Subsampling Approach for Least Squares Fitting of Semi-parametric Accelerated Failure Time Models to Massive Survival Data

Zehan Yang[1*], HaiYing Wang[1*] and Jun Yan[1]

[1*]Department of Statistics, University of Connecticut, Storrs, Connecticut, 06269–4120, USA.

*Corresponding author(s). E-mail(s): zehan.yang@uconn.edu; haiying.wang@uconn.edu;
Contributing authors: jun.yan@uconn.edu;

Abstract

Massive survival data are increasingly common in many research fields, and subsampling is a practical strategy for analyzing such data. Although optimal subsampling strategies have been developed for Cox models, little has been done for semiparametric accelerated failure time (AFT) models due to the challenges posed by non-smooth estimating functions for the regression coefficients. We develop optimal subsampling algorithms for fitting semi-parametric AFT models using the least-squares approach. By efficiently estimating the slope matrix of the non-smooth estimating functions using a resampling approach, we construct optimal subsampling probabilities for the observations. For feasible point and interval estimation of the unknown coefficients, we propose a two-step method, drawing multiple subsamples in the second stage to correct for overestimation of the variance in higher censoring scenarios. We validate the performance of our estimators through a simulation study that compares single and multiple subsampling methods and apply the methods to analyze the survival time of lymphoma patients in the Surveillance, Epidemiology, and End Results program.

Keywords: A-optimality; Non-smooth estimating function; Survival analysis

1 Introduction

The proliferation of storage and surveillance technologies has led to the emergence of large-scale datasets with survival outcomes in a variety of domains such as healthcare. The size of these datasets, however, often exceeds the computational capacity of an analyst's computer, posing significant challenges for their analysis. To tackle this issue, several strategies have been proposed. The divide-and-conquer strategy divides massive data into groups, processes them separately, and aggregates the results. This strategy has been applied to Cox models (Wang et al., 2021, 2022) and accelerated failure time (AFT) models (Su et al., 2023). Carefully devised, the strategy facilitates the full LASSO path through a batch screening approach in the case of the ultrahigh-dimensional Cox model with sparse solutions at all predefined regularization parameters in Li et al. (2022). Another strategy is the online updating strategy, which handles massive survival data arriving in a stream by batches and updates the cumulative estimators. Examples of this approach include testing for the proportional hazards assumption (Xue et al., 2020) and fitting Cox models (Wu et al., 2021).

Our focus here is the subsampling strategy, which selects a significantly smaller yet optimal subsample for analysis instead of using the full data. This concept was developed for linear regression in the form of leverage sampling by Drineas et al. (2006) and Mahoney et al. (2011). Ma et al. (2015) examined the statistical aspects of this method, referring to it as algorithmic leveraging. In this method, non-uniform subsampling probabilities are based on the empirical statistical leverage scores derived from the input covariate matrix. The asymptotic properties of the leverage sampling estimator were further explored by Ma et al. (2022). Wang et al. (2018) introduced an optimal subsampling algorithm for logistic regression based on the A-optimality criterion, which minimizes the trace of the asymptotic variance matrix of the resulting estimator. This approach has been extended to a variety of statistical models such as generalized linear models (Ai et al., 2021) and quantile regression models (Wang and Ma, 2021). In the field of survival analysis, this approach has been developed for Cox models (Zhang et al., 2023), Cox models with rare events (Keret and Gorfine, 2022), additive hazard rate models (Zuo et al., 2021), and parametric AFT models (Yang et al., 2022). To the best of our knowledge, however, no prior work has explored its application to semi-parametric AFT models for massive survival data.

Developing optimal subsampling strategies for semi-parametric AFT models can be a challenging task. Two commonly used approaches for fitting semi-parametric AFT models are the rank-based approach (Tsiatis, 1990; Jin et al., 2003; Chiou et al., 2014, 2015) and the least squares approach (Buckley and James, 1979; Jin et al., 2006; Chiou et al., 2014). In the presence of censoring, the key challenge is to derive the optimal subsampling probabilities (SSP) for censored observations. The SSP of an observation is proportional to its contribution to the estimating functions in standard approaches (Zhang et al., 2023; Yang et al., 2022). For the rank-based method, it is tempting to assign a zero SSP to censored observations since they do not contribute as individual terms to the estimating functions for the regression coefficients. This is also true for the estimating equation approaches for the additive hazard models and Cox models. A general approach is to express the estimating equations in terms of appropriately defined martingales, as was done by Zhang et al. (2023) to the partial likelihood score function for the Cox proportional hazards model. For the least squares method, however, the contribution of a censored observation to the estimating equations has an explicit form (Tsiatis, 1990). Conceptually, the optimal SSPs are expected to behave similarly to those in a parametric AFT model (Yang et al., 2022). The extra challenge comes from the evaluation of these contributions.

Here we address the challenge of developing optimal subsampling strategies for semi-parametric AFT models using the least-squares approach. Specifically, we focus on two types of optimal SSPs as discussed in Wang et al. (2022). The first type depends on the estimating functions and their slope matrices. For a censored observation, we define its contribution to the estimating function with the conditional expectation of the event time in place of the censored time (Buckley and James, 1979). Since the resulting estimating function depends on the Kaplan-Meier estimator of the residuals, which is non-smooth, we use a resampling procedure proposed by Zeng and Lin (2008) to evaluate the slope matrix. The second type of optimal SSPs only depends on the estimating function, which is computationally simpler and faster to calculate. For both types, since the true optimal SSPs are based on the unknown full data estimator, we propose a two-step method for practical implementation. In the first step, we approximate the optimal SSPs using a pilot estimator obtained from a small pilot subsample. In the second step, we use multiple subsamples selected by the approximated optimal SSPs to obtain the point estimator and its standard error. We demonstrate the effectiveness of this method through extensive simulation studies and a real data example, confirming the utility of our proposed optimal subsampling strategies for semi-parametric AFT models. Our implementation is part of an R package `aftosmac`, which is publicly available at https://github.com/YEnthalpy/aftosmac.

The remainder of the paper is structured as follows. In Section 2, we present a general subsampling procedure for semiparametric AFT models with least-squares using given SSPs. Section 3 focuses on deriving the optimal SSPs based on two criteria motivated by experiment design. Since the optimal SSPs depend on the unknown full-data estimator, in Section 4, we propose a feasible two-step approach and derive an estimator of the asymptotic variance. In Section 5, we evaluate the performance of the estimator through a simulation study. Section 6 illustrates the application of the proposed method to analyze the survival time of lymphoma patients in the Surveillance, Epidemiology, and End Results (SEER) program. Finally, we conclude with a discussion in Section 7.

2 Preliminaries

Consider a semi-parametric AFT model for a log-transformed failure time T with a p-dimensional covariate vector $\mathbf{X}$:

$$T = \alpha + \mathbf{X}^\top \boldsymbol{\beta} + \epsilon, \tag{1}$$

where α is an intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and ϵ is a random error with mean zero and an unspecified distribution. Due to right censoring, the observed time is $Y = \min(T, C)$, where C is a log-transformed censoring time, and C and T are conditionally independent given $\mathbf{X}$. Also observed is the event indicator $\delta = I(T < C)$ with $I(\cdot)$ being the indicator function. Suppose that a random sample of size n is available: $\{\mathbf{X}_i, Y_i, \delta_i\}_{i=1}^n$, which are independent and identically distributed copies of $\{\mathbf{X}, Y, \delta\}$.

The least squares estimation of $\boldsymbol{\beta}$ has the same principle as the classical least squares for non-censored data. In the case where $\{T_i\}_{i=1}^n$ are all observed (i.e., no censoring), the classical least-squares estimator of $\boldsymbol{\beta}$ can be obtained by solving the equation

$$\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(T_i - \mathbf{X}_i^\top \boldsymbol{\beta}) = 0,$$

where $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i / n$. In the presence of censoring, however, the true failure time T_i is unknown for those individuals with $\delta_i = 0$, in which case, the equation cannot be evaluated. Buckley and James (1979) proposed replacing each T_i with its conditional expectation given the observed data $(\mathbf{X}_i, Y_i, \delta_i)$,

$$\hat{T}_i(\boldsymbol{\beta}) = \delta_i Y_i + (1 - \delta_i)\left[\kappa_i(\boldsymbol{\beta}) + \mathbf{X}_i^\top \boldsymbol{\beta} + \alpha\right],$$

where

$$\kappa_i(\boldsymbol{\beta}) = \frac{\int_{e_i(\boldsymbol{\beta})}^\infty u \mathrm{d}\hat{F}_{\boldsymbol{\beta}}(u)}{1 - \hat{F}_{\boldsymbol{\beta}}\{e_i(\boldsymbol{\beta})\}},$$

and $\hat{F}_{\boldsymbol{\beta}}(\cdot)$ is the estimated cumulative distribution function for $e_i(\boldsymbol{\beta}) = Y_i - \mathbf{X}_i^\top \boldsymbol{\beta} - \alpha$, via the Kaplan-Meier estimator. The Buckley–James least squares estimator $\hat{\boldsymbol{\beta}}_n$ is the root of

$$\mathbf{U}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{U}_{n,i}(\boldsymbol{\beta}) = 0, \tag{2}$$

where

$$\mathbf{U}_{n,i}(\boldsymbol{\beta}) = (\mathbf{X}_i - \bar{\mathbf{X}})\left\{\hat{T}_i(\boldsymbol{\beta}) - \mathbf{X}_i^\top \boldsymbol{\beta}\right\}.$$

Finding the solution to Equation (2) is time-consuming. Jin et al. (2006) proposed an iterative procedure $\hat{\boldsymbol{\beta}}_n^{(m)} = L_n(\hat{\boldsymbol{\beta}}_n^{(m-1)})$ with an initial estimator

$\hat{\boldsymbol{\beta}}_n^{(0)}$ to calculate $\hat{\boldsymbol{\beta}}_n$, where

$$L_n(\boldsymbol{\beta}) = \left[\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top\right]^{-1}\left[\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})\left(\hat{T}_i(\boldsymbol{\beta}) - \bar{T}(\boldsymbol{\beta})\right)\right],$$

and $\bar{T}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\hat{T}_i(\boldsymbol{\beta})$. In practice, the zero vector is an appropriate initial value. In each iteration, multiple steps are needed to calculate $L_n(\boldsymbol{\beta})$. The expression of $L_n(\boldsymbol{\beta})$ with a given $\hat{T}_i(\boldsymbol{\beta})$ is similar to that of the traditional least-squares estimator with time complexity $O(np^2)$. Evaluating $\hat{T}_i(\boldsymbol{\beta})$ involves multiple steps. Sorting $\{e_i(\boldsymbol{\beta})\}_{i=1}^n$ is of complexity $O\{n\log(n)\}$. Getting the Kaplan-Meier-type estimator $\hat{F}_{\boldsymbol{\beta}}(\cdot)$ using the sorted $e_i(\boldsymbol{\beta})$'s takes $O(n)$ time. Calculating the numerators of $\{\kappa_i(\boldsymbol{\beta})\}_{i=1}^n$, which are cumulative summations with sorted $\{e_i(\boldsymbol{\beta})\}_{i=1}^n$, costs $O(n)$ time. Finally, computing $\{\hat{T}_i(\boldsymbol{\beta})\}_{i=1}^n$ with known $\{\kappa_i(\boldsymbol{\beta})\}_{i=1}^n$ takes $O(np)$ time. The overall time complexity of one iteration is $O\{np^2 + n\log(n) + np + n\} = O\{np^2 + n\log(n)\}$.

This procedure is computing intensive because it requires sorting $\{e_i(\boldsymbol{\beta})\}_{i=1}^n$ in each iteration, which becomes infeasible when dealing with large datasets that exceed the computer's memory. The overall time complexity of the iterative process is $O\{\xi_n[np^2 + n\log(n)]\}$, where ξ_n represents the average number of iterations required to obtain $\hat{\boldsymbol{\beta}}_n$. The value of ξ_n is primarily dependent on the censoring rate and not on n. With the simulated datasets in Section 5, ξ_n was approximately 20 for censoring rate 0.25, 45 for censoring rate 0.5, and 100 for censoring rate 0.75. For this situation, the divide-and-conquer strategy and the online updating strategy cannot be easily adopted because calculating $\hat{T}_i(\boldsymbol{\beta})$ for a censored observation relies on the residuals of the full dataset.

Now we consider the subsampling strategy. Draw a subsample of size r with replacement according to pre-assigned SSPs $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$. Denote the subsample by $\{\mathbf{X}_i^*, Y_i^*, \delta_i^*, \pi_i^*\}_{i=1}^r$, where $\mathbf{X}_i^*$ is the covariates, Y_i^* is the observed log-transformed time, δ_i^* is the censoring indicator, and π_i^* is the SSP of the ith observation in the subsample. We approximate the full data estimator $\hat{F}_{\boldsymbol{\beta}}$ by the subsample estimator

$$\hat{F}^*_{\boldsymbol{\beta}}(t) = 1 - \prod_{i:T_i^*\le T}\left(1 - \frac{\sum_{j=1}^r (\pi_j^*)^{-1}\delta_j{}^* I\left\{e_j^*(\boldsymbol{\beta}) = e_i^*(\boldsymbol{\beta})\right\}}{\sum_{j=1}^r (\pi_j^*)^{-1} I\left\{e_j^*(\boldsymbol{\beta}) \ge e_i^*(\boldsymbol{\beta})\right\}}\right),$$

where $e_i^*(\boldsymbol{\beta}) = Y_i^* - (\mathbf{X}_i^*)^\top\boldsymbol{\beta} - \alpha$.

Based on the subsample, we estimate $\boldsymbol{\beta}$ with a weighted estimating function

$$\mathbf{U}_r^*(\boldsymbol{\beta}) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\mathbf{U}_{r,i}^*(\boldsymbol{\beta}), \tag{3}$$

where

$$\mathbf{U}_{r,i}^*(\boldsymbol{\beta}) = \frac{1}{n}(\mathbf{X}_i^* - \tilde{\mathbf{X}}^*)\left\{\hat{T}_i^*(\boldsymbol{\beta}) - \mathbf{X}_i^{*\top}\boldsymbol{\beta}\right\}.$$

In the above formula, $\tilde{\mathbf{X}}^* = (nr)^{-1}\sum_{i=1}^r \mathbf{X}_i^*/\pi_i^*$ and

$$\hat{T}_i^*(\boldsymbol{\beta}) = \delta_i^* T_i^* + (1-\delta_i^*)\left[\kappa_i^*(\boldsymbol{\beta}) + \mathbf{X}_i^*\boldsymbol{\beta} + \alpha\right],$$

where

$$\kappa_i^*(\boldsymbol{\beta}) = \frac{\int_{e_i^*(\boldsymbol{\beta})}^\infty u \, \mathrm{d}\hat{F}_{\boldsymbol{\beta}}^*(u)}{1 - \hat{F}_{\boldsymbol{\beta}}^*\{e_i^*(\boldsymbol{\beta})\}}.$$

The solution to Equation (3) can be derived from the iterative procedure $\tilde{\boldsymbol{\beta}}_r^{(m)} = L_r^*\left[\tilde{\boldsymbol{\beta}}_r^{(m-1)}\right]$, with an initial value $\tilde{\boldsymbol{\beta}}_n^{(0)}$, where

$$L_r^*(\boldsymbol{\beta}) = \left[\sum_{i=1}^r \frac{1}{\pi_i^*}(\mathbf{X}_i^* - \tilde{\mathbf{X}}^*)(\mathbf{X}_i^* - \tilde{\mathbf{X}}^*)^\top\right]^{-1} \sum_{i=1}^r \frac{1}{\pi_i^*}(\mathbf{X}_i^* - \tilde{\mathbf{X}}^*)\left[\hat{T}_i^*(\boldsymbol{\beta}) - \tilde{T}^*(\boldsymbol{\beta})\right], \tag{4}$$

and $\tilde{T}^*(\boldsymbol{\beta}) = (nr)^{-1}\sum_{i=1}^r \hat{T}_i^*(\boldsymbol{\beta})/\pi_i^*$. We suggest using a zero vector as the initial value in practice. By similar arguments to the full data, the time complexity of the subsample estimator is $O\{\xi_r[rp^2 + r\log(r)]\}$, where ξ_r is the number of iterations to get a converging result based on the subsample. Again, it is worth noting that ξ_r depends more on the censoring rate than on r and π_i.

A subsample of size $r \ll n$ allows for obtaining the estimator $\tilde{\boldsymbol{\beta}}_r$ in a computationally feasible manner. However, the statistical efficiency of the estimator heavily relies on the selection of the SSPs.

3 Optimal Subsampling Probabilities

We determine the SSPs using procedures introduced by Wang et al. (2022) which depend on the norms of the summands in an estimating equation. Specifically for our estimating equation (2), the SSPs under the A-optimality criterion are $\boldsymbol{\pi}^{\mathrm{optA}} = \left\{\pi_i^{\mathrm{optA}}\right\}_{i=1}^n$ with

$$\pi_i^{\mathrm{optA}} = \frac{\left\|\mathbf{M}_n^{-1}\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|}{\sum_{i=1}^n \left\|\mathbf{M}_n^{-1}\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|}, \qquad i = 1, 2, \ldots, n \tag{5}$$

where $\mathbf{M}_n$ is the slope of $\mathbf{U}_n(\hat{\boldsymbol{\beta}}_n)$ and

$$\left\|\mathbf{M}_n^{-1}\mathbf{U}_{n,i}(\boldsymbol{\beta})\right\| = \left\|\mathbf{M}_n^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})\right\|\{(1-\delta_i)|\kappa_i(\boldsymbol{\beta})| + \delta_i|e_i(\boldsymbol{\beta})|\}. \tag{6}$$

Since the estimating function is non-smooth, we estimate $\mathbf{M}_n$ by an efficient resampling method proposed in Zeng and Lin (2008). In the resampling method, $\{Z_i\}_{i=1}^R$ are generated in the first step where Z_i's are zero-mean random vectors of dimension p and are independent of the data. In the second step, $n^{-1/2}\mathbf{U}_n(\hat{\boldsymbol{\beta}}_n + n^{-1/2}Z_i)$'s are calculated for $i = 1, \ldots, R$. In the third step, we calculate the least squares estimate of $n^{-1/2}\mathbf{U}_{jn}(\hat{\boldsymbol{\beta}}_n + n^{-1/2}Z_i)$'s on Z_i's for $j = 1, \ldots, p$, where $\mathbf{U}_{jn}$ denotes the jth component of $\mathbf{U}_n$. The jth row of $\mathbf{M}_n$ is estimated by the jth least squares estimates.

In practice, we use a small pilot subsample of size r_0 where $r_0 \ll n$ to estimate $\hat{\boldsymbol{\beta}}_n$ and $\{\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\}_{i=1}^n$ in order to approximate the optimal SSPs. Let $\tilde{\boldsymbol{\beta}}_{r_0}$ be the pilot estimator derived from the pilot subsample. We calculate $\{e_i^*(\tilde{\boldsymbol{\beta}}_{r_0})\}_{i=1}^{r_0}$ which are prediction errors of the selected pilot sample. Centering $\mathbf{X}$ is required in estimating $\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)$ which takes $O(np)$ time. Estimating $\kappa_i(\hat{\boldsymbol{\beta}}_n)$ dominates the computing time of estimating $\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)$ and it takes multiple steps. We sort $e_i^*(\tilde{\boldsymbol{\beta}}_{r_0})$'s in the first step which takes $O\{r_0 \log(r_0)\}$ time. In the second step, the denominator of $\kappa_i^*(\tilde{\boldsymbol{\beta}}_{r_0})$'s are calculated by the Kaplan-Meier type cumulative distribution function using sorted $e_i^*(\tilde{\boldsymbol{\beta}}_{r_0})$'s which costs $O(r_0)$ time. In the third step, the numerators of $\kappa_i^*(\tilde{\boldsymbol{\beta}}_{r_0})$'s that are cumulative summations are calculated with a cost of $O(r_0)$ time. In the fourth step, we calculate $\{e_i(\tilde{\boldsymbol{\beta}}_{r_0})\}_{i=1}^n$ which takes $O(np)$ time. In the last step, we estimate $\kappa_i(\hat{\boldsymbol{\beta}}_n)$ using constant interpolation. Specifically, we employ binary search to locate the position of $e_i(\tilde{\boldsymbol{\beta}}_{r_0})$ in the sorted $e_i^*(\tilde{\boldsymbol{\beta}}_{r_0})$'s, which takes $O\{\log(r_0)\}$ time. We assume that $e_{(k-1)}^*(\tilde{\boldsymbol{\beta}}_{r_0}) \le e_i(\tilde{\boldsymbol{\beta}}_{r_0}) \le e_{(k)}^*(\tilde{\boldsymbol{\beta}}_{r_0})$, where $e_{(k)}^*(\tilde{\boldsymbol{\beta}}_{r_0})$ is the kth element in the sorted $e_i^*(\tilde{\boldsymbol{\beta}}_{r_0})$'s. We estimate $\kappa_i(\hat{\boldsymbol{\beta}}_n)$ using $\kappa_{(k)}^*(\tilde{\boldsymbol{\beta}}_{r_0})$, which corresponds to $e_{(k)}^*(\tilde{\boldsymbol{\beta}}_{r_0})$. Since we have n observations in the full sample, the time complexity to estimate $\{\kappa_i(\hat{\boldsymbol{\beta}}_n)\}_{i=1}^n$ is $O\{n\log(r_0)\}$. In conclusion, the overall time complexity to estimate $\{\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\}_{i=1}^n$ is $O\{r_0 \log(r_0) + r_0 + n\log(r_0)\} = O\{n\log(r_0)\}$.

The slope matrix $\mathbf{M}_n$ is estimated using the pilot subsample only. Thus, the interpolation procedure is no longer needed in estimating $\mathbf{M}_n$. The time complexity for calculating R estimating equations is $O\{r_0 R\log(r_0)\}$ and solving the least squares estimate with a $R \times p$ design matrix for p times takes $O(Rp^3)$ time. In practice, $R = 100$ is enough to derive a good estimate of $\mathbf{M}_n$. The matrix multiplication of $\mathbf{M}_n^{-1}$ and $\mathbf{U}_{n,i}(\boldsymbol{\beta})$'s take $O(np^2)$ time with given $\mathbf{M}_n$. Calculating the norm of a p dimensional vector for n times takes $O(np)$ time. Thus, the time complexity of calculating $\boldsymbol{\pi}^{\text{optA}}$ is $O\{np^2 + np + n\log(r_0) + r_0\log(r_0) + Rp^3 + r_0 R\log(r_0)\} = O\{np^2 + n\log(r_0)\}$.

To avoid estimating $\mathbf{M}_n$ and matrix multiplications, we propose another version of SSPs based on the L-optimality $\boldsymbol{\pi}^{\text{optL}} = \{\pi_i^{\text{optL}}\}_{i=1}^n$, where

$$\pi_i^{\text{optL}} = \frac{\left\|\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|}{\sum_{i=1}^n \left\|\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|}, \tag{7}$$

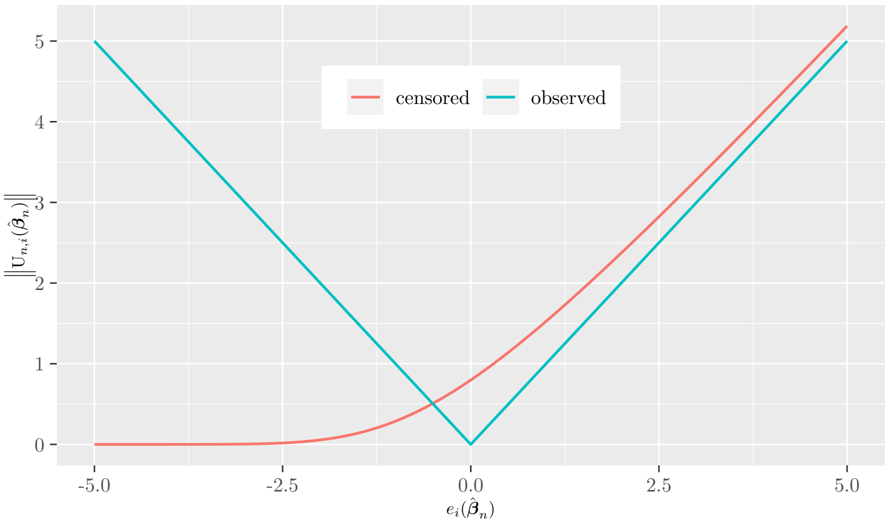

Fig. 1 Influence of prediction errors on $\boldsymbol{\pi}^{\mathrm{optL}}$.

and

$$\|\mathbf{U}_{n,i}(\boldsymbol{\beta})\| = \left\|\mathbf{X}_i - \bar{\mathbf{X}}\right\| \{(1-\delta_i)\,|\kappa_i(\boldsymbol{\beta})| + \delta_i\,|e_i(\boldsymbol{\beta})|\}\,. \tag{8}$$

For $\boldsymbol{\pi}^{\mathrm{optL}}$, we only need to estimate $\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)$'s which take $O\{n\log(r_0)\}$ time and calculating norms of n vectors of dimension p takes $O(np)$ time. Thus the time complexity to calculate $\boldsymbol{\pi}^{\mathrm{optL}}$ is $O\{np + n\log(r_0)\}$ which is less time-consuming than $\boldsymbol{\pi}^{\mathrm{optA}}$. It should be noted that there are several steps involved in estimating $\{\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\}_{i=1}^n$, each of which takes $O(np)$ time. When p is small and comparable to $\log(r)$, the time complexity of the interpolations required, which take $O\{n\log(n)\}$ time, is similar to $O(np)$. As a result, estimating $\{\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\}_{i=1}^n$ takes only slightly less than $O(np^2)$ time. Nevertheless, as p increases, the computational efficiency of $\boldsymbol{\pi}^{\mathrm{optL}}$ becomes more apparent.

The effect of $e_i(\hat{\boldsymbol{\beta}}_n)$ on the optimal SSPs is interesting. For parametric models without censoring, observations with residuals of large magnitude have large optimal SSPs in existing investigations. This is not true for censored observations. Note that $\mathbb{E}(\epsilon) = \int_{-\infty}^{\infty} u\mathrm{d}F_\epsilon(u)$ is 0 in model (1), where $F_\epsilon(u)$ is the cumulative distribution function of ϵ. Thus, $\left\|\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|$ converges to 0 as $e_i(\hat{\boldsymbol{\beta}}_n) \to -\infty$ for censored observations. When $e_i(\hat{\boldsymbol{\beta}}_n) \to +\infty$, the numerator of $\kappa_i(\hat{\boldsymbol{\beta}}_n)$ converges to zero slower than the denominator. Thus, $\left\|\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|$ converges to $+\infty$ as $e_i(\boldsymbol{\beta}) \to +\infty$. Note that $\{\pi_i^{\mathrm{optA}}\}_{i=1}^n$ are proportional to $\left\{\left\|\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|\right\}_{i=1}^n$ and $\{\pi_i^{\mathrm{optL}}\}_{i=1}^n$ are proportional to $\left\{\left\|\mathbf{M}_n^{-1}\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|\right\}_{i=1}^n$ where $\mathbf{M}_n^{-1}$ does not change for different i's. Thus, π_i^{optA} and π_i^{optL} have the same trend as $\left\|\mathbf{U}_{n,i}(\hat{\boldsymbol{\beta}}_n)\right\|$ with respect to $e_i(\hat{\boldsymbol{\beta}}_n)$. Nevertheless, it does not contradict the fact that optimal SSPs prefer data points whose event time

Table 1 Means and summations of uniform SSPs and $\boldsymbol{\pi}^{\mathrm{optA}}$ for censored and observed observations with Gumbel (G), Logistic (L) and Normal (N) distributions as the error distributions and different censoring rates c_r.

	c_r: 25%				c_r: 50%				c_r: 75%			
	uniform	G	L	N	uniform	G	L	N	uniform	G	L	N
					summation							
Censored	0.25	0.253	0.309	0.272	0.50	0.443	0.459	0.447	0.75	0.544	0.545	0.545
Event	0.75	0.747	0.691	0.728	0.50	0.557	0.541	0.553	0.25	0.456	0.455	0.455
					mean ($\times n$)							
Censored	1.00	1.013	1.235	1.087	1.00	0.884	0.919	0.892	1.00	0.728	0.729	0.725
Event	1.00	0.996	0.922	0.971	1.00	1.117	1.081	1.109	1.00	1.807	1.802	1.832

is harder to predict. A censored observation means $C_i \leq T_i$, and a negative $e_i(\hat{\boldsymbol{\beta}}_n)$ means $C_i < \hat{T}_i$. Thus, for a censored observation, a larger magnitude of a negative $e_i(\hat{\boldsymbol{\beta}}_n)$ does not mean a larger prediction error, $|T_i - \hat{T}_i|$. On the other hand, a positive $e_i(\hat{\boldsymbol{\beta}}_n)$ means $C_i > \hat{T}_i$, and thus a larger magnitude of a positive $e_i(\hat{\boldsymbol{\beta}}_n)$ means a larger prediction error, $|T_i - \hat{T}_i|$. For uncensored observations, clearly a large absolute $e_i(\hat{\boldsymbol{\beta}}_n)$ means that the event time is hard to predict, thus both π_i^{optL} and π_i^{optA} are large when $e_i(\hat{\boldsymbol{\beta}}_n)$ is far away from zero. For the same magnitude of a positive error, the optimal SSPs are higher for censored observations than uncensored ones since the event time of a censored observation is harder to predict than an uncensored observation. This pattern is shown in Figure 1.

To investigate which types of observation are preferred by optimal SSPs in given datasets, we used simulated datasets of size n whose detailed information is stated in the first paragraph of Section 5 to generate the optimal SSPs. Table 1 presents the means and sums of the SSPs for censored and uncensored observations separately. We calculated $\boldsymbol{\pi}^{\mathrm{optA}}$ and derived their sums and means for each dataset. When we compared the mean optimal SSPs with the mean of the uniform SSPs, we observed that $\boldsymbol{\pi}^{\mathrm{optA}}$ prefers uncensored observations at high censoring rates but prefers censored observations at low censoring rates. At a censoring rate of 0.5, the mean of $\boldsymbol{\pi}^{\mathrm{optA}}$ was higher than n^{-1} (the uniform SSP) for uncensored observations but smaller than n^{-1} for censored observations, indicating that optimal SSPs prefer uncensored observations that provide more information beyond the influence of the censoring rate. This preference can also be observed in the summation of $\boldsymbol{\pi}^{\mathrm{optA}}$. The summations of $\boldsymbol{\pi}^{\mathrm{optA}}$ for both types of observations were similar to the summations of uniform SSPs at a censoring rate of 0.25, but significantly different at a censoring rate of 0.75. These two preferences likely involve trade-offs that require further investigation.

4 A Two-Step Subsampling Approach

In this section, a feasible two-step method is proposed to derive the subsampling estimator. Note that the optimal SSPs in Section 3 are dependent on the full sample estimator $\hat{\boldsymbol{\beta}}_n$ which cannot be used directly. To resolve this issue, in the first step, we approximate $\boldsymbol{\pi}^{\text{optL}}$ and $\boldsymbol{\pi}^{\text{optA}}$ based on a pilot estimator $\tilde{\boldsymbol{\beta}}_{r_0}$ of $\hat{\boldsymbol{\beta}}_n$, which is derived from a small, pilot subsample of size r_0 in the first step. Denote the estimated optimal SSPs as $\boldsymbol{\pi}^{\text{optA}}(\tilde{\boldsymbol{\beta}}_{r_0})$ and $\boldsymbol{\pi}^{\text{optL}}(\tilde{\boldsymbol{\beta}}_{r_0})$. In the second step, a subsample of size r is drawn according to the estimated SSPs in the first step. The subsampling estimator $\check{\boldsymbol{\beta}}_r$ is obtained by the second step subsample in combination with the pilot subsample. Following the idea of Zeng and Lin (2008), the variance of $\check{\boldsymbol{\beta}}_r$ is estimated based on a sandwich formula $\mathbf{M}_r^{-1}\mathbf{V}_r\mathbf{M}_r^{-1}$, where $\mathbf{M}_r$ is the estimator of $\mathbf{M}_n$ based on the combined subsample and

$$\mathbf{V}_r = \frac{1}{n^2(r_0+r)}\left\{\sum_{i=1}^{r}\frac{\mathbf{U}^*_{r,i}(\check{\boldsymbol{\beta}}_r)\{\mathbf{U}^*_{r,i}(\check{\boldsymbol{\beta}}_r)\}^\top}{\{\pi_i^{\text{opt}}(\check{\boldsymbol{\beta}}_r)\}^2} + n^2\sum_{i=1}^{r_0}\mathbf{U}^*_{r_0,i}(\tilde{\boldsymbol{\beta}}_{r_0})\{\mathbf{U}^*_{r_0,i}(\tilde{\boldsymbol{\beta}}_{r_0})\}^\top\right\},$$

with $\mathbf{U}^*_{r_0,i}(\boldsymbol{\beta})$ being the estimating function for the ith element in the pilot subsample.

Now we consider the time complexity of the two-step approach. As discussed in Section 3, the first step involves obtaining either $\boldsymbol{\pi}^{\text{optA}}$ or $\boldsymbol{\pi}^{\text{optL}}$, which takes $O\{np^2+n\log(r_0)\}$ time and $O\{np+n\log(r_0)\}$ time, respectively. In the second step, calculating the subsample estimator takes $O\{\xi_r[rp^2+r\log(r)]\}$ time. For the sandwich variance estimator, calculating $\mathbf{M}_r$ takes $O\{Rr\log(r)\}$ time, as discussed in Section 3; calculating $\mathbf{V}_r$ costs $O(rp^2+r_0p^2)$ time. Therefore, the overall time complexity of the two-step method using $\boldsymbol{\pi}^{\text{optA}}$ is $O\{np^2+n\log(r_0)+\xi_r[rp^2+r\log(r)]+Rr\log(r)+rp^2+r_0p^2\} = O\{np^2+n\log(r_0)+\xi_r[rp^2+r\log(r)]+Rr\log(r)\}$. The time complexity of $\boldsymbol{\pi}^{\text{optL}}$ is similar this formula except that the np^2 term is replaced by np.

Note that the approximate optimal SSPs, denoted by $\boldsymbol{\pi}^{\text{opt}}(\tilde{\boldsymbol{\beta}}_{r_0})$, are derived from a random pilot estimator which may cause additional disturbance. Based on (6), for uncensored observations with $e_i(\hat{\boldsymbol{\beta}})$ more approaching zero, their exact SSPs are closer to zero and the additional disturbances may get amplified. To protect the subsample estimator, we adopt the idea of defensive sampling and mix the approximated $\boldsymbol{\pi}^{\text{opt}}(\tilde{\boldsymbol{\beta}}_{r_0})$ with the uniform SSP denoted by $\boldsymbol{\pi}^{\text{Uni}}_{r_0}$ (Hesterberg, 1995). That is, we use adjusted optimal SSPs $\boldsymbol{\pi}^{\text{opt}}_{\alpha}(\tilde{\boldsymbol{\beta}}_{r_0}) = \{\pi^{\text{opt}}_{\alpha i}(\tilde{\boldsymbol{\beta}}_{r_0})\}_{i=1}^{n}$ instead of $\boldsymbol{\pi}^{\text{opt}}(\tilde{\boldsymbol{\beta}}_{r_0})$ to do subsampling, where

$$\pi^{\text{opt}}_{\alpha i}(\tilde{\boldsymbol{\beta}}_{r_0}) = (1-\alpha)\pi^{\text{opt}}_{i}(\tilde{\boldsymbol{\beta}}_{r_0}) + \frac{\alpha}{n}, \quad 0<\alpha<1, \quad i=1,2,\ldots,n.$$

In the simulation study and the real data analysis, we set $\alpha = 0.2$.

At high censoring rates, the sandwich estimator in Zeng and Lin (2008) overestimates the empirical variance; see Section 5. We resolve this issue by selecting B subsamples of size r to estimate B subsampling estimators

$\{\breve{\boldsymbol{\beta}}_{b,r}\}_{b=1}^{B}$ in the second step. In this scenario, the resultant estimator takes the form

$$\breve{\boldsymbol{\beta}}_r = \frac{1}{B}\sum_{b=1}^{B}\breve{\boldsymbol{\beta}}_{b,r}, \tag{9}$$

and its variance estimator is

$$\breve{\mathbf{V}}_r = \frac{1}{B(B-1)}\sum_{b=1}^{B}\left(\breve{\boldsymbol{\beta}}_{b,r} - \breve{\boldsymbol{\beta}}_r\right)\left(\breve{\boldsymbol{\beta}}_{b,r} - \breve{\boldsymbol{\beta}}_r\right)^{\top}. \tag{10}$$

Note that $\breve{\mathbf{V}}_r$ can be used for statistical inferences on the true regression coefficients if the subsample size is much smaller than the full data size (Wang et al., 2022). This requires that rB/n is close to zero in practice since the actual size of the subsample is $r \times B$. The dimension of $\breve{\boldsymbol{\beta}}_r$ is p, thus B should be larger than p in order to get a reliable variance estimator. In practice, the choice of B should be much smaller than n/r but greater than p. Since we do not need to estimate $\mathbf{M}_r$ when $B > 1$, the time complexity in this case using $\boldsymbol{\pi}^{\mathrm{optA}}$ is $O\{np^2 + n\log(r_0) + \xi_r B[rp^2 + r\log(r)] + Bp^2\} = O\{np^2 + \xi_r B[rp^2 + r\log(r)]\}$. Similarly, the time complexity when using $\boldsymbol{\pi}^{\mathrm{optL}}$ is $O\{np + \xi_r B[rp^2 + r\log(r)]\}$. Nevertheless, it is important to note that the computation time is dominated by the derivation of $\breve{\boldsymbol{\beta}}_r$ when $\xi_r Br \geq n$.

5 Simulation

The performances of the estimator from the two-step procedure were assessed in a simulation study. We used three different error distributions: standard normal, standard logistic, and centered Gumbel distribution with shape parameter zero and scale parameter one. The covariates follow multivariate normal with mean zero and covariance matrix $\boldsymbol{\Sigma}_{ij} = 0.5^{I(i \neq j)}$. The dimension of $\boldsymbol{\beta}$ was seven and all coefficients were set to be 1 including the intercept. The censoring times were generated from the Uniform distribution with the minimum and maximum values equal to 0 and c, respectively, where c was tuned to achieve censoring rates $c_r \in \{0.25, 0.50, 0.75\}$.

For each of the nine configurations, 1000 large datasets of size $n = 500,000$ were generated. In our simulation, the pilot sample size was $r_0 = 3000$. The second-step subsample sizes considered were $r \in \{4000, 8000, 16000\}$ and $B \in \{1, 10\}$. For the ith dataset in each of the nine configurations, we derived $\breve{\boldsymbol{\beta}}_r^{(i)}$ by the two-step subsampling method using $\boldsymbol{\pi}^{\mathrm{optA}}$, $\boldsymbol{\pi}^{\mathrm{optL}}$ and the uniform SSPs. We compared the performance of the two-step method using different SSPs by the root mean square error (RMSE) of $\breve{\boldsymbol{\beta}}_r^{(i)}$, where the RMSE is calculated by

$$\mathrm{RMSE} = \left(\frac{1}{s}\sum_{i=1}^{s}\|\breve{\boldsymbol{\beta}}_r^{(i)} - \hat{\boldsymbol{\beta}}_n\|^2\right)^{1/2}. \tag{11}$$

Note that for each replicate, the pilot subsample is different.

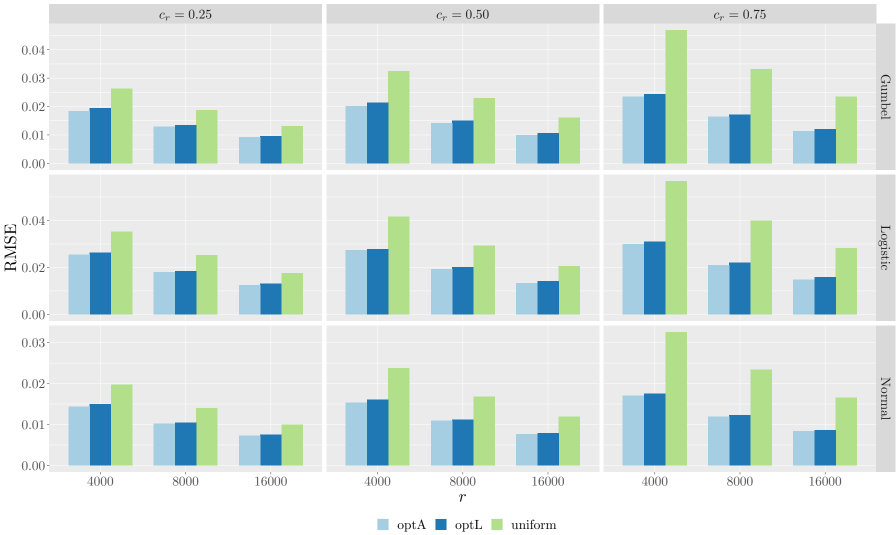

Fig. 2 Empirical RMSEs for different SSPs, error distribution, subsample sizes r and censoring rates when covariates follow multivariate normal distribution based on the two-step procedure with $B = 10$.

The estimation efficiency of our method is shown in Figure 2. It shows the RMSEs of $\breve{\boldsymbol{\beta}}_r$ based on the uniform SSPs, $\boldsymbol{\pi}^{\text{optA}}$, and $\boldsymbol{\pi}^{\text{optL}}$ for the two-step method when $B = 10$. Note that the actual subsample sizes we use to estimate the resultant estimator are $B \times r$. As expected, in all data configurations, $\boldsymbol{\pi}^{\text{optL}}$ and $\boldsymbol{\pi}^{\text{optA}}$ give smaller RMSE than uniform SSP and $\boldsymbol{\pi}^{\text{optA}}$ give the smallest RMSE. As the censoring rate increases, there will be fewer informative observations. So the RMSEs of all methods increase as less information is available. In all configurations, the RMSEs decrease as the subsample size r increases.

We evaluated the accuracy of the variance estimator by comparing its average over 1000 subsamples with the empirical variance. The upper panel of Figure 3 presents the results for the formula-based variance estimator when $B = 1$. The figure reveals that the estimated and empirical RMSEs are close at censoring rates 0.25 and 0.5, indicating that the sandwich estimator estimates the true variance well at low to moderate censoring rates. However, the sandwich estimator noticeably overestimates the true variance differences at the censoring rate 0.75, which leads to conservative conclusions and loss of power in inferences. To correct the bias for high-censoring cases, we set $B = 10$ and estimate the variances using Equation (10). The lower panel of Figure 3 demonstrates that this provides accurate variance estimates for high censoring rates. Hence, we suggest using $B = 10$ in the second step and estimating the standard error by (10) for high censoring rates.

Finally, we evaluate the computational efficiency of the two-step methods. We compared the computing time when $B = 10$ and $B = 1$. To ensure a fair comparison, we increased the subsample size for $B = 1$ to $10r$. We performed

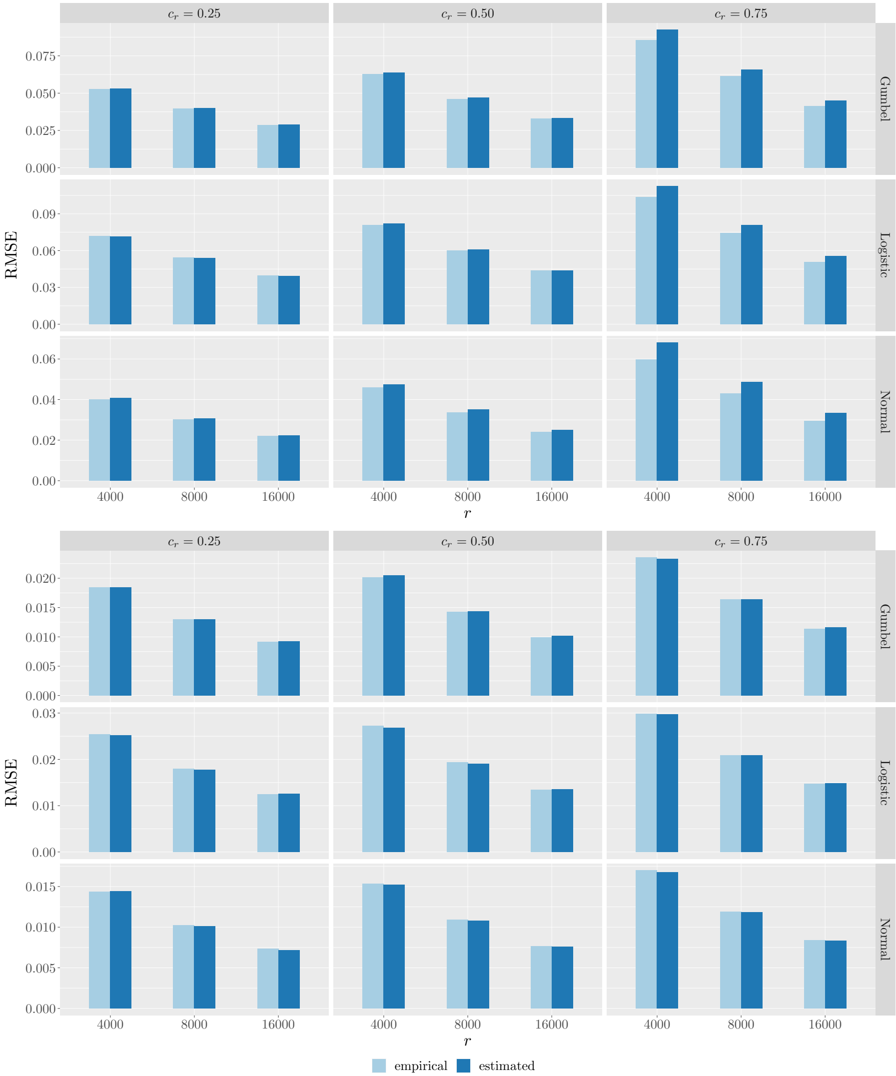

Fig. 3 Empirical and estimated RMSEs with $\boldsymbol{\pi}^{\mathrm{optA}}$ for different error distribution, subsample sizes r and 0.75 censoring rates when covariates follow multivariate normal distribution based on the two-step procedure with $B = 1$ (upper) and $B = 10$ (lower).

the computation on a laptop running Windows 11 with an Intel Core (TM) i7–8650U @ 1.90GHz processor and 16 GB memory. Figure 4 summarizes the computational and estimation efficiency of both methods. It shows that using $B = 10$ subsamples of size r is less time-consuming than using $B = 1$ subsample of size $10r$. This is due to the fact that the formula-based variance estimation when $B = 1$ takes a considerable amount of time, as discussed in Section 4. Using $\boldsymbol{\pi}^{\mathrm{optL}}$, $\boldsymbol{\pi}^{\mathrm{optA}}$, and uniform SSPs take similar CPU time because the

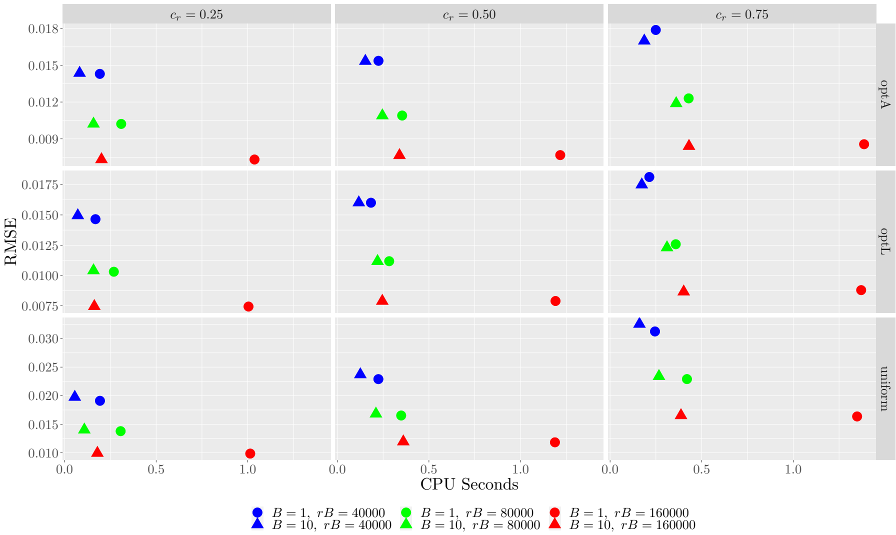

Fig. 4 Comparison of estimation efficiency and computational efficiency for different subsample sizes r, censoring rates, subsampling methods, and values of B when the errors are generated from a standard normal distribution.

value of r is large enough so that calculating the SSPs does not dominate the computing time.

6 Survival of Lymphoma

The two-step procedure was applied to model the survival time of lymphoma patients in the SEER program. The dataset contains 159,149 patients that were diagnosed with lymphoma from 1973 to 2012 and the censoring rate is 58.3%. We considered four risk factors, including age with the unit of year, nonwhite race indicator (1 =nonwhite), male indicator (1 = male) and the diagnostic year. We also included the interaction between age with the male indicator and age with the non-white indicator. The pilot sample size was set to be $r_0 = 2000$ and the subsample sizes were $r \in \{2000, 4000, 8000\}$. We considered three kinds of SSPs, uniform SSPs, the L-optimal SSPs ($\boldsymbol{\pi}^{\text{optL}}$) and the A-optimal SSPs ($\boldsymbol{\pi}^{\text{optA}}$). In the real data analysis, we set $B = 10$.

Figure 5 shows the empirical RMSEs from 1000 replicates of the two-step method with $B = 10$ based on different SSPs and different second-step subsample sizes. The RMSE decreases as r increases which indicates the consistency of our method. As expected, both optimal SSPs perform better than the uniform SSPs. It should be noted that $\boldsymbol{\pi}^{\text{optA}}$ does not result in universally smaller RMSEs for all parameters. As shown in Figure 2, for the interaction term 'Age×Male' and the risk factor 'Diagnostic Year', the 'optA' estimates have higher RMSE than the 'optL' estimates. This is because $\boldsymbol{\pi}^{\text{optA}}$ are designed to minimize the overall RMSEs for all risk factors and interactions, rather than specifically targeting individual risk factors or interactions.

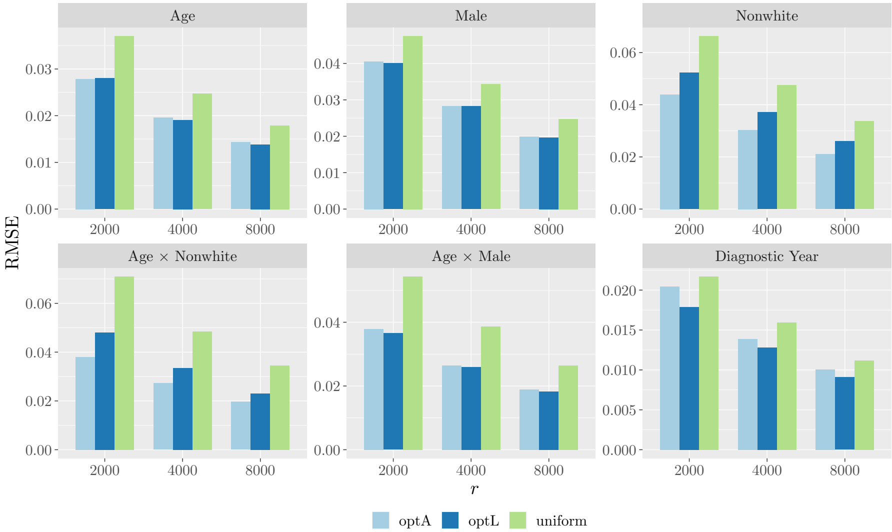

Fig. 5 Empirical RMSEs of different risk factors for different SSPs and different second-step subsample sizes r when fixing the pilot sample size $r_0 = 2000$ over 1000 replicates of the two-step method with $B = 10$.

Table 2 Estimates (EST) and their empirical standard errors (ESE) and average estimated standard errors (ASE) from different subsampling approaches for $r = 4000$ and r_0=2000 over 1000 replicates of the two-step method with $B = 10$.

	optL			optA			uniform			Full	
	EST	ESE	ASE	EST	ESE	ASE	EST	ESE	ASE	EST	SE
Age	−1.029	0.019	0.020	−1.030	0.020	0.020	−1.029	0.025	0.026	−1.030	0.013
Male	0.701	0.028	0.028	0.702	0.028	0.028	0.701	0.034	0.035	0.700	0.017
Nonwhite	−0.665	0.037	0.036	−0.667	0.030	0.030	−0.666	0.048	0.046	−0.666	0.023
Age×Nonwhite	0.303	0.034	0.033	0.306	0.027	0.027	0.299	0.048	0.049	0.306	0.026
Age×Male	−0.486	0.026	0.026	−0.486	0.026	0.026	−0.488	0.039	0.038	−0.486	0.018
Diagnostic Year	0.478	0.013	0.013	0.478	0.014	0.014	0.479	0.016	0.015	0.478	0.008

Table 2 summarizes the average estimates and their corresponding average empirical standard errors (EST) and average pooled standard error estimators (ASE) for all subsampling methods when $r = 4000$ and $B = 10$ over 1000 subsamples. We compared the subsample estimates with the full data estimates whose standard errors were derived from the non-parametric bootstrap of 1000 samples. The optimal subsampling methods yield one-third less standard errors than the uniform subsampling method The empirical and estimated standard errors are similar which indicates the subsampling methods are suitable for statistical inference. The empirical standard errors for both optimal subsampling methods are small which shows that using a small subsample is sufficient in practice to estimate the full data estimates. The results indicate that elder, female, nonwhite, and earlier-diagnosed patients had shorter survival times. For white patients and male patients, the slope of age was steeper.

7 Discussion

The optimal subsampling method for the least square fitting of semiparametric AFT model for massive survival data is challenging due to non-smooth estimating functions. The crucial element of this approach is determining the SSP, which we addressed by a resampling method (Zeng and Lin, 2008). We proposed two types of optimal SSPs, induced by the A-optimality and the L-optimality from design of experiments. Optimal SSPs prefer extreme observations, but for censored observations, only those with positive residuals of large magnitudes are considered extreme, while those with negative residuals of large magnitudes are not. Moreover, for positive residuals with the same magnitude, optimal SSPs prefer censored observations over uncensored observations. This preference for extreme observations does not contradict the accepted notion that optimal SSPs tend to choose observations that are harder to predict. We conducted a simulation study and a real data analysis to demonstrate the feasibility and effectiveness of the proposed methods, which provide good approximations of full data inferences while being computationally feasible.

Further investigation is warranted for optimal subsampling methods in fitting semiparametric AFT models with the rank-based approach. In rank-based estimation, censored observations do not contribute to the estimating function, but they contribute to the ranking. Simply assigning a zero SSP to censored observations would not properly account for their contributions. A possible solution is to express the estimating functions using some martingales, which facilitates the evaluations of the contributions of censored observations. This approach has been successfully applied in Cox models (Zhang et al., 2023). Additionally, the induced smoothing approach, which improves computational efficiency (Chiou et al., 2014, 2015), remains important. This method replaces the non-smooth estimating equations with a smoothed version whose solutions are asymptotically equivalent to those of the non-smooth version. Ongoing investigation in this direction will be reported elsewhere.

References

Ai, M., J. Yu, H. Zhang, and H. Wang. 2021. Optimal subsampling algorithms for big data generalized linear models. *Statistica Sinica 31*(2): 749–772 .

Buckley, J. and I. James. 1979. Linear regression with censored data. *Biometrika 66*(3): 429–436 .

Chiou, S., S. Kang, and J. Yan. 2015. Rank-based estimating equations with general weight for accelerated failure time models: An induced smoothing approach. *Statistics in Medicine 34*(9): 1495–1510 .

Chiou, S.H., S. Kang, and J. Yan. 2014. Fitting accelerated failure time models in routine survival analysis with R package aftgee. *Journal of Statistical Software 61*(11): 1–23 .

Drineas, P., M.W. Mahoney, and S. Muthukrishnan 2006. Sampling algorithms for L_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 1127–1136. Association of Computing Machinary.

Hesterberg, T. 1995. Weighted average importance sampling and defensive mixture distributions. *Technometrics 37*(2): 185–194 .

Jin, Z., D. Lin, L. Wei, and Z. Ying. 2003. Rank-based inference for the accelerated failure time model. *Biometrika 90*(2): 341–353 .

Jin, Z., D. Lin, and Z. Ying. 2006. On least-squares regression with censored data. *Biometrika 93*(1): 147–161 .

Keret, N. and M. Gorfine. 2022. Optimal Cox regression subsampling procedure with rare events. arXiv preprint: https://arxiv.org/abs/2012.02122.

Li, R., C. Chang, J.M. Justesen, Y. Tanigawa, J. Qian, T. Hastie, M.A. Rivas, and R. Tibshirani. 2022. Fast lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK biobank. *Biostatistics 23*(3): 522–540 .

Ma, P., Y. Chen, X. Zhang, X. Xing, J. Ma, and M.W. Mahoney. 2022. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. *The Journal of Machine Learning Research 23*(1): 7970–8014 .

Ma, P., M.W. Mahoney, and B. Yu. 2015. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research 16*(27): 861–911 .

Mahoney, M.W. et al. 2011. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning 3*(2): 123–224 .

Su, W., G. Yin, J. Zhang, and X. Zhao. 2023. Divide and conquer for accelerated failure time model with massive time-to-event data. *The Canadian Journal of Statistics 51*(2): 400–419 .

Tsiatis, A.A. 1990. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics 18*(1): 354–372 .

Wang, H. and Y. Ma. 2021. Optimal subsampling for quantile regression in big data. *Biometrika 108*(1): 99–112 .

Wang, H., R. Zhu, and P. Ma. 2018. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association 113*(522): 829–844 .

Wang, J., J. Zou, and H. Wang. 2022. Sampling with replacement vs Poisson sampling: A comparative study in optimal subsampling. *IEEE Transactions on Information Theory 68*(10): 6605–6630 .

Wang, W., S.E. Lu, J.Q. Cheng, M. Xie, and J.B. Kostis. 2022. Multivariate survival analysis in big data: A divide-and-combine approach. *Biometrics 78*(3): 852–866 .

Wang, Y., C. Hong, N. Palmer, Q. Di, J. Schwartz, I. Kohane, and T. Cai. 2021. A fast divide-and-conquer sparse Cox regression. *Biostatistics 22*(2): 381–401 .

Wu, J., M.H. Chen, E.D. Schifano, and J. Yan. 2021. Online updating of survival analysis. *Journal of Computational and Graphical Statistics 30*(4): 1209–1223 .

Xue, Y., H. Wang, J. Yan, and E.D. Schifano. 2020. An online updating approach for testing the proportional hazards assumption with streams of survival data. *Biometrics 76*(1): 171–182 .

Yang, Z., H. Wang, and J. Yan. 2022. Optimal subsampling for parametric accelerated failure time models with massive survival data. *Statistics in Medicine 41*(27): 5421–5431 .

Zeng, D. and D. Lin. 2008. Efficient resampling methods for nonsmooth estimating functions. *Biostatistics 9*(2): 355–363 .

Zhang, H., L. Zuo, H. Wang, and L. Sun. 2023. Approximating partial likelihood estimators via optimal subsampling. *Journal of Computational and Graphical Statistics* .

Zuo, L., H. Zhang, H. Wang, and L. Liu. 2021. Sampling-based estimation for massive survival data with additive hazards model. *Statistics in Medicine 40*(2): 441–450 .