^{-1/2} \right) \leq c \exp\left(-\frac{Q_1}{c^2} \right). \tag{15}$$

Finally, noting that $\widehat{G}(y) \in \mathcal{F}$ suggests that $\widehat{\varphi}(V_{i,(k)}, V_{j,(k)}) \in \mathcal{H}$, we obtain that $|\mathbb{P}\widehat{\varphi} - \mathcal{P}\widehat{\varphi}| \leq \sup_{\varphi \in \mathcal{H}, \|\varphi\|_{2,\mathcal{P}} \leq n^{-1/(2+\alpha)}} |\mathbb{P}\varphi - \mathcal{P}\varphi|$. Moreover, $\mathcal{P}\widehat{\varphi} = 0$ due to Condition (C1). Consequently, we have that

$$P\left(\left| A_1 \right| \geq Q_1 n^{-2/(2+\alpha)} \right) \leq c \exp\left\{ -\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2} \right\}. \tag{16}$$

Examine A_2:

In this step, we derive the probabilistic inequality for A_2. Specifically, by the Minkowski inequality, for any $m \geq 2$, we have that

$$\begin{aligned} & E\left(\left| \left\{G_0(T_i)\right\}^{-2} \mathbb{I}(S_{ij}) \left(X_{i,(k)} - X_{j,(k)}\right) - E\left[\left\{G_0(T_i)\right\}^{-2} \mathbb{I}(S_{ij}) \left(X_{i,(k)} - X_{j,(k)}\right) \right] \right|^m \right) \\ \leq\ & 2^m E\left[\left| \left\{G_0(T_i)\right\}^{-2} \mathbb{I}(S_{ij}) \left(X_{i,(k)} - X_{j,(k)}\right) \right|^m \right] \\ \leq\ & 2^m M E\left[\left| X_{i,(k)} - X_{j,(k)} \right|^m \right] \\ \leq\ & 2e(4K_1)^m m! \end{aligned}$$

with $K_1^m \triangleq MR^m$, where the third step is due to the boundness of $\left\{G_0(T_i)\right\}^{-2}$ with positive upper bound M, and the last inequality is obtained by Lemma A.2. Therefore, by Lemma A.3, for any $\delta > 0$, we have that

$$P\left(\left| A_2 \right| \geq \frac{\delta}{n} \right) \leq 2 \exp\left\{ -\frac{d_n \delta^2}{2n\left(64neK_1^2 + 4K_1\delta\right)} \right\}. \tag{17}$$

With $\frac{\delta}{n}$ in (17) replaced by $Q_1 n^{-2/(2+\alpha)}$, we can obtain that

$$P\left(\left| A_2 \right| \geq Q_1 n^{-2/(2+\alpha)} \right) \leq 2 \exp\left\{ -\frac{d_n Q_1^2 n^{2\alpha/(2+\alpha)}}{2n\left(64neK_1^2 + 4K_1 Q_1 n^{-2/(2+\alpha)} \right)} \right\}. \tag{18}$$

Derive the desired result:

Together with (16) and (18), we have that

$$\begin{aligned} P\left(\left| \widehat{\rho}_k - \rho_k \right| \geq Q_1 n^{-2/(2+\alpha)} \right) \leq\ & 2 \exp\left\{ -\frac{d_n Q_1^2 n^{2\alpha/(2+\alpha)}}{2n\left(64neK_1^2 + 4K_1 Q_1 n^{-2/(2+\alpha)}\right)} \right\} \\ & + c \exp\left\{ -\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2} \right\}. \end{aligned} \tag{19}$$

Moreover, by (19), we obtain that

$$\begin{aligned} & P\left(\max_{1 \leq k \leq p} \left| \widehat{\rho}_k - \rho_k \right| \geq Q_1 n^{-2/(2+\alpha)} \right) \\ \leq\ & pP\left(\left| \widehat{\rho}_k - \rho_k \right| \geq Q_1 n^{-2/(2+\alpha)} \right) \\ \leq\ & p \left[c \exp\left\{ -\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2} \right\} + 2 \exp\left\{ -\frac{d_n Q_1^2 n^{2\alpha/(2+\alpha)}}{2n\left(64neK_1^2 + 4K_1 Q_1 n^{-2/(2+\alpha)}\right)} \right\} \right]. \end{aligned} \tag{20}$$

Table 1
Simulation results: feature screening for model M1 under $\widetilde{A}$ following the uniform distribution and Scenario I. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.970	0.980	0.970	0.900	0.840	0.615	0.611	0.611	0.613	0.600	0.660	0.850	0.840	0.730	0.610
			1500	0.940	1.000	1.000	0.960	0.900	0.611	0.605	0.613	0.610	0.600	0.620	0.840	0.810	0.620	0.614
			3000	0.870	0.990	0.990	0.870	0.730	0.570	0.609	0.609	0.615	0.560	0.610	0.850	0.790	0.610	0.602
		400	1000	1.000	1.000	1.000	1.000	1.000	0.651	0.658	0.656	0.638	0.635	0.792	0.790	0.797	0.797	0.790
			1500	1.000	1.000	1.000	1.000	1.000	0.635	0.648	0.646	0.634	0.630	0.788	0.792	0.794	0.790	0.783
			3000	1.000	1.000	1.000	1.000	1.000	0.625	0.641	0.640	0.630	0.623	0.786	0.794	0.789	0.788	0.782
		600	1.000	1.000	1.000	1.000	1.000	1.000	0.673	0.681	0.674	0.620	0.615	0.794	0.790	0.789	0.797	0.785
			1500	1.000	1.000	1.000	1.000	1.000	0.630	0.681	0.672	0.580	0.577	0.792	0.796	0.798	0.797	0.791
			3000	1.000	1.000	1.000	1.000	1.000	0.530	0.670	0.630	0.644	0.642	0.795	0.798	0.798	0.792	0.790
	50%	150	1000	0.970	0.980	0.980	0.930	0.910	0.613	0.613	0.615	0.613	0.610	0.780	0.792	0.798	0.786	0.762
			1500	0.960	0.990	0.990	0.990	0.960	0.611	0.615	0.615	0.610	0.600	0.780	0.794	0.789	0.795	0.776
			3000	0.940	0.900	0.890	0.870	0.870	0.605	0.606	0.609	0.608	0.600	0.789	0.794	0.795	0.790	0.784
		400	1000	1.000	1.000	1.000	0.990	0.990	0.648	0.630	0.654	0.634	0.614	0.789	0.799	0.788	0.789	0.788
			1500	1.000	1.000	1.000	1.000	1.000	0.628	0.647	0.653	0.635	0.629	0.794	0.791	0.788	0.799	0.787
			3000	1.000	1.000	1.000	1.000	1.000	0.625	0.636	0.633	0.620	0.620	0.788	0.791	0.783	0.789	0.780
		600	1000	1.000	1.000	1.000	1.000	1.000	0.640	0.674	0.667	0.615	0.606	0.791	0.785	0.789	0.788	0.785
			1500	1.000	1.000	1.000	1.000	1.000	0.610	0.677	0.670	0.579	0.572	0.773	0.781	0.790	0.784	0.770
			3000	1.000	1.000	1.000	1.000	1.000	0.400	0.670	0.625	0.635	0.638	0.791	0.793	0.789	0.787	0.785
50%	15%	150	1000	0.960	0.920	0.920	0.950	0.960	0.600	0.579	0.585	0.567	0.530	0.550	0.780	0.750	0.600	0.530
			1500	0.940	0.950	0.930	0.940	0.940	0.570	0.579	0.575	0.530	0.524	0.520	0.720	0.760	0.540	0.535
			3000	0.930	0.950	0.910	0.940	0.900	0.538	0.578	0.572	0.545	0.531	0.533	0.760	0.740	0.744	0.513
		400	1000	0.998	1.000	1.000	1.000	0.998	0.592	0.593	0.596	0.587	0.587	0.780	0.783	0.785	0.789	0.779
			1500	0.997	1.000	1.000	1.000	0.997	0.579	0.575	0.572	0.585	0.570	0.786	0.783	0.795	0.781	0.780
			3000	0.973	1.000	1.000	0.987	0.960	0.533	0.564	0.571	0.567	0.560	0.780	0.790	0.773	0.784	0.770
		600	1000	1.000	1.000	1.000	1.000	1.000	0.593	0.594	0.595	0.587	0.587	0.787	0.783	0.772	0.783	0.768
			1500	1.000	1.000	1.000	1.000	1.000	0.581	0.579	0.576	0.565	0.560	0.768	0.759	0.760	0.779	0.754
			3000	1.000	1.000	1.000	1.000	1.000	0.403	0.574	0.574	0.562	0.400	0.773	0.778	0.772	0.756	0.751
	50%	150	1000	0.928	0.937	0.969	0.963	0.928	0.563	0.568	0.578	0.560	0.554	0.548	0.776	0.749	0.579	0.546
			1500	0.922	0.935	0.953	0.956	0.920	0.490	0.563	0.570	0.520	0.478	0.517	0.780	0.787	0.780	0.767
			3000	0.921	0.923	0.924	0.944	0.917	0.440	0.564	0.571	0.517	0.436	0.531	0.751	0.732	0.729	0.726
		400	1000	0.990	1.000	0.990	0.970	0.960	0.499	0.564	0.570	0.578	0.497	0.770	0.778	0.776	0.771	0.766
			1500	1.000	1.000	1.000	0.988	0.988	0.562	0.569	0.567	0.576	0.554	0.781	0.779	0.789	0.773	0.769
			3000	0.990	1.000	1.000	0.950	0.940	0.526	0.534	0.569	0.557	0.513	0.780	0.789	0.769	0.778	0.761
		600	1000	1.000	1.000	1.000	1.000	1.000	0.591	0.589	0.579	0.579	0.568	0.778	0.779	0.769	0.780	0.765
			1500	1.000	1.000	1.000	1.000	1.000	0.579	0.576	0.571	0.558	0.552	0.759	0.753	0.754	0.768	0.753
			3000	1.000	1.000	1.000	1.000	1.000	0.388	0.574	0.569	0.557	0.383	0.771	0.777	0.767	0.750	0.745

4.2. Proof of (b):

Recall that $\mathcal{I}$ and $\widehat{\mathcal{I}}$ are defined in Section 3.2. By (20) and similar derivations in Li et al. (2012), we can show that

$$\begin{aligned} P\left(\widehat{\mathcal{I}} \supseteq \mathcal{I}\right) &\geq P\left(\min_{1\leq k\leq p} \left|\widehat{\rho}_k - \rho_k\right| > Q_1 n^{-2/(2+\alpha)}\right) \\ &\geq 1 - qP\left(\left|\widehat{\rho}_k - \rho_k\right| > Q_1 n^{-2/(2+\alpha)}\right) \\ &\geq 1 - q\left[c\exp\left\{-\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2}\right\}\right. \\ &\quad \left. + 2\exp\left\{-\frac{d_n Q_1^2 n^{2\alpha/(2+\alpha)}}{2n\left(64neK_1^2 + 4K_1Q_1 n^{-2/(2+\alpha)}\right)}\right\}\right], \end{aligned} \tag{21}$$

where q is the cardinality of $\mathcal{I}$. Moreover, with $\alpha > 0$, when $n \to \infty$, we have that $P\left(\widehat{\mathcal{I}} \supseteq \mathcal{I}\right) \to 1$. It indicates that the estimated active set $\widehat{\mathcal{I}}$ includes the true active set that contains truly important predictors with probability approaching one. Therefore, the proof is completed. □

Table 2
Simulation results: feature screening for model M2 under $\widetilde{A}$ following the uniform distribution and Scenario I. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.985	0.998	0.995	0.997	0.985	0.630	0.657	0.697	0.686	0.626	0.784	0.795	0.795	0.787	0.780
			1500	0.989	0.998	0.997	0.998	0.987	0.630	0.660	0.693	0.684	0.628	0.782	0.793	0.793	0.780	0.779
			3000	0.986	0.993	0.997	0.997	0.984	0.645	0.671	0.692	0.692	0.644	0.772	0.789	0.785	0.786	0.770
		400	1000	1.000	1.000	1.000	1.000	1.000	0.670	0.691	0.697	0.705	0.664	0.811	0.803	0.799	0.820	0.797
			1500	1.000	1.000	1.000	1.000	1.000	0.680	0.689	0.687	0.679	0.677	0.805	0.801	0.796	0.812	0.794
			3000	1.000	1.000	1.000	1.000	1.000	0.671	0.679	0.676	0.666	0.665	0.799	0.789	0.783	0.795	0.780
		600	1000	1.000	1.000	1.000	1.000	1.000	0.697	0.694	0.697	0.710	0.694	0.826	0.820	0.807	0.833	0.800
			1500	1.000	1.000	1.000	1.000	1.000	0.688	0.689	0.693	0.695	0.687	0.810	0.821	0.810	0.824	0.808
			3000	1.000	1.000	1.000	1.000	1.000	0.696	0.680	0.689	0.684	0.681	0.805	0.814	0.803	0.819	0.800
	50%	150	1000	0.990	0.996	0.996	0.998	0.990	0.632	0.643	0.645	0.621	0.620	0.761	0.787	0.784	0.769	0.761
			1500	0.995	0.996	0.998	0.998	0.995	0.626	0.638	0.639	0.624	0.623	0.764	0.782	0.788	0.764	0.759
			3000	1.000	1.000	1.000	1.000	1.000	0.622	0.637	0.634	0.616	0.610	0.765	0.688	0.736	0.751	0.683
		400	1000	1.000	1.000	1.000	1.000	1.000	0.683	0.691	0.691	0.686	0.679	0.801	0.798	0.799	0.799	0.798
			1500	1.000	1.000	1.000	1.000	1.000	0.659	0.686	0.682	0.681	0.653	0.792	0.787	0.798	0.795	0.781
			3000	0.998	1.000	1.000	1.000	0.998	0.652	0.678	0.680	0.655	0.648	0.794	0.796	0.793	0.792	0.788
		600	1000	1.000	1.000	1.000	1.000	1.000	0.691	0.699	0.696	0.693	0.688	0.797	0.802	0.806	0.799	0.796
			1500	1.000	1.000	1.000	1.000	1.000	0.681	0.694	0.690	0.687	0.680	0.794	0.798	0.798	0.796	0.793
			3000	1.000	1.000	1.000	1.000	1.000	0.679	0.681	0.694	0.691	0.676	0.786	0.789	0.799	0.777	0.774
50%	15%	150	1000	0.975	0.990	0.992	0.977	0.975	0.622	0.627	0.622	0.625	0.620	0.749	0.733	0.753	0.739	0.728
			1500	0.950	0.975	0.986	0.960	0.950	0.590	0.626	0.630	0.622	0.588	0.730	0.714	0.759	0.763	0.727
			3000	0.945	0.972	0.975	0.949	0.944	0.553	0.614	0.625	0.615	0.550	0.753	0.784	0.774	0.750	0.750
		400	1000	1.000	1.000	1.000	1.000	1.000	0.621	0.626	0.631	0.620	0.617	0.788	0.796	0.794	0.789	0.781
			1500	1.000	1.000	1.000	1.000	1.000	0.617	0.614	0.621	0.618	0.611	0.793	0.799	0.796	0.795	0.789
			3000	0.939	1.000	1.000	0.959	0.898	0.590	0.610	0.612	0.619	0.583	0.796	0.794	0.789	0.784	0.782
		600	1000	1.000	1.000	1.000	1.000	1.000	0.645	0.647	0.654	0.641	0.636	0.786	0.809	0.792	0.797	0.780
			1500	1.000	1.000	1.000	1.000	1.000	0.635	0.642	0.640	0.637	0.628	0.791	0.789	0.797	0.786	0.785
			3000	1.000	1.000	1.000	1.000	1.000	0.624	0.632	0.620	0.619	0.616	0.785	0.785	0.776	0.779	0.774
	50%	150	1000	0.994	0.994	0.996	0.995	0.991	0.561	0.570	0.573	0.566	0.560	0.720	0.697	0.749	0.800	0.695
			1500	0.984	0.990	0.996	0.991	0.980	0.564	0.573	0.554	0.553	0.550	0.721	0.729	0.722	0.731	0.720
			3000	0.986	0.995	0.997	0.996	0.984	0.558	0.566	0.532	0.560	0.555	0.729	0.718	0.719	0.727	0.714
		400	1000	1.000	1.000	1.000	1.000	1.000	0.523	0.529	0.527	0.524	0.514	0.740	0.790	0.792	0.770	0.735
			1500	0.997	0.996	1.000	1.000	0.996	0.514	0.524	0.529	0.518	0.512	0.762	0.783	0.781	0.764	0.732
			3000	1.000	1.000	1.000	0.998	0.998	0.490	0.517	0.512	0.498	0.474	0.747	0.790	0.767	0.754	0.745
		600	1000	0.999	1.000	1.000	1.000	0.999	0.634	0.649	0.648	0.641	0.631	0.799	0.795	0.803	0.789	0.785
			1500	0.997	1.000	0.990	0.997	0.997	0.636	0.641	0.633	0.625	0.623	0.791	0.783	0.779	0.791	0.773
			3000	0.996	0.997	1.000	1.000	0.996	0.615	0.624	0.638	0.622	0.612	0.789	0.773	0.781	0.787	0.770

5. Numerical studies

5.1. Simulation design

In this section we conduct simulation studies to evaluate the finite sample performance of the proposed method, where we set the sample size n to be 150, 400 or 600 and the dimension of covariates p is given by 1000, 1500, or 3000. For $i = 1, \ldots, n$, we independently generate $\mathbf{X}_i \triangleq (X_{i,(1)}, X_{i,(2)}, \ldots, X_{i,(p)})^\top$ from the following two scenarios:

I. the multivariate normal distribution $\mathbf{X}_i \sim N(\mathbf{0}_p, \boldsymbol{\Sigma}_X)$, where $\mathbf{0}_p$ is the p-dimensional zero vector, $\boldsymbol{\Sigma}_X$ is the covariance matrix with entry (i, j) specified as $0.5^{|i-j|}$ for $i, j = 1, \ldots, p$.

II. $X_{i,(1)}$ follows the Bernoulli distribution with probability 0.5, $X_{i,(2)}$ follows the Poisson distribution with mean 2, and the remaining covariates $X_{i,(j)}$ follow the standard normal distribution for $j = 3, \ldots, p$.

Scenario I says that the covariates $\mathbf{X}_i$ are all continuous random variables, and Scenario II says that the covariates contain mixed distributions, where the first two covariates are discrete and the remaining ones are continuous. In our study, we let the true vector of the parameters be $\boldsymbol{\beta}_0 = (1, 1, 1, 1, \mathbf{0}_{p-4}^\top)^\top$, which indicates that the first four covariates are informative to the response. Our goal is to detect the first four covariates by feature screening methods under Scenarios I and II.

Given the covariates and $\boldsymbol{\beta}_0$, we examine two frequently used survival models:

Table 3
Simulation results: feature screening for model M1 under $\widetilde{A}$ following the exponential distribution and Scenario I. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.987	0.981	0.981	0.985	0.984	0.609	0.619	0.624	0.616	0.600	0.791	0.787	0.791	0.792	0.785
			1500	0.967	0.988	0.989	0.971	0.965	0.622	0.620	0.614	0.613	0.610	0.790	0.785	0.785	0.796	0.783
			3000	0.966	0.987	0.985	0.949	0.943	0.613	0.610	0.618	0.612	0.608	0.791	0.789	0.795	0.794	0.785
		400	1000	0.999	1.000	1.000	1.000	0.999	0.675	0.680	0.681	0.661	0.661	0.892	0.896	0.896	0.895	0.890
			1500	1.000	1.000	1.000	0.998	0.998	0.675	0.672	0.674	0.653	0.652	0.890	0.895	0.895	0.879	0.868
			3000	0.999	1.000	1.000	0.997	0.996	0.675	0.666	0.665	0.654	0.653	0.869	0.853	0.864	0.826	0.823
		600	1000	1.000	1.000	1.000	1.000	1.000	0.684	0.686	0.687	0.678	0.671	0.896	0.898	0.897	0.896	0.895
			1500	1.000	1.000	1.000	1.000	1.000	0.677	0.679	0.678	0.673	0.672	0.894	0.897	0.895	0.896	0.892
			3000	1.000	1.000	1.000	1.000	1.000	0.672	0.679	0.678	0.673	0.671	0.886	0.858	0.879	0.847	0.845
	50%	150	1000	0.994	0.996	0.998	0.996	0.992	0.636	0.648	0.637	0.631	0.628	0.720	0.722	0.726	0.714	0.713
			1500	0.992	0.996	0.998	0.991	0.990	0.625	0.633	0.631	0.628	0.627	0.714	0.716	0.727	0.718	0.711
			3000	0.994	0.997	0.997	0.995	0.994	0.617	0.619	0.623	0.618	0.613	0.711	0.717	0.718	0.707	0.705
		400	1000	0.998	1.000	1.000	0.999	0.998	0.672	0.761	0.725	0.683	0.670	0.858	0.852	0.831	0.865	0.831
			1500	1.000	1.000	1.000	1.000	1.000	0.672	0.764	0.720	0.671	0.670	0.859	0.847	0.825	0.856	0.825
			3000	0.986	1.000	1.000	1.000	0.986	0.649	0.689	0.671	0.662	0.643	0.842	0.844	0.812	0.842	0.811
		600	1000	0.998	0.997	0.998	0.998	0.997	0.678	0.679	0.679	0.678	0.676	0.882	0.891	0.896	0.882	0.882
			1500	0.999	1.000	1.000	0.998	0.998	0.679	0.677	0.679	0.678	0.677	0.877	0.886	0.863	0.878	0.859
			3000	0.977	0.977	1.000	1.000	0.977	0.679	0.674	0.678	0.675	0.673	0.827	0.864	0.839	0.850	0.825
50%	15%	150	1000	0.941	0.934	0.937	0.939	0.934	0.567	0.578	0.582	0.575	0.567	0.769	0.806	0.868	0.866	0.764
			1500	0.917	0.923	0.919	0.916	0.916	0.567	0.579	0.589	0.576	0.566	0.781	0.723	0.813	0.835	0.780
			3000	0.906	0.918	0.910	0.906	0.906	0.566	0.578	0.587	0.575	0.565	0.754	0.794	0.897	0.852	0.754
		400	1000	1.000	1.000	1.000	1.000	1.000	0.565	0.566	0.571	0.565	0.564	0.792	0.793	0.794	0.791	0.785
			1500	1.000	1.000	1.000	0.997	0.997	0.565	0.565	0.568	0.565	0.565	0.788	0.796	0.796	0.792	0.787
			3000	0.996	1.000	1.000	0.999	0.995	0.563	0.565	0.565	0.563	0.563	0.791	0.776	0.754	0.785	0.749
		600	1000	1.000	1.000	1.000	1.000	1.000	0.566	0.568	0.569	0.568	0.565	0.893	0.898	0.879	0.854	0.852
			1500	1.000	1.000	1.000	1.000	1.000	0.568	0.569	0.569	0.567	0.566	0.884	0.879	0.872	0.850	0.833
			3000	1.000	1.000	1.000	1.000	1.000	0.561	0.578	0.570	0.568	0.527	0.872	0.869	0.871	0.839	0.836
	50%	150	1000	0.986	0.997	0.996	0.991	0.981	0.508	0.505	0.508	0.509	0.501	0.673	0.871	0.844	0.709	0.385
			1500	0.991	0.994	0.995	0.988	0.982	0.503	0.504	0.503	0.503	0.500	0.620	0.850	0.840	0.650	0.370
			3000	0.984	0.994	0.995	0.976	0.972	0.503	0.501	0.501	0.500	0.500	0.603	0.778	0.825	0.508	0.270
		400	1000	1.000	1.000	1.000	1.000	1.000	0.503	0.509	0.506	0.508	0.500	0.798	0.797	0.798	0.797	0.796
			1500	1.000	1.000	1.000	0.999	0.999	0.507	0.507	0.502	0.504	0.500	0.788	0.799	0.789	0.798	0.785
			3000	1.000	1.000	1.000	1.000	1.000	0.505	0.505	0.508	0.500	0.500	0.783	0.775	0.764	0.762	0.760
		600	1000	1.000	1.000	1.000	1.000	1.000	0.513	0.515	0.517	0.515	0.505	0.793	0.814	0.796	0.828	0.780
			1500	1.000	1.000	1.000	1.000	1.000	0.509	0.514	0.511	0.512	0.500	0.774	0.808	0.762	0.809	0.765
			3000	1.000	1.000	1.000	1.000	1.000	0.508	0.512	0.508	0.504	0.500	0.684	0.743	0.705	0.761	0.670

(M1) The Cox proportional hazards (PH) model

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t)\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_0) \tag{22}$$

with the baseline hazard function being specified as $\lambda_0(t) = t$, where $\lambda(t|\mathbf{X}_i)$ represents the conditional hazard function of $\widetilde{T}_i$ given $\mathbf{X}_i$.

(M2) The accelerated failure time (AFT) model

$$\log \widetilde{T}_i = (\mathbf{X}_i^\top \boldsymbol{\beta}_0) + W_i,$$

where W_i is the error term following the standard logistic distribution of the probability density function

$$f_W(w) = \frac{\exp(-w)}{\{1+\exp(-w)\}^2}.$$

For a given i, let U_i be generated from the uniform distribution in an interval $[0,1]$. Then survival times $\widetilde{T}_i$ from models M1 and M2 can be, respectively, generated by

$$\widetilde{T}_i = \sqrt{-2\exp(-\mathbf{X}_i^\top \boldsymbol{\beta}_0)\log\left(1-U_i\right)}$$

and

$$\widetilde{T}_i = \exp\left\{(\mathbf{X}_i^\top \boldsymbol{\beta}_0) + W_i\right\}.$$

Table 4
Simulation results: feature screening for model M2 under $\widetilde{A}$ following the exponential distribution and Scenario I. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.987	0.997	0.997	0.994	0.986	0.612	0.610	0.612	0.598	0.596	0.784	0.779	0.789	0.778	0.775
			1500	0.989	0.996	0.998	0.995	0.987	0.610	0.612	0.612	0.606	0.600	0.785	0.789	0.784	0.779	0.773
			3000	0.987	0.996	0.991	0.994	0.984	0.605	0.605	0.597	0.607	0.594	0.782	0.772	0.791	0.781	0.768
		400	1000	1.000	1.000	1.000	1.000	1.000	0.643	0.656	0.655	0.643	0.638	0.798	0.799	0.801	0.798	0.796
			1500	1.000	1.000	1.000	1.000	1.000	0.631	0.640	0.634	0.633	0.625	0.797	0.798	0.806	0.795	0.794
			3000	1.000	1.000	1.000	1.000	1.000	0.634	0.648	0.650	0.645	0.632	0.790	0.787	0.799	0.796	0.785
		600	1000	1.000	1.000	1.000	1.000	1.000	0.697	0.699	0.701	0.703	0.696	0.796	0.799	0.799	0.798	0.795
			1500	1.000	1.000	1.000	1.000	1.000	0.699	0.708	0.710	0.705	0.698	0.798	0.808	0.809	0.799	0.796
			3000	1.000	1.000	1.000	1.000	1.000	0.687	0.699	0.694	0.683	0.680	0.794	0.809	0.797	0.811	0.793
	50%	150	1000	0.996	0.998	0.997	0.996	0.994	0.608	0.609	0.614	0.610	0.601	0.789	0.788	0.783	0.767	0.738
			1500	0.996	0.997	0.997	0.996	0.993	0.604	0.610	0.608	0.606	0.603	0.758	0.759	0.766	0.754	0.740
			3000	0.996	0.993	0.995	0.998	0.990	0.603	0.608	0.606	0.602	0.600	0.789	0.768	0.763	0.748	0.736
		400	1000	0.994	0.998	0.996	0.995	0.994	0.641	0.651	0.647	0.642	0.639	0.793	0.795	0.789	0.797	0.786
			1500	1.000	1.000	1.000	1.000	1.000	0.634	0.650	0.644	0.633	0.630	0.793	0.797	0.799	0.795	0.790
			3000	1.000	1.000	1.000	1.000	1.000	0.625	0.630	0.634	0.623	0.618	0.784	0.787	0.791	0.788	0.780
		600	1000	1.000	1.000	1.000	1.000	1.000	0.702	0.674	0.670	0.654	0.646	0.810	0.803	0.797	0.799	0.795
			1500	1.000	1.000	1.000	1.000	1.000	0.692	0.681	0.671	0.657	0.667	0.798	0.800	0.802	0.795	0.791
			3000	1.000	1.000	1.000	1.000	1.000	0.654	0.659	0.651	0.667	0.651	0.790	0.796	0.792	0.791	0.788
50%	15%	150	1000	0.995	0.997	0.996	0.994	0.994	0.610	0.638	0.616	0.614	0.604	0.670	0.785	0.781	0.766	0.670
			1500	0.994	0.996	0.996	0.994	0.994	0.636	0.648	0.596	0.603	0.593	0.700	0.780	0.770	0.760	0.700
			3000	0.994	0.995	0.996	0.995	0.992	0.622	0.644	0.604	0.615	0.602	0.705	0.720	0.700	0.758	0.703
		400	1000	0.998	1.000	1.000	0.997	0.997	0.639	0.649	0.655	0.641	0.635	0.769	0.780	0.797	0.790	0.768
			1500	0.997	0.998	0.996	0.998	0.997	0.637	0.640	0.647	0.634	0.632	0.788	0.793	0.793	0.795	0.785
			3000	0.997	1.000	0.998	0.998	0.997	0.622	0.637	0.621	0.631	0.620	0.787	0.790	0.792	0.783	0.781
		600	1000	1.000	1.000	1.000	0.999	0.999	0.680	0.672	0.674	0.665	0.658	0.797	0.797	0.795	0.794	0.792
			1500	0.998	1.000	1.000	1.000	0.998	0.680	0.679	0.664	0.667	0.660	0.795	0.799	0.789	0.796	0.783
			3000	1.000	1.000	1.000	0.998	0.998	0.649	0.653	0.620	0.652	0.616	0.789	0.796	0.797	0.792	0.785
	50%	150	1000	0.997	0.997	0.999	0.995	0.995	0.596	0.598	0.586	0.584	0.579	0.781	0.786	0.782	0.723	0.715
			1500	0.995	0.995	1.000	0.992	0.992	0.589	0.606	0.596	0.582	0.579	0.754	0.753	0.771	0.746	0.725
			3000	0.990	0.994	0.996	0.997	0.990	0.586	0.595	0.599	0.588	0.571	0.780	0.721	0.720	0.714	0.711
		400	1000	0.994	1.000	0.997	1.000	0.994	0.627	0.644	0.649	0.627	0.624	0.793	0.789	0.795	0.789	0.784
			1500	1.000	1.000	1.000	1.000	1.000	0.632	0.656	0.644	0.628	0.625	0.788	0.797	0.799	0.793	0.785
			3000	0.998	0.999	0.998	0.997	0.997	0.625	0.627	0.631	0.617	0.612	0.786	0.784	0.785	0.785	0.782
		600	1000	1.000	1.000	1.000	1.000	1.000	0.653	0.664	0.677	0.658	0.654	0.801	0.797	0.812	0.798	0.798
			1500	1.000	1.000	1.000	0.998	0.998	0.649	0.654	0.663	0.641	0.638	0.773	0.788	0.790	0.776	0.770
			3000	0.998	1.000	1.000	0.997	0.997	0.638	0.654	0.656	0.645	0.636	0.780	0.779	0.784	0.792	0.773

To generate the biased sample with observed failure time T_i and truncation time A_i, we repeatedly generate $(\widetilde{T}_i, \widetilde{A}_i)$ and only recruit subjects whenever $\widetilde{T}_i \geq \widetilde{A}_i$ is satisfied, and we stop the recruitment procedure when the desired sample size n is achieved. Suppose that the number of repetition of data generation before achieving the desired sample size n is denoted by N_0, then the *truncation rate* $P(\widetilde{T} < \widetilde{A})$ is defined as $1 - \frac{n}{N_0}$ (e.g., Chen, 2019b). In our numerical studies, we generate the truncation time $\widetilde{A}_i$ from the exponential distribution with mean η_e or the uniform distribution under an interval $[0, \eta_u]$, where $\eta_e > 0$ and $\eta_u > 0$ are pre-specified constants so that the truncation rate is approximated 15% or 50%. Higher values of the truncation rate imply the more severe biased sampling.

For $i = 1, \ldots, n$, the censoring time C_i is generated independently from the uniform distribution in an interval $[0, \eta_c]$, where η_c is specified as a value so that the censoring rate is approximately 15% or 50%. Let $Y_i = \min\{T_i, A_i + C_i\}$ and $\Delta_i = \mathbb{I}(T_i \leq A_i + C_i)$. As a result, we have the sample of data $\{(Y_i, A_i, \mathbf{X}_i, \Delta_i) : i = 1, \ldots, n\}$. For each setting, we run 1000 simulations. To determine the size of (7), we can specify a value γ_n such that the number of selected covariates $\left|\widehat{\mathcal{I}}\right|$ is equal to $\left[\frac{n}{\log(n)}\right]$. This approach is commonly used in the framework of feature screening (e.g., Fan and Lv, 2008; Li et al., 2012; Chen, 2021, 2023).

5.2. Simulation results

To compare with the proposed method and see the impact of ignoring the effects of left-truncation and/or right-censoring, we conduct two existing methods: CSS (Ma et al., 2017) and DC-RC (Chen, 2021), where the former signal for $k = 1, \ldots, p$ is defined as

$$\widehat{\rho}_{k,\mathrm{CSS}} = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j)(X_{i,(k)} - X_{j,(k)}), \tag{23}$$

Table 5
Simulation results: feature screening for model M1 under $\widetilde{A}$ following the uniform distribution and Scenario II. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.000	0.938	0.966	0.980	0.000	0.000	0.638	0.670	0.663	0.000	0.030	0.761	0.746	0.743	0.020
			1500	0.020	0.930	0.965	0.954	0.020	0.000	0.628	0.661	0.660	0.000	0.031	0.758	0.737	0.740	0.031
			3000	0.017	0.900	0.961	0.958	0.015	0.000	0.625	0.630	0.656	0.000	0.024	0.743	0.732	0.727	0.020
		400	1000	0.180	1.000	0.989	0.995	0.160	0.000	0.679	0.725	0.723	0.000	0.190	0.860	0.840	0.820	0.180
			1500	0.104	1.000	0.980	0.994	0.104	0.000	0.677	0.721	0.717	0.000	0.190	0.863	0.761	0.816	0.184
			3000	0.010	1.000	1.000	0.990	0.010	0.000	0.640	0.713	0.716	0.000	0.080	0.800	0.690	0.730	0.070
		600	1000	0.350	1.000	1.000	1.000	0.350	0.000	0.694	0.745	0.747	0.000	0.220	0.892	0.891	0.830	0.220
			1500	0.250	1.000	1.000	1.000	0.250	0.010	0.687	0.738	0.737	0.000	0.200	0.850	0.879	0.878	0.200
			3000	0.100	1.000	1.000	1.000	0.100	0.000	0.678	0.725	0.726	0.000	0.160	0.839	0.850	0.855	0.160
	50%	150	1000	0.019	0.930	0.961	0.965	0.018	0.000	0.639	0.671	0.660	0.000	0.143	0.790	0.755	0.757	0.140
			1500	0.019	0.919	0.946	0.946	0.017	0.000	0.629	0.670	0.655	0.000	0.120	0.770	0.766	0.751	0.115
			3000	0.018	0.908	0.939	0.937	0.000	0.000	0.623	0.671	0.652	0.000	0.116	0.720	0.745	0.747	0.114
		400	1000	0.110	1.000	0.991	0.968	0.090	0.000	0.672	0.729	0.730	0.000	0.390	0.791	0.794	0.784	0.390
			1500	0.060	1.000	0.996	0.994	0.060	0.000	0.670	0.723	0.721	0.000	0.250	0.792	0.788	0.788	0.246
			3000	0.021	1.000	0.992	0.989	0.020	0.000	0.637	0.713	0.721	0.000	0.244	0.789	0.779	0.805	0.244
		600	1000	0.131	1.000	1.000	0.996	0.131	0.000	0.693	0.738	0.744	0.000	0.430	0.798	0.797	0.794	0.421
			1500	0.125	1.000	1.000	1.000	0.123	0.000	0.689	0.739	0.743	0.000	0.310	0.799	0.789	0.772	0.310
			3000	0.109	1.000	1.000	1.000	0.100	0.000	0.682	0.726	0.722	0.000	0.200	0.792	0.793	0.785	0.200
50%	15%	150	1000	0.018	0.890	0.948	0.943	0.014	0.002	0.578	0.646	0.645	0.000	0.050	0.648	0.643	0.694	0.042
			1500	0.013	0.933	0.982	0.982	0.013	0.000	0.589	0.642	0.643	0.000	0.029	0.642	0.693	0.681	0.027
			3000	0.015	0.928	0.983	0.986	0.009	0.000	0.583	0.629	0.618	0.000	0.020	0.630	0.679	0.662	0.017
		400	1000	0.020	1.000	0.994	0.997	0.020	0.020	0.583	0.649	0.660	0.010	0.170	0.688	0.791	0.783	0.160
			1500	0.017	1.000	0.995	0.994	0.000	0.013	0.578	0.684	0.683	0.000	0.195	0.685	0.781	0.770	0.193
			3000	0.010	1.000	0.986	0.993	0.005	0.000	0.600	0.674	0.690	0.000	0.150	0.710	0.765	0.751	0.140
		600	1000	0.011	1.000	1.000	1.000	0.011	0.011	0.677	0.699	0.673	0.011	0.290	0.795	0.792	0.793	0.290
			1500	0.011	1.000	0.990	1.000	0.011	0.010	0.620	0.697	0.689	0.010	0.210	0.794	0.789	0.771	0.210
			3000	0.010	1.000	0.990	0.980	0.010	0.000	0.633	0.685	0.693	0.000	0.210	0.718	0.783	0.784	0.200
	50%	150	1000	0.013	0.921	0.942	0.952	0.010	0.000	0.500	0.620	0.624	0.000	0.130	0.650	0.695	0.665	0.125
			1500	0.009	0.921	0.940	0.949	0.009	0.000	0.580	0.644	0.641	0.000	0.024	0.640	0.656	0.648	0.020
			3000	0.007	0.924	0.943	0.947	0.006	0.001	0.582	0.537	0.531	0.000	0.027	0.659	0.643	0.646	0.015
		400	1000	0.012	1.000	0.986	0.985	0.012	0.010	0.547	0.538	0.546	0.010	0.340	0.784	0.788	0.785	0.332
			1500	0.010	1.000	0.984	0.985	0.010	0.000	0.535	0.531	0.534	0.000	0.270	0.756	0.787	0.790	0.268
			3000	0.009	1.000	0.990	0.982	0.009	0.000	0.533	0.529	0.531	0.000	0.230	0.729	0.793	0.769	0.225
		600	1000	0.050	1.000	1.000	1.000	0.050	0.020	0.661	0.687	0.679	0.010	0.270	0.796	0.794	0.793	0.264
			1500	0.010	1.000	0.980	1.000	0.010	0.010	0.596	0.661	0.653	0.003	0.267	0.780	0.791	0.759	0.262
			3000	0.018	1.000	0.997	0.992	0.012	0.005	0.591	0.687	0.631	0.000	0.228	0.794	0.788	0.769	0.224

which implements the observed survival time Y_i with the ignorance of biased sampling and incomplete responses; and the latter signal for $k = 1, \ldots, p$ is defined as

$$\widehat{\rho}_{k,\mathrm{DC}} = \frac{\widehat{\mathrm{dcov}}(\widehat{Y}_{\mathrm{BJ}}, X_{(k)})}{\sqrt{\widehat{\mathrm{dcov}}(\widehat{Y}_{\mathrm{BJ}}, \widehat{Y}_{\mathrm{BJ}})\widehat{\mathrm{dcov}}(X_{(k)}, X_{(k)})}}, \tag{24}$$

where $\widehat{Y}_{\mathrm{BJ}}$ is the pseudo response derived by the Buckley-James estimator (e.g., Buckley and James, 1979) and is used to adjust the censoring effect, and $\widehat{\mathrm{dcov}}(\widehat{Y}_{\mathrm{BJ}}, X_{(k)}) = \widehat{J}_{1,k} + \widehat{J}_{2,k} - 2\widehat{J}_{3,k}$ with

$$\begin{aligned}
\widehat{J}_{1,k} &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \widehat{Y}_{\mathrm{BJ},i} - \widehat{Y}_{\mathrm{BJ},j} \right\|_1 \left\| X_{i,(k)} - X_{j,(k)} \right\|_1, \\
\widehat{J}_{2,k} &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \widehat{Y}_{\mathrm{BJ},i} - \widehat{Y}_{\mathrm{BJ},j} \right\|_1 \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| X_{i,(k)} - X_{j,(k)} \right\|_1, \\
\widehat{J}_{3,k} &= \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \left\| \widehat{Y}_{\mathrm{BJ},i} - \widehat{Y}_{\mathrm{BJ},l} \right\|_1 \left\| X_{j,(k)} - X_{l,(k)} \right\|_1.
\end{aligned}$$

(24) is only used to deal with the censoring effects but does not take left-truncation into account.

Table 6
Simulation results: feature screening for model M2 under $\widetilde{A}$ following the uniform distribution and Scenario II. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.005	0.990	0.995	0.996	0.004	0.000	0.675	0.615	0.612	0.000	0.013	0.800	0.747	0.747	0.013
			1500	0.002	0.991	0.994	0.995	0.000	0.000	0.660	0.617	0.610	0.000	0.024	0.791	0.747	0.746	0.020
			3000	0.002	0.986	0.994	0.993	0.000	0.000	0.640	0.609	0.612	0.000	0.018	0.788	0.746	0.739	0.015
		400	1000	0.013	1.000	0.997	1.000	0.013	0.000	0.699	0.640	0.620	0.000	0.012	0.795	0.760	0.768	0.011
			1500	0.010	1.000	0.990	0.998	0.010	0.000	0.678	0.653	0.655	0.000	0.016	0.795	0.757	0.780	0.015
			3000	0.000	1.000	0.997	0.994	0.000	0.000	0.678	0.635	0.645	0.000	0.013	0.774	0.753	0.769	0.013
		600	1000	0.021	1.000	1.000	1.000	0.021	0.000	0.681	0.684	0.679	0.000	0.014	0.798	0.781	0.790	0.013
			1500	0.020	1.000	1.000	1.000	0.020	0.000	0.677	0.672	0.671	0.000	0.035	0.799	0.787	0.791	0.035
			3000	0.013	1.000	1.000	1.000	0.010	0.000	0.671	0.666	0.673	0.000	0.016	0.789	0.767	0.793	0.014
	50%	150	1000	0.002	0.991	0.995	0.995	0.000	0.000	0.630	0.609	0.608	0.000	0.047	0.724	0.748	0.749	0.045
			1500	0.011	0.990	0.993	0.993	0.010	0.000	0.623	0.610	0.603	0.000	0.009	0.779	0.736	0.732	0.005
			3000	0.005	0.993	0.993	0.993	0.003	0.000	0.617	0.603	0.611	0.000	0.010	0.730	0.724	0.728	0.010
		400	1000	0.013	1.000	0.989	0.991	0.012	0.000	0.650	0.620	0.619	0.000	0.020	0.786	0.740	0.742	0.015
			1500	0.010	1.000	0.992	0.994	0.010	0.000	0.648	0.616	0.613	0.000	0.019	0.789	0.749	0.728	0.016
			3000	0.010	1.000	0.995	0.990	0.008	0.000	0.638	0.613	0.603	0.000	0.014	0.776	0.759	0.730	0.013
		600	1000	0.017	1.000	1.000	1.000	0.017	0.000	0.679	0.626	0.635	0.000	0.036	0.795	0.793	0.780	0.033
			1500	0.018	1.000	1.000	0.997	0.015	0.000	0.653	0.627	0.632	0.000	0.012	0.794	0.730	0.770	0.011
			3000	0.014	1.000	0.997	1.000	0.010	0.000	0.654	0.616	0.621	0.000	0.024	0.765	0.765	0.783	0.019
50%	15%	150	1000	0.003	0.990	0.935	0.963	0.003	0.000	0.672	0.609	0.617	0.000	0.010	0.785	0.727	0.728	0.010
			1500	0.000	0.986	0.956	0.962	0.000	0.000	0.626	0.609	0.603	0.000	0.013	0.710	0.720	0.716	0.010
			3000	0.001	0.990	0.949	0.972	0.000	0.000	0.609	0.610	0.604	0.000	0.000	0.706	0.716	0.705	0.000
		400	1000	0.008	1.000	0.997	0.973	0.006	0.000	0.647	0.625	0.616	0.000	0.060	0.794	0.769	0.754	0.054
			1500	0.010	1.000	0.997	0.986	0.005	0.007	0.643	0.616	0.619	0.004	0.063	0.764	0.755	0.756	0.061
			3000	0.010	1.000	0.996	0.990	0.003	0.000	0.646	0.611	0.611	0.000	0.051	0.748	0.754	0.737	0.050
		600	1000	0.017	1.000	0.998	0.996	0.015	0.002	0.620	0.649	0.645	0.000	0.050	0.778	0.803	0.771	0.041
			1500	0.012	1.000	0.996	0.997	0.010	0.001	0.614	0.625	0.631	0.000	0.043	0.757	0.766	0.773	0.042
			3000	0.000	1.000	0.992	0.995	0.000	0.000	0.618	0.621	0.618	0.000	0.018	0.762	0.738	0.709	0.017
	50%	150	1000	0.000	0.978	0.986	0.992	0.000	0.001	0.617	0.606	0.608	0.000	0.017	0.673	0.665	0.669	0.012
			1500	0.003	0.977	0.984	0.989	0.003	0.000	0.620	0.611	0.615	0.000	0.011	0.698	0.675	0.678	0.008
			3000	0.002	0.972	0.984	0.984	0.001	0.000	0.570	0.605	0.606	0.000	0.012	0.696	0.680	0.672	0.010
		400	1000	0.004	0.973	0.985	0.994	0.003	0.002	0.598	0.610	0.608	0.000	0.018	0.688	0.697	0.695	0.017
			1500	0.003	0.974	0.986	0.993	0.003	0.000	0.607	0.615	0.605	0.000	0.016	0.681	0.676	0.679	0.016
			3000	0.001	0.971	0.984	0.993	0.001	0.000	0.597	0.600	0.606	0.000	0.018	0.669	0.664	0.675	0.016
		600	1000	0.001	1.000	0.998	0.996	0.000	0.001	0.599	0.582	0.579	0.001	0.018	0.735	0.708	0.717	0.017
			1500	0.003	1.000	0.997	0.995	0.000	0.000	0.610	0.574	0.609	0.000	0.016	0.712	0.699	0.720	0.016
			3000	0.015	0.985	0.996	0.993	0.010	0.000	0.599	0.618	0.618	0.000	0.017	0.709	0.696	0.715	0.017

Noting that our purpose is to identify important covariates, i.e., $X_{(1)} - X_{(4)}$ in the models M1 and M2, from ultrahigh-dimensional data. To evaluate the finite sample performance of the proposed method, we follow the presentation in the relevant literature (e.g., Li et al., 2012; Chen, 2021) and measure the frequency of retaining those important covariates. Specifically, we examine the proportion that *each active covariate is selected* and the proportion that *all active covariates are selected* out of 1000 simulations, which are denoted as $\mathcal{P}_s$ and $\mathcal{P}_a$, respectively. Higher proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ indicate higher possibility that truly informative covariates could be detected. All numerical results derived by existing and proposed methods under all settings are summarized in Tables 1–8.

In general, we observe from Tables 1–8 that the CI-LTRC method is able to correctly retain important covariates in most cases, and its performance is robust with stable numerical results regardless of the choice of regression models or distributions of left-truncation times. Moreover, various truncation rates and censoring rates seem not have significant impacts, which suggest that the adjustment based on (1) is valid and the resulting signal (6) is useful to detect the continuous covariates that are highly corrected to the failure time $\widetilde{T}$ with the proportions $\mathcal{P}_s$ and $\mathcal{P}_a$ approaching 1. Since the estimated active set $\widehat{\mathcal{I}}$ is determined by the first $\left[\frac{n}{\log(n)}\right]$ largest values of (6), smaller sample size n incurs smaller size of $\widehat{\mathcal{I}}$, then it is possible to miss important covariates if the corresponding signal (6) is not large enough. On the contrary, when the sample size n is increasing, the selection result becomes more accurate.

Compared with the proposed method, numerical results show severe impact of ignoring biased sampling or censoring effects. Specifically, with the ignorance of truncation and censoring, the CSS method has the lowest proportion of correctly detecting the important covariates, and the result becomes worse when the sample size is small or the truncation rate becomes large. The possible reason is that the CSS method is implemented when the data is complete, but Y_i is observed survival time with biased and incomplete structure. The DC-RC method is better than the CSS method since it implements the Buckley-James estimator to derive the pseudo

Table 7
Simulation results: feature screening for model M1 under $\widetilde{A}$ following the exponential distribution and Scenario II. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.015	0.900	0.950	0.943	0.012	0.002	0.650	0.711	0.770	0.000	0.250	0.758	0.786	0.796	0.242
			1500	0.010	0.941	0.990	0.996	0.010	0.000	0.693	0.745	0.739	0.000	0.230	0.756	0.774	0.785	0.230
			3000	0.013	0.928	0.992	0.994	0.010	0.001	0.630	0.760	0.730	0.000	0.220	0.739	0.793	0.776	0.220
		400	1000	0.030	1.000	0.998	0.997	0.030	0.010	0.712	0.725	0.721	0.000	0.280	0.795	0.890	0.900	0.278
			1500	0.021	1.000	0.997	1.000	0.020	0.000	0.680	0.728	0.718	0.000	0.273	0.760	0.877	0.874	0.272
			3000	0.014	1.000	0.994	0.996	0.014	0.000	0.630	0.720	0.714	0.000	0.250	0.780	0.874	0.873	0.250
		600	1000	0.029	1.000	1.000	1.000	0.029	0.005	0.724	0.737	0.741	0.000	0.320	0.899	0.896	0.891	0.320
			1500	0.026	1.000	0.990	1.000	0.026	0.003	0.760	0.752	0.751	0.001	0.260	0.859	0.895	0.897	0.260
			3000	0.024	1.000	0.987	1.000	0.024	0.000	0.707	0.767	0.727	0.000	0.207	0.788	0.840	0.883	0.207
	50%	150	1000	0.011	0.953	0.994	0.992	0.010	0.003	0.595	0.578	0.598	0.001	0.170	0.720	0.767	0.775	0.165
			1500	0.010	0.951	0.991	0.993	0.010	0.000	0.548	0.560	0.580	0.000	0.105	0.721	0.753	0.757	0.105
			3000	0.010	0.947	0.994	0.994	0.008	0.000	0.520	0.571	0.578	0.000	0.163	0.659	0.748	0.735	0.162
		400	1000	0.030	0.982	0.991	0.989	0.030	0.002	0.623	0.650	0.630	0.000	0.212	0.844	0.898	0.895	0.212
			1500	0.036	0.982	0.995	0.988	0.034	0.000	0.602	0.648	0.621	0.000	0.204	0.870	0.895	0.884	0.204
			3000	0.025	0.978	0.896	0.897	0.020	0.000	0.602	0.637	0.626	0.000	0.190	0.900	0.897	0.891	0.180
		600	1000	0.039	0.995	0.991	0.995	0.038	0.003	0.624	0.679	0.663	0.000	0.231	0.865	0.893	0.884	0.230
			1500	0.031	1.000	1.000	1.000	0.030	0.001	0.617	0.673	0.655	0.001	0.340	0.795	0.878	0.890	0.340
			3000	0.010	1.000	0.990	1.000	0.010	0.000	0.610	0.651	0.620	0.000	0.300	0.784	0.821	0.854	0.300
50%	15%	150	1000	0.017	0.868	0.995	0.995	0.016	0.000	0.545	0.612	0.606	0.000	0.053	0.670	0.704	0.681	0.051
			1500	0.012	0.854	0.996	0.994	0.012	0.000	0.542	0.540	0.561	0.000	0.049	0.620	0.642	0.643	0.043
			3000	0.016	0.837	0.993	0.993	0.015	0.000	0.534	0.506	0.556	0.000	0.042	0.616	0.643	0.644	0.042
		400	1000	0.017	0.950	0.989	0.988	0.014	0.000	0.610	0.660	0.646	0.000	0.180	0.783	0.783	0.788	0.180
			1500	0.015	0.959	0.988	0.988	0.015	0.000	0.607	0.636	0.643	0.000	0.183	0.740	0.750	0.792	0.178
			3000	0.010	0.944	0.978	0.984	0.010	0.000	0.586	0.619	0.622	0.000	0.170	0.719	0.762	0.769	0.167
		600	1000	0.022	0.978	0.996	0.996	0.021	0.010	0.760	0.766	0.724	0.010	0.260	0.796	0.869	0.885	0.258
			1500	0.027	0.974	0.993	0.997	0.027	0.008	0.721	0.731	0.740	0.007	0.310	0.789	0.863	0.872	0.310
			3000	0.026	0.974	0.991	0.996	0.025	0.005	0.753	0.741	0.743	0.005	0.298	0.780	0.794	0.879	0.295
	50%	150	1000	0.011	0.882	0.962	0.950	0.006	0.000	0.350	0.310	0.260	0.000	0.036	0.654	0.613	0.681	0.033
			1500	0.015	0.885	0.954	0.959	0.015	0.000	0.245	0.172	0.189	0.000	0.020	0.410	0.520	0.566	0.019
			3000	0.005	0.869	0.951	0.949	0.004	0.000	0.250	0.180	0.084	0.000	0.027	0.434	0.503	0.504	0.025
		400	1000	0.040	0.876	0.987	0.962	0.038	0.000	0.520	0.510	0.460	0.000	0.123	0.716	0.746	0.731	0.120
			1500	0.035	0.874	0.990	0.962	0.033	0.000	0.461	0.504	0.420	0.000	0.130	0.711	0.734	0.737	0.127
			3000	0.034	0.970	0.991	0.961	0.032	0.000	0.405	0.422	0.427	0.000	0.139	0.733	0.725	0.722	0.135
		600	1000	0.042	1.000	0.994	0.960	0.040	0.020	0.640	0.630	0.660	0.015	0.223	0.780	0.756	0.755	0.220
			1500	0.039	0.998	0.990	0.957	0.039	0.010	0.615	0.633	0.650	0.006	0.238	0.783	0.757	0.743	0.231
			3000	0.038	0.980	0.989	0.952	0.038	0.011	0.570	0.634	0.661	0.010	0.227	0.776	0.734	0.739	0.220

response with the adjustment of the censoring effect. However, since the implementation of the DC-RC method is based on the biased sample without suitable adjustment, the performance of the DC-RC method is slightly worse than the CI-LTRC method.

A common situation among three methods is the detection of binary covariates in Scenario II. When the covariate is binary, all methods have unsatisfactory performance in detecting $X_{(1)}$ due to the possible tie for the C-index approaches, such as CI-LTRC and CSS (e.g., Yan and Greene, 2008), or the distance correlation, but the CI-LTRC method is better than the CSS method and is comparable with the DC-RC method. In contrast, it is interesting to see that detection of the categorical variable $X_{(2)}$ has the similar performance with that for the detection of continuous variables. In summary, numerical results verify the validity of the CI-LTRC method under the continuous covariates and justify the theoretical property in Section 3.2. The CI-LTRC method also shows the importance of adjusting biased and incomplete effects when the dataset is subject to LTRC.

6. Analysis of breast cancer data

In this section, we implement the proposed method to analyze the NKI breast cancer data, which has been available in the Kaggle website[1]. The original dataset in the Kaggle website contains $p = 1554$ continuous gene expressions. The goal is to use gene expressions to characterize survival time `survival` and the status `eventdeath` of the breast cancer. In addition, due to ultrahigh-dimensional gene expressions, it is also crucial to detect informative gene expressions that can be used to characterize the survival

[1] The source of the public data: https://www.kaggle.com/datasets/nancyalaswad90/cancer-statistics-in-us-states

Table 8
Simulation results: feature screening for model M2 under $\widetilde{A}$ following the exponential distribution and Scenario II. $\mathcal{P}_s$ and $\mathcal{P}_a$ record the frequency of retaining informative covariates. CI-LTRC is the proposed method, CSS was proposed by Ma et al. (2017), and DC-RC was proposed by Chen (2021). 'Trun' represents the truncation rate; 'Cen' indicates the censoring rate; n and p are sample size and dimension of covariates, respectively.

Trun	Cen	n	p	CI-LTRC					CSS					DC-RC				
				$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
				$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$		$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	
15%	15%	150	1000	0.006	1.000	0.994	0.995	0.005	0.000	0.630	0.622	0.634	0.000	0.007	0.787	0.744	0.734	0.002
			1500	0.005	0.979	0.993	0.993	0.003	0.000	0.622	0.624	0.619	0.000	0.004	0.768	0.733	0.739	0.001
			3000	0.005	0.990	0.993	0.992	0.003	0.000	0.627	0.630	0.633	0.000	0.004	0.750	0.730	0.732	0.003
		400	1000	0.012	1.000	0.991	0.992	0.010	0.000	0.671	0.682	0.683	0.000	0.013	0.796	0.759	0.780	0.012
			1500	0.013	1.000	0.990	0.996	0.012	0.010	0.657	0.673	0.670	0.005	0.012	0.793	0.750	0.769	0.010
			3000	0.015	1.000	0.988	0.992	0.013	0.000	0.654	0.655	0.677	0.000	0.017	0.792	0.763	0.754	0.016
		600	1000	0.017	1.000	0.998	1.000	0.017	0.010	0.698	0.697	0.692	0.007	0.014	0.799	0.781	0.789	0.011
			1500	0.016	1.000	0.996	0.996	0.015	0.000	0.701	0.696	0.693	0.000	0.018	0.787	0.784	0.782	0.015
			3000	0.016	1.000	1.000	0.997	0.013	0.000	0.678	0.679	0.685	0.000	0.012	0.767	0.733	0.733	0.010
	50%	150	1000	0.008	0.983	0.992	0.993	0.008	0.000	0.631	0.622	0.610	0.000	0.018	0.748	0.758	0.761	0.017
			1500	0.005	0.981	0.995	0.993	0.005	0.000	0.622	0.623	0.613	0.000	0.019	0.750	0.760	0.754	0.014
			3000	0.006	0.978	0.995	0.992	0.005	0.000	0.618	0.618	0.628	0.000	0.017	0.744	0.769	0.778	0.011
		400	1000	0.014	0.984	0.996	0.996	0.013	0.000	0.620	0.679	0.673	0.000	0.016	0.781	0.797	0.786	0.016
			1500	0.010	0.986	0.996	0.997	0.010	0.000	0.608	0.675	0.622	0.000	0.016	0.789	0.796	0.787	0.016
			3000	0.010	0.984	0.994	0.995	0.010	0.000	0.615	0.640	0.625	0.000	0.017	0.783	0.785	0.790	0.017
		600	1000	0.012	1.000	0.998	0.998	0.012	0.004	0.687	0.691	0.697	0.002	0.021	0.795	0.798	0.799	0.018
			1500	0.018	1.000	0.998	0.997	0.018	0.000	0.689	0.688	0.686	0.000	0.018	0.790	0.794	0.788	0.018
			3000	0.011	1.000	0.997	0.998	0.010	0.000	0.661	0.672	0.666	0.000	0.016	0.778	0.782	0.795	0.016
50%	15%	150	1000	0.005	0.981	0.991	0.994	0.005	0.000	0.588	0.608	0.632	0.000	0.005	0.770	0.730	0.731	0.004
			1500	0.006	0.979	0.989	0.987	0.005	0.001	0.589	0.616	0.613	0.000	0.007	0.757	0.725	0.718	0.005
			3000	0.005	0.983	0.993	0.985	0.004	0.000	0.584	0.618	0.620	0.000	0.005	0.764	0.712	0.713	0.003
		400	1000	0.010	0.995	0.993	0.996	0.000	0.007	0.612	0.679	0.677	0.000	0.008	0.797	0.764	0.766	0.008
			1500	0.007	0.994	0.993	0.995	0.006	0.000	0.601	0.637	0.675	0.000	0.005	0.796	0.750	0.762	0.003
			3000	0.008	0.993	0.995	0.996	0.000	0.000	0.603	0.616	0.652	0.000	0.006	0.790	0.742	0.751	0.005
		600	1000	0.016	0.980	0.996	0.997	0.016	0.004	0.682	0.655	0.652	0.002	0.019	0.799	0.767	0.757	0.015
			1500	0.014	1.000	0.994	0.997	0.012	0.000	0.671	0.643	0.648	0.000	0.016	0.798	0.760	0.761	0.014
			3000	0.010	0.992	0.996	0.998	0.009	0.000	0.633	0.623	0.630	0.000	0.013	0.786	0.744	0.751	0.010
	50%	150	1000	0.006	0.933	0.952	0.956	0.006	0.000	0.581	0.607	0.624	0.000	0.008	0.693	0.720	0.719	0.003
			1500	0.007	0.928	0.945	0.953	0.005	0.000	0.583	0.608	0.601	0.000	0.007	0.696	0.716	0.705	0.002
			3000	0.006	0.923	0.943	0.953	0.006	0.000	0.577	0.604	0.615	0.000	0.007	0.785	0.704	0.705	0.004
		400	1000	0.008	0.964	0.991	0.992	0.008	0.000	0.599	0.639	0.623	0.000	0.014	0.730	0.764	0.759	0.014
			1500	0.007	0.958	0.994	0.995	0.006	0.000	0.605	0.624	0.621	0.000	0.013	0.699	0.729	0.746	0.012
			3000	0.008	0.959	0.995	0.996	0.006	0.000	0.598	0.617	0.619	0.000	0.014	0.698	0.736	0.739	0.013
		600	1000	0.012	0.996	0.999	0.998	0.010	0.005	0.635	0.639	0.647	0.004	0.016	0.740	0.738	0.775	0.016
			1500	0.010	0.995	0.997	0.997	0.010	0.004	0.645	0.645	0.638	0.002	0.015	0.727	0.735	0.769	0.015
			3000	0.009	0.994	0.996	0.999	0.009	0.000	0.641	0.642	0.640	0.000	0.014	0.721	0.736	0.749	0.010

time. From the Kaggle website, however, this dataset also contains the variable of the time of recruitment `timerecurrence`, which may incur truncation if values of `survival` are smaller than values of `timerecurrence`. As a result, we recruit the patients whose values in `survival` are greater than values in `timerecurrence`. It yields the left-truncated and right-censored data, denoted as $\mathcal{D}$, with sample size $n = 105$, suggesting the truncation rate 61.4%. In addition, the censoring rate of $\mathcal{D}$ is around 30.5%. We follow the notation in Section 2.1 to define `timerecurrence`, `survival`, `eventdeath`, and all gene expressions in $\mathcal{D}$ as A_i, Y_i, δ_i, and $\mathbf{X}_i$, respectively.

We now apply the proposed method to do feature screening and detect informative gene expressions. In addition, we follow simulation studies in Section 5 to examine several existing methods which ignore the feature of censoring and/or truncation. We retain 20 gene expressions in the estimated active set $\widehat{\mathcal{I}}$, and the detailed list is summarized in Table 9.

After that, based on retained gene expressions in Table 9, we further use them to fit the Cox model (22). The estimation for $\boldsymbol{\beta}$ can be carried out by the partial likelihood method (e.g., Lawless). More specifically, in the presence of left-truncation, to adjust the biased sampling effect, we adopt the pseudo likelihood method proposed by Chen and Yi (2021) to model gene expressions selected by the CI-LTRC method. In contrast, with the ignorance of left-truncation, we use the R package `coxph` to fit the Cox model for gene expressions selected by the CSS and DC-RC methods. To evaluate the standard error (S.E.), we perform the bootstrap procedure with repetitions $B = 1000$. Based on the estimators and the standard errors, we compute the p-values to examine the significance of selected gene expressions. The analysis results are summarized in Table 9.

From feature screening results, it is interesting to see that no gene expressions are commonly selected by three different methods. It might be due to the effects of incorporating or ignoring left-truncation and/or right-censoring. Regarding the model construction, when the likelihood function is accommodated with the adjustment of left-truncation, we can see that gene expressions determined

Table 9
Analysis of breast cancer data: results of feature screening and model construction. Columns 'genes' reflect gene expressions selected by different feature screening methods. Columns $\widehat{\beta}$ are estimates derived under the Cox model for selected gene expressions. Columns S.E. are standard errors, and columns p-value are p-values derived by $\widehat{\beta}$ and S.E.s.

#	CI-LTRC				CSS				DC-RC			
	genes	$\widehat{\beta}$	S.E.	p-value	genes	$\widehat{\beta}$	S.E.	p-value	genes	$\widehat{\beta}$	S.E.	p-value
1	Contig30480_RC	0.058	0.019	0.002	Contig35814_RC	−0.027	0.021	0.193	NM_019013	0.815	0.059	0.000
2	Contig46435_RC	0.151	0.022	0.000	NM_003500	−0.289	0.019	0.000	NM_016109	1.212	0.030	0.000
3	NM_002989	0.962	0.027	0.000	NM_001185	−0.476	0.030	0.000	AL117418	−1.281	0.038	0.000
4	NM_002001	−0.282	0.028	0.000	NM_012319	1.003	0.036	0.000	NM_001333	0.413	0.038	0.000
5	NM_002652	0.132	0.020	0.000	NM_005375	−1.433	0.044	0.000	NM_004701	−0.373	0.074	0.000
6	NC_001807	1.228	0.031	0.000	NM_014668	1.459	0.031	0.000	NM_004143	−1.494	0.026	0.000
7	Contig44010_RC	1.151	0.022	0.000	Contig46937_RC	−0.259	0.018	0.000	Contig41530_RC	−0.288	0.047	0.000
8	NM_007191	−0.117	0.028	0.000	Contig58301_RC	−0.241	0.034	0.000	AB037836	−0.404	0.047	0.000
9	NM_001150	0.589	0.028	0.000	AB020689	−1.122	0.036	0.000	U79293	−1.154	0.047	0.000
10	AK000451	0.501	0.022	0.000	NM_004143	−0.892	0.027	0.000	Contig37562_RC	−1.640	0.036	0.000
11	NM_001074	0.379	0.046	0.000	NM_004496	0.284	0.029	0.000	NM_003981	2.213	0.074	0.000
12	NM_000854	0.107	0.020	0.000	Contig56390_RC	−0.637	0.027	0.000	Contig37571_RC	3.364	0.062	0.000
13	NM_000477	−0.954	0.027	0.000	Contig14284_RC	0.059	0.020	0.003	NM_005733	−0.309	0.062	0.000
14	NM_002343	−0.353	0.016	0.000	NM_003226	0.346	0.034	0.000	NM_004456	0.615	0.064	0.000
15	AK000345	1.473	0.087	0.000	NM_002614	−0.075	0.020	0.000	Contig56390_RC	−0.323	0.031	0.000
16	NM_000909	−1.113	0.015	0.000	Contig46934_RC	−0.310	0.027	0.000	NM_003226	−0.036	0.028	0.197
17	NM_000353	−1.206	0.023	0.000	NM_020974	−0.221	0.021	0.000	NM_001168	0.692	0.054	0.000
18	NM_006551	0.236	0.045	0.000	NM_000909	−0.071	0.017	0.000	Contig55725_RC	0.299	0.030	0.000
19	NM_005794	−0.919	0.082	0.000	esr1	0.310	0.031	0.000	NM_003225	−0.149	0.026	0.000
20	NM_002411	−0.540	0.044	0.000	NM_000125	0.002	0.036	0.945	Contig58301_RC	0.107	0.040	0.008

by the CI-LTRC method are significant with p-values smaller than 0.05. On the contrary, if the R package `coxph` is implemented with the ignorance of left-truncation effect, we can observe that some of gene expressions selected by the CSS or DC-RC methods are insignificant with p-values greater than 0.05. It might reflect the impact of biased samples.

7. Discussion

Analysis of LTRC data is an attractive topic in survival analysis and it has been discussed under various models or complex structures in recent years. In the era of big data, datasets are collected easily, and undoubtedly, ultrahigh-dimensional data become ubiquitous. As a result, existing methods that focus on low-dimensional data are no longer valid. Motivated by this, we aim to deal with LTRC survival data subject to ultrahigh-dimensional covariates. Our key idea is to adopt the C-index estimator and transform it as a signal to do feature screening. To show the validity of the feature screening procedure, we rigorously establish the sure screening property. Numerical studies also verify that the proposed method successfully retains informative covariates.

In addition to LTRC data, sometimes more complex settings are accommodated, such as cure model (e.g., Chen, 2019a) or measurement error in covariates (e.g., Chen, 2020; Chen and Yi, 2021, 2022). It is interesting to extend the current development to address those complex settings. On the other hand, besides the C-index approach, it is expected to propose alternative strategies, such as model-free development (e.g., Chen, 2021), to address feature screening for LTRC data. For example, as commented by a referee, the C-index approach might encounter an issue of ties when covariates are binary, which is also reflected in our numerical experiments. To address this concern, a possible solution is the rank based estimation (e.g., Chen, 2023), but it is required to deal with the LTRC structure carefully.

Acknowledgments

The author would like to thank the editorial team for useful comments to improve the initial manuscript. Chen's research was supported by National Science and Technology Council.

Appendix. Technical Lemmas

Before stating the first lemma, we introduce several notation for empirical process.

Let $\mathbb{P}$ and $\mathcal{P}$ denote empirical and probability measures and let Ψ denote a class of real-valued functions, $\psi : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$. Applying any function $\psi \in \Psi$ to a sequence of independent random variables $\left\{\{Z_i, Z_j\} : i, j = 1, \ldots, n, i \neq j\right\}$, we write

$$\mathcal{P}\psi \triangleq E\left\{\psi(Z_i, Z_j)\right\},$$

where the expectation is taken with respect to the probability measure $\mathcal{P}$ for $\{Z_i, Z_j\}$, and the corresponding estimate is defined as

$$\mathbb{P}\psi \triangleq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} \psi(Z_i, Z_j).$$

For $q \geq 1$, let $\|\psi\|_{q,\mathcal{P}}$ denote the L_q-norm of $\psi \in \Psi$ under $\mathcal{P}$. Let $\phi_0 \in \Psi$ represent a given function. We review the following definitions related to empirical process theory (e.g., van der Vaart and Wellner, 1996, p.83; van der Vaart, 1998, p. 270).

Definition A.1. Given $\epsilon > 0$ and $\phi_0 \in \Psi$, the "covering number", denoted by $N(\epsilon, \Psi, \|\cdot\|_{q,\mathcal{P}})$, is defined as the minimal number of balls $\left\{\psi : \left\|\psi - \phi_0\right\|_{q,\mathcal{P}} < \epsilon\right\}$ of radius ϵ that are needed to cover the set Ψ, and the "entropy" is defined as the logarithm of the covering number.

Definition A.2. For any given functions $\psi^l, \psi^u \in \Psi$ and $\epsilon > 0$, an "*ϵ-bracket*" $[\psi^l, \psi^u]$ is defined to be the set of all functions $\psi \in \Psi$ satisfying $\psi^l \leq \psi \leq \psi^u$ and $\int \left|\psi^u - \psi^l\right|^q d\mathcal{P} < \epsilon^q$ for some $q \geq 1$.

Definition A.3. For $\epsilon > 0$, the smallest number of ϵ-bracket needed to cover Ψ is called "bracketing number", denoted $N_{[\ ]}(\epsilon, \Psi, \|\cdot\|_{q,\mathcal{P}})$. The logarithm of the bracketing number is defined as the "entropy with bracketing".

Based on definitions above, we now state the following lemma that is related to the bracketing number.

Lemma A.1. *Assume that*

$$\sup_{\psi \in \Psi} \|\psi\|_\infty \leq 1 \quad \text{and} \quad N_{[\ ]}\left(\epsilon, \Psi, \|\cdot\|_{q,\mathcal{P}}\right) \leq A\epsilon^{-\alpha} \tag{A.1}$$

hold for every $\epsilon > 0$ and some $\alpha > 0$ and some constant A. Then for some constants c and n_0 which may depend on α and A, we have that for all $Q_1 \geq c$ and $n \geq n_0$,

$$P\left(\sup_{\psi \in \Psi, \|\psi\|_{2,\mathcal{P}} \leq n^{-1/(2+\alpha)}} |\mathbb{P}\psi - \mathcal{P}\psi| \geq Q_1 n^{-2/(2+\alpha)}\right) \leq c \exp\left\{-\frac{Q_1 n^{\alpha/(2+\alpha)}}{c^2}\right\} \tag{A.2}$$

and

$$P\left(\sup_{\psi \in \Psi, \|\psi\|_{2,\mathcal{P}} > n^{-1/(2+\alpha)}} \frac{|\mathbb{P}\psi - \mathcal{P}\psi|}{\|\psi\|_{2,\mathcal{P}}^{1-\alpha/2}} \geq Q_1 n^{-1/2}\right) \leq c \exp\left(-\frac{Q_1}{c^2}\right). \tag{A.3}$$

It is a direct application of Lemma 5.13 in van de Geer (2000) whose proof can be found in van de Geer (2000, p.79).

Next, we state two useful lemmas that show upper bounds of the expectation and the probabilistic inequality.

Lemma A.2. *Let Z_1 and Z_2 denote independent random variables satisfying $P(|Z_i| > t) < \exp\left\{1 - \left(\frac{t}{K}\right)^r\right\}$ with $i = 1, 2$ for all $t \geq 0$, where $K > 0$ and $r \geq 1$ are two constants. Then for all $m \geq 2$,*

$$E\left(|Z_1 + Z_2|^m\right) \leq 2e(2K)^m m!.$$

Lemma A.3. *Let $h(\cdot, \cdot)$ denote a kernel function of the U-statistics*

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} h(Z_i, Z_j).$$

If $E\left\{h(Z_1, Z_2)\right\} = \mu$ and $E\left\{\left|h(Z_1, Z_2) - \mu\right|^m\right\} \leq m! R^{m-2}\zeta/2$ for some constant $R > 0$, $\zeta > 0$, and $m \geq 2$. Then for any $\delta > 0$,

$$P\left(\left|U_n - \mu\right| > \delta\right) \leq 2\exp\left\{-\frac{d_n \delta^2}{2(\zeta + R\delta)}\right\},$$

where $d_n = \left[\frac{n}{2}\right]$ is the greatest integer less than $n/2$.

Lemmas A.2 and A.3 are from Lemmas 2 and 5 in the supplementary material of Ma et al. (2017), respectively.

References

Buckley, J., James, I., 1979. Linear regression with censored data. Biometrika 66, 429–436.

Chen, L.-P., 2019a. Semiparametric estimation for cure survival model with left-truncated and right-censored data and covariate measurement error. Statist. Probab. Lett. 154, 108547.

Chen, L.-P., 2019b. Pseudo likelihood estimation for the additive hazards model with data subject to left-truncation and right-censoring. Stat. Interface 12, 135–148.

Chen, L.-P., 2020. Variable selection and estimation for the additive hazards model subject to left-truncation, right-censoring and measurement error in covariates. J. Stat. Comput. Simul. 90, 3261–3300.

Chen, L.-P., 2021. Feature screening based on distance correlation for ultrahigh-dimensional censored data with covariates measurement error. Comput. Statist. 36, 857–884.

Chen, L.-P., 2023. A note of feature screening via rank-based coefficient of correlation. Biom. J. 65, 2100373.

Chen, L.-P., Qiu, B., 2023. Analysis of length-biased and partly interval-censored survival data with mismeasured covariates. Biometrics 79, 3929–3940.

Chen, L.-P., Yi, G.Y., 2020. Model selection and model averaging for analysis of truncated and censored data with measurement error. Electron. J. Stat. 14, 4054–4109.

Chen, L.-P., Yi, G.Y., 2021. Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error. Ann. Inst. Statist. Math. 73, 481–517.

Chen, L.-P., Yi, G.Y., 2022. Robust feature screening for ultrahigh-dimensional censored data subject to measurement error. In: He, W., Wang, L., Chen, J., Lin, C.D. (Eds.), Advances and Innovations in Statistics and Data Science (ICSA Book Series in Statistics). Springer, Cham, pp. 23–53.

Fan, J., Lv, J., 2008. Sure independence screening for ultra high dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 849–911.

Hartman, N., Kim, S., He, K., Kalbfleisch, J.D., 2023. Concordance indices with left-truncated and right-censored data. Biometrics 79, 1624–1634.

Hu, Q., Zhu, L., Liu, Y., Sun, J., Srivastava, D.K., 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, C.-Y., Qin, J., 2013. Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. Biometrika 100, 877–888.

Lawless, J.F., Statistical Models and Methods for Lifetime Data. Wiley, New York.

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

Ma, Y., Li, Y., Lin, H., Li, Y., 2017. Concordance measure-based feature screening and variable selection. Statist. Sinica 27, 1967–1985.

McGough, S.F., Incerti, D., Lyalina, S., Copping, R., Narasimhan, B., Tibshirani, R., 2021. Penalized regression for left-truncated and right-censored survival data. Stat. Med. 40, 5487–5500.

van de Geer, S.A., 2000. Empirical Processes in M-Estimation. Cambridge University Press, New York.

van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, New York.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer, New York.

Wang, J.G., 1987. A note on the uniform consistency of the Kaplan–Meier estimator. Ann. Statist. 15, 1313–1316.

Wang, M.C., 1991. Nonparametric estimation from cross-sectional survival data. J. Amer. Statist. Assoc. 86, 130–143.

Yan, G., Greene, T., 2008. Investigating the effects of ties on measures of concordance. Stat. Med. 27, 4190–4206.

Zhao, S.D., Li, Y., 2014. Score test variable screening. Biometrics 70, 862–871.