Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Optimal subsampling for semi-parametric accelerated failure time models with massive survival data using a rank-based approach

Zehan Yang | HaiYing Wang | Jun Yan

[1]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA

Correspondence
HaiYing Wang, Department of Statistics, University of Connecticut, Storrs, 06269, Connecticut. Email: haiying.wang@uconn.edu

Abstract

Subsampling is a practical strategy for analyzing vast survival data, which are progressively encountered across diverse research domains. While the optimal subsampling method has been applied to inferences for Cox models and parametric accelerated failure time (AFT) models, its application to semi-parametric AFT models with rank-based estimation have received limited attention. The challenges arise from the non-smooth estimating function for regression coefficients and the seemingly zero contribution from censored observations in estimating functions in the commonly seen form. To address these challenges, we develop optimal subsampling probabilities for both event and censored observations by expressing the estimating functions through a well-defined stochastic process. Meanwhile, we apply an induced smoothing procedure to the non-smooth estimating functions. As the optimal subsampling probabilities depend on the unknown regression coefficients, we employ a two-step procedure to obtain a feasible estimation method. An additional benefit of the method is its ability to resolve the issue of underestimation of the variance when the subsample size approaches the full sample size. We validate the performance of our estimators through a simulation study and apply the methods to analyze the survival time of lymphoma patients in the Surveillance, Epidemiology, and End Results program.

KEYWORDS:
A-optimality; stochastic process; survival analysis

1 | INTRODUCTION

The rapid growth of storage and surveillance technologies, along with advancements in data collection, have empowered the medical industry to gather and utilize extensive datasets containing survival outcomes for their research and development activities. Nevertheless, the size of these datasets often surpasses the computational capacities of researchers' computers. To tackle the computational burden that arises due to large datasets, various subsampling methods have been proposed. In the context of logistic models, Wang et al. [1] introduced an optimal subsampling technique, which aimed to approximate the inferences derived from the entire dataset by utilizing a carefully weighted subsample. For each observation, the optimal subsampling probability (SSP) is proportional to its contribution to the estimating function[2]. In survival models, this method has been applied to, for example, the additive hazard model[3], the Cox model[4,5], and the Cox model when dealing with rare event data[6].

Accelerated failure time (AFT) models characterize the survival times directly, where the regression coefficients correspond to multiplicative effects on the survival time. As a useful and more intuitive alternative to the Cox model[7], AFT models have gained more popularity recently with the advancement in inferences, computational strategies, and software packages[8]. For big survival data, Yang et al.[9] investigated the optimal subsampling method with parametric AFT models, where the optimal subsampling procedure is similar to that of a generalized linear model[10]. Semi-parametric AFT models with unspecified error distributions are more desired in practice. Two commonly used estimation approaches for semi-parametric AFT models are the least-squares approach[11,12,13,14] and the rank-based approach[15,16,17,18,19]. For the least-squares approach, Yang et al.[20] studied the optimal subsampling method with the optimal SSPs intuited from Wang et al.[2].

Despite its widespread use, the least-squares approach requires a consistent estimate as the initial value for optimization, and the rank-based approach is often used for this purpose. Both the rank-based approach and the least-squares approach face challenges in optimization due to the non-smooth nature of their estimating functions. Nevertheless, the induced smoothed method can be applied to smooth the estimating function for the rank-based approach[21]. In contrast, no solutions have been proposed to smooth the estimating function for the least-squares approach. Additionally, the rank-based approach outperforms the least-squares approach when the error distribution has a heavy tail. This corresponds to the empirical observation that the median (or more general quantile) regression outperforms mean regression with heavy-tailed error distributions in non-censored scenarios. Moreover, a significant gap exists in the literature regarding subsampling for the rank-based approach. This method involves a time complexity of $O(\xi_n n^2 p)$ to derive the estimator from a full sample of size n with p covariates, where ξ_n represents the number of iterations needed for convergence. Given this computational burden, the development of an optimal subsampling method for rank-based estimation is imperative.

Developing optimal subsampling probabilities for the rank-based AFT modeling is challenging. The optimal SSP of an observation depends on its contribution to the estimation function[2]. The rank-based estimating functions in their most commonly used form[15] seemingly suggest zero weight for censored observations. Of course, a careful investigation reveals that censored observations contribute implicitly. We address this challenge by expressing the estimating functions in terms of a well-defined stochastic process[17,22]. The contributions of censored observations can then be explicitly assessed. Further, rank-based estimating functions are non-smooth in regression coefficients, which present general computational challenges in finding their root. We employ an induced smoothing procedure[8,21,23,24] that effectively renders the non-smooth part of the estimating function smooth without altering the asymptotic properties of the resulting estimator. The variance matrix of the resulting estimator is estimated by a sandwich estimator that accounts for both the uncertainty of the subsampling process and the uncertainty of the full-data estimator. This is in contrast to existing literature[9,20] where the uncertainty in the full-data estimator has been discarded as negligible. Our implementation is part of an R package `aftosmac`, which is publicly available at https://github.com/YEnthalpy/aftosmac.

The rest of the paper is organized as follows. Section 2 introduces the model and the general subsampling procedure for semi-parametric AFT models based on the rank-based approach. Section 3 first presents two optimal SSPs based on two criteria that are motivated by the optimal design of experiments, and then proposes a feasible two-step procedure along with a bias-corrected sandwich estimator for the asymptotic variance. Section 4 reports the performance of the proposed estimator through a simulation study. In section 5, we illustrate the usage of the proposed method in a case study of the survival time of lymphoma patients in the Surveillance, Epidemiology, and End Results (SEER) program. Section 6 concludes with a discussion.

2 | SUBSAMPLING ESTIMATION FOR RANK-BASED AFT MODELING

2.1 | Full Sample Estimation

Consider a full sample consisting of n subjects. For subject $i = 1, \dots, n$, let T_i, C_i, and $\mathbf{X}_i$ represent the log-transformed failure time, the log-transformed censoring time, and a $p \times 1$ covariate vector, respectively. We assume that T_i and C_i are independent conditional on $\mathbf{X}_i$. The semi-parametric accelerated failure time model specifies that

$$T_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, 2, \dots, n,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients and ϵ_i's are independent error terms with identical but unspecified distribution. Due to right censoring, the observed data are $\mathcal{D}_n = (Y_i, \delta_i, \mathbf{X}_i)_{i=1}^n$, where $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i < C_i)$, with $I(\cdot)$ denoting the indicator function. Observations across subjects are independent and identically distributed. Let $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$ be a subsampling distribution so that $\sum_{i=1}^n \pi_i = 1$ and $\pi_i > 0$ for all i's.

The estimating function induced by the linear rank test[16,17] is defined based on the ranks of $\{e_i(\boldsymbol{\beta})\}_{i=1}^n$, where $e_i(\boldsymbol{\beta}) = Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}$. Let $N_i(t, \boldsymbol{\beta}) = \delta_i I\left\{e_i(\boldsymbol{\beta}) \le t\right\}$ be the counting process on the time scale of the residual. Define

$$S^{(0)}(t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I\{t \le e_i(\boldsymbol{\beta})\} \quad \text{and} \quad S^{(1)}(t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I\{t \le e_i(\boldsymbol{\beta})\}\mathbf{X}_i.$$

According to Tsiatis[17], the rank-based estimating function of $\boldsymbol{\beta}$ for the semi-parametric AFT model is

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} \varphi(t; \boldsymbol{\beta})\left[\mathbf{X}_i - \bar{\mathbf{X}}(t; \boldsymbol{\beta})\right] \mathrm{d}N_i(t; \boldsymbol{\beta}), \tag{1}$$

where $\varphi(t; \boldsymbol{\beta})$ is a possibly data-dependent weight function. $\bar{\mathbf{X}}(t; \boldsymbol{\beta}) = S^{(1)}(t; \boldsymbol{\beta})/S^{(0)}(t; \boldsymbol{\beta})$.

Among the various options of the weight $\varphi(t; \boldsymbol{\beta})$, we focus on Gehan's weight[25] $\varphi(t; \boldsymbol{\beta}) = S^{(0)}(t; \boldsymbol{\beta})$. This weight has the advantage of canceling the denominator of $\bar{\mathbf{X}}$ Equation (1). The resulting estimating function takes the form

$$\mathbf{U}_G(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\frac{1}{n}\sum_{j=1}^{n} I\{t \le e_j(\boldsymbol{\beta})\}(\mathbf{X}_i - \mathbf{X}_j)\mathrm{d}N_i(t; \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}\delta_i(\mathbf{X}_i - \mathbf{X}_j)I\{e_i(\boldsymbol{\beta}) \le e_j(\boldsymbol{\beta})\} \tag{2}$$

This estimating function is discontinuous in $\boldsymbol{\beta}$, so finding its root is computationally challenge and sometimes convergence in iterative root-finding algorithms may not be possible. The form of this estimating function, however, facilitates the application of the induced smoothing approach[21,23].

The induced smooth approach replaces the non-smooth estimating function (2) with a smooth version whose solution is asymptotically equivalent to the direct solution to (2). Define a $p \times 1$ standard normal random vector $\mathbf{Z}$ that is independent of the data. The induced smoothing procedure replaces $\mathbf{U}_G(\boldsymbol{\beta})$ with $\mathbb{E}\left[\mathbf{U}_G(\boldsymbol{\beta} + n^{-1/2}\mathbf{Z})\right]$, where the expectation is taken concerning $\mathbf{Z}$. The smoothed version of Equation (2) is $\tilde{\mathbf{U}}_G(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{U}}_{G,i}(\boldsymbol{\beta})$, where

$$\tilde{\mathbf{U}}_{G,i}(\boldsymbol{\beta}) = \frac{\delta_i}{n}\sum_{j=1}^{n}(\mathbf{X}_i - \mathbf{X}_j)\Phi[\kappa_{ij}(\boldsymbol{\beta})], \tag{3}$$

and $\kappa_{ij}(\boldsymbol{\beta}) = \sqrt{n}[e_j(\boldsymbol{\beta}) - e_i(\boldsymbol{\beta})]/r_{ij}$, with $r_{ij}^2 = (\mathbf{X}_i - \mathbf{X}_j)^\top(\mathbf{X}_i - \mathbf{X}_j)/n$. The slope matrix of $\tilde{\mathbf{U}}_G(\boldsymbol{\beta})$ takes the form

$$\mathbf{M}_n(\boldsymbol{\beta}) = \frac{\partial \tilde{\mathbf{U}}_G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{n}\sum_{j=1}^{n}\frac{\phi[\kappa_{ij}(\boldsymbol{\beta})]}{r_{ij}}(\mathbf{X}_i - \mathbf{X}_j)^{\otimes 2}, \tag{4}$$

where $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^\top$ for vector $\mathbf{A}$ and $\phi(\cdot)$ is the probability density function of the standard normal distribution.

2.2 | Subsampling Estimation

Finding the solution $\hat{\boldsymbol{\beta}}_n$ to $\tilde{\mathbf{U}}_G(\boldsymbol{\beta}) = 0$ is time-consuming, because it requires evaluating $\tilde{\mathbf{U}}_G(\boldsymbol{\beta})$ which takes $O(n^2 p)$ time in each iteration of traditional optimization methods. In situations where the dataset's enormity is truly massive, this endeavor might even be unattainable. Therefore, it is imperative to utilize subsampling methods to reduce the time complexity. Let $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$ be a subsampling distribution so that $\sum_{i=1}^n \pi_i = 1$ and $\pi_i > 0$ for all i's. Suppose we draw a subsample of size r with replacement through $\boldsymbol{\pi}$. The subsample is denoted by $\mathcal{D}_r^* = \left\{Y_i^*, \delta_i^*, \mathbf{X}_i^*, \pi_i^*\right\}_{i=1}^r$, where Y_i^*, δ_i^*, $\mathbf{X}_i^*$, and π_i^* are the responses, censoring indicators, covariates, and subsampling probabilities (SSPs) of the subsample, respectively. Define $e_i^*(\boldsymbol{\beta}) = Y_i^* - (\mathbf{X}_i^*)^\top\boldsymbol{\beta}$. With the subsample $\mathcal{D}_r^*$, the smoothed estimating function of the subsample under Gehan's weight takes the form

$$\tilde{\mathbf{U}}_G^*(\mathcal{D}_r^*, \boldsymbol{\beta}) = \frac{1}{r}\sum_{i=1}^{r}\frac{\delta_i^*}{rn\pi_i^*}\sum_{j=1}^{r}\frac{1}{n\pi_j^*}(\mathbf{X}_i^* - \mathbf{X}_j^*)\Phi[\kappa_{ij}^*(\boldsymbol{\beta})] \tag{5}$$

and $\kappa_{ij}^*(\boldsymbol{\beta}) = \sqrt{r}[e_j^*(\boldsymbol{\beta}) - e_i^*(\boldsymbol{\beta})]/r_{ij}^*$ with $r_{ij}^{*2} = (\mathbf{X}_i^* - \mathbf{X}_j^*)^\top(\mathbf{X}_i^* - \mathbf{X}_j^*)/r$. The slope matrix of (5) is

$$\mathbf{M}^*(\mathcal{D}_r^*, \boldsymbol{\beta}) = \frac{\partial \tilde{\mathbf{U}}_G^*(\boldsymbol{\beta}; F_r^*)}{\partial \boldsymbol{\beta}} = \frac{1}{r}\sum_{i=1}^{r}\frac{\delta_i^*}{rn\pi_i^*}\sum_{j=1}^{r}\frac{\phi[\kappa_{ij}^*(\boldsymbol{\beta})]}{n\pi_i^* r_{ij}^*}(\mathbf{X}_i^* - \mathbf{X}_j^*)^{\otimes 2}, \tag{6}$$

which plays an important role in estimating the variance and defining optimal SSPs.

Let ξ_r represent the number of iterations required to compute the subsample estimator. The time complexity of the subsample estimator $\tilde{\boldsymbol{\beta}}_r$ is $O(\xi_r r^2 p)$ when using given SSPs, which is much more computationally efficient than obtaining the full sample estimator $\hat{\boldsymbol{\beta}}_n$ when $r \ll n$. Nevertheless, the estimating efficiency of $\tilde{\boldsymbol{\beta}}_r$ heavily depends on SSPs.

3 | FEASIBLE OPTIMAL SUBSAMPLING

We consider two types of optimal SSPs based on criteria from optimal design of experiments[2]. The first type of SSP is based on the A-optimal criteria which seeks to minimize the trace of the asymptotic variance of the subsample estimator. Wang et al.[2] showed a general form to define the A-optimal SSPs. For the ith observation, the A-optimal SSP is proportional to the Euclidean norm of the full data slope matrix multiplied by the ith observation's contribution to the full data estimating function. For the rank-based semi-parametric AFT model, the A-optimal SSP for the ith observation takes the form

$$\frac{\left\|\mathbf{M}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\tilde{\mathbf{U}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|}{\sum_{i=1}^n \left\|\mathbf{M}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\tilde{\mathbf{U}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|}.$$

That is, the above A-optimal SSP is proportional to the observation's contribution to $\tilde{\mathbf{U}}_G(\hat{\boldsymbol{\beta}})$. Since $\tilde{\mathbf{U}}_{G,i}(\hat{\boldsymbol{\beta}}) = 0$ when $\delta_i = 0$, the formula above seemingly suggests that censored observations should have zero optimal SSPs which is not true. To reveal the contributions of censored observations to $\tilde{\mathbf{U}}_G(\hat{\boldsymbol{\beta}})$, we adopt the standard approach where the estimating function is expressed by a well-defined counting process. Tsiatis[17] used this approach to prove the asymptotic normality of the estimator derived from the linear rank test for censored data. The detailed derivation of the A-optimal SSPs is shown below.

Given $\boldsymbol{\beta}$, let $\hat{H}(\cdot)$ be the Nelson-Aalen-type estimator of the cumulative hazard function for $\{e_i(\boldsymbol{\beta})\}_{i=1}^n$, where

$$\hat{H}(t;\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \int_{-\infty}^{t} \frac{\mathrm{d}N_i(u;\boldsymbol{\beta})}{S^{(0)}(u;\boldsymbol{\beta})} = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i I\left\{e_i(\boldsymbol{\beta}) \le t\right\}}{S^{(0)}[e_i(\boldsymbol{\beta});\boldsymbol{\beta}]}.$$

By some algebraic manipulations[22], Equation (2) can be written as

$$\begin{aligned}\mathbf{U}_G(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n \int_{-\infty}^{\infty} \frac{1}{n}\sum_{j=1}^n I\{t \le e_i(\boldsymbol{\beta})\}(\mathbf{X}_i - \mathbf{X}_j)\mathrm{d}\hat{M}_i(t;\boldsymbol{\beta}) \\ &= \frac{1}{n}\sum_{i=1}^n \left\{ \frac{1}{n}\sum_{j=1}^n \delta_i(\mathbf{X}_i - \mathbf{X}_j) I\{e_i(\boldsymbol{\beta}) \le e_j(\boldsymbol{\beta})\} - \frac{1}{n}\sum_{j=1}^n \delta_j I\{e_j(\boldsymbol{\beta}) \le e_i(\boldsymbol{\beta})\}\left[\mathbf{X}_i - \bar{\mathbf{X}}[e_j(\boldsymbol{\beta});\boldsymbol{\beta}]\right] \right\},\end{aligned} \tag{7}$$

where $\hat{M}_i(t;\boldsymbol{\beta}) = N_i(t;\boldsymbol{\beta}) - \int_{-\infty}^{t} I\{u \le e_i(\boldsymbol{\beta})\}\mathrm{d}\hat{H}(u;\boldsymbol{\beta})$. The smoothed version of (7) takes the form of $n^{-1}\sum_{i=1}^n \tilde{\mathbf{V}}_{G,i}(\boldsymbol{\beta})$, where

$$\tilde{\mathbf{V}}_{G,i}(\boldsymbol{\beta}) = \frac{1}{n}\left\{ \delta_i \sum_{j=1}^n \left(\mathbf{X}_i - \mathbf{X}_j\right)\Phi[\kappa_{ij}(\boldsymbol{\beta})] - \sum_{j=1}^n \delta_j \Phi[\kappa_{ji}(\boldsymbol{\beta})]\left[\mathbf{X}_i - \frac{\sum_{k=1}^n \mathbf{X}_k \Phi[\kappa_{jk}(\boldsymbol{\beta})]}{\sum_{k=1}^n \Phi[\kappa_{jk}(\boldsymbol{\beta})]}\right] \right\},$$

and it can be shown that $\tilde{\mathbf{U}}_G(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{U}}_{G,i}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{V}}_{G,i}(\boldsymbol{\beta})$. Similarly, $\tilde{\mathbf{U}}_G^*(\mathcal{D}_r^*, \boldsymbol{\beta})$ in (5) can be written as

$$\tilde{\mathbf{U}}_G^*(\mathcal{D}_r^*, \boldsymbol{\beta}) = \frac{1}{r}\sum_{i=1}^r \frac{1}{rn\pi_i^*}\tilde{\mathbf{V}}_{G,i}^*(\mathcal{D}_r^*, \boldsymbol{\beta}), \tag{8}$$

where

$$\tilde{\mathbf{V}}_{G,i}^*(\mathcal{D}_r^*, \boldsymbol{\beta}) = \sum_{j=1}^r \frac{\delta_i^*}{n\pi_j^*}\left(\mathbf{X}_i^* - \mathbf{X}_j^*\right)\Phi[\kappa_{ij}^*(\boldsymbol{\beta})] - \sum_{j=1}^r \frac{\delta_j^*}{n\pi_j^*}\Phi[\kappa_{ji}^*(\boldsymbol{\beta})]\left[\mathbf{X}_i^* - \frac{\sum_{k=1}^r (\pi_k^*)^{-1}\mathbf{X}_k^*\Phi[\kappa_{jk}^*(\boldsymbol{\beta})]}{\sum_{k=1}^r (\pi_k^*)^{-1}\Phi[\kappa_{jk}^*(\boldsymbol{\beta})]}\right]. \tag{9}$$

Note that the expression in Equation (8) helps us find an appropriate estimator of the variance matrix of the subsample estimator. We will still use Equation (5) to calculate the subsample point estimate.

Since $\tilde{\mathbf{V}}_{G,i}(\boldsymbol{\beta}) \neq 0$ for all observations, we should use it to define the contribution of the ith observation to $\tilde{\mathbf{U}}_G(\boldsymbol{\beta})$ and the optimal SSP of the ith observation. The A-optimal SSPs, denoted by $\boldsymbol{\pi}^{\text{optA}} = \{\pi_i^{\text{optA}}\}_{i=1}^n$, takes the form

$$\pi_i^{\text{optA}} = \frac{\left\|\mathbf{M}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|}{\sum_{i=1}^n \left\|\mathbf{M}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|}.$$

The A-optimal SSP $\boldsymbol{\pi}^{\text{optA}}$ dependents on $\hat{\boldsymbol{\beta}}_n$ which is not feasible in practice. To resolve this issue, we used $\tilde{\boldsymbol{\beta}}_{r_0}$, a subsample estimator derived from a small pilot sample $\mathcal{D}^*_{r_0}$ of size r_0 where $r_0 \ll n$, to replace $\hat{\boldsymbol{\beta}}_n$. The pilot sample is derived by sampling with replacement through uniform SSPs. The time complexity of calculating $\tilde{\boldsymbol{\beta}}_{r_0}$ is $O(\xi_{r_0} r_0^2 p)$ with ξ_{r_0} being the iteration for convergence. The slope matrix $\mathbf{M}_n(\hat{\boldsymbol{\beta}}_n)$ is approximated by $\mathbf{M}^*(\mathcal{D}^*_{r_0}, \tilde{\boldsymbol{\beta}}_{r_0})$ with a time complexity of $O(r_0^2 p^2)$. The time complexity to calculate the inverse of $\mathbf{M}^*(\mathcal{D}^*_{r_0}, \tilde{\boldsymbol{\beta}}_{r_0})$ is $O(p^3)$. Instead of using the full data to calculate $\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)$, we use the pilot subsample to approximate it by

$$\frac{1}{r_0}\sum_{j=1}^{r_0} \delta_i \left(\mathbf{X}_i - \mathbf{X}^*_j\right) \Phi[\kappa^{**}_{ij}(\tilde{\boldsymbol{\beta}}_{r_0})] - \frac{1}{r_0}\sum_{j=1}^{r_0} \delta^*_j \Phi[\kappa^{**}_{ji}(\tilde{\boldsymbol{\beta}}_{r_0})] \left[\mathbf{X}_i - \frac{\sum_{k=1}^{r_0} \mathbf{X}^*_k \Phi[\kappa^*_{jk}(\tilde{\boldsymbol{\beta}}_{r_0})]}{\sum_{k=1}^{r_0} \Phi[\kappa^*_{jk}(\tilde{\boldsymbol{\beta}}_{r_0})]}\right],$$

where $\kappa^{**}_{ij} = \sqrt{n}[e^*_j(\boldsymbol{\beta}) - e_i(\boldsymbol{\beta})]/\sqrt{(\mathbf{X}_i - \mathbf{X}^*_j)^\top(\mathbf{X}_i - \mathbf{X}^*_j)}$. The above formula is equivalent to the evaluation of (9), considering $\mathcal{D}^*_{r_0}$ and $\tilde{\boldsymbol{\beta}}_{r_0}$, while substituting $\mathbf{X}^*_i$ and Y^*_i for $\mathbf{X}_i$ and Y_i. We need to calculate $\sum_{j=1}^{r_0} \Phi[\kappa^{**}_{ij}(\tilde{\boldsymbol{\beta}}_{r_0})]$ and $\sum_{j=1}^{r_0} \mathbf{X}_j \Phi[\kappa^{**}_{ij}(\tilde{\boldsymbol{\beta}}_{r_0})]$ which both take $O(r_0 p)$ time to approximate $\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)$. The matrix multiplication between a $p \times p$ matrix and a $p \times 1$ vector takes $O(p^2)$ time. Calculating the norm of a $p \times 1$ vector takes $O(p)$ time. The overall time complexity to approximate $\left\|\mathbf{M}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|$ with given $\tilde{\boldsymbol{\beta}}_{r_0}$ and $\mathbf{M}_n^{-1}(\hat{\boldsymbol{\beta}}_n)$ is $O(r_0 p + p^2 + p) = O(r_0 p)$. Since we have n observations, approximating $\{\pi_i^{\text{optA}}\}_{i=1}^n$ takes $O(n r_0 p + \xi_{r_0} r_0^2 p + r_0^2 p^2 + p^3) = O(n r_0 p + \xi_{r_0} r_0^2 p)$ time.

To avoid approximating $\mathbf{M}_n$ and reduce the computing time, the second optimal SSPs are based on the L-optimal criteria, which is denoted by $\boldsymbol{\pi}^{\text{optL}} = \{\pi_i^{\text{optL}}\}_{i=1}^n$, with

$$\pi_i^{\text{optL}} = \frac{\left\|\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|}{\sum_{i=1}^n \left\|\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|}.$$

Since $\hat{\boldsymbol{\beta}}_n$ is not feasible in practice, it needs to be substituted with the pilot estimator $\tilde{\boldsymbol{\beta}}_{r_0}$ which takes $O(\xi_{r_0} r_0^2 p)$ time to derive. It takes $O(r_0 p + p) = O(r_0 p)$ time to approximate $\left\|\tilde{\mathbf{V}}_{G,i}(\hat{\boldsymbol{\beta}}_n)\right\|$. The overall time complexity to approximate $\{\pi_i^{\text{optL}}\}_{i=1}^n$ is $O(n r_0 p + \xi_{r_0} r_0^2 p)$.

Since the approximated SSPs are derived by a random pilot subsample, there might exist additional disturbance. For instance, the approximated SSP of a censored observation i will be zero if $e_i(\tilde{\boldsymbol{\beta}}_{r_0})$ is smaller than $e^*_j(\tilde{\boldsymbol{\beta}}_{r_0})$ for all j in the pilot subsample. Furthermore, the variance of the subsample estimator could be inflated by observations whose approximated optimal SSPs are close to zero[2]. To resolve these issues, we adopt the idea of defensive sampling[26,27]. That is, the practically used adjusted optimal SSPs, denoted by $\boldsymbol{\pi}^{\text{opt}}_\alpha(\tilde{\boldsymbol{\beta}}_{r_0}) = \left\{\pi^{\text{opt}}_{\alpha i}(\tilde{\boldsymbol{\beta}}_{r_0})\right\}_{r=1}^n$, is a weighted average of the approximated optimal SSPs, denoted by $\boldsymbol{\pi}^{\text{opt}}(\tilde{\boldsymbol{\beta}}_{r_0}) = \left\{\pi_i^{\text{opt}}(\tilde{\boldsymbol{\beta}}_{r_0})\right\}_{i=1}^n$, and the uniform SSP, with α controlling the weight of the uniform SSP. The adjusted optimal SSPs take the following form:

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

where $0 < \alpha < 1$. This adjustment aims to prevent $\boldsymbol{\pi}^{\text{opt}}_\alpha(\tilde{\boldsymbol{\beta}}_{r_0})$ from being too close to zero, which can occur in practice and result in excessively high inverse probability weights. A smaller α results in less distortion of the $\boldsymbol{\pi}^{\text{opt}}_\alpha(\tilde{\boldsymbol{\beta}}_{r_0})$ but increases the risk of explosive weights. We chose $\alpha = 0.2$ in the simulation and real data analysis as a conservative value to ensure that the adjusted inverse probability weights remain within a reasonable range, which already led to better results relative to those from the uniform SSP. We by no means want to indicate that $\alpha = 0.2$ is optimal. It is possible that a smaller α yields better results.

To explore the types of observations favored by optimal SSPs, we used simulated datasets of size 100,000 to calculate adjusted optimal SSPs. The covariates in the simulated datasets follow a multivariate t-distribution with 3 degrees of freedom. Detailed information about the simulated datasets is provided in Section 4. Nine configurations involving three censoring rates and three error distributions were considered. For each of the nine configurations, we generated 1000 different datasets and we calculated adjusted A-optimal SSPs based on the Weibull parametric AFT model, the semi-parametric AFT model by the rank-based approach and the least-squares approach. The pilot samples are different for different datasets. Table 1 displays the average means and average sums of the adjusted A-optimal SSPs for both censored and uncensored observations over 1000 datasets for each configuration. The table indicates that the least-squares approach has less preference for uncensored observations compared to the rank-based approach. These differences in preference for uncensored observations are more significant at higher censoring rates. For the Weibull parametric AFT model, the performance of A-optimal SSPs aligns closely with that of the semi-parametric AFT model by the rank-based approach.

Based on the adjusted optimal SSPs derived in the first step, a subsample of size r, denoted by $\mathcal{D}^*_r$, is selected by sampling with replacement in the second step. The second-step subsample estimator denoted as $\tilde{\boldsymbol{\beta}}_r$ is derived by solving (5). The information from the pilot sample should not be wasted. We make use of it by borrowing insights from the aggregation step in the divide-and-conquer strategy[28] and the online updating approach[29]. The aggregated estimator $\check{\boldsymbol{\beta}}_r$ is derived by combining $\tilde{\boldsymbol{\beta}}_{r_0}$ and $\tilde{\boldsymbol{\beta}}_r$ through a linear combination, where

$$\check{\boldsymbol{\beta}}_r = (r + r_0)\mathbf{M}^{*-1}_{r,r_0} \left\{ r_0\mathbf{M}^*(\mathcal{D}^*_{r_0}, \tilde{\boldsymbol{\beta}}_{r_0})\tilde{\boldsymbol{\beta}}_{r_0} + r\mathbf{M}^*(\mathcal{D}^*_r, \tilde{\boldsymbol{\beta}}_r)\tilde{\boldsymbol{\beta}}_r \right\},$$

and $\mathbf{M}^*_{r,r_0} = [r_0\mathbf{M}^*(\mathcal{D}^*_{r_0}, \tilde{\boldsymbol{\beta}}_{r_0}) + r\mathbf{M}^*(\mathcal{D}^*_r, \tilde{\boldsymbol{\beta}}_r)]/(r + r_0)$. In contrast to the optimal subsampling procedure employed in Yang et al.[9],[20] where the pilot subsample and the second-step subsample are combined to obtain the final estimator, aggregating $\tilde{\boldsymbol{\beta}}_{r_0}$ and $\tilde{\boldsymbol{\beta}}_r$ is less time-consuming since this procedure avoids using the pilot subsample twice. Since the final estimator is aggregated by the pilot and second-step estimators and we aim for the second-step estimator to play a dominant role, we favor a significantly smaller pilot sample size r_0 compared to the second-step subsample size r. The pilot subsample should not be too small either. A sufficient amount of data is necessary to derive good estimates of the optimal subsampling probabilities. In our simulation study, we selected $r_0 = 500$. In practical applications, users are advised to select larger pilot samples when dealing with higher censoring rates to obtain more accurate estimates of optimal subsampling probabilities.

Most existing subsampling studies focus on using $\check{\boldsymbol{\beta}}_r$ to approximate $\hat{\boldsymbol{\beta}}_n$. The asymptotic variance matrix of the approximation error $\check{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_n$ given the full data can be estimated by

$$\frac{1}{r + r_0}\mathbf{M}^{*-1}_{r,r_0}\check{\Lambda}_{r,r_0}\mathbf{M}^{*-1}_{r,r_0}, \tag{10}$$

where

$$\check{\Lambda}_{r,r_0} = \frac{1}{(r_0 + r)^3}\sum_{i=1}^{r_0+r}\frac{1}{(n\pi^*_i)^2}\tilde{\mathbf{V}}^{*\otimes 2}_{G,i}(D^*_{r,r_0}, \check{\boldsymbol{\beta}}_r),$$

is a moment estimator of $n^{-2}\sum_{i=1}^n \tilde{\mathbf{V}}^{\otimes 2}_{G,i}(\boldsymbol{\beta})/\pi_i$ and π^*_i is the corresponding SSP of the ith observation in the combined subsample which is denoted by D^*_{r,r_0}. The formula in (10) does not take into account the variation of the full data estimator $\hat{\boldsymbol{\beta}}_n$, so it is not appropriate to use it for inference on the true regression coefficient $\boldsymbol{\beta}_0$. In this scenario, we proposed an estimator for the asymptotic variance of $\check{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_0$:

$$\frac{1}{r + r_0}\mathbf{M}^{*-1}_{r,r_0}\left(\frac{r + r_0}{n}\tilde{\Lambda}_r + \check{\Lambda}_{r,r_0}\right)\mathbf{M}^{*-1}_{r,r_0}, \tag{11}$$

where

$$\tilde{\Lambda}_r = \frac{1}{(r_0 + r)^3}\sum_{i=1}^{r_0+r}\frac{1}{n\pi^*_i}\tilde{\mathbf{V}}^{*\otimes 2}_{G,i}(D^*_{r,r_0}, \check{\boldsymbol{\beta}}_r),$$

and $\tilde{\Lambda}_r$ is the estimator of $n^{-1}\sum_{i=1}^n \tilde{\mathbf{V}}^{\otimes 2}_{G,i}(\boldsymbol{\beta})$ based on the combined subsample. Equation (11) is constructed by adding (10) with the estimated asymptotic variance of $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$ which is $n^{-1}\mathbf{M}^{*-1}_{r,r_0}\tilde{\Lambda}_r\mathbf{M}^{*-1}_{r,r_0}$. Note that when the subsampling ratio $(r + r_0)/n$ is not close to zero, the impact of $\hat{\boldsymbol{\beta}}_n$ to (11) becomes substantial.

Now we consider the time complexity of the two-step procedure. As mentioned in Section 3, it takes $O(nr_0p + \xi_{r_0}r_0^2p)$ time to derive the optimal SSPs in step one. Calculating the second step subsample estimator costs $O\{\xi_r r^2 p\}$ time. Evaluating $\mathbf{M}^{*-1}_{r,r_0}$ takes $O\{r^2p^2\}$ time and calculating $\tilde{\Lambda}_r$ and $\check{\Lambda}_r$ both take $O\{(r + r_0)^2 p\}$ time. The overall time complexity of the two-step procedure is $O\{nr_0p + \xi_r r^2 p + (r + r_0)^2 p + r^2p^2\}$.

4 | SIMULATION STUDY

The performances of the two-step procedure were evaluated through a simulation study. In this investigation, we employed three distinct error distributions, the standard normal distribution, the standard logistic distribution, and the centered Gumbel distribution with a shape parameter of zero and a scale parameter of one. The covariates followed a multivariate normal distribution with a mean of zero and a covariance matrix denoted by $\boldsymbol{\Sigma}_{ij} = 0.5^{I(i \neq j)}$. Additionally, we incorporated a multivariate t distribution with 3 degrees of freedom and the same covariance matrix as the multivariate normal distribution. The dimension of covariates was seven, and the true coefficients, including the intercept, were set to values of ones. To emulate censoring in our study, we generated censoring times from a Uniform distribution, with the minimum and maximum values set at 0 and c respectively. The value of c was tuned to achieve three levels of censoring rates $c_r \in \{0.25, 0.50, 0.95\}$.

The simulation design led to eighteen configurations, each involving the generation of 1000 large datasets with the sample size of $n = 10,000$. It is worth noticing that the rank-based approach requires less subsample size to get a converging estimator compared to the least-squares approach. This arises from the non-smooth nature of the least-squares approach's estimating function, which is harder to solve than the smoothed estimating function of the rank-based approach. In analyzing each dataset, we used a pilot sample size of $r_0 = 500$ and explored different second-step subsample sizes of $r \in \{1000, 2000, 4000\}$. Three SSP schemes were applied: $\boldsymbol{\pi}^{\mathrm{optA}}$, $\boldsymbol{\pi}^{\mathrm{optL}}$, and the uniform SSPs. To assess and compare the performance of the two-step procedure across different SSPs, we calculated the root mean square error (RMSE) from $s = 1000$ estimators:

$$\mathrm{RMSE} = \left(\frac{1}{s} \sum_{i=1}^{s} \|\check{\boldsymbol{\beta}}_r^{(i)} - \boldsymbol{\beta}_0\|^2 \right)^{1/2},$$

where $\check{\boldsymbol{\beta}}_r^{(i)}$ is the estimate from the ith replicate. We omitted some simulation results when covariates followed the multivariate t distribution with 3 degrees of freedom as they showed similar patterns to those observed when covariates followed the multivariate normal distribution.

The RMSEs of the final estimator under three SSP schemes are compared in Figure 1. Across all the configurations, both $\boldsymbol{\pi}^{\mathrm{optL}}$ and $\boldsymbol{\pi}^{\mathrm{optA}}$ exhibit lower RMSEs compared to uniform SSPs. The A-optimal SSPs, $\boldsymbol{\pi}^{\mathrm{optA}}$, yielded the smallest RMSE, which is as expected since the A-optimality minimizes the summand of asymptotic variances. As the censoring rate increases, the number of informative observations decreases, resulting in higher RMSE values for all methods due to a reduction in information. At the 0.95 censoring rate, the advantage of optimal subsampling methods in terms of RMSE compared to the uniform subsampling method was more significant than at low censoring rates. Regardless of the configuration, the RMSE values decrease as the subsample size r increases. Note that for covariates with heavier tails, the optimal SSPs demonstrated a more pronounced advantage in terms of estimation at low censoring rates. This observation echoes the results obtained from optimal subsampling in the context of the quantile regression model[30], which could be seen as the extreme case of our model when the censoring rate is 0.

Figure 2 presents the results of the variance estimator given by equation (10) and (11) when the covariates followed a multivariate normal distribution. To illustrate the accuracy, we calculated the average of the square root of the trace for the estimated variance matrix over 1000 replicates and compared it with the empirical RMSE based on $\boldsymbol{\pi}^{\mathrm{optA}}$. They demonstrated close agreement across all six settings for 0.25 and 0.50 censoring rates, indicating that the formula in (11) fixed the underestimating issue and offers a reliable estimate of the variance. For the 0.95 censoring rate, the underestimating issue persisted when $r = 1000$ but gradually diminished as r increased to 4000. This could be due to the limited informative observations with a small subsample at a high censoring rate.

To further evaluate the performance of the proposed method in statistical inference, we considered the coverage probabilities of confidence intervals using the estimated variance matrix in (11). Figure 3 shows the average coverage probabilities for regression coefficients at different subsample sizes, censoring rates, and error distributions when covariates followed the multivariate normal distribution. The confidence interval for each regression coefficient was calculated by $\check{\beta}_r^{(i)} \pm 1.96 \times se(\check{\beta}_r^{(i)})$, where $\check{\beta}_r^{(i)}$ is the ith element of $\check{\boldsymbol{\beta}}_r$ and $se(\check{\beta}_r^{(i)})$ is the corresponding standard error. The confidence interval worked well for all three error distributions in our consideration. The coverage rates for the 0.95 censoring rates when $r = 1000$ were lower than 0.95, because (11) underestimated the standard errors at very high censoring rates and low subsample sizes. This issue would disappear as the subsample size increases.

Finally, we evaluated the computational efficiency of the optimal subsampling methods. We performed the computation on a Mac Studio with 32GB memory and M2 Max chip. Figure 4 summarizes the average CPU time in seconds of the second-step procedure and the average number of iterations to derive the final estimator over 50 experiments for different error distributions, covariate distributions, censoring rates, and subsample sizes, when covariates followed the multivariate normal distribution. For the 0.25 and 0.50 censoring rates, the CPU time is mainly affected by the subsample size, rather than other factors. The CPU times for both the uniform subsampling method and optimal subsampling methods are similar. This is because solving the second-step estimator took a longer time than calculating the subsampling probabilities, given the full sample size of 10,000. Nevertheless, the optimal subsampling methods have a significantly higher computing efficiency than the uniform subsampling method at the 0.95 censoring rate. The lower plot of Figure 4 and Table 1 in Section 3 help to explain this observation. They show that the optimal subsampling methods had a higher preference for selecting uncensored observations at the 0.95 censoring rate, which makes deriving the second-step estimator require much fewer iterations. Table 2 shows the CPU time for obtaining full sample estimates under each configuration. Deriving the full sample estimator takes half the time for cases with censoring rates of 0.25 and 0.50 compared to a censoring rate of 0.95. This indicates the difficulty of solving the estimating function at

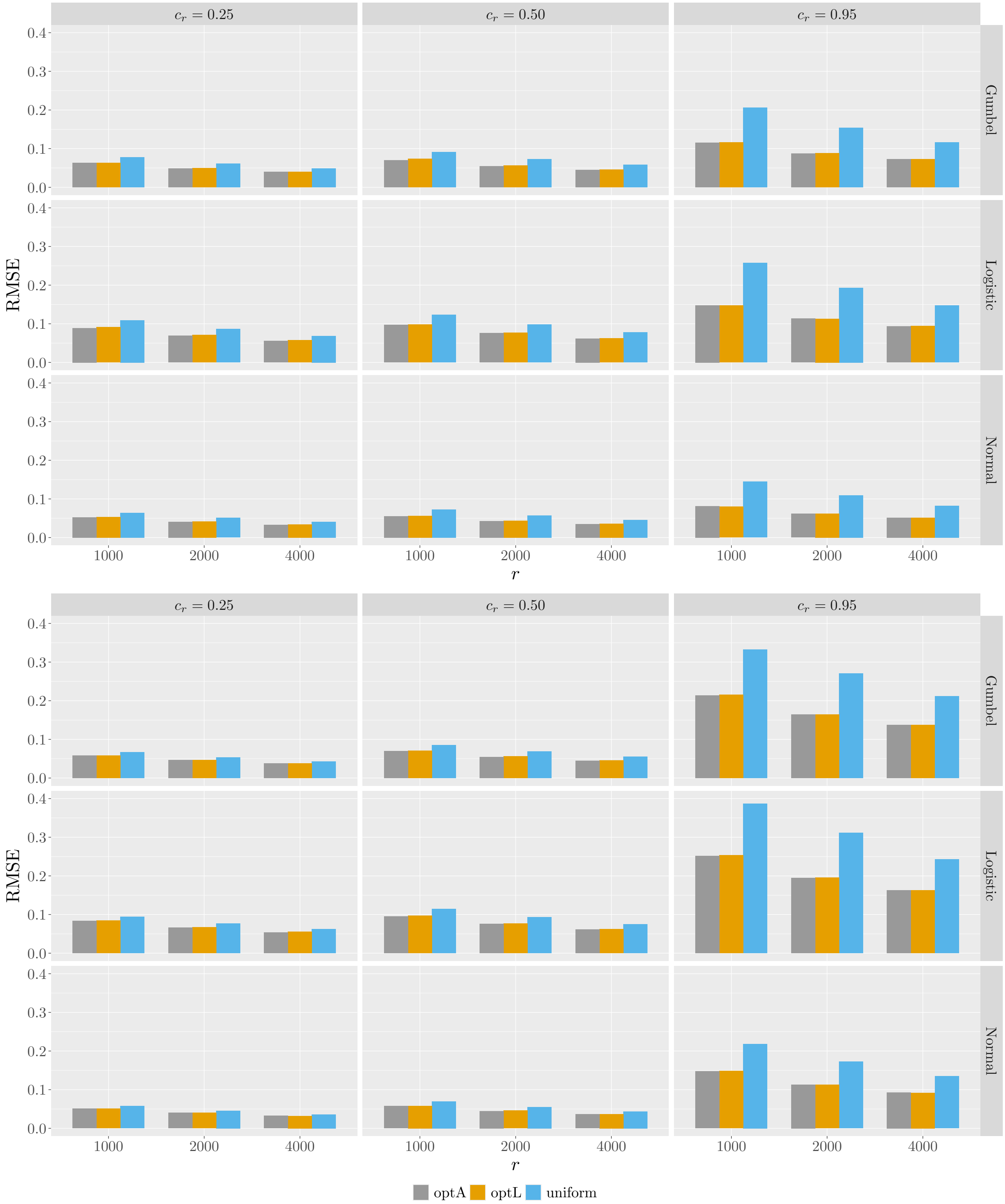

FIGURE 1 Empirical RMSEs for different SSPs, error distribution, subsample sizes r and censoring rates when covariates follow the multivariate t distribution with 3 degrees of freedom (upper) and the multivariate normal distribution (lower) based on the two-step procedure.

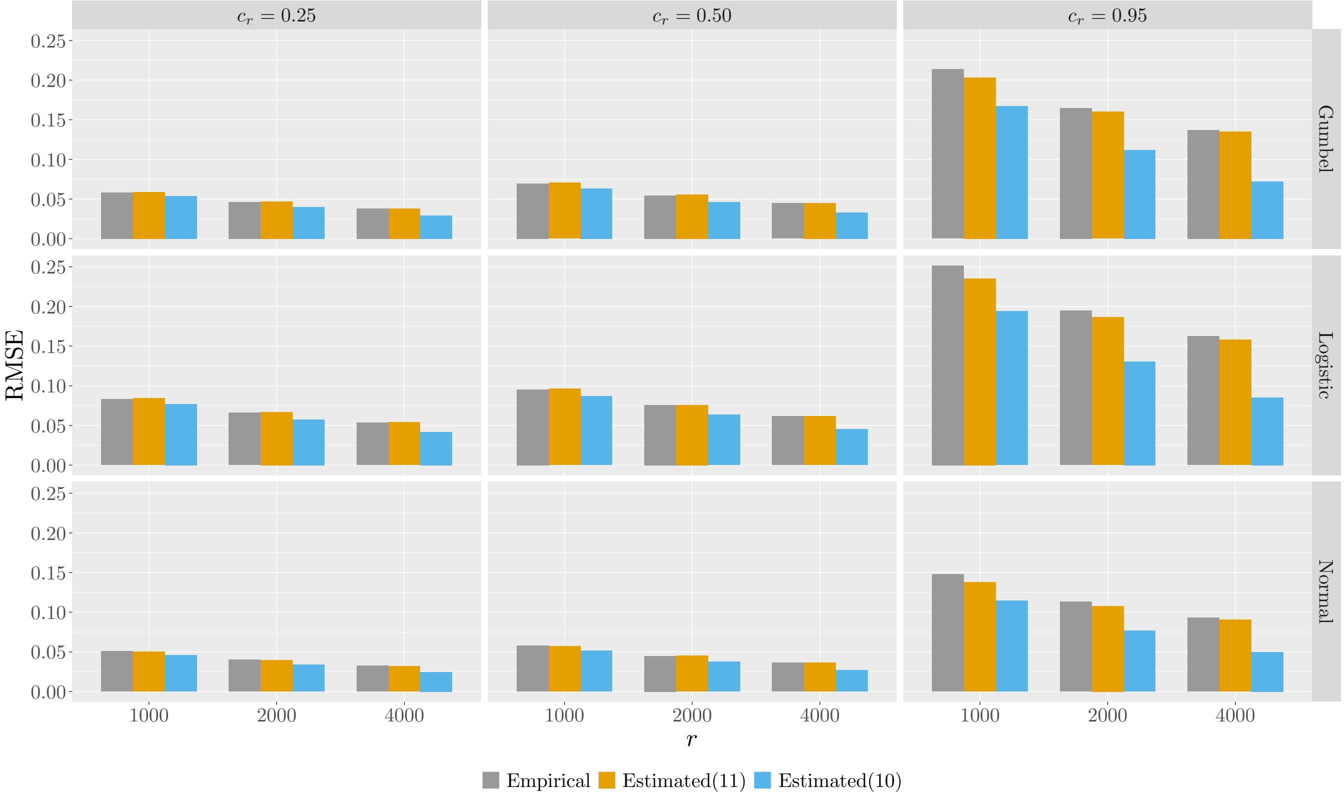

FIGURE 2 Comparison between the empirical RMSE and square roots of the trace for the estimated variance matrix calculated by formula (10) and (11) based on $\boldsymbol{\pi}^{\mathrm{optA}}$ for different error distribution, subsample sizes r and censoring rates when covariates follow the multivariate normal distribution using the two-step procedure.

high censoring rates. Compared with optimal subsampling methods, full sample estimates take a significantly longer time to calculate, which shows the advantage of optimal subsampling in terms of computational efficiency.

5 | SURVIVAL OF LYMPHOMA

We employed the subsampling procedure to model the survival time of patients diagnosed with lymphoma in the SEER program. The dataset contained information on 159,149 lymphoma patients diagnosed between 1973 and 2012, with a censoring rate of 58.3%. The event time was the survival time of lymphoma patients after being diagnosed with cancer. Four risk factors were considered, including age which was measured in years, nonwhite race indicator (1 = nonwhite), male indicator (1 = male), and the diagnostic year. Additionally, interactions between age with the male indicator, and age with the nonwhite indicator were included. The pilot sample size was set as $r_0 = 500$, and second-step subsample sizes were chosen from $r \in \{1000, 2000, 4000\}$. Three types of SSPs were used, the uniform SSPs, the L-optimal SSPs ($\boldsymbol{\pi}^{\mathrm{optL}}$), and the A-optimal SSPs ($\boldsymbol{\pi}^{\mathrm{optA}}$).

Figure 5 displays the RMSEs obtained from 1000 replicates under three subsample sizes and three SSP types. $\boldsymbol{\pi}^{\mathrm{optA}}$ and $\boldsymbol{\pi}^{\mathrm{optL}}$, as well as the uniform SSPs. It is observed that the RMSEs decrease as the subsample size r increases, indicating the consistency of the two-step procedure. As expected, both optimal SSPs exhibit higher estimation efficiency compared to the uniform SSPs. Nevertheless, for risk factors such as ‘Age’ and ‘Diagnostic Year’ and the interaction term ‘Age×Male’, the A-optimal subsampling method does not yield lower RMSEs compared to the L-optimal method. This is because $\boldsymbol{\pi}^{\mathrm{optA}}$ is designed to minimize overall RMSEs for all risk factors and interactions, rather than specifically targeting individual risk factors or interactions.

Table 3 summarizes the average estimates (EST) and their average empirical standard errors (ESE) and average estimated standard error (ASE) for all subsampling methods when $r = 4000$ over 1000 replicates. The estimated standard errors were

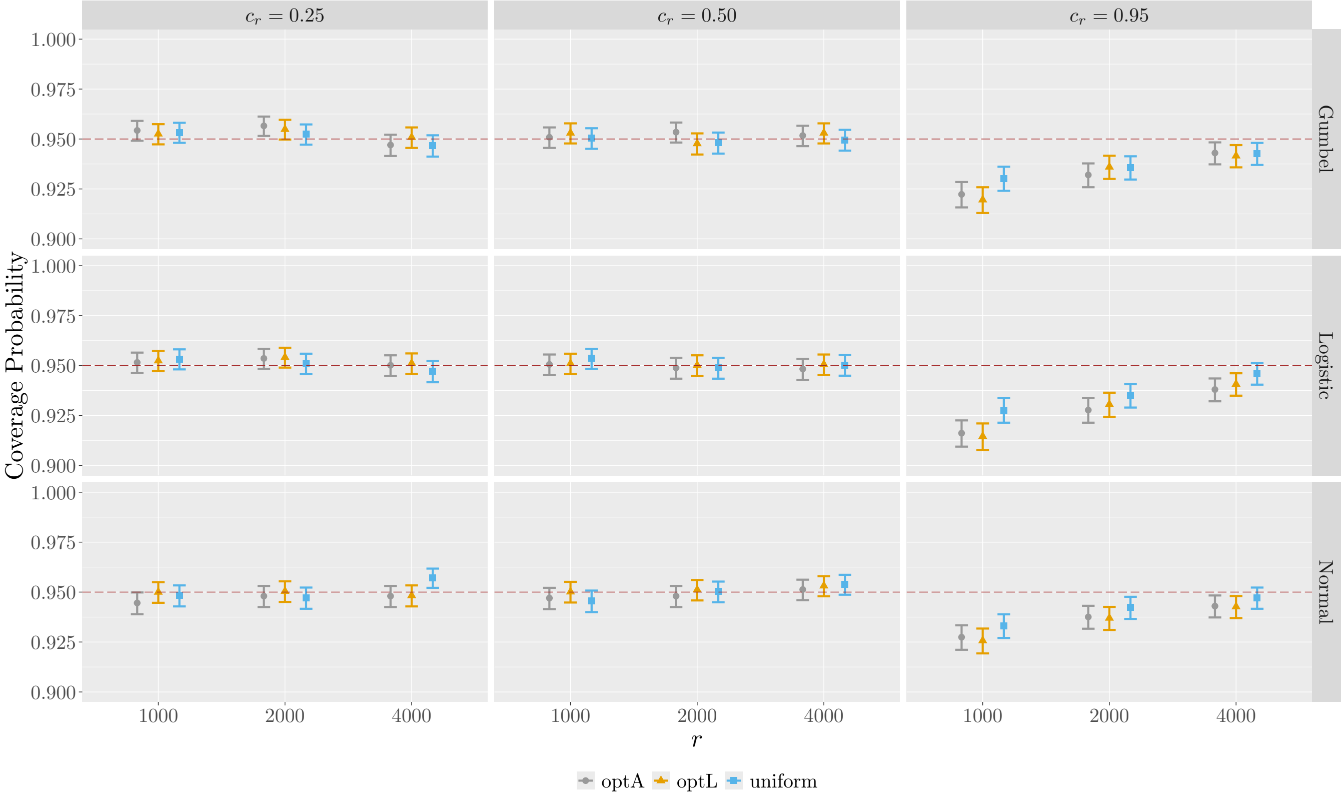

FIGURE 3 Empirical coverage probabilities with confidence intervals for different second step subsample size r, subsampling probabilities and error distributions when covariates follow the multivariate normal distribution.

calculated based on formula (10) since the 1000 replicates were carried out on a single full sample. We also included the full sample estimator in the table. The subsample estimators are close to the full sample estimator which shows that a small subsample is sufficient. The standard errors of the full data estimates are smaller than those of the subsample estimators. This is because the standard errors of the full data estimates are of order $O(n^{-1/2})$, while the standard errors of subsample estimators are of order $O\{(r+r_0)^{-1/2}\}$. Compared to the uniform subsampling method, the optimal subsampling methods yield a quarter smaller standard errors. The estimated and empirical standard errors are close, indicating that the variance estimator (10) is accurate. The results show that males and patients who were diagnosed later lived longer, while elder and nonwhite patients had less survival time. Moreover, the slope of age for white patients and male patients was steeper.

Table 4 presents the CPU times for using different subsampling probabilities and subsample sizes. The computations were done on a Mac Studio with 32 GB of memory and M2 Max CPU. The uniform subsampling method has the shortest computing time since it requires no additional calculations for subsampling probabilities. The computing time for the uniform subsampling method and optimal subsampling methods are getting closer as the subsample size increases. This is because deriving the second-step estimator dominates the computing time when the subsample size becomes large. Nevertheless, computing the full sample estimator requires 8.55 hours, with an additional 1.82 hours needed for the variance calculation on the same computer.

6 | DISCUSSION

The optimal subsampling method for the rank-based fitting of the semi-parametric AFT model for massive survival data has not been extensively studied. We expressed the estimating function by a well-defined stochastic process, which manifests non-zero optimal SSPs for censored observations. To overcome the numerical convergence issue when solving a non-smooth estimating function, we used the induced smoothing approach[8,21,23,24] to smooth the estimating function. For the variance estimation, we introduced a new sandwich estimator that accounts for the uncertainty of the full-data estimator, such that it can be used for

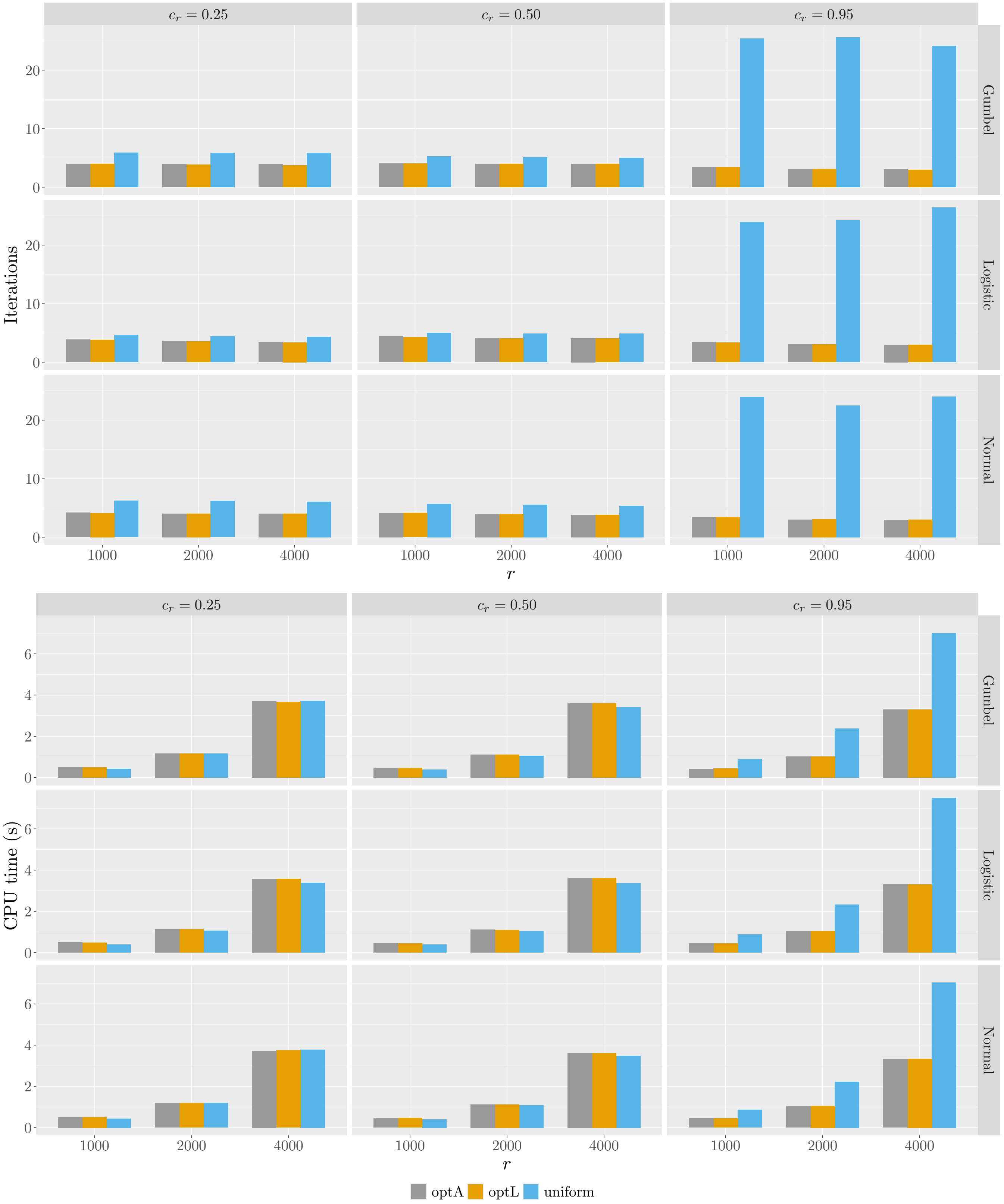

FIGURE 4 Average CPU time in seconds (upper) and average number of iterations to derive the second-step estimator (lower) obtained by different SSPs for different subsample sizes, error distributions and censoring rates when covariates follow the multivariate normal distribution over 50 experiments.

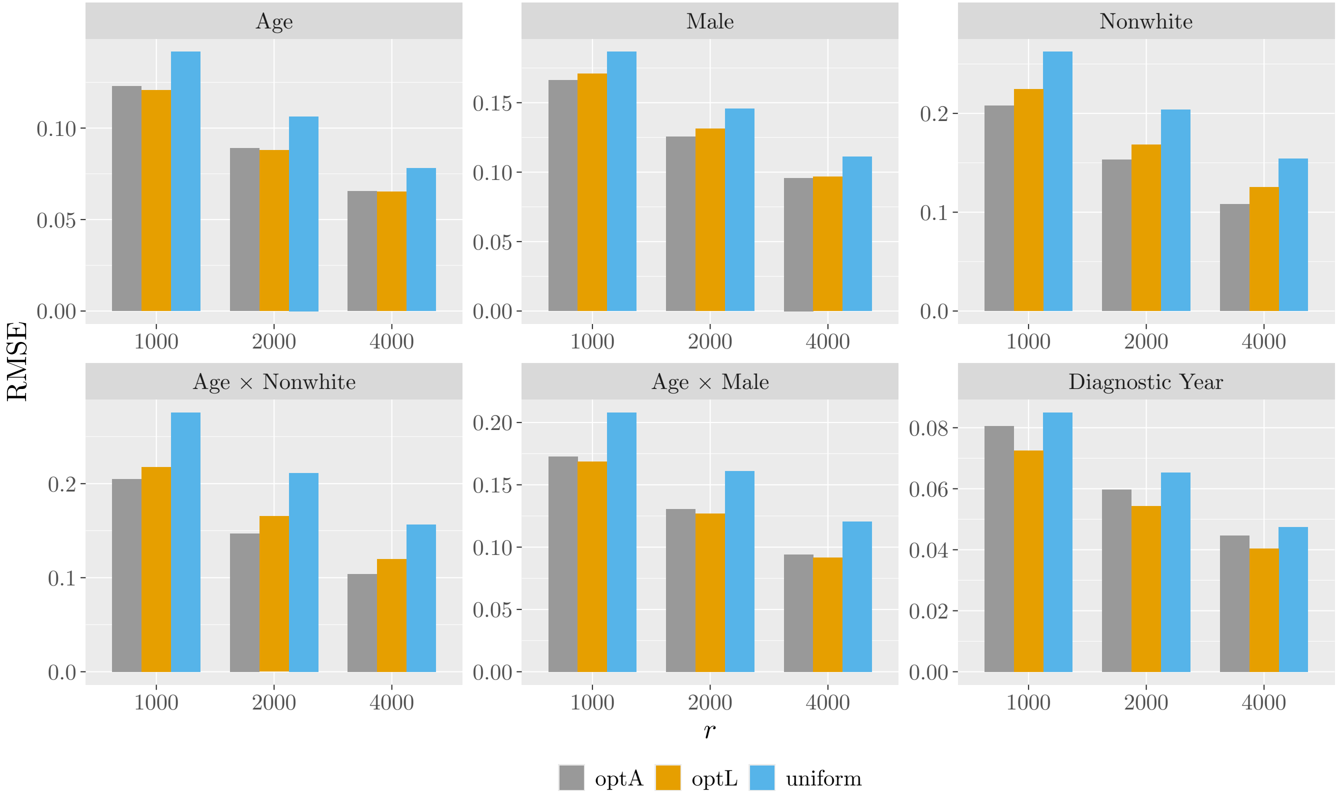

FIGURE 5 Empirical RMSEs of different risk factors for different SSPs and different second-step subsample sizes r when fixing the pilot sample size $r_0 = 500$ over 1000 replicates of the two-step procedure.

inferences about the true regression coefficients. This is in contrast to the estimator in most existing works where the inference target is the full-data estimator instead of the true parameters. The effectiveness of the proposed methods is validated in a comprehensive simulation study and a real data analysis, providing close approximations to the inferences obtained based on the full data with much more feasible computational resources.

Further investigation is in need for optimal subsampling methods with semi-parametric AFT models using Poisson sampling. Sampling without replacement avoids duplicate observations in the resulting subsample and may have a higher estimation efficiency when the subsampling ratio is high[31]. Nevertheless, with nonuniform subsampling probabilities, sampling without replacement becomes time-consuming due to the need to re-calculate subsampling probabilities after each selection. Recent literature on subsampling for big data focuses on sampling with replacement[1,5,10,30]. Poisson sampling can resolve both the problem of duplicate observations in sampling with replacement and the time-consuming issue of sampling without replacement[2]. This sampling approach considers each data point in one pass of the data and determines its inclusion in the subsample by generating a random number from a uniform distribution. Compared with sampling with replacement, Poisson sampling does not require calculating SSPs for all observations simultaneously. This means that the data can be read and processed line-by-line or chunk-by-chunk, which reduces the memory requirements and is more computationally efficient for big data. Unlike sampling with replacement, which allows for a predetermined subsample size, the subsample size from Poisson sampling is random. Wang et al.[2] show that Poisson sampling is more efficient than sampling with replacement for models with uncensored data. An optimal subsampling procedure via Poisson sampling for censored data is expected to be more efficient than that via sampling with replacement.

ACKNOWLEDGEMENT

We would like to thank the editor, associate editor and reviewers for taking the time and effort to review our manuscript. We sincerely appreciate all their valuable comments and suggestions, which have helped us improve its quality.

□

APPENDIX

COMPARISON WITH THE LEAST-SQUARES APPROACH

We dove into comparing the rank-based approach and the least-squares approach by a simulation study. The covariates distributions have two levels: the multivariate normal distribution and the multivariate t distribution with 3 degrees of freedom. The mean and variance matrix of the error distributions adhere to the configuration outlined in Section 4. Three levels of censoring rates are considered, 0.25, 0.5, and 0.95. The censoring distribution aligns with the specifications detailed in Section 4. For error distributions, we considered the standard normal distribution, standard logistic distribution, centered Gumbel distribution with shape parameter 0 and scale parameter 1, and the t distribution with 3 degrees of freedom. The four error distributions are ordered in terms of kurtosis, with the first distribution having the least kurtosis and the subsequent distributions exhibiting larger kurtosis.

The simulation design led to twenty-four configurations, each involving the generation of 1000 large datasets with the sample size of $n = 100,000$. For each configuration, we used a pilot sample size of $r_0 = 4000$ and explored different second-step subsample sizes of $r \in \{4000, 8000\}$. We chose large sizes of the pilot sample and second-step subsample to ensure the convergence of the least-squares approach. Two types of optimal SSPs and the uniform SSPs are considered. We collected the subsample estimator estimated by the rank-based approach and least-squares approach and compare the estimation efficiency via RMSE.

The RMSEs of the final estimator under the rank-based approach and the least-squares approach by different error distributions and censoring rates when covariates followed the multivariate normal distribution are shown in Figure A.1. The plot illustrates the superiority of the rank-based approach over the least-squares approach when errors follow the t distribution with 3 degrees of freedom. This aligns with empirical findings that the mean regression outperforms the quantile regression for heavy-tailed error distributions in non-censored scenarios[30]. The optimal subsampling methods for the least-squares approach perform stably at low censoring rates. Nevertheless, the estimates generated by the least-squares approach exhibit some outliers at the 0.95 censoring rate.

Figure A.2 illustrates the numerical stability of the rank-based approach compared to the least-squares approach. Notably, when the covariate distribution has heavier tails, the estimates generated by the least-squares approach exhibit instability, especially under optimal subsampling methods. This instability is shown by the boxplot of Euclidean errors, with 1.5% of the largest values trimmed for clarity. The boxplot reveals that there exist more outliers for the least-squares method in comparison to the rank-based approach. The prevalence of outliers in the least-squares approach can be attributed to the non-smooth nature of its estimating function which is hard to solve.

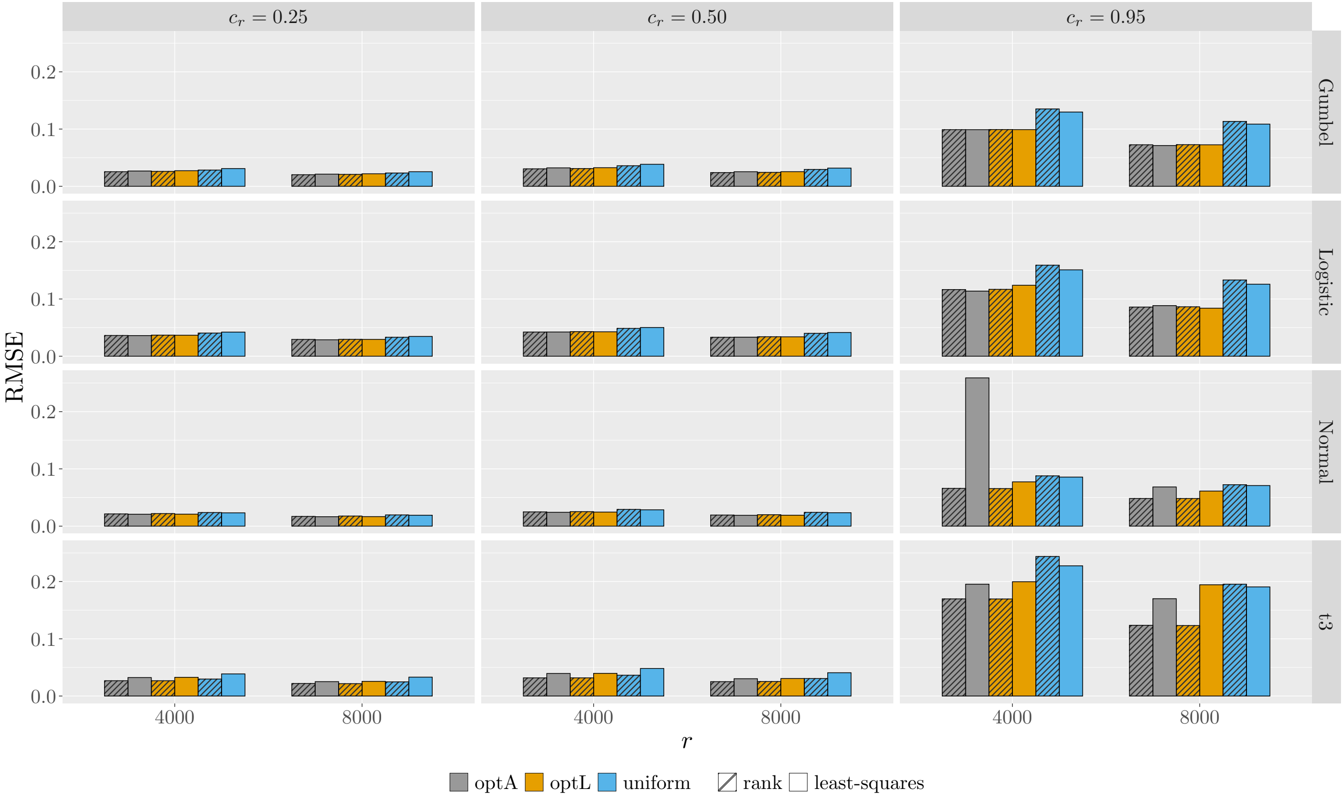

FIGURE A.1 Bar charts of RMSEs obtained from the least-squares approach and the rank-based approach when covariates follow the multivariate normal distribution and the error terms follow different error distributions over different censoring rates.

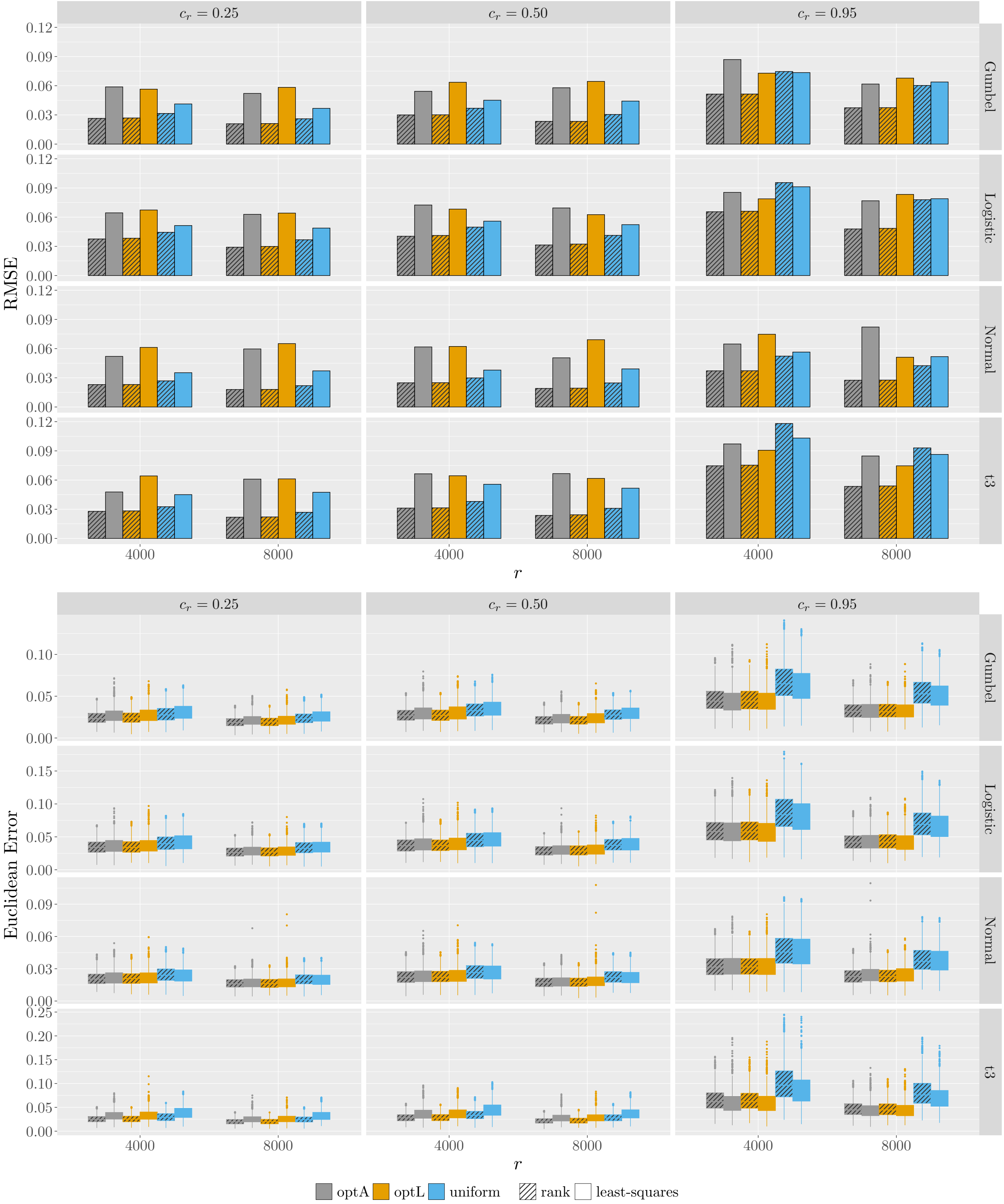

FIGURE A.2 Bar charts of RMSEs (Upper) and trimmed boxplot of Euclidean errors (lower) obtained from the least-squares approach and the rank-based approach when covariates follow multivariate *t* distribution with 3 degrees of freedom and the error terms follow different distributions over different censoring rates.

References

[1] Wang H, Zhu R, Ma P. Optimal Subsampling for Large Sample Logistic Regression. *Journal of the American Statistical Association* 2018; 113(522): 829–844.

[2] Wang J, Zou J, Wang H. Sampling with Replacement vs Poisson Sampling: A comparative Study in Optimal Subsampling. *IEEE Transactions on Information Theory* 2022; 68(10): 6605–6630.

[3] Zuo L, Zhang H, Wang H, Liu L. Sampling-based Estimation for Massive Survival data with Additive hazards model. *Statistics in Medicine* 2021; 40(2): 441–450.

[4] Zhang H, Zuo L, Wang H, Sun L. Approximating Partial Likelihood Estimators via Optimal Subsampling. *Journal of Computational and Graphical Statistics* 2023; 33(1): 276–288.

[5] Qiao N, Li W, Xiao F, Lin C, Zhou Y. Optimal Subsampling for the Cox Proportional Hazards Model with Massive Survival Data. *Journal of Statistical Planning and Inference* 2024; 231: 106136.

[6] Keret N, Gorfine M. Analyzing Big EHR Data—Optimal Cox Regression Subsampling Procedure with Rare Events. *Journal of the American Statistical Association* 2023; 118(544): 2262–2275.

[7] Wei LJ. The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis. *Statistics in medicine* 1992; 11(14-15): 1871–1879.

[8] Chiou SH, Kang S, Yan J. Fitting Accelerated Failure Time Models in Routine Survival Analysis with R Package aftgee. *Journal of Statistical Software* 2014; 61(11): 1–23.

[9] Yang Z, Wang H, Yan J. Optimal Subsampling for Parametric Accelerated Failure Time Models with Massive Survival Data. *Statistics in Medicine* 2022; 41(27): 5421–5431.

[10] Ai M, Yu J, Zhang H, Wang H. Optimal Subsampling Algorithms for Big Data Generalized Linear Models. *Statistica Sinica* 2021; 31(2): 749–772.

[11] Buckley J, James I. Linear Regression with Censored Data. *Biometrika* 1979; 66(3): 429–436.

[12] Jin Z, Lin D, Ying Z. On Least-squares Regression with Censored Data. *Biometrika* 2006; 93(1): 147–161.

[13] Lai TL, Ying Z. Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data. *The Annals of Statistics* 1991; 19(3): 1370–1402.

[14] Ritov Y. Estimation in a Linear Regression Model with Censored Data. *The Annals of Statistics* 1990; 18(1): 303–328.

[15] Jin Z, Lin D, Wei L, Ying Z. Rank-based Inference for the Accelerated Failure Time Model. *Biometrika* 2003; 90(2): 341–353.

[16] Prentice RL. Linear Rank Tests with Right Censored Data. *Biometrika* 1978; 65(1): 167–179.

[17] Tsiatis AA. Estimating Regression Parameters Using Linear Rank Tests for Censored Data. *The Annals of Statistics* 1990; 18(1): 354–372.

[18] Ying Z. A Large Sample Study of Rank Estimation for Censored Regression Data. *The Annals of Statistics* 1993; 21(1): 76 – 99.

[19] Lai TL, Ying Z. Rank Regression Methods for Left-truncated and Right-censored Data. *The Annals of Statistics* 1991; 19(2): 531–556.

[20] Yang Z, Wang H, Yan J. Subsampling Approach for Least Squares Fitting of Semi-parametric Accelerated Failure Time Models to Massive Survival Data. *Statistics and Computing* 2024; 34: 77.

[21] Brown BM, Wang YG. Induced Smoothing for Rank Regression with Censored Survival Times. *Statistics in Medicine* 2007; 26(4): 828–836.

[22] Lin D, Wei L, Ying Z. Accelerated Failure time Models for Counting Processes. *Biometrika* 1998; 85(3): 605–618.

[23] Brown BM, Wang YG. Standard Errors and Covariance Matrices for Smoothed Rank Estimators. *Biometrika* 2005; 92(1): 149–158.

[24] Chiou S, Kang S, Yan J. Rank-based Estimating Equations with General Weight for Accelerated Failure Time Models: An Induced Smoothing Approach. *Statistics in Medicine* 2015; 34(9): 1495–1510.

[25] Gehan EA. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples. *Biometrika* 1965; 52(1-2): 203–224.

[26] Hesterberg T. Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics* 1995; 37(2): 185–194.

[27] Owen A, Zhou Y. Safe and effective importance sampling. *Journal of the American Statistical Association* 2000; 95(449): 135–143.

[28] Lin N, Xi R. Aggregated Estimating Equation Estimation. *Statistics and Its Interface* 2011; 4(1): 73–83.

[29] Schifano ED, Wu J, Wang C, Yan J, Chen MH. Online Updating of Statistical Inference in the Big Data Setting. *Technometrics* 2016; 58(3): 393–403.

[30] Wang H, Ma Y. Optimal Subsampling for Quantile Regression in Big Data. *Biometrika* 2021; 108(1): 99–112.

[31] Basu D. On sampling with and without replacement. *Sankhyā: The Indian Journal of Statistics* 1958; 20(3/4): 287–294.

TABLE 1 Means and summations of uniform SSPs and adjusted A-optimal SSPs for censored and uncensored observations with Gumbel (G), Logistic (L) and Normal (N) distributions as the error distributions and different censoring rates c_r when covariates follow the multivariate t distribution with 3 degrees of freedom and using different AFT models.

Observation		c_r: 25%				c_r: 50%				c_r: 95%		
	uniform	G	L	N	uniform	G	L	N	uniform	G	L	N
Semiparametric AFT Model - Rank-based Approach												
summation												
Censored	0.250	0.121	0.149	0.112	0.500	0.226	0.261	0.207	0.950	0.349	0.379	0.305
Uncensored	0.750	0.879	0.851	0.888	0.500	0.774	0.739	0.793	0.050	0.651	0.621	0.695
mean ($\times n$)												
Censored	1.000	0.495	0.606	0.456	1.000	0.457	0.527	0.417	1.000	0.368	0.401	0.322
Uncensored	1.000	1.164	1.128	1.177	1.000	1.534	1.465	1.573	1.000	12.208	11.332	13.405
Semiparametric AFT Model - Least-squares Approach												
summation												
Censored	0.250	0.319	0.305	0.346	0.500	0.443	0.433	0.471	0.950	0.748	0.718	0.780
Uncensored	0.750	0.681	0.695	0.654	0.500	0.557	0.567	0.529	0.050	0.252	0.282	0.220
mean ($\times n$)												
Censored	1.000	1.299	1.244	1.410	1.000	0.895	0.874	0.950	1.000	0.790	0.759	0.823
Uncensored	1.000	0.903	0.921	0.866	1.000	1.103	1.124	1.049	1.000	4.724	5.157	4.234
Weibullc parametric AFT Model												
summation												
Censored	0.250	0.127	0.186	0.126	0.500	0.242	0.301	0.233	0.950	0.335	0.330	0.302
Uncensored	0.750	0.873	0.814	0.874	0.500	0.758	0.699	0.767	0.050	0.665	0.670	0.698
mean ($\times n$)												
Censored	1.000	0.515	0.757	0.513	1.000	0.488	0.607	0.470	1.000	0.354	0.349	0.318
Uncensored	1.000	1.158	1.079	1.158	1.000	1.502	1.386	1.521	1.000	12.460	12.226	13.466

TABLE 2 Average CPU time (s) obtained by full sample estimates for different censoring rates, error distributions when covariates follow the multivariate normal distribution over 10 different full samples for each setting.

	Gumbel	Normal	Logistic
0.25	17.98	16.43	18.47
0.50	16.57	16.23	16.57
0.95	33.66	32.63	32.61

TABLE 3 Estimates (EST) and their empirical standard errors (ESE) and average estimated standard errors (ASE) from different subsampling approaches for $r = 4000$ and r_0=500 over 1000 replicates. The standard errors (SE) of the full sample estimates are estimated by the sandwich form.

	uniform			optL			optA			Full	
	EST	ESE	ASE	EST	ESE	ASE	EST	ESE	ASE	EST	SE
Age	−1.075	0.078	0.081	−1.071	0.066	0.068	−1.070	0.065	0.067	−1.076	0.013
Male	0.724	0.111	0.110	0.720	0.096	0.097	0.723	0.097	0.096	0.724	0.018
Nonwhite	−0.711	0.154	0.151	−0.707	0.108	0.112	−0.711	0.125	0.128	−0.709	0.025
Age×Nonwhite	0.297	0.157	0.160	0.297	0.104	0.107	0.296	0.120	0.122	0.298	0.027
Age×Male	−0.517	0.120	0.121	−0.512	0.094	0.095	−0.515	0.092	0.094	−0.516	0.020
Diagnostic Year	0.517	0.047	0.049	0.514	0.045	0.046	0.514	0.040	0.042	0.515	0.008

TABLE 4 Average CPU time (s) obtained by different subsampling methods for different subsample sizes with $r_0 = 500$ over 50 experiments.

	r : 1000	r : 2000	r : 4000
optA	1.45	1.84	3.21
optL	1.43	1.83	3.25
uniform	0.43	0.87	2.27