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

DOI: xxx/xxxx

RESEARCH ARTICLE

Optimal subsampling for parametric accelerated failure time models with massive survival data

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

With increasing availability of massive survival data, researchers need valid statistical inferences for survival modeling whose computation is not limited by computer memories. Existing works focus on relative risk models using the online updating and divide-and-conquer strategies. The subsampling strategy has not been available due to challenges in developing the asymptotic properties of the estimator under semiparametric models with censored data. This paper tackles optimal subsampling algorithms to fast approximate the maximum likelihood estimator for parametric accelerate failure time (AFT) models with massive survival data. We derive the asymptotic distributions of the subsampling estimator and the optimal sampling probabilities that minimize the asymptotic mean squared error of the estimator. A feasible two-step algorithm is proposed where the optimal sampling probabilities in the second step are estimated based on a pilot sample in the first step. The asymptotic properties of the two-step estimator are established. The performance of the estimator is validated in a simulation study. A real data analysis illustrates the usefulness of the methods.

KEYWORDS:
A-optimality; censoring; L-optimality; survival analysis

1 | INTRODUCTION

Massive survival data, which are increasingly available with the rapid advancement of surveillance and storage technologies, call for novel methodologies for fitting regression models for them. The key challenge is that the data can be so big that they exceed the memory of even super computers, rendering traditional computational methods in fitting them infeasible. General strategies to address the challenge can be grouped into three categories[1]: divide and conquer approaches[2,3,4], online updating approaches[5,6], and subsampling approaches[7,8]. Specifically for big survival data, the divide-and-conquer strategy has been developed for the Cox model[4] and for frailty models with multivariate failure times[9]. The online updating strategy has been applied to testing the proportional hazards assumption[10] and fitting the Cox model[11]. Fewer works, nevertheless, have been available using the subsampling strategy..

The subsampling strategy is a straightforward and efficient approach to approximate the full data inferences by inferences based on a subsample where observations are appropriately weighted. Subsampling probabilities (SSPs) are constructed with certain statistical leverage score or variate of the data. SSPs of non-informative subsampling schemes are based on covariates instead of responses[7,8]. In contrase, more recent works use informative SSPs, which depend on both responses and covariates. The optimal SSPs are oftentimes dependent on the maximum likelihood estimator (MLE). Wang et al. [12], for example, proposed

an optimal subsampling algorithm for logistic regression based on the A-optimality which minimizes the trace of the variance matrix of the resultant estimator. This method has been extended to many statistical models such as generalized linear model [13] and quantile regression model [14]. For survival analysis, the asymptotic properties of subsample estimators under semiparametric models with censored data are difficult to access and they may no provide useful guidance to define optimal sampling probabilities. One exception is Zuo et al. [15] for the additive hazard model, where the sampling probabilities were derived in an ad hoc way.

The accelerate failure time (AFT) model is one of the most popular models for survival data, such as generalized Gamma AFT model [16]. No existing work has investigated the optimal subsampling algorithms for AFT models with massive data. Understanding the challenging issues from a semiparametric model, we consider a less ambitious problem and develop optimal subsampling algorithms for parametric AFT models. The parametric setting facilitates the derivation of the optimal SSPs. We establish the optimal SSPs based on the A-optimality and the L-optimality, which depend on the full data MLE. As the full data MLE is not available due to the computational barrier imposed by the large size of the data, we propose a two-step procedure where the optimal SSPs is estimated by a pilot subsample first and then the subsampling estimator based on a subsample selected by the optimal SSPs is obtained. The asymptotic normality of the two-step estimator are derived, and the asymptotic variance is estimated by the method of moments. The method is validated through extensive simulation studies and illustrated by a real data example. Our implementation in R is publicly available in GitHub repo: https://github.com/YEnthalpy/osmac-parametric-aft-models.

The rest of the paper is organized as follows. A general subsampling procedure with given SSPs is presented and the asymptotic properties of the resulting estimator are established for parametric AFT models in Section 2. The optimal SSPs are derived under two criteria motivated from experiment design in Section 3. As the optimal SSPs depend on unknown full-data MLE, a feasible two-step algorithm is proposed in Section 4, with the asymptotic properties of the resulting estimator established and an estimator of the asymptotic variance derived. The performance of the estimator is assessed in a simulation study in Section 5. The method is applied to analyzing the survival time of lymphoma patients in the Surveillance, Epidemiology, and End Results (SEER) program in Section 6. Section 7 concludes with a discussion. Proofs of the theoretical results are relegated to the Supplementary Material.

2 | SUBSAMPLING FOR PARAMETRIC AFT MODELING

For subject i, $i = 1, 2, ..., n$, let t_i, c_i, and $\mathbf{x}_i$ be the log-transformed failure time, the log-transformed censoring time, and a $p \times 1$ covariate vector, respectively. Given $\mathbf{x}_i$, assume that t_i is independent of c_i. A general form of Parametric AFT models is

$$t_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma \epsilon_i, \qquad i = 1, 2, ..., n, \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, σ is the scale parameter, and ϵ_i's are independent and identically distributed (i.i.d.) random variables with probability density function $f_\epsilon(x)$, cumulative distribution function $F_\epsilon(x)$, survival function $S_\epsilon(x) = 1 - F_\epsilon(x)$ and hazard function $h_\epsilon(x) = f_\epsilon(x)/S_\epsilon(x)$, $x \in \mathbb{R}$. Due to right censoring, the observed data are i.i.d. copies of $(y_i, \delta_i, \mathbf{x}_i)$, where $y_i = \min(t_i, c_i)$, $\delta_i = I(t_i < c_i)$, and $I(\cdot)$ is the indicator function. Denote the full data matrix as $\mathcal{F}_n = \{y_i, \delta_i, \mathbf{x}_i, i = 1, ..., n\}$.

The target of inferences is the parameters of parametric AFT models $\boldsymbol{\theta} = (\sigma, \boldsymbol{\beta}^\top)^\top$. The MLE is the maximizer of the log-likelihood function,

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}),$$

where

$$l_i(\boldsymbol{\theta}) = (1 - \delta_i) \log\{S_i(\boldsymbol{\theta})\} + \delta_i \log\{f_i(\boldsymbol{\theta})\},$$

$S_i(\boldsymbol{\theta}) = S_\epsilon\{e_i(\boldsymbol{\theta})\}$, $f_i(\boldsymbol{\theta}) = f_\epsilon\{e_i(\boldsymbol{\theta})\}/\sigma$, and $e_i(\boldsymbol{\theta}) = (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})/\sigma$. In the sequel, we denote the gradient and Hessian matrix of $l_i(\boldsymbol{\theta})$ as $\dot{l}_i(\boldsymbol{\theta})$ and $\ddot{l}_i(\boldsymbol{\theta})$, respectively. We use $\|\mathbf{B}\|$ for the Frobenius norm of a matrix or vector $\mathbf{B}$.

For massive data with large n, the subsampling strategy makes inferences about the MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}} = \left[\hat{\sigma}_{\text{MLE}}, \hat{\boldsymbol{\beta}}_{\text{MLE}}^\top\right]^\top$ based on an appropriately formed subsample of a much smaller size. Suppose that the SSPs are given as $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_n)$ for all observations. We first draw a subsample of size r with replacement denoted by $\{y_i^*, \delta_i^*, \mathbf{x}_i^*, \pi_i^*, i = 1, ..., r\}$, where y_i^*, δ_i^*, $\mathbf{x}_i^*$, and π_i^* are the responses, censoring indicators, covariates, and subsampling probabilities of the subsample, respectively. We obtain

the subsample estimator $\tilde{\theta}_r$ by maximizing the following target function

$$l^*(\theta) = \sum_{i=1}^{r} \frac{l_i^*(\theta)}{\pi_i^*},$$

where

$$l_i^*(\theta) = (1-\delta_i^*)\log\{S_i^*(\theta)\} + \delta_i^* \log\{f_i^*(\theta)\},$$

$S_i^*(\theta) = S_\epsilon\{e_i^*(\theta)\}$, $f_i^*(\theta) = f_\epsilon\{e_i^*(\theta)\}/\sigma$, and $e_i^*(\theta) = (y_i^* - \beta^\top \mathbf{x}_i^*)/\sigma$. In particular, the maximization can be approached by a block coordinate decent method, where β is treated as one block and σ as the other.

The following assumptions are needed to derive the asymptotic properties of $\tilde{\theta}_r$.

Assumption 1. The true value θ_0 of θ is an interior point of the compact parameter space Θ in which $\|\beta\| \le B < \infty$ and $0 < A \le \sigma$ for some positive constants A and B.

Assumption 2. As $n \to \infty$,

$$(i) \quad n^{-2} \sup_{\theta\in\Theta} \left\{ \sum_{i=1}^{n} \pi_i^{-1} \left\| \dot{l}_i(\theta) \right\|^2 \right\} = O_P(1),$$

$$(ii) \quad n^{-2} \sup_{\theta\in\Theta} \left\{ \sum_{i=1}^{n} \pi_i^{-1} \left\| \ddot{l}_i(\theta) \right\|^2 \right\} = O_P(1),$$

$$(iii) \quad n^{-1} \sup_{\theta\in\Theta} \left\{ \sum_{i=1}^{n} \left\| \frac{\partial^2 \dot{l}_i^{(k)}(\theta)}{\partial\theta\partial\theta^\top} \right\| \right\} = O_P(1),$$

$$(iv) \quad n^{-1} \sum_{i=1}^{n} \left\| \dot{l}_i(\hat{\theta}_{\text{MLE}}) \right\|^4 = O_P(1),$$

where $\dot{l}_i^{(k)}(\theta)$ is the kth component of $\dot{l}_i(\theta)$, $k = 1, 2, ..., p+1$. Specifically, conditions (i) and (ii) hold when $\pi_i = 1/n$.

Assumption 3. The maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ is unique and $\mathbf{M}_n = n^{-1}\sum_{i=1}^{n} \ddot{l}_i(\hat{\theta}_{\text{MLE}})$ goes to a negative definite matrix as $n \to \infty$.

Assumption 4. There exists some $\xi > 0$ such that

$$\frac{1}{n^{2+\xi}} \sum_{i=1}^{n} \frac{1}{\pi_i^{1+\xi}} \left\| \dot{l}_i(\hat{\theta}_{\text{MLE}}) \right\|^{2+\xi} = O_P(1).$$

Assumption 1 assures that the regression coefficients are finite and the scale σ is bounded away from zero. Assumption 2 puts moment conditions on the derivatives of the log-likelihood function to ensure the consistency and asymptotic normality of the estimator based on the subsample. Assumption 3 imposes a condition on the Hessian matrix. Assumption 4 is required by the Lindeberg–Feller central limit theorem.

The theorem below establishes the consistency and asymptotic normality of $\tilde{\theta}_r$.

Theorem 1. If Assumptions 1–4 hold, as $r \to \infty$, $n \to \infty$, $r/n \to 0$, given $\mathcal{F}_n$ in probability,

$$r^{1/2}\mathbf{\Gamma}_n^{-1/2}(\tilde{\theta}_r - \hat{\theta}_{\text{MLE}}) \to N(\mathbf{0}, \mathbf{I})$$

in distribution, where $\mathbf{\Gamma}_n = \mathbf{M}_n^{-1}\mathbf{V}_n\mathbf{M}_n^{-1} = O_P(1)$, and

$$\mathbf{V}_n = \frac{1}{n^2} \sum_{i=1}^{n} \frac{\dot{l}_i(\hat{\theta}_{\text{MLE}})\dot{l}_i^\top(\hat{\theta}_{\text{MLE}})}{\pi_i}.$$

The simplest subsampling method is to use the uniform SSPs where $\boldsymbol{\pi} = \{n^{-1}\}_{i=1}^n$ but it is far from optimal. Thus, we will consider a more efficient subsampling procedure which intend to "minimize" the asymptotic variance-covariance matrix $\mathbf{\Gamma}_n/r$ mentioned in Theorem 1.

3 | OPTIMAL SUBSAMPLING PROBABILITIES

In this section, we consider the idea of A-optimality from optimal design of experiment which seeks to minimize the trace of the asymptotic variance-covariance matrix $\mathbf{\Gamma}_n/r$. From Theorem 1, this is the same as minimizing the asymptotic MSE of the resultant estimator $\tilde{\boldsymbol{\theta}}_r$. The following theorem gives the specific expression of the A-optimal SSP.

Theorem 2. Given $\mathcal{F}_n$, the optimal SSP denoted as $\boldsymbol{\pi}^{\text{mMSE}}(\hat{\boldsymbol{\theta}}_{\text{MLE}}) = \{\pi_i^{\text{mMSE}}(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}_{i=1}^n$ which minimizes $\text{tr}(\mathbf{\Gamma}_n)$ satisfies

$$\pi_i^{\text{mMSE}}(\hat{\boldsymbol{\theta}}_{\text{MLE}}) = \frac{\|\mathbf{M}_n^{-1}\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|}{\sum_{i=1}^n \|\mathbf{M}_n^{-1}\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|}, \quad i = 1, 2, \ldots, n,$$

where

$$\|\mathbf{M}_n^{-1}\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\| = \frac{1}{\hat{\sigma}_{\text{MLE}}}\left\|\mathbf{M}_n^{-1}\left[e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}}) + \delta_i\frac{f_\epsilon\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}{f_\epsilon'\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}, \mathbf{x}_i^\top\right]^\top\right\|\left[(1-\delta_i)h_\epsilon\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\} + \delta_i\frac{f_\epsilon'\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}{f_\epsilon\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}\right],$$

and $f_\epsilon'(x)$ is the first derivative of $f_\epsilon(x)$.

Since multiplying a p dimensional vector by a $p\times p$ matrix takes $O(p^2)$ time, the time complexity for calculating $\boldsymbol{\pi}^{\text{mMSE}}(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ is $O(np^2)$. In order to save the computing time, we also consider the L-optimality criterion from optimal design of experiment[17]. Specfically for our problem, consider minimizing $\text{tr}(\mathbf{V}_n) = \text{tr}(\mathbf{\Gamma}_n\mathbf{M}_n^2)$ where $\mathbf{\Gamma}_n/r$ is the asymptotic variance-covariance matrix of $\tilde{\boldsymbol{\theta}}_r$. Note that given $\mathcal{F}_n$,

$$r^{1/2}\mathbf{M}_n(\tilde{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}_{\text{MLE}}) \to N(\mathbf{0}, \mathbf{V}_n)$$

in distribution, which indicates that minimizing $\text{tr}(\mathbf{V}_n)$ is minimizing the asymptotic MSE of $\mathbf{M}_n\tilde{\boldsymbol{\theta}}_r$. The following theorem presents the SSP for this criterion.

Theorem 3. The optimal SSP denoted as $\boldsymbol{\pi}^{\text{mVc}}(\hat{\boldsymbol{\theta}}_{\text{MLE}}) = \{\pi_i^{\text{mVc}}(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}_{i=1}^n$ which minimizes $\text{tr}(\mathbf{V}_n)$ is

$$\pi_i^{\text{mVc}}(\hat{\boldsymbol{\theta}}_{\text{MLE}}) = \frac{\|\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|}{\sum_{i=1}^n \|\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|}, \quad i = 1, 2, \ldots, n,$$

where

$$\|\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\| = \frac{1}{\hat{\sigma}_{\text{MLE}}}\sqrt{\|\mathbf{x}_i\|^2 + \left(e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}}) + \delta_i\frac{f_\epsilon\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}{f_\epsilon'\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}\right)^2\left[(1-\delta_i)h_\epsilon\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\} + \delta_i\frac{f_\epsilon'\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}{f_\epsilon\{e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\}}\right]}.$$

The time complexity for deriving $\pi_i^{\text{mVc}}(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ is $O(np)$ because calculating the norm of a p dimentional vector takes $O(p)$ time. This shows that calculating $\boldsymbol{\pi}^{\text{mVc}}(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ is faster than calculating $\boldsymbol{\pi}^{\text{mMSE}}(\hat{\boldsymbol{\theta}}_{\text{MLE}})$.

The sensitivity of the optimal SSP in response to $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ is interesting. For parametric models without censoring, observations with residuals of large magnitude have large optimal SSPs in existing investigations[13,14]. This is not true for censored observations. Nevertheless, it does not contradict the fact that optimal SSPs prefer data points that are harder to predict. Since the influences of $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ on π_i^{mMSE} and π_i^{mVc} are complicated as seen in Theorems 2 and 3, respectively, we use a specific example of Weibull parametric AFT model, and plot $\|\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|$ and $\|\mathbf{M}_n^{-1}\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|$ against $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ in Figure 1 for a fixed covariate $\mathbf{x}_i$. Here $\mathbf{M}_n$ and $\hat{\sigma}_{\text{MLE}}$ were calculated from a simulated full data set where the covariates followed a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}_{ij} = 0.5^{I(i\neq j)}$. Figure 1 shows that $\|\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|$ and $\|\mathbf{M}_n^{-1}\dot{l}_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})\|$ both approach zero as $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ approaches $-\infty$ for censored observations. This indicates that π_i^{mVc} and π_i^{mMSE} are smaller with a larger negative $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$. We can explain this result based on the definition of censoring. A censored observation means $c_i \le t_i$, and a negative $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ means $c_i < \hat{t}_i$. Thus, for a censored observation, a larger magnitude of a negative $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ does not mean a larger prediction error, $|t_i - \hat{t}_i|$. On the other hand, a positive $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ means $c_i > \hat{t}_i$, and thus a larger magnitude of a positive $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ means a larger prediction error, $|t_i - \hat{t}_i|$. For uncensored observations, clearly a large absolute $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ means hard to predict, thus both π_i^{mVc} and π_i^{mMSE} are large when $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$ is far away from zero. However, the minimum SSP may not be achieved at zero $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$, which is different from the results in existing investigations[13,14]. The reason is these investigations considered estimating regression coefficient $\boldsymbol{\beta}$ only, while our SSPs are optimal when both the scale parameter σ and regression coefficients $\boldsymbol{\beta}$ are of interest.

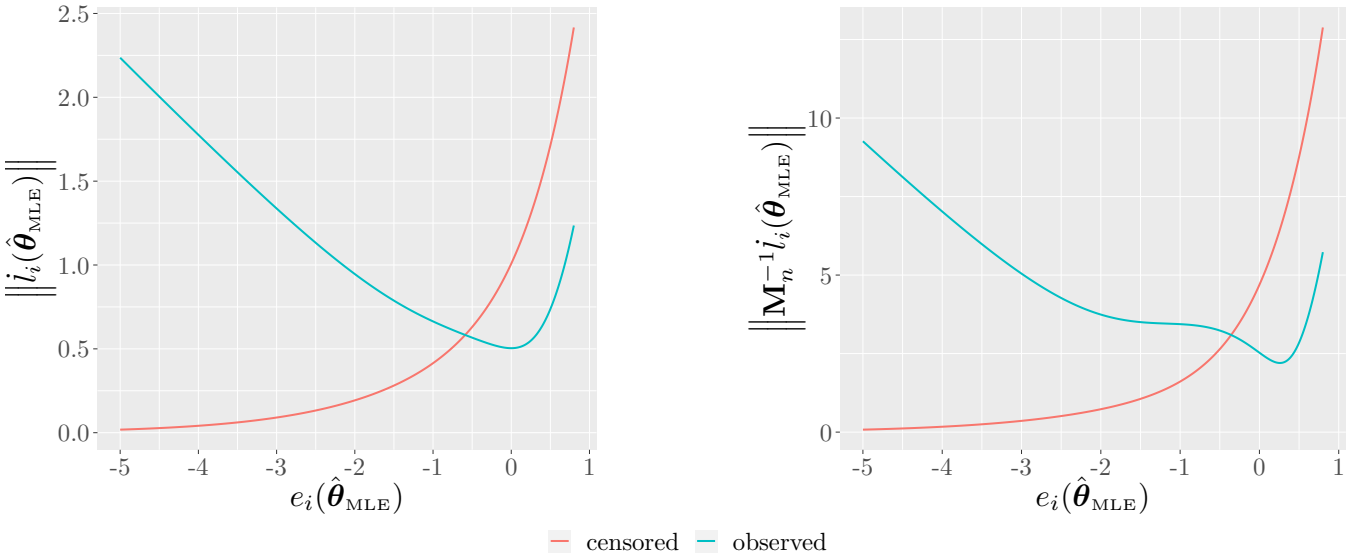

FIGURE 1 The influence of $e_i(\hat{\theta}_{\text{MLE}})$ on $\|\dot{l}_i(\hat{\theta}_{\text{MLE}})\|$ (left panel) and $\|\mathbf{M}_n^{-1}\dot{l}_i(\hat{\theta}_{\text{MLE}})\|$ (right panel) when $\mathbf{x}_i$ is fixed for Weibull AFT model.

4 | A TWO-STEP PROCEDURE

Note that the SSPs in Theorem 2 and 3 depend on $\hat{\theta}_{\text{MLE}}$, so they are not feasible in practice. Therefore, we will propose a workable, two-step procedure in this section. In the first step, we approximate $\boldsymbol{\pi}^{\text{mVc}}(\hat{\theta}_{\text{MLE}})$ and $\boldsymbol{\pi}^{\text{mMSE}}(\hat{\theta}_{\text{MLE}})$ based on a pilot estimator $\tilde{\theta}_r^0$ which is obtained from a small, pilot subsample of size r_0. In the second step, a subsample of size r is drawn according to the SSPs derived in the first step and, in combination with the pilot subsample, is used to obtain the subsampling estimator as presented in Section 2 with combined subsample of size $r + r_0$. In practice, the subsample size is typically restricted by the computing resources we have. We recommend choosing as many observations as allowed by the capacity of the computing facility in order to extract maximum amount of information.

Note that the approximate optimal SSPs, denoted by $\boldsymbol{\pi}^{\text{opt}}(\tilde{\theta}_r^0)$, are derived from a random pilot estimator which may cause additional disturbance. For those data points whose exact optimal SSPs denoted by $\boldsymbol{\pi}^{\text{opt}}(\hat{\theta}_{\text{MLE}})$ are more closer to zero, this additional disturbance may be amplified. In Theorem 1, the matrix $\mathbf{V}_n$ is dominated by these SSPs and thus the asymptotic variance of the subsample estimator will be inflated by these data points. To protect the subsample estimator, we adopt the idea of defensive sampling[18] and mix the approximated $\boldsymbol{\pi}^{\text{opt}}(\tilde{\theta}_r^0)$ with the uniform SSP denoted by $\boldsymbol{\pi}^{\text{Uni}}$. That is, we use adjusted optimal SSPs $\boldsymbol{\pi}_\alpha^{\text{opt}}(\tilde{\theta}_r^0) = \{\pi_{\alpha i}^{\text{opt}}(\tilde{\theta}_r^0)\}_{i=1}^n$ instead of $\boldsymbol{\pi}^{\text{opt}}(\tilde{\theta}_r^0)$ to do subsampling, where

$$\pi_{\alpha i}^{\text{opt}}(\tilde{\theta}_r^0) = (1-\alpha)\pi_i^{\text{opt}}(\tilde{\theta}_r^0) + \frac{\alpha}{n}, \quad 0 < \alpha < 1, \quad i = 1, 2, ..., n.$$

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

The asymptotic properties of the estimator $\check{\theta}_r$ obtained from the two-step procedure based on $\boldsymbol{\pi}_\alpha^{\text{opt}}(\tilde{\theta}_r^0)$ are summarized by the following theorem.

Theorem 4. If Assumptions 1–4 hold and the estimate $\tilde{\theta}_r^0$ from the first step exists, then, as $r \to \infty$, $n \to \infty$, $r_0 \to \infty$, $r/n \to 0$, $r_0/r \to 0$, conditional on $\mathcal{F}_n$ and $\tilde{\theta}_r^0$,

$$r^{1/2}(\boldsymbol{\Gamma}_n^{\text{opt}})^{-1/2}(\check{\theta}_r - \hat{\theta}_{\text{MLE}}) \to N(\mathbf{0}, \mathbf{I})$$

in distribution, where $\boldsymbol{\Gamma}_n^{\text{opt}} = \mathbf{M}_n^{-1}\mathbf{V}_n^{\text{opt}}\mathbf{M}_n^{-1} = O_P(1)$ and

$$\mathbf{V}_n^{\text{opt}} = \frac{1}{n^2}\sum_{i=1}^n \frac{\dot{l}_i(\hat{\theta}_{\text{MLE}})\dot{l}_i^\top(\hat{\theta}_{\text{MLE}})}{\pi_{\alpha i}^{\text{opt}}(\tilde{\theta}_r^0)}.$$

Note that when $r = o(n)$, we can directly use $\boldsymbol{\Gamma}_n^{\text{opt}}$ to discuss the statistical inference on the true parameter θ_0[14]. Based on Theorem 4, we construct an estimator of the variance and covariance matrix of $\check{\theta}_r$, called $\check{\boldsymbol{\Gamma}}_{nr}$. Consider $\dot{l}_{r,i}(\check{\theta}_r)$, $\ddot{l}_{r,i}(\check{\theta}_r)$

and $\pi_{\alpha i}^{\text{opt}_r}(\tilde{\boldsymbol{\theta}}_r^0)$ as the gradient, Hessian matrix of $l_i(\check{\boldsymbol{\theta}}_r)$ and the adjusted optimal SSP for the ith observation of the second-step subsample in the two-step procedure, respectively. Also, $\dot{l}_{r_0,i}(\check{\boldsymbol{\theta}}_r)$ and $\ddot{l}_{r_0,i}(\check{\boldsymbol{\theta}}_r)$ are the gradient and Hessian matrix of $l_i(\check{\boldsymbol{\theta}}_r)$ for the ith observation of the pilot subsample in the two-step procedure, respectively. We can calculate $\check{\boldsymbol{\Gamma}}_{nr}$ by the following formula,

$$\check{\boldsymbol{\Gamma}}_{nr} = \check{\mathbf{M}}_{nr}^{-1}\check{\mathbf{V}}_{nr}\check{\mathbf{M}}_{nr}^{-1}, \tag{2}$$

where

$$\check{\mathbf{M}}_{nr} = \frac{1}{n(r_0+r)}\left(\sum_{i=1}^{r}\frac{\ddot{l}_{r,i}(\check{\boldsymbol{\theta}}_r)}{\pi_{\alpha i}^{\text{opt}_r}(\tilde{\boldsymbol{\theta}}_r^0)} + n\sum_{i=1}^{r_0}\ddot{l}_{r_0,i}(\check{\boldsymbol{\theta}}_r)\right),$$

$$\check{\mathbf{V}}_{nr} = \frac{1}{n^2(r_0+r)}\left(\sum_{i=1}^{r}\frac{\dot{l}_{r,i}(\check{\boldsymbol{\theta}}_r)\left[\dot{l}_{r,i}(\check{\boldsymbol{\theta}}_r)\right]^\top}{\left\{\pi_{\alpha i}^{\text{opt}_r}(\tilde{\boldsymbol{\theta}}_r^0)\right\}^2} + n^2\sum_{i=1}^{r_0}\dot{l}_{r_0,i}(\check{\boldsymbol{\theta}}_r)\left[\dot{l}_{r_0,i}(\check{\boldsymbol{\theta}}_r)\right]^\top\right).$$

In the above formulas, $\check{\mathbf{M}}_{nr}$ and $\check{\mathbf{V}}_{nr}$ are obtained by method of moment. If we replace $\check{\boldsymbol{\theta}}_r$ by $\hat{\boldsymbol{\theta}}_{\text{MLE}}$, then $\check{\mathbf{M}}_{nr}$ and $\check{\mathbf{V}}_{nr}$ are unbiased estimators of $\mathbf{M}_n$ and $\mathbf{V}_n$, respectively. The variance of $(\check{\boldsymbol{\theta}}_r)_i$ is the i^{th} diagonal component of $\check{\boldsymbol{\Gamma}}_{nr}$. We can obtain the estimated MSE of $\check{\boldsymbol{\theta}}_r$ by calculating $\text{tr}(\check{\boldsymbol{\Gamma}}_{nr})$.

5 | SIMULATION STUDY FOR WEIBULL AFT MODEL

The performance of the estimator from the two-step procedure was assessed in a simulation study based on the Weibull AFT model. In the Weibull AFT model, the error terms are i.i.d standard Gumbel variables with probability density function $f_\epsilon(x) = \exp\{x - \exp(x)\}$, $x \in \mathbb{R}$. We generated data from model (1) with seven covariates and an intercept, where the coefficients were all set to be 0.01. The distribution of the covariates had two levels, multivariate normal and multivariate t with 5 degrees of freedom, denoted by "Normal" and "T5", both had mean zero and covariance matrix $\boldsymbol{\Sigma}_{ij} = 0.5^{I(i\neq j)}$. The true scale parameter in (1) was set to be $\sigma \in \{1.0, 2.0\}$ (i.e. the true Weibull shape parameter was set in $\{0.5, 1.0\}$). The censoring distribution was set to be Weibull with the same shape parameters as that of the survival time and the scale parameter was tuned to achieve censoring rates $c_r \in \{0.25, 0.50, 0.75\}$. For each of the 12 configurations, a large dataset of size $n = 100,000$ was generated. For each configuration, the pilot sample size was $r_0 = 1000$ and the subsample size considered were $r \in \{1000, 2000, 3000, 4000\}$. In each setting, we compared the empirical MSE of $\check{\boldsymbol{\theta}}_r$ from $s = 1000$ replicates of the subsampling process from the given dataset

$$\text{MSE} = s^{-1}\sum_{i=1}^{s}\|\check{\boldsymbol{\theta}}_r^{(i)} - \hat{\boldsymbol{\theta}}_{\text{MLE}}\|^2, \tag{3}$$

where $\check{\boldsymbol{\theta}}_r^{(i)}$ is the estimate from the i^{th} subsample. Note that for each replicate, the pilot subsample is different. We report the results for $\sigma = 1$, in the sequel; the results for $\sigma = 2.0$, which are similar, are summarized in the supplement material.

Figure 2 shows the MSEs of $\check{\boldsymbol{\theta}}_r$ based on the uniform, mMse, and mVc SSPs. As expected, in all 6 data configurations, $\boldsymbol{\pi}^{\text{mVc}}(\tilde{\boldsymbol{\theta}}_r^0)$ and $\boldsymbol{\pi}^{\text{mMSE}}(\tilde{\boldsymbol{\theta}}_r^0)$ give smaller MSE than uniform SSP. In particular, in the case of censoring rate 0.25, T5 covariates, and $r = 4000$, the MSE of $\boldsymbol{\pi}^{\text{mMSE}}(\tilde{\boldsymbol{\theta}}_r^0)$ is less than a quarter of that from the uniform SSP. This is a striking reduction; four times of the sample size would be needed for the uniform SSP to achieve this. Covariates with a heavier-tail T5 distribution are likely to yield subsamples with higher variance under $\boldsymbol{\pi}^{\text{mVc}}(\tilde{\boldsymbol{\theta}}_r^0)$ and $\boldsymbol{\pi}^{\text{mMSE}}(\tilde{\boldsymbol{\theta}}_r^0)$, which leads to slightly smaller MSE in comparison to those from normally distributed covariates. As the censoring rate increases, the MSEs of all methods increase as less information is available. In all configurations, the MSE decreases as the subsample size r increases.

The accuracy of the variance estimator (2) is assessed by comparing its average over the 1000 subsamples with the empirical variance. Because the biases are virtually zero, the comparison of the variances can be done with the MSE, which simplifies the comparison over all the parameters to a comparison of the normed version of the MSE in Equation (3). Figure 3 shows the results of the comparison with $\boldsymbol{\pi}^{\text{mVc}}(\tilde{\boldsymbol{\theta}}_r^0)$. The estimated and empirical MSEs are close in all 6 settings. As there is little bias, this close agreement indicates that the variance formula estimates the true variance well. Consequently, when the variance estimate is used construct 95% confidence intervals for the true MLE, the empirical coverage rate matches closely the nominal level (not shown). The result for the other optimal SSP $\boldsymbol{\pi}^{\text{mMSE}}(\tilde{\boldsymbol{\theta}}_r^0)$ is similar and thus omitted.

Finally, we assess the computational efficiency of the proposed methods. We recorded the computing times for the two-step procedure and the uniform subsampling method implemented in R for the 'Normal' data set with scale parameter $\sigma = 1.0$. The computing was carried out on a laptop running Window 10 with an Intel i7-8650U processor and 16 GB memory. Table 1

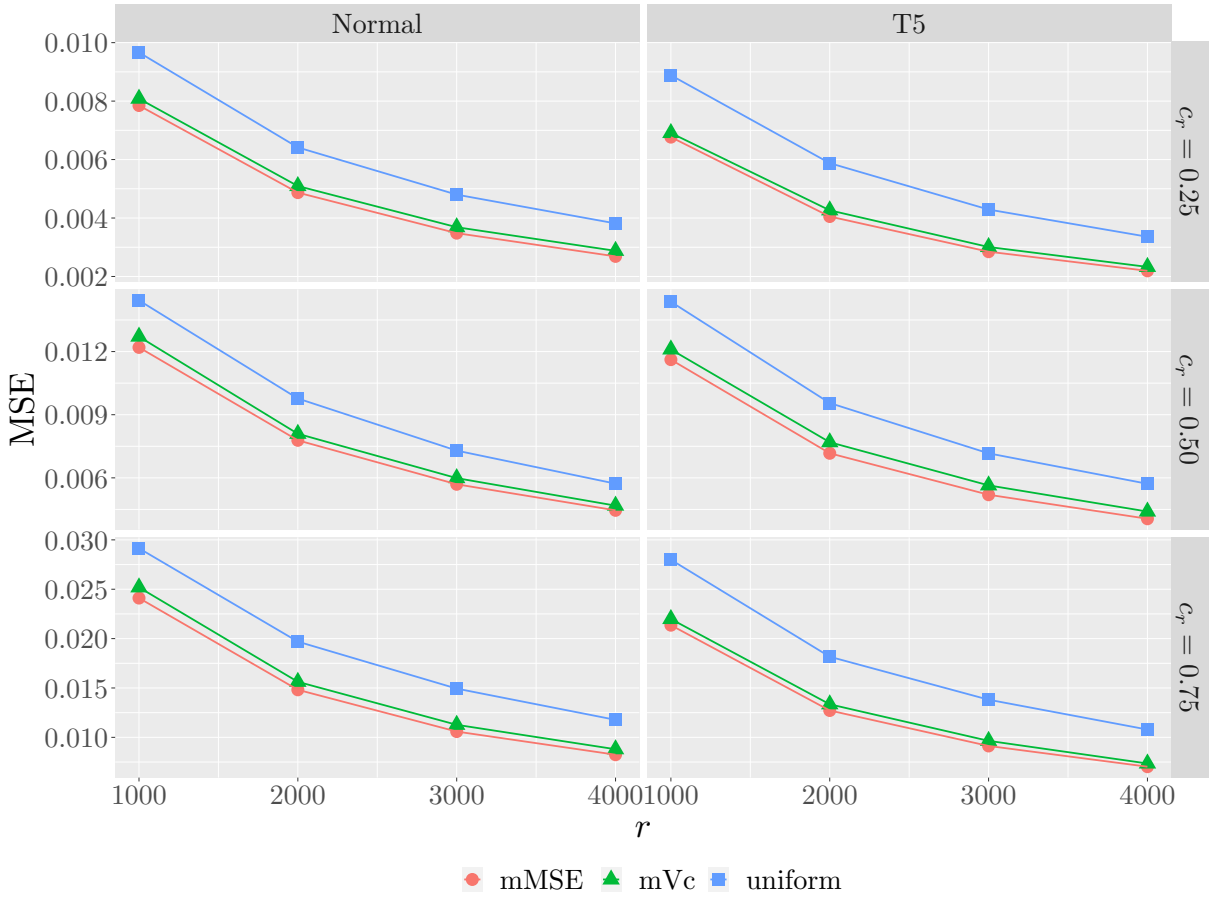

FIGURE 2 MSE for different second step subsample size r and different censoring rate with the first being fixed at $r_0 = 1000$ for different distributions when $\sigma = 1.0$.

summarizes the results based on 1000 replicates. We scaled the computing times using the time for the uniform subsampling method when $r = 1000$ and censoring rate $c_r = 0.25$ as the unit 1. The mVc method used much less time than the mMSE method as expected. Both the mVc and the mMSE methods used more CPU times than the uniform subsampling method because the latter does not need the extra step of calculating the SSPs.

To further investigate the computational gain of the subsampling approach for massive data volume, we collected computing time of the two-step procedure and the uniform subsampling from an implementation in R for the 'Normal' covariate case with $\sigma = 1.0$ and $c_r = 0.25$. We set $r = 2000$ and $r_0 = 1000$ along with dimension increasing to $d = 50$ and all coefficients were set to be 0.01 and the full sample sizes was designed as $n \in \{5 \times 10^6, 10^7, 2 \times 10^7\}$ so that the data took 1/8, 1/4, and 1/2 of the physical memory (RAM), respectively. Note that it was infeasible to get the full data MLE even when $n = 5 \times 10^6$ since R makes multiple copies of the data internally. Table 2 summarizes the results in seconds. We set the computing time of uniform subsampling method when $n = 5 \times 10^6$ as the unit 1. As expected, the computing time using $\boldsymbol{\pi}^{\text{mVc}}(\tilde{\boldsymbol{\theta}}_r^0)$ is less than that using $\boldsymbol{\pi}^{\text{mMSE}}(\tilde{\boldsymbol{\theta}}_r^0)$ and both optimal subsampling methods are more computing-intensive than the uniform subsampling method.

6 | SURVIVAL OF LYMPHOMA BASED ON WEIBULL AFT MODEL

We applied the two-step procedure to AFT modeling of the survival time of lymphoma patients in the SEER program. This data set contained 159,149 patients diagnosed with lymphoma from 1973 to 2012. The censoring rate was 58.3%. Available risk factors included age, nonwhite race indicator (1 = nonwhite), male indicator (1 = male), and the diagnostic year. Interactions between age with gender and age with nonwhite indicator were included also. All the covariates were standardized so that the

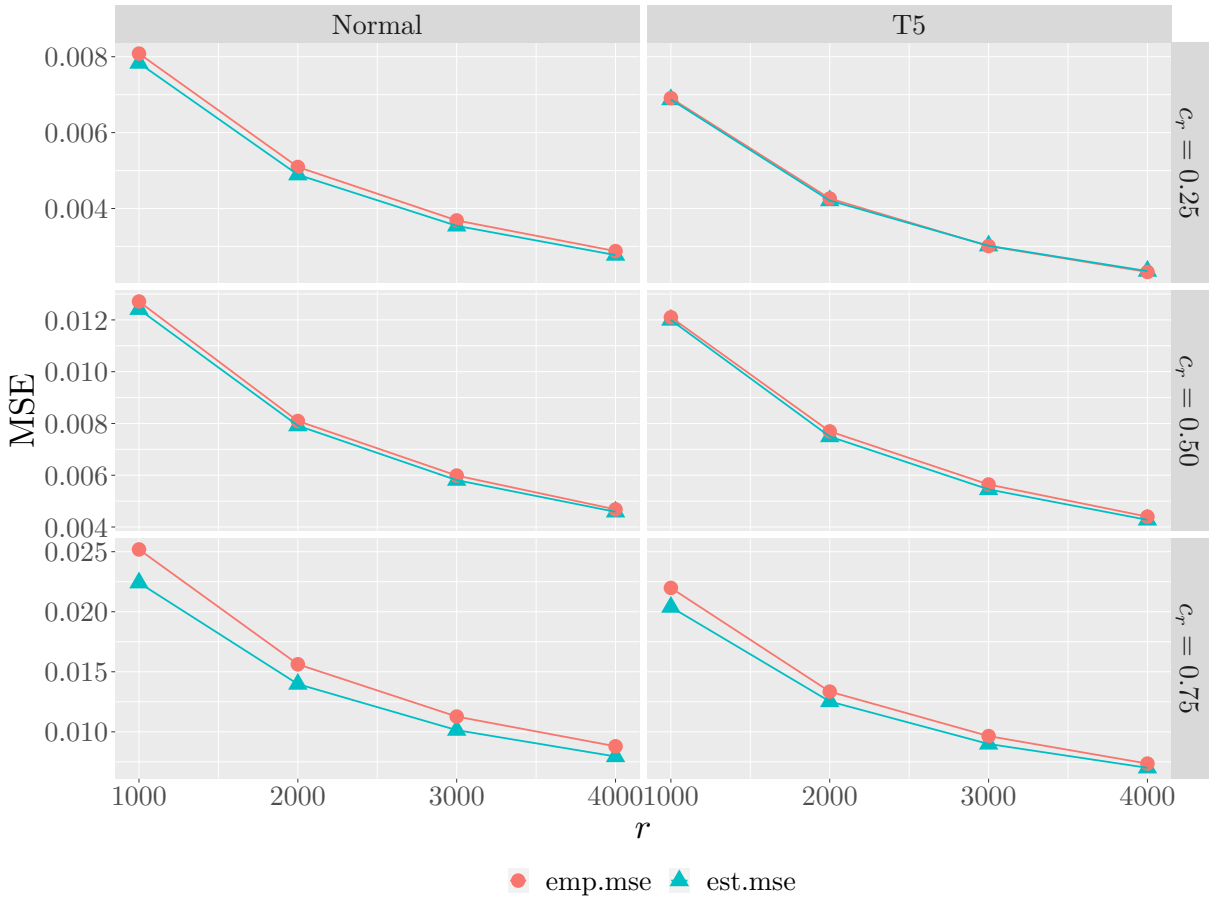

FIGURE 3 Estimated and empirical MSEs with $\boldsymbol{\pi}^{\mathrm{mVc}}(\tilde{\boldsymbol{\theta}}_r^0)$. The first step subsample size is fixed at $r_0 = 1000$ and the second step subsample size r and censoring rate is changing when $\sigma = 1.0$.

TABLE 1 CPU time for 'Normal' data sets when $\sigma = 1.0$ with $r_0 = 1000$ and different second subsample sizes and censoring rates (c_r) over 1000 experiments.

r	c_r: 25%			c_r: 50%			c_r: 75%		
	mVc	mMSE	uni	mVc	mMSE	uni	mVc	mMSE	uni
1000	1.99	5.78	1.00	2.49	7.50	1.19	2.69	7.92	1.36
2000	2.64	7.94	1.34	2.90	8.90	1.48	3.26	9.97	1.93
3000	2.82	9.66	1.81	3.22	9.96	2.04	3.79	11.82	2.47
4000	3.35	11.27	2.45	4.06	11.84	2.76	4.46	12.75	3.25

TABLE 2 CPU time for the selected 'Normal' data set with $r_0 = 1000$, $r = 2000$ for different sample sizes when the dimension of covariates is 50.

Method	Full sample size: n		
	5×10^6	10^7	2×10^7
mVc	21.71	55.71	116.12
mMSE	70.94	180.24	300.24
uniform	1.00	1.53	1.82

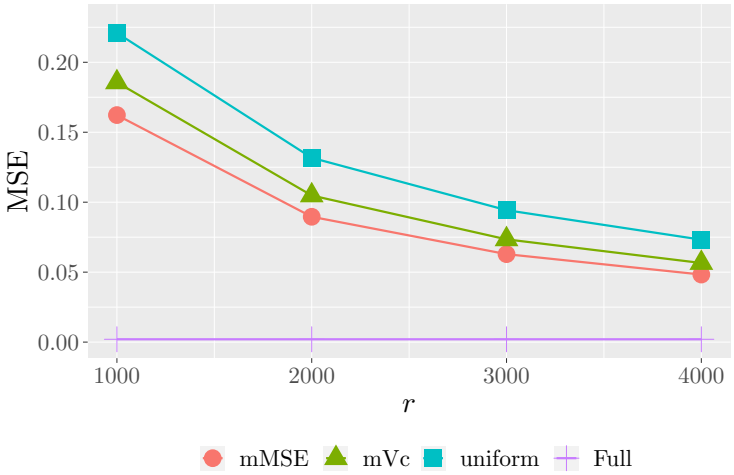

FIGURE 4 Empirical MSEs for different SSPs when fixing pilot sample size $r_0 = 500$ and 4 different second step subsample sizes r over 1000 replicates.

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 experiments and Bootstrap standard errors (BSE) for the full data.

	mVc			mMse			uniform			Full	
	EST	ESE	ASE	EST	ESE	ASE	EST	ESE	ASE	EST	BSE
Scale	1.756	0.027	0.023	1.755	0.031	0.025	1.756	0.031	0.027	1.757	0.005
Intercept	5.056	0.069	0.061	5.056	0.067	0.060	5.055	0.068	0.063	5.054	0.011
Age	−1.115	0.067	0.067	−1.114	0.063	0.063	−1.114	0.075	0.074	−1.116	0.012
Year	0.501	0.037	0.036	0.502	0.040	0.040	0.503	0.041	0.041	0.501	0.007
Male	0.691	0.096	0.092	0.691	0.096	0.090	0.692	0.104	0.101	0.694	0.017
Nonwhite	−0.564	0.114	0.113	−0.563	0.098	0.097	−0.564	0.129	0.127	−0.564	0.022
Age×Nonwhite	0.333	0.120	0.119	0.332	0.100	0.099	0.331	0.141	0.143	0.333	0.023
Age×Male	−0.463	0.093	0.096	−0.462	0.092	0.093	−0.463	0.114	0.113	−0.462	0.019

norm of covariates does not affect the calculation of SSPs. We set initial pilot subsample size to be $r_0 = 500$, and the subsample size $r \in \{1000, 2000, 3000, 4000\}$ for three kinds of SSPs (uniform, mMSE, and mVc).

Figure 4 shows the empirical MSEs from 1000 subsamples of size $r_0 + r$ with fixed $r_0 = 500$. The MSEs based on all SSPs are going to 0 as r increasing which indicates the consistency of the subsampling method for the real data. The optimal SSPs $\boldsymbol{\pi}^{\text{MSE}}(\tilde{\boldsymbol{\theta}}_r^0)$ and $\boldsymbol{\pi}^{\text{mVc}}(\tilde{\boldsymbol{\theta}}_r^0)$ always yield the smaller MSE than the uniform SSPs which shows the efficiency of our method. In particular, the optimal SSP $\boldsymbol{\pi}^{\text{MSE}}(\tilde{\boldsymbol{\theta}}_r^0)$ always has the smallest MSE which meets the theoretical result.

Table 3 summarizes the average estimates obtained by different SSPs and their corresponding empirical and estimated SEs over 1000 replicates with $r_0 = 500$ and $r = 4000$. The full data MLE and the standard errors from nonparametric bootstrap of 1000 replicates are included as they are the target of the subsampling methods. All three subsampling methods produced reliable estimates, but the two optimal SSPs methods yield smaller standard errors than the uniform SSP, especially for Nonwhite and the two interactions. The empirical standard errors based on optimal subsampling approaches are small which indicates that using a smaller subsample instead of the full data is sufficient in practice if computational resources are limited. For all three methods, the estimated standard errors are close to the empirical ones, confirming that the subsampling method is suitable for inference. The results suggest that patients who were elder, female, nonwhite, and diagnosed earlier had shorter survival times. The slope of age was steeper for white patients than nonwhite patients and for male patients than female patients.

TABLE 4 Overall computing time obtained by different SSPs for different subsample sizes with $r_0 = 500$ over 1000 experiments.

	r : 1000	r : 2000	r : 3000
mVc	3.52	5.15	6.69
mMSE	16.02	29.95	32.38
uni	1.00	2.08	2.62

Table 4 shows the total computing time of 1000 replicates with $r_0 = 500$ and different r. Again, we scaled the computing times so that the time for the uniform subsampling method with $r = 1000$ is unit one. The computing time for the mVc method is much shorter than that for the mMSE method. The benefit of the mMSE method relative to the mVc method is to be judged by considering jointly the computing time here and the gain in standard errors in Table 3.

7 | DISCUSSION

The subsampling method for big survival data modeling is challenging due to censoring. Unlike the divide-and-conquer or the online updating methods, which process the whole data, the subsampling method attempts to approximate the whole-data-inference by one or multiple appropriately chosen subsamples. The ultimately essential component of the method is the determination of the optimal SSPs. Parametric AFT models provide insights about the impact of censoring on optimal SSPs, which has not been investigated in existing works [15]. Two optimal SSPs based on A-optimality and L-optimality from optimal design of experiment were proposed under parametric AFT models. For uncensored observations, the impact of $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$'s on the subsampling probabilities are similar to the results in the existing literature that larger absolute $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$'s result in larger subsampling probabilities. For censored observations, however, positive $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$'s with larger magnitude lead to higher subsampling probabilities while negative $e_i(\hat{\boldsymbol{\theta}}_{\text{MLE}})$'s with larger magnitude lead to smaller probabilities. As shown in the simulation study and real data analysis, our method is computationally feasible for big survival data with good approximation to the results based on the full data for the Weibull AFT model. In principle, the subsampling procedure applies to other censoring cases, such as interval censoring [19,20] and left censoring [21]. The framework could be made more flexible, for example, by finite mixture construction for parametric distributions, by allowing nonlinear smooth covariate effects, or by adding random effects [22].

The likelihood based development of the optimal SSPs does not work well for more advanced survival models. For semiparametric Cox relative risk or additive models, the contribution of an observation to the partial likelihood involves information not only from this observation but also from other observations at risk. For semiparametric AFT models, the estimating equations in rank-based or least squares inferences also need information from all observations to compute the contribution from each observation. They are the same challenge as faced by the additive hazard model [15]. New theories and methodologies are needed to address the challenge.

Data Availability Statement

The lymphoma survival data were obtained from the SEER program website (https://seer.cancer.gov/data/access).

References

[1] Wang C, Chen MH, Schifano E, Wu J, Yan J. Statistical Methods and Computing for Big Data. *Statistics and Its Interface* 2016; 9(4): 399–414.

[2] Chen X, Xie Mg. A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data. *Statistica Sinica* 2014; 24(4): 1655–1684.

[3] Song Q, Liang F. A Split-and-Merge Bayesian Variable Selection Approach for Ultrahigh Dimensional Regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 2015; 77(5): 947–972.

[4] Wang Y, Hong C, Palmer N, et al. A Fast Divide-and-Conquer Sparse Cox Regression. *Biostatistics* 2021; 22(2): 381–401.

[5] 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.

[6] Wang C, Chen MH, Wu J, Yan J, Zhang Y, Schifano E. Online Updating Method with New Variables for Big Data Streams. *Canadian Journal of Statistics* 2018; 46(1): 123–146.

[7] Drineas P, Mahoney MW, Muthukrishnan S. Sampling Algorithms for L_2 Regression and Applications. In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm. Society for Industrial and Applied Mathematics; 2006: 1127–1136.

[8] Ma P, Mahoney MW, Yu B. A Statistical Perspective on Algorithmic Leveraging. *Journal of Machine Learning Research* 2015; 16(27): 861–911.

[9] Wang W, Lu SE, Cheng JQ, Xie M, Kostis JB. Multivariate Survival Analysis in Big Data: A Divide-and-Combine Approach. *Biometrics* 2021. Forthcoming, https://doi.org/10.1111/biom.13469.

[10] Xue Y, Wang H, Yan J, Schifano ED. An Online Updating Approach for Testing the Proportional Hazards Assumption with Streams of Survival Data. *Biometrics* 2020; 76(1): 171–182.

[11] Wu J, Chen MH, Schifano ED, Yan J. Online Updating of Survival Analysis. *Journal of Computational and Graphical Statistics* 2021; 30(4): 1209–1223.

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

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

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

[15] 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.

[16] Cox C, Chu H, Schneider MF, Munoz A. Parametric Survival Analysis and Taxonomy of Hazard Functions for the Generalized Gamma Distribution. *Statistics in Medicine* 2007; 26(23): 4352–4374.

[17] Atkinson A, Donev A, Tobias R. *Optimum Experimental Designs, with SAS*. Oxford University Press Inc., New York . 2007.

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

[19] Li J, Ma S. Interval-Censored Data with Repeated Measurements and a Cured Subgroup. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2010; 59(4): 693–705.

[20] Komárek A, Lesaffre E. Bayesian Accelerated Failure Time Model for Correlated Interval-censored Data with A Normal Mixture as Error Distribution. *Statistica Sinica* 2007; 17(2): 549–569.

[21] Polpo A, Coque M, Pereira C. Statistical Analysis for Weibull Distributions in Presence of Right and Left censoring. In: The 8th International Conference on Reliability, Maintainability and Safety. Institute of Electrical and Electronics Engineers; 2009: 219–223.

[22] Lambert P, Collett D, Kimber A, Johnson R. Parametric Accelerated Failure Time Models with Random Effects and An Application to Kidney Transplant Survival. *Statistics in Medicine* 2004; 23(20): 3177–3192.