The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–25
La revue canadienne de statistique

Divide-and-Conquer for Accelerated Failure Time Model with Massive Time-to-Event Data

Wen Su[1]*, Guosheng Yin[1], Jing Zhang[2] and Xingqiu Zhao[3]

[1]*Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong*
[2]*School of Statistics and Mathematics, Zhongnan University of Economics and Law, China*
[3]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong*

Key words and phrases: Accelerated failure time model; Adaptive LASSO; Divide-and-conquer; Oracle property; Survival data.

MSC 2010: Primary 62N01; secondary 62N02

Abstract: Big data present new theoretical and computational challenges as well as tremendous opportunities in many fields. In health care research, we develop a novel divide-and-conquer (DAC) approach to deal with massive and right-censored data under the accelerated failure time model, where the sample size is extraordinarily large and the dimension of predictors is large but smaller than the sample size. Specifically, we construct a penalized loss function through approximating the weighted least squares loss function by combining estimation results without penalization from all subsets. The resulting adaptive LASSO penalized DAC estimator enjoys the oracle property. Simulation studies demonstrate that the proposed DAC procedure performs well and also reduces the computation time with satisfactory performance compared to estimation results using the full data. Our proposed DAC approach is applied to a massive dataset from the Chinese Longitudinal Healthy Longevity Survey.

The Canadian Journal of Statistics xx: 1–25; 20??

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–25; 20??

1. INTRODUCTION

Statistical analysis for big data has become increasingly important with the rapid advance in technologies and the corresponding application in many diverse fields of science and humanities, including e-commerce, finance, engineering, genomics, and biomedical imaging. In this digital era, we can gain access to massive data collected in various locations and explore the potential for turning "big data" to "big information". It is impractical and unnecessary to centrally store and process millions, and sometimes billions of data records. For this reason, traditional statistical methods and computational algorithms are no longer applicable. In medical research, the main goals of analyzing big data are to offer insights into the possible relationships between predictors and response variables

* *Author to whom correspondence may be addressed.*
E-mail: jenna.wen.su@connect.hku.hk

of interest and to accurately predict future outcomes. Big data analysis enables us to perform in-depth and wide-ranging analysis deemed impossible a decade ago, but also presents new challenges such as high-dimensionality, heterogeneity and complexity of data structures (Fan et al, 2014).

In recent years, two methods have commonly been used to tackle the challenges arising due to massive data. One is the divide-and-conquer (DAC) algorithm (e.g., Zhang et al., 2013; Chen & Xie, 2014; Battey et al., 2018; Chan & Peng, 2018); the other is the resampling-based method (e.g., Kleiner et al., 2014; Sengupta et al., 2016; Wang et al., 2018). By the DAC method, a massive dataset is partitioned into small subsamples, and estimators obtained from each subsample are then aggregated to form the final estimator. For example, Zhang et al. (2013) and Huang & Huo (2019) used DAC for M-estimators; Chen & Xie (2014) and Lee et al. (2017) applied DAC to the linear and generalized linear models. Battey et al. (2018) used DAC to study hypothesis testing and parameter estimation in a general likelihood-based framework in both low-dimensional and sparse high-dimensional settings; Chen & Peng (2018) applied DAC to U-statistics and M-estimators. Chen & Peng (2018) pointed out that the resampling-based method has some limitations such as high computational cost when the massive data are stored at different locations. Additionally, the DAC strategy has been extended to a sparse Cox regression by Wang et al. (2019); Xue et al. (2020) developed a DAC algorithm that updates test statistics for hypothesis testing of the proportional hazards assumption under the Cox model as blocks of data are received sequentially. Moreover, the DAC algorithm has been incorporated in many existing techniques from various fields to improve precision and efficiency, namely, the evolutionary algorithm for large-scale optimization (Yang et el., 2019), information-based optimal subdata selection algorithm (Wang, 2019), an coevolutionary algorithm to enhance resource allocation for better control of a spreading virus (Zhao et al., 2020), and precision oncology for subtypes of sarcoma (Pestana et al., 2020), among others.

The DAC approach has been extensively used to develop statistical inferences for massive data when the sample size is exceedingly large and the predictor dimension is not small but smaller than the sample size. Wang et al. (2019) developed a fast and efficient DAC algorithm under the sparse Cox model for use with massive datasets. To the best of our knowledge, there is no existing alternative for massive survival data with censoring when the proportional hazards assumption is violated and the Cox model is no longer appropriate. To fill the gap, we propose a novel DAC approach for an accelerated failure time (AFT) model: (i) Partition massive data into small subsets; (ii) Fit an AFT model to each subset by the weighted least squares (WLS) method without penalization; (iii) Approximate the WLS function using the exact Taylor expansion of the loss function at the WLS estimator and combination of the estimation results based on all subsets; (iv) Obtain the DAC estimator using the approximated WLS loss function

the adaptive LASSO penalization.

A remarkable feature of this proposed DAC estimator is that we apply the adaptive least absolute shrinkage and selection operator (aLASSO) penalization to parameters only once. The proposed methodology provides a useful alternative for fitting massive data with censored response variables when the proportional hazards assumption of the Cox model is violated. Our proposed DAC estimators enjoy the oracle property and outperform the competing full sample-based estimators with respect to required computational time.

The remainder of this paper is organized as follows. Section 2 describes our proposed DAC approach under the AFT model with the aLASSO penalty, and Section 3 establishes its asymptotic properties. In Section 4 we report the results of simulation studies to evaluate the performance of this proposed DAC methodology and apply it to a practical question arising from the Chinese Longitudinal Healthy Longevity Survey (CLHLS) in Section 5. Finally, we discuss our results and possible future work in Section 6. Proofs of our theoretical results may be found in the Appendix.

2. METHODS

2.1. The AFT model with an adaptive LASSO

Consider a survival study that consists of n independent subjects. Let T denote the logarithm of the failure time, and $(X_1, \ldots, X_p)^\top$ be a p-dimensional covariate vector. The AFT model (e.g., Buckley & James, 1979; Jin et al., 2003; Huang, Ma, & Xie, 2006) takes the form,

$$T = \boldsymbol{X}^\top \boldsymbol{\beta} + \epsilon,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is a $(p+1)$-dimensional vector of unknown regression parameters, $\boldsymbol{X} = (1, X_1 \ldots, X_p)^\top$, and ϵ is a random error. Due to censoring, we only observe $(Y, \delta, \boldsymbol{X})$, where $Y = \min(T, C)$, $\delta = I(T \leq C)$, C is the logarithm of the censoring time, and $I(\cdot)$ is the indicator function. The observed data $\{(Y_i, \delta_i, \boldsymbol{X}_i); i = 1, \ldots, n\}$ are i.i.d. copies of $(Y, \delta, \boldsymbol{X})$.

Let $F_0(\cdot)$ denote the distribution function of T, and $\widehat{F}_n(\cdot)$ be the corresponding Kaplan–Meier estimator. Furthermore, let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ denote the order statistics of Y_i $(i = 1, \ldots, n)$, $\delta_{(1)}, \ldots, \delta_{(n)}$ and $\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(n)}$ be the corresponding censoring indicators and covariate vectors. Following Stute (1993), $\widehat{F}_n(\cdot)$ can be expressed as $\widehat{F}_n(y) = \sum_{i=1}^n w_i I(Y_{(i)} \leq y)$, where w_i $(i = 1, \ldots, n)$ are the Kaplan–Meier weights, defined as the jumps in the Kaplan–Meier estimator,

$$w_1 = \frac{\delta_{(1)}}{n}, \;\; w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} \;\; (i = 2, \ldots, n).$$

The weighted least squares (WLS) loss function is defined as

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i (Y_{(i)} - \boldsymbol{X}_{(i)}^\top \boldsymbol{\beta})^2. \tag{1}$$

When the dimension p is small, one can obtain the WLS estimator by directly minimizing $\ell_n(\boldsymbol{\beta})$. Under some regularity conditions, Stute (1993, 1996) proved the WLS estimator is $\sqrt{n}$-consistent and asymptotically normal. Unfortunately, this method does not perform well when p is large. Under the sparsity assumption, only a small number of covariates influence the response variable. To simultaneously identify these contributing covariates and obtain parameter estimates, we consider the aLASSO penalized objective function,

$$Q_n(\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) + \lambda \sum_{j=0}^{p} \frac{|\beta_j|}{|\widetilde{\beta}_j|}, \tag{2}$$

where $\lambda > 0$ is a tuning parameter, and $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^\top$ is an initial estimator. A simple choice of $\widetilde{\boldsymbol{\beta}}$ is the WLS estimator, i.e., $\widetilde{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta})$. The aLASSO penalized estimator is given by

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}). \tag{3}$$

However, when n is extraordinarily large, directly minimizing $Q_n(\boldsymbol{\beta})$ is computationally infeasible. To overcome these difficulties, we propose a novel DAC approach to the sparse AFT model.

2.2. The DAC procedure

Let $\mathcal{D}_{\text{full}} = \{(Y_i, \delta_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ denote the full data, and $\mathcal{I}_{\text{full}} = \{1, \ldots, n\}$ the corresponding index set. Assume n is exceedingly large and p is also large but $n \gg p$. First, we randomly divide the full data $\mathcal{D}_{\text{full}}$ into K subsets $\mathcal{D}_k$ $(k = 1, \ldots, K)$. Without loss of generality, we assume $n^* = n/K$ is an integer and these subsets have the same sample size. Let $\mathcal{I}_k = \{(k-1)n^* + 1, (k-1)n^* + 2, \ldots, kn^*\}$ denote the index of the kth subset, $\mathcal{D}_k = \{(Y_i, \delta_i, \boldsymbol{X}_i), i \in \mathcal{I}_k\}$ denote the data in the kth subset. We assume that $K = O(n^\alpha), 0 \le \alpha < 1$.

A standard DAC method is to obtain the aLASSO penalized estimator based on each subset $\mathcal{D}_k$ $(k = 1, \ldots, K)$, and then combine the subset-specific estimators into an aggregated estimator by arithmetic averaging. Specifically, the DAC estimator is

$$\widehat{\boldsymbol{\beta}}^*_{\text{DAC}} = K^{-1} \sum_{k=1}^{K} \widehat{\boldsymbol{\beta}}_{\mathcal{I}_k}, \tag{4}$$

where $\widehat{\boldsymbol{\beta}}_{\mathcal{I}_k} = \text{argmin}_{\boldsymbol{\beta}} Q_{\mathcal{I}_k}(\boldsymbol{\beta})$, and $Q_{\mathcal{I}_k}(\boldsymbol{\beta})$ is similarly defined as in Equation (2), based on the kth subset $\mathcal{D}_k$. This method can overcome the computational difficulty of $Q_n(\boldsymbol{\beta})$ when n is extraordinarily large. However, seeking the optimal tuning parameter and enumerating the corresponding parameter estimates K times can still occupy considerable computational time.

To reduce the computational burden, we propose a novel DAC method from another perspective. We first apply the Taylor expansion to $\ell_n(\boldsymbol{\beta})$ at $\widetilde{\boldsymbol{\beta}}$, where $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}^{-1} \sum_{i=1}^n \omega_i Y_{(i)} \boldsymbol{X}_{(i)}$ and $\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^n \omega_i \boldsymbol{X}_{(i)} \boldsymbol{X}_{(i)}^\top$. Note that the third-order derivative of $\ell_n(\boldsymbol{\beta})$ equals zero, we obtain the exact expression

$$\ell_n(\boldsymbol{\beta}) = \ell_n(\widetilde{\boldsymbol{\beta}}) + \dot{\ell}_n(\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^\top \ddot{\ell}_n(\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}), \tag{5}$$

where $\dot{\ell}_n(\cdot)$ and $\ddot{\ell}_n(\cdot)$ denote the first- and second-order derivatives of $\ell_n(\boldsymbol{\beta})$, respectively. By the definition of $\widetilde{\boldsymbol{\beta}}$, we have $\dot{\ell}_n(\widetilde{\boldsymbol{\beta}}) = 0$. Thus Equation (5) can be simplified to

$$\ell_n(\boldsymbol{\beta}) = \ell_n(\widetilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^\top \ddot{\ell}_n(\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}),$$

where $\ddot{\ell}_n(\widetilde{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\Sigma}}$. By ignoring the constant $\ell_n(\widetilde{\boldsymbol{\beta}})$, the objective function $Q_n(\boldsymbol{\beta})$ in Equation (2) can be simplified to

$$Q_n^*(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^\top \widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) + \lambda \sum_{j=0}^p \frac{|\beta_j|}{|\widetilde{\beta}_j|}.$$

When n is extraordinary large, we construct the DAC approximations for $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$. To this end, we let $\widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}$ be calculated based on the kth subset $\mathcal{D}_k$, and define

$$\widetilde{\boldsymbol{\beta}}_{\text{DAC}} = K^{-1} \sum_{k=1}^K \widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k}, \quad \widehat{\boldsymbol{\Sigma}}_{\text{DAC}} = K^{-1} \sum_{k=1}^K \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}.$$

Using $\widetilde{\boldsymbol{\beta}}_{\text{DAC}}$ and $\widetilde{\boldsymbol{\Sigma}}_{\text{DAC}}$ to approximate $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ respectively, we obtain the following approximation to $Q_n^*(\widetilde{\boldsymbol{\beta}})$:

$$Q_n^\dagger(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}_{\text{DAC}})^\top \widehat{\boldsymbol{\Sigma}}_{\text{DAC}}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}_{\text{DAC}}) + \lambda \sum_{j=0}^p \frac{|\beta_j|}{|\widetilde{\beta}_{\text{DAC},j}|},$$

where $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} = (\widetilde{\beta}_{\mathrm{DAC},0}, \widetilde{\beta}_{\mathrm{DAC},1}, \ldots, \widetilde{\beta}_{\mathrm{DAC},p})^\top$. Hence we propose the aLASSO penalized DAC estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$ as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} = \mathrm{argmin}_{\boldsymbol{\beta}} Q_n^\dagger(\boldsymbol{\beta}). \tag{6}$$

Let $\widetilde{Y}^* = \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{1/2}\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}$ and $\widetilde{\boldsymbol{X}}^* = \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{1/2}$, where $\widetilde{\boldsymbol{X}}^*$ is a $(p+1)\times(p+1)$ matrix. The optimization problem identified in Equation (6) is equivalent to

$$\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} = \mathrm{argmin}_{\boldsymbol{\beta}} \left\{ (\widetilde{Y}^* - \widetilde{\boldsymbol{X}}^*\boldsymbol{\beta})^\top (\widetilde{Y}^* - \widetilde{\boldsymbol{X}}^*\boldsymbol{\beta}) + \lambda \sum_{j=0}^{p} \frac{|\beta_j|}{|\widetilde{\beta}_{\mathrm{DAC},j}|} \right\}. \tag{7}$$

Intuitively, the computation time involved in solving the problem posed in Equation (7) reduces substantially compared to solving the version posed in Equation (3) when $n \gg p$. Moreover, this proposed DAC method for computing $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$ as outlined in Equation (6) is much faster than the standard DAC method for computing $\widehat{\boldsymbol{\beta}}^*_{\mathrm{DAC}}$ using Equation (4) since it only runs one round for the penalization with the optimal selection of tuning parameter λ, while the standard DAC method needs to run K rounds for the penalized estimation.

To solve the optimization problem posed in Equation (7), we apply the local quadratic approximation (Fan & Li, 2001) to the aLASSO penalty function,

$$\frac{|\beta_j|}{|\widetilde{\beta}_{\mathrm{DAC},j}|} \approx \frac{|\beta_j^*|}{|\widetilde{\beta}_{\mathrm{DAC},j}|} + \frac{1}{2}\frac{1}{|\widetilde{\beta}_{\mathrm{DAC},j}| \cdot |\beta_j^*|}, (\beta_j^2 - (\beta_j^*)^2),$$

where $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^\top$ is a nonzero initial value that is close to $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$. Similar to the arguments in Fan & Li (2001), $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$ can be obtained by iteratively computing the ridge regression,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} = \left(\widetilde{\boldsymbol{X}}^{*\top} \widetilde{\boldsymbol{X}}^* + \frac{1}{2}\lambda \boldsymbol{\Omega}(\boldsymbol{\beta}^*) \right)^{-1} \widetilde{\boldsymbol{X}}^{*\top} \widetilde{Y}^*, \tag{8}$$

where $\boldsymbol{\Omega}(\boldsymbol{\beta}^*) = \mathrm{diag}(\frac{1}{|\widetilde{\beta}_{\mathrm{DAC},0}|\cdot|\beta_0^*|}, \frac{1}{|\widetilde{\beta}_{\mathrm{DAC},1}|\cdot|\beta_1^*|}, \ldots, \frac{1}{|\widetilde{\beta}_{\mathrm{DAC},p}|\cdot|\beta_p^*|})$.

We summarize this new DAC algorithm as follows.

Algorithm 1 The DAC algorithm

Step 1. For each subset $\mathcal{D}_k$ $(k = 1, \ldots, K)$, compute $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}$ and $\widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k}$.

Step 2. Construct the DAC approximations to $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ as $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} = K^{-1}\sum_{k=1}^{K}\widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}} = K^{-1}\sum_{k=1}^{K}\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}$.

Step 3. Compute the aLASSO penalized DAC estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} = \mathrm{argmin}_{\boldsymbol{\beta}} Q_n^{\dagger}(\boldsymbol{\beta})$.

Step 4. To solve the optimization problem in Step 3, we propose the following iterative steps:

(i) Set a nonzero initial value $\boldsymbol{\beta}^*$, and let $m = 1$.

(ii) Compute $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}^{(m)} = \left(\widetilde{\boldsymbol{X}}^{*\top}\widetilde{\boldsymbol{X}}^{*} + \frac{1}{2}\lambda\boldsymbol{\Omega}(\boldsymbol{\beta}^*)\right)^{-1}\widetilde{\boldsymbol{X}}^{*\top}\widetilde{Y}^{*}$.

(iii) Let $\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}^{(m)}$. Update $\boldsymbol{\Omega}(\boldsymbol{\beta}^*)$ and compute the new $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}^{(m+1)}$.

(iv) Repeat Steps (ii) and (iii) until a prespecified convergence criterion is met, i.e., $\|\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}^{(m+1)} - \widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}^{(m)}\| < \epsilon$, with $\epsilon = 10^{-7}$.

The key advantage of this DAC strategy is that it retains the precision of variable selection and parameter estimation while significantly reducing the computational time, especially for massive data where sample size n is extraordinarily large and the number of covariates p is also large.

2.3. Selection of the tuning parameter

Selection of the tuning parameter λ is crucial to the performance of the proposed method. Various choices have been proposed to select the optimal λ in the aLASSO regularization. Here we adopt the Bayesian information criterion (BIC), which was developed by Schwarz (1978). Specifically, for any tuning parameter λ, $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},\lambda}$ denotes its corresponding aLASSO penalized DAC estimator. We define BIC as

$$\mathrm{BIC}(\lambda) = n(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},\lambda} - \widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}})^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},\lambda} - \widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}) + m\log(n),$$

where m is the number of nonzero elements of $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},\lambda}$. The optimal λ equals

$$\hat{\lambda} = \mathrm{argmin}_{\lambda}\mathrm{BIC}(\lambda). \tag{9}$$

3. THEORETICAL PROPERTIES

Before we establish the asymptotic properties of our proposed DAC estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$, we first introduce some notation. Let $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p})^{\top}$ represent the unknown true value of $\boldsymbol{\beta}$. Let H denote the distribution function of Y, and G denote the distribution function of C. The endpoints of Y, T and C are denoted by τ_Y, τ_T, and τ_C, respectively. Let F^0 represent the joint distribution of $(\boldsymbol{X}, T)$.

Define

$$\widetilde{F}^0(\boldsymbol{x},t)=\begin{cases}F^0(\boldsymbol{x},t), & t<\tau_Y,\\ F^0(\boldsymbol{x},\tau_Y-)+F^0(\boldsymbol{x},\tau_Y)I(\tau_Y\in A), & t\geq\tau_Y,\end{cases}$$

where A is the set of axioms of H. Define two sub-distribution functions,

$$\widetilde{H}^{11}(\boldsymbol{x},y)=P(\boldsymbol{X}\leq\boldsymbol{x},Y\leq y,\delta=1),\quad \widetilde{H}^0(y)=P(Y\leq y,\delta=0).$$

For $j=1,\ldots,p$, define

$$\gamma_0(y)=\exp\left\{\int_0^{y-}\frac{\widetilde{H}^0(\mathrm{d}s)}{1-H(s)}\right\},$$

$$\gamma_{1,j}(y;\boldsymbol{\beta})=\frac{1}{1-H(y)}\int I(s>y)(s-\boldsymbol{x}^\top\boldsymbol{\beta})x_j\gamma_0(s)\widetilde{H}^{11}(\mathrm{d}\boldsymbol{x},\mathrm{d}s),$$

$$\gamma_{2,j}(y;\boldsymbol{\beta})=\iint\frac{I(v<y,v<s)(s-\boldsymbol{x}^\top\boldsymbol{\beta})x_j\gamma_0(s)}{\{1-H(v)\}^2}\widetilde{H}^0(\mathrm{d}v)\widetilde{H}^{11}(\mathrm{d}\boldsymbol{x},\mathrm{d}s).$$

For $l=1,2$, let $\gamma_l(y;\boldsymbol{\beta})=(\gamma_{l,0}(y;\boldsymbol{\beta}),\ldots,\gamma_{l,p}(y;\boldsymbol{\beta}))^\top$.

According to Huang, Ma, & Xie (2006), we need the following regularity conditions:

(A1) $E(\epsilon|\boldsymbol{X})=0$.
(A2) T and C are independent and $P(T\leq C|T,\boldsymbol{X})=P(T\leq C|T)$.
(A3) $\boldsymbol{\Sigma}_0=E(\boldsymbol{X}\boldsymbol{X}^\top)$ is finite and nonsingular.
(A4) $\tau_T<\tau_C$ or $\tau_T=\tau_C=\infty$.
(A5) $E\{(Y-\boldsymbol{X}^\top\boldsymbol{\beta}_0)^2\boldsymbol{X}\boldsymbol{X}^\top\delta\}<\infty$; $\int|(y-\boldsymbol{x}^\top\boldsymbol{\beta}_0)x_j|D^{1/2}(y)\widetilde{F}^0(\mathrm{d}\boldsymbol{x},\mathrm{d}y)<\infty$ for $j=0,1,\ldots,p$, where $D(y)=\int_0^{y^-}\{1-H(y)\}^{-1}\{1-G(y)\}^{-1}G(\mathrm{d}y)$.

We now state the theoretical properties of our proposed DAC estimator.

Theorem 1. *Suppose that Assumptions (A1)–(A5) hold. If* $\sqrt{n}\lambda\to0$ *as* $n\to\infty$*, then we have* $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}-\boldsymbol{\beta}_0=O_P(n^{-1/2})$.

Theorem 1 shows that for a properly chosen λ, the proposed DAC estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$ is root-n consistent. Under the sparsity assumption, only a small number of covariates are related to the response variable. Without loss of generality, we assume that the first d_0 covariates are relevant; that is, $\beta_{0j}\neq0$ for $0\leq j\leq d_0$ and $\beta_{0j}=0$ for $d_0<j\leq p$. Define $\boldsymbol{\beta}_0=(\boldsymbol{\beta}_{0,a}^\top,\boldsymbol{\beta}_{0,b}^\top)^\top$, where $\boldsymbol{\beta}_{0,a}=(\beta_{00},\beta_{01},\ldots,\beta_{0d_0})^\top$ and $\boldsymbol{\beta}_{0,b}=(\beta_{0(d_0+1)},\ldots,\beta_{0p})^\top$. Correspondingly, we write $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}=(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},a}^\top,\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},b}^\top)^\top$. Decompose the matrix $\boldsymbol{\Sigma}_0$ into a block

matrix form,

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{11}$ is the first $(d_0 + 1) \times (d_0 + 1)$ submatrix.

Theorem 2. *Suppose that Assumptions (A1)–(A5) hold. If* $\sqrt{n}\lambda \to 0$ *and* $n\lambda \to \infty$ *as* $n \to \infty$*, then we have*

(i) Consistency in variable selection: $\lim_{n\to\infty} P\left(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},b} = 0\right) = 1$*;*

(ii) Asymptotic normality:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},a} - \boldsymbol{\beta}_{0,a}) \xrightarrow{d} \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{W}_a,$$

where $\xrightarrow{d}$ *stands for convergence in distribution,* $\boldsymbol{W}_a$ *is the first* $(d_0 + 1)$ *part of* $\boldsymbol{W}$*,* $\boldsymbol{W} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$*, and* $\boldsymbol{\Sigma}_{\boldsymbol{W}} = \mathrm{Var}\{\delta\gamma_0(Y)(Y - \boldsymbol{X}^\top\boldsymbol{\beta}_0)\boldsymbol{X} + (1 - \delta)\gamma_1(Y; \boldsymbol{\beta}_0) - \gamma_2(Y; \boldsymbol{\beta}_0)\}$.

Theorem 2 shows that the proposed the DAC estimator possesses the oracle property. The unimportant predictors can be excluded with probability approaching to 1, and the estimated nonzero coefficients have the same asymptotic normality as the ideal estimator.

4. SIMULATION STUDIES

We conducted extensive simulation studies to evaluate the finite-sample performance of our proposed method. First, we carried out a large sample simulation to demonstrate its accuracy and efficiency. Second, we carried out a comparison study to assess the proposed DAC method against the full sample-based adaptive lasso estimator, denoted as $\widehat{\boldsymbol{\beta}}_{\mathrm{Full}}$.

We generated the logarithm of the failure time from an AFT model

$$T = \boldsymbol{X}^\top\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, for $\sigma = (0.1, 0.5)$, and $\boldsymbol{X} = (1, \boldsymbol{Z}^\top)^\top$, $\boldsymbol{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\rho^{|i-j|})$ for $\rho = (0.3, 0.5, 0.7), i, j = 1, \ldots, p$. We considered three choices of $\boldsymbol{\beta}$ that represent different signal strengths and sparsity: (i) $\boldsymbol{\beta}^{(1)} = (0, 0.8, 0.7, 0.6, 0.5, 0.4, \mathbf{0}_{p-5}^\top)$, (ii) $\boldsymbol{\beta}^{(2)} = (0, 0.35, 0.3, 0.2, 0.1, 0.07, \mathbf{0}_{p-5}^\top)$, and (iii) $\boldsymbol{\beta}^{(3)} = (0, 0.8_2^\top, 0.7_2^\top, 0.6_2^\top, 0.5_2^\top, 0.4_2^\top, \mathbf{0}_{p-10}^\top)$, where $a_2^\top = (a, a)^\top$. The censoring times were generated from a uniform distribution $\mathrm{U}(0, \tau)$, where τ was chosen to achieve the intended censoring rates 20%, 50% and 70%, respectively. We set the full sample size $n = 10^6$, the number of subsets $K = 100$, where each subset has sample size $n^* = 10^4$. For each configuration, we considered the number of covariates $p = (50, 100)$ and each combination of study

settings $M = 500$ times. Two additional scenarios were also considered, $(p = 50, K = 2000)$ and $(p = 200, K = 100)$. We used the R packages *condSURV* to compute the Kaplan–Meier weights and *glmnet* to obtain the aLASSO penalized estimator.

Tables 1–4 summarize the proportion of covariates correctly identified as having zero and nonzero regression coefficient estimates. Even when the number of noncontributing covariates increased from $p = 50$ to $p = 200$, the proposed model was still capable of correctly classifying covariates as either associated, or not associated, with the response variable. The performance was not affected if we drastically increased the number of subsets or covariates. Furthermore, it is worth mentioning that for a small covariate effect such as 0.07 in $\boldsymbol{\beta}^{(2)}$, the model was able to distinguish such a small value from 0 with 100% accuracy, even with a censoring rate as high as 70%.

TABLE 1: Proportion of parameters estimated as zero by the proposed DAC method under AFT model with $p = 50$ and $K = 100$ over 500 repetitions

			$\boldsymbol{\beta}^{(1)}$						$\boldsymbol{\beta}^{(2)}$						$\boldsymbol{\beta}^{(3)}$					
σ	ρ	CR	0.8	0.7	0.6	0.5	0.4	**0**	0.35	0.3	0.2	0.1	0.07	**0**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0**
0.1	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1

Note: "CR" denotes censoring rate.

We also conducted an additional simulation study to compare the finite-sample performance of our proposed DAC estimator with the full data-based estimator. We used the bootstrap method to estimate the standard errors of the

TABLE 2: Proportion of parameters estimated as zero by the proposed DAC method under an AFT model with $p = 50$ and $K = 2000$ over 500 repetitions

			$\boldsymbol{\beta}^{(1)}$						$\boldsymbol{\beta}^{(2)}$						$\boldsymbol{\beta}^{(3)}$					
σ	ρ	CR	0.8	0.7	0.6	0.5	0.4	**0**	0.35	0.3	0.2	0.1	0.07	**0**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0**
0.1	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1

Note: "CR" denotes censoring rate.

estimators obtained from the DAC approach and the full sample. In this study we fixed $\boldsymbol{\beta} = (0, 0.8, 0.7, 0.6, 0.5, 0.4, \mathbf{0}_{p-5}^{\top})$ with $p = 10$, $\sigma = 0.5$, $\rho = 0.5$ and three target censoring rates of 20%, 50%, and 70%. To shorten the computational time, we used the sample size $n = 10^5$ and the number of subsets $K = 10$. For each censoring rate, we ran 500 simulations, and used 500 bootstrap samples based on each simulated dataset for estimating the standard deviations. To further reduce the computational burden, we used the optimal tuning parameters $\widehat{\lambda}_{\mathrm{DAC}}$ and $\widehat{\lambda}_{\mathrm{Full}}$ determined from the proposed DAC algorithm and the full dataset-based penalized estimation based on each simulated dataset as the optimal tuning parameters for computing the corresponding aLASSO penalized estimates, over 500 bootstrap repetitions.

In Table 5 we report the observed values of the following performance measures:

- Bias: the average deviation of $\widehat{\boldsymbol{\beta}}$ from the true value;
- SSE: sample standard error;
- ESE: the average estimated standard error;

TABLE 3: Proportion of parameters estimated as zero by the proposed DAC method under an AFT model with $p = 100$ and $K = 100$ over 500 repetitions

			$\boldsymbol{\beta}^{(1)}$						$\boldsymbol{\beta}^{(2)}$						$\boldsymbol{\beta}^{(3)}$					
σ	ρ	CR	0.8	0.7	0.6	0.5	0.4	**0**	0.35	0.3	0.2	0.1	0.07	**0**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0**
0.1	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1

Note: "CR" denotes censoring rate.

- GMSE: global mean squared error, defined as

$$\text{GMSE} = \frac{1}{500}\sum_{\text{m}=1}^{500}\left(\widehat{\boldsymbol{\beta}}_{\text{DAC}_\text{m}} - \boldsymbol{\beta}_0\right)\Sigma\left(\widehat{\boldsymbol{\beta}}_{\text{DAC}_\text{m}} - \boldsymbol{\beta}_0\right)^T.$$

According to the summary results found in Table 5, our proposed DAC approach yields remarkably similar observed values to the corresponding full sample-based estimation. Moreover, the SSEs and ESEs in all cases are nearly identical, indicating the appropriateness of the bootstrap methodology.

Next, we compared the computational time required to obtain parameter estimates for our proposed DAC method and for full data-based evaluation. To ensure fair comparison, the simulation was carried out as a single-core job on an Intel Xeon Gold 6248R, 24C/48T, CPU @ 3.0GHz. Table 6 summarizes the observed computational times based on 10 replications for $n = (10^5, 2\times10^5, 5\times10^5, 10^6, 2\times10^6)$ and $p = (50, 100, 200)$ with $\sigma = 0.5, \rho = 0.5$, and censoring rate 20%. Clearly, our proposed DAC method required less computational time than the competing full data-based evaluation in all cases, and the improvement

TABLE 4: Proportion of parameters estimated as zero by the proposed DAC method under an AFT model with $p = 200$ and $K = 100$ over 500 repetitions

			$\boldsymbol{\beta}^{(1)}$						$\boldsymbol{\beta}^{(2)}$						$\boldsymbol{\beta}^{(3)}$					
σ	ρ	CR	0.8	0.7	0.6	0.5	0.4	**0**	0.35	0.3	0.2	0.1	0.07	**0**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0**
0.1	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.1	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.3	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.5	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
0.5	0.7	20%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		50%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
		70%	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1

Note: "CR" denotes censoring rate.

in speed increases with larger sample sizes and higher dimensionality. In practice, each subset would be evaluated in parallel on separate machines with the proposed DAC method, which further reduces the associated computational burden.

5. APPLICATION TO CLHLS DATA

We applied our proposed DAC method of estimation to the problem of identifying risk factors associated with human mortality based on the CLHLS data. During the period from 1998 to 2014, the CLHLS study collected health and quality of life-related information among elderly people aged 65 or older from 22 provinces in China (Yi et al., 2017). The survey was conducted in seven waves, and carried out every two years with new participants entering the study to replace the deceased and individuals lost to follow-up. A total of 44,576 individuals were interviewed during the seven waves of the study. See to Yi et al. (2008) for a detailed description of the CLHLS data. After excluding subjects with missing data, those participants younger than or deceased at 65 as well as the last two waves, 40,530 subjects remained for this analysis, with 8,611 in the first wave,

TABLE 5: Comparison of the proposed DAC estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$ vs the full sample-based estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{Full}}$

CR	$\boldsymbol{\beta}_0$	0.8	0.7	0.6	0.5	0.4
				$\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$		
20%	BIAS	-0.0166	-0.0144	-0.0124	-0.0103	-0.0084
	SSE	0.0025	0.0029	0.0028	0.0027	0.0024
	ESE	0.0025	0.0028	0.0027	0.0027	0.0025
	GMSE	0.0018				
50%	BIAS	-0.0483	-0.0426	-0.0365	-0.0307	-0.0243
	SSE	0.0038	0.0042	0.0041	0.0041	0.0035
	ESE	0.0037	0.0039	0.0039	0.0039	0.0036
	GMSE	0.0152				
70%	BIAS	-0.0804	-0.0698	-0.0598	-0.0501	-0.0401
	SSE	0.0048	0.0057	0.0052	0.0050	0.0045
	ESE	0.0048	0.0053	0.0052	0.0051	0.0048
	GMSE	0.0412				
				$\widehat{\boldsymbol{\beta}}_{\mathrm{Full}}$		
20%	BIAS	-0.0166	-0.0142	-0.0124	-0.0102	-0.0083
	SSE	0.0027	0.0036	0.0030	0.0028	0.0029
	ESE	0.0026	0.0029	0.0028	0.0028	0.0026
	GMSE	0.0018				
50%	BIAS	-0.0482	-0.0425	-0.0364	-0.0306	-0.0242
	SSE	0.0041	0.0045	0.0045	0.0044	0.0040
	ESE	0.0043	0.0044	0.0044	0.0045	0.0043
	GMSE	0.0153				
70%	BIAS	-0.0803	-0.0696	-0.0597	-0.0499	-0.0401
	SSE	0.0053	0.0065	0.0059	0.0058	0.0051
	ESE	0.0054	0.0058	0.0056	0.0057	0.0054
	GMSE	0.0414				

Note: "CR" denotes censoring rate.

6,180 in the second wave, 9,289 in the third wave, 7,376 in the fourth wave, and 9,074 in the fifth wave. Waves six and seven, consisting of 1,360 and 1,124 newly-added interviewees respectively, were not included due to their significantly smaller sample sizes. Moreover, wave six involved a high proportion of missing data and wave seven was interviewed only once, making both subsets inappropriate for this analysis. Datasets from each interview wave were treated as natural subsets for use of our proposed DAC estimation.

Since the survey required all participating seniors to be 65 years or older, survival time was defined as the age at death in years in excess of 65. Age at death was calculated as the difference between the validated birth year and month and the reported death year and month, so that the age calculation was accurate

TABLE 6: Comparison of computational time to obtain parameter estimates from the proposed DAC approach and full data-based evaluation

	p=50			p=100			p=200		
N	DAC	Full	Relative	DAC	Full	Relative	DAC	Full	Relative
100000	0.554	0.753	1.360	0.905	1.410	1.557	1.884	4.216	2.237
200000	0.629	1.177	1.870	1.091	2.552	2.340	2.275	8.121	3.570
500000	0.913	2.333	2.556	1.604	6.242	3.891	3.432	19.794	5.768
1000000	1.396	4.289	3.072	2.650	12.083	4.559	5.633	39.368	6.989
2000000	2.533	8.607	3.398	4.971	24.308	4.890	10.630	79.296	7.459

Note: Average run time in seconds based on 10 replications as single-core jobs. Additional settings: $\sigma = 0.5, \rho = 0.5$, and censoring rate 20%. Relative $= \text{run time}_{\text{Full}}/\text{run time}_{\text{DAC}}$.

to months. For the study subjects who died between two interview periods but whose exact death times were unknown, the midpoint between the previous interview date and the current interview date was interpolated to be the surrogate death time. The date of the final interview conducted for each wave was taken to be the censoring times for subjects in that wave who were lost to follow-up during the current set of interviews. In other words, the censoring indicator was 1 when the individual died and 0 when the individual was known to be alive during the last interview session but the current survival status was unknown. In total, 25,274 subjects died and 15,256 were censored, yielding an observed censoring rate 37.64%.

Poston & Min (Yi et al., 2008, Chapter 7) fitted a Cox proportional hazards model to the 1998–2000 CLHLS dataset and concluded that sociodemographic characteristics such as age, gender and marital status were strong predictors of the hazard of death. For our analysis, we selected 43 covariates from survey sections including Basic Information, Lifestyle, Katz Activities of Daily Living (ADL), Personal Background, and Objective Examination and Illnesses. Under Basic Information, we chose gender, ethnic group (Han/other), place born (rural/urban), residence (rural/urban), co-residence (family/nursing home/alone), and number of household members if residing with family. For Lifestyle, personal living habits chosen included main food (rice/other), fresh fruit/vegetables intake, physical labour, as well as personal habits such as smoking, drinking and exercising. Furthermore, from a list of daily activities, we selected housework, play cards/mahjong, and watch TV/listen to radio as these are more representative and applicable to the majority of the Chinese population. Katz ADL criteria included whether or not an individual received assistance with bathing, dressing, toilet, transfer, continence, and feeding. We also added three covariates from Personal Background, if an individual received adequate medical service when sick at present, 80 years old, and 60 years old. Since the survey covered the most elderly segment of the Chinese population, the majority (up to 70%) of those surveyed were widowed and thus marital status could potentially serve as a

confounder that masks the true relationship between the survival time and other covariates such as gender. Lastly, from Objective Examination and Illnesses, we chose two sets of covariates: the number of times a subject had experienced serious illness in the past two years and if currently suffering from hypertension, diabetes, heart disease, stroke, cerebrovacular disease, bronchitis, pulmonary emphysema, asthma, pneumonia, pulmonary tuberculosis, cataract, glaucoma, cancer, gastric or duodenal ulcer, Parkinson's disease and bedsores. Prostate cancer was excluded as it only applies to male subjects, and many subjects indicated the value of that particular covariate was unknown or missing.

The proportional hazards test (Grambsch & Therneau, 1994) for subset 1 yielded a p-value $< 2 \times 10^{-16}$, indicating that the Cox model is not an appropriate fit. Moreover, Figure 1 shows the residuals from fitting an AFT model to subset 1; they appear to be randomly scattered around zero, indicating that the AFT assumption is reasonable. Thus, we chose to fit the AFT model to the data, using our proposed DAC method to estimate the model parameters, with five subsets consisting of the five remaining waves. Out of 43 preselected covariates, the proposed model identified 10 nonzero predictors: 1) suffering from diabetes, 2) stroke, cerebrovascular disease, 3) hypertension, 4) cancer, 5) received assistance during bathing, 6) being male, 7) did housework, 8) watching TV/listening to radio, 9) playing cards/mahjong, and 10) being a smoker. Here, we utilized the bootstrap method with 500 repetitions to estimate the standard errors of the various parameter estimates. At the 95% confidence level, except for the covariate suffering from cancer, the remaining nine covariates are significant. Detailed results are summarized in Table 7. A possible explanation for the apparent absence of an association between the covariate suffering from cancer and study subject mortality could be the recent advances in medicine and cancer treatment that significantly prolonged patient survival and hence reduced study subject mortality among individuals identified as suffering from cancer. In particular, some of the less severe cancers are curable if they are diagnosed early enough, such as breast cancer, stomach cancer, and nasopharyngeal carcinoma. Hence depending on the stage and type of cancer, this particular explanatory variate may represent a lesser threat to the continued survival of an individual compared to other chronic illnesses. Figure 2 shows the Kaplan–Meier survival curves derived from different groups for nine covariates based on subset 1, which has the largest sample size; corresponding plots based on other subsets have a similar appearance, and thus are not shown. Like other results that we have observed after fitting an AFT model with aLASSO penalization using the DAC algorithm, we can easily observe that apart from the covariate that distinguishes a study subject who receives assistance for bathing, all the other covariates are negatively associated with subject survival. Receiving assistance during bathing could reduce the risk of accidents for elders, and hence be associated with reduced overall mortality. Although moderate physical and mental activities could be beneficial for main-

taining a healthy brain and body, heavy labour such as doing housework and excessive sedentary entertainment could impose stress and cause harm. For instance, prolonged sitting from playing cards/mahjong and watching TV/listening to radio, as well as heavy lifting from household chores could adversely affect the spine and joints.

TABLE 7: Estimation results of the AFT model fitted to CLHLS Data using the proposed DAC approach

Variable	$\hat{\beta}$	ESE	p-value
Suffer from diabetes (yes = 1, no = 0)	-0.1715	0.0291	3.8432e-09
Gender (male = 1, female =0)	-0.1142	0.0055	2.5515e-95
Suffer from stroke, cerebrovascular disease (yes = 1, no = 0)	-0.1093	0.0134	4.2380e-16
Do housework (yes = 1, no = 0)	-0.1023	0.0053	1.7787e-83
Suffer from cancer (yes = 1, no = 0)	-0.0750	0.0656	2.5318e-01
Bathing Assistance (receives assistance = 1, otherwise = 0)	0.0705	0.0055	1.0506e-37
Watch TV/listen to radio (yes = 0, no = 0)	-0.0600	0.0054	3.2242e-28
Suffer from hypertension (yes = 1, no = 0)	-0.0550	0.0091	1.5942e-09
Play cards/mahjong (yes = 1, no = 0)	-0.0549	0.0123	7.9645e-06
Smoker (yes = 1, no = 0)	-0.0451	0.0142	1.5270e-03

Note: Standard errors are estimated using the bootstrap method with 500 replications.

6. CONCLUDING REMARKS

To deal with massive survival data, we have proposed a novel DAC method of estimating an accelerated failure time model. This method involves constructing an approximate WLS loss function, thereby efficiently reducing the dimension of the estimation problem. A remarkable advantage of this approach is that the penalized estimation procedure is implemented only once, while the standard DAC approach to the same estimation problem requires the penalized estimation procedure to run K times, seeking an optimal tuning parameter λ for each subset, where K is the number of subsets. The DAC estimator that we have derived possesses the oracle property. Our simulation studies demonstrate that our proposed DAC approach is able to correctly identify important and unimportant predictors; in addition, our method of estimation achieves a level of accuracy and efficiency that is comparable to the accuracy and efficiency of the full sample-based estimator as expected, suggesting that our DAC method of estimating an accelerated failure time model also may have potential applicability in massive data analytics.

Note that the condition concerning K identified in Section 2.3 is crucial with respect to the large-sample properties of the DAC estimators. While many other published methods require $K = o(n^{1/2})$ – for example, see Tang et al. (2020), Chen & Zhou (2020), Volgushev et al. (2019), and Wang et al. (2019) – we

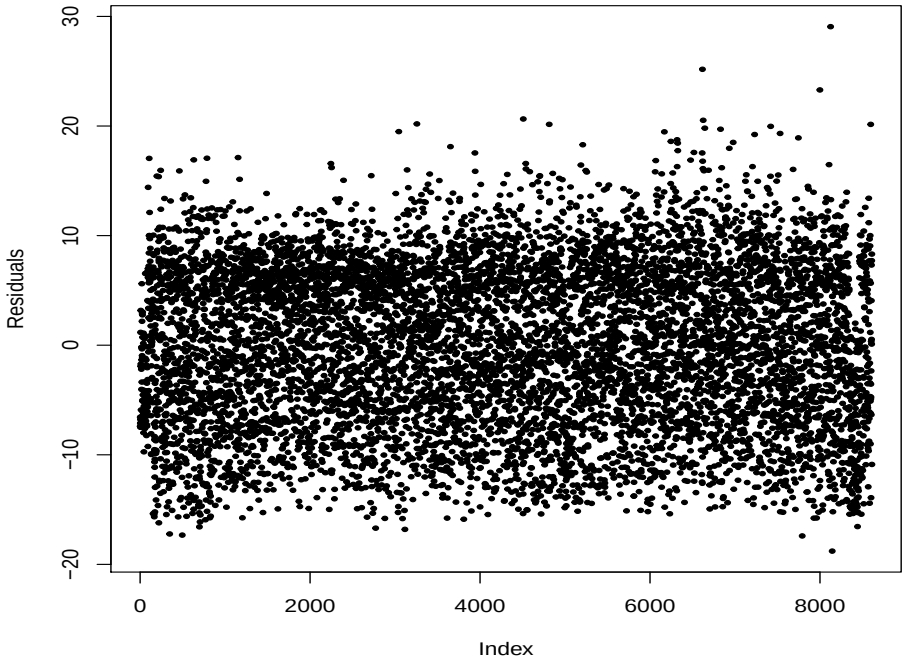

FIGURE 1: Residuals plot for checking AFT assumption based on subset 1.

only need $K = O(n^{\alpha})$ with $0 \leq \alpha < 1$. The reason is that the Taylor expansion of $\ell_n(\boldsymbol{\beta})$ in Equation (5) is not an approximation but an equality. However, in the abovementioned literature, various approximations are used to construct the DAC estimators, so they need strong conditions to obtain the expected theoretical properties. For example, in Wang et al. (2019) who studied the Cox model, the condition concerning K is $K = o(n^{1/2})$. They proved

$$\widetilde{\beta}_{\mathcal{I}_k,\mathrm{lin}} - \widetilde{\beta}_{\mathcal{I}_k} = O((n^*)^{-1}) = o(n^{-1/2}),$$

where $\widetilde{\beta}_{\mathcal{I}_k}$ is obtained from the likelihood function and $\widetilde{\beta}_{\mathcal{I}_k,\mathrm{lin}}$ via a Taylor approximation of the likelihood function. Therefore, Wang et al. (2019) need this condition to ensure $\widetilde{\beta}_{\mathcal{I}_k,\mathrm{lin}} - \widetilde{\beta}_{\mathcal{I}_k} = o(n^{-1/2})$, while $\widetilde{\beta}_{\mathcal{I}_k,\mathrm{lin}} - \widetilde{\beta}_{\mathcal{I}_k} = 0$ in our method of estimation so that the condition on K is weaker.

Recall that $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}$ is defined as

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} = K^{-1} \sum_{k=1}^{K} \widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k} = K^{-1} \sum_{k=1}^{K} \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1} \sum_{i=1}^{n^*} \omega_{\mathcal{I}_k,i} \boldsymbol{Y}_{\mathcal{I}_k,(i)} X_{\mathcal{I}_k,(i)}.$$

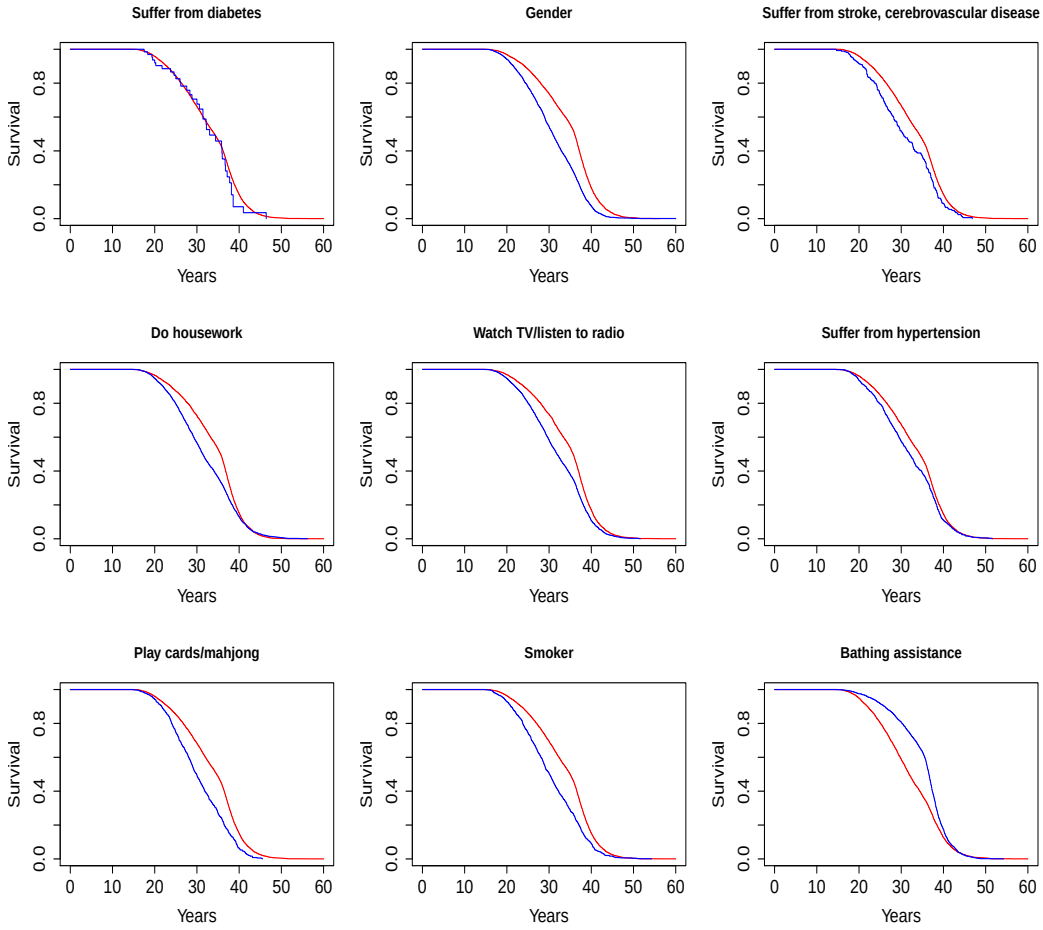

FIGURE 2: Kaplan-Meier curves of survival functions based on subset 1. Blue lines represent the corresponding covariate values $= 1$, while red lines represent the corresponding covariate values $= 0$. *Years* is the actual age of individuals minus 65.

Here we need to compute $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1}$ for each $\mathcal{D}_k, k = 1, \ldots, K$. According to one referee's suggestion, all $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1}$ $(k = 1, \ldots, K)$ are replaced by $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{-1}$, where $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}} = K^{-1} \sum_{k=1}^{K} \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}$. Then $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}^{\mathrm{New}}$ is defined as

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}^{\mathrm{New}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{-1} K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n^*} \omega_{\mathcal{I}_k,i} \boldsymbol{Y}_{\mathcal{I}_k,(i)} X_{\mathcal{I}_k,(i)}.$$

By the definition of $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}$ and $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}^{\mathrm{New}}$, we have

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}^{\mathrm{New}} = K^{-1} \sum_{k=1}^{K} (\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{-1}) \sum_{i=1}^{n^*} \omega_{\mathcal{I}_k,i} \boldsymbol{Y}_{\mathcal{I}_k,(i)} X_{\mathcal{I}_k,(i)}.$$

By Lemma A.1, which we state and prove in the Appendix, we know that $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k} \xrightarrow{P} \boldsymbol{\Sigma}_0$, $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}} \xrightarrow{P} \boldsymbol{\Sigma}_0$, and so

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}^{\mathrm{New}} \xrightarrow{P} 0.$$

While Lemma A.2 – again, see Appendix – plays a key role in the proofs of Theorems 1–2, the expression of Equation (1) that appears in the proof of Lemma A.2 can be simplified via this new construction, thereby reducing the complexity of our proofs of the theoretical results. In theory, the alternative estimator $\tilde{\boldsymbol{\beta}}_{\mathrm{DAC}}^{\mathrm{New}}$ would yield comparable statistical performance as well as reduce computational time since it eliminates matrix inversion $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1}$ for each subset $\mathcal{D}_k$. We conducted simulation studies for $p = (50, 100, 200)$, $n = (10^5, 2 \times 10^5, 5 \times 10^5, 10^6, 2 \times 10^6)$ with $\sigma = 0.5, \rho = 0.5, \boldsymbol{\beta} = (0.8, 0.7, 0.6, 0.5, 0.4, \mathbf{0}_{p-5}^{\top})$, and censoring rate 20%. Both methods consumed nearly equal amounts of computing time with practically no difference. Furthermore, simulations show that for $p = 50, 200$, both methods yielded similar bias values for the bias and sample standard error based on 500 Monte Carlo replications.

Note that the weighted least squares method of estimating the AFT model (Suite, 1993, 1996) requires the condition that T and C are independent, which may not hold in certain circumstances. To relax this condition, we can employ the local Kaplan–Meier estimator (Dabrowska, 1989; Wang & Wang, 2009) to construct local weights instead of the Kaplan–Meier weights in Equation (1). Furthermore, future research could consider massive interval-censored survival data. In the CLHLS dataset, for participants who died between two interview sessions but whose exact time of death was not recorded, the midpoint between the two interview sessions was used as a surrogate death time. Hence, extending the current DAC approach to include the possibility of interval censoring is warranted.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor-in-Chief, Dr. Fang Yao, the Associate Editor, and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper. Yin's research is partly supported by the Research Grant Council of Hong Kong (17308321). Zhang's research is partly supported by the National Natural Science Foundation of China (11901581) and the National Science Foundation of Hubei Province (2021CFB502). Zhao's research is partly supported by the Research Grant Coun-

cil of Hong Kong (15303319).

BIBLIOGRAPHY

attey, H., Fan, J., Liu, H., Lu, J., & Zhu, Z. (2018). Distributed testing and parameter estimation under sparse high dimensional models. *Annals of Statistics* **46**, 1352–1382.

uckley, J. & James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

hen, S. & Peng, L. (2018). Distributed statistical inference for massive data. arXiv:1805.11214v1.

hen, X. & Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **24**, 1655–1684.

abrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. **19** 1157–1167.

an, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review* **1**, 293–314.

an, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

rambsch, P. M. & Therneau, T. M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika 81*, 515–526.

uang, C. & Huo, X. (2019). A distributed one-step estimator. *Mathematical Programming* **174**, 41–76.

uang, J., Ma, S., & Xie, M. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820.

in, Z., Lin, D. Y., Wei, L. J., & Ying, Z. L. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341–353.

leiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B* **76**, 795–816.

ee, J. D., Liu, Q., Sun, Y., & Taylor, J. E. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research* **18**, 1–30.

estana, R.C., Groisberg, R., Roszik, J., & Subbiah, V. (2020). Precision oncology in sarcomas: divide and conquer. *JCO Precision Oncology*, DOI: 10.1200/PO.18.00247.

chwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

engupta, S., Volgushev, S., & Shao, X. (2016). A subsampled double bootstrap for massive data. *Journal of the American Statistical Association* **111**, 1222–1232.

tute, W. (1993). Consistent estimation under random censorship when covariables are available. *Journal of Multivariate Analysis* **45**, 89–103.

tute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* **23**, 461–471.

an der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, US.

an der Vaart, A. W., & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

ang, H. J., & Wang, L (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association* **104**, 1117–1128.

ang, H. Y. (2019). Divide-and-conquer information-based optimal subdata selection algorithm. *Journal of Statistical Theory and Practice* **13**, 46. https://doi.org/10.1007/s42519-019-0048-5.

ang, H., Yang, M., & Stufken, J. (2018). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* **114**, 393–405.

ang, Y., Palmer, N., Di, Q., Schwartz, J. Kohane, I., & Cai, T. (2019). A fast divide-and-conquer sparse Cox regression. *Biostatistics*, doi.org/10.1093/biostatistics/kxz036.

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

ang, P., Tang, K., & Yao, X. (2019). A parallel divide-and-conquer-based evolutionary algorithm for large-sale optimization. *IEEE Access* **7**, 163105–163118.

i, Z., Vaupel, J., Xiao, Z.Y., Liu, Y.Z., & Zhang, C.Y. Chinese longitudinal healthy longevity survey (CLHLS), 1998-2014. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2017-04-11. https://doi.org/10.3886/ICPSR36692.v1

i, Z., Poston D.L., Vlosky, D.A., & Gu, D. (2008). Healthy Longevity in China: Demographic, Socioeconomic, and Psychological Dimensions. Springer.

hang, Y., Duchi, J. C., & Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research* **14**, 3321–3363.

hao, T.F., Chen, W.N., Kwong, S., Gu, T.L., Yuan, H.Q., Zhang, J., & Zhang, J. (2020). Evolutionary divide-and-conquer algorithm for virus spreading control over networks. *IEE Transactions on Cybernetics* **99**, 1–15.

APPENDIX

We first state several lemmas.

Lemma A.1. *Suppose that Assumptions (A1) and (A2) hold. Then, for* $k = 1, \ldots, K$*, we have* $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k} \xrightarrow{P} \boldsymbol{\Sigma}_0$ *and* $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}} \xrightarrow{P} \boldsymbol{\Sigma}_0$.

The first result appears in Stute (1993) and the second result is directly deduced by noting that $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}} = K^{-1} \sum_{k=1}^{K} \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}$.

Lemma A.2. *Suppose that Assumptions (A1)–(A5) hold. Then we have*

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_{\boldsymbol{W}} \boldsymbol{\Sigma}_0^{-1}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ *is defined in Theorem 2.*

F i. rst, we consider the case with fixed $K < \infty$. For all $k = 1, \ldots, K$, Stute (1996) proved that $\sqrt{n/K}(\widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_{\boldsymbol{W}} \boldsymbol{\Sigma}_0^{-1})$. Note that $\widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k}$ $(k = 1, \ldots, K)$ are mutually independent and $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} = K^{-1} \sum_{k=1}^{K} \widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k}$. Direct calculations entail that

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_{\boldsymbol{W}} \boldsymbol{\Sigma}_0^{-1}).$$

Next, we consider the case $K \to \infty$ as $n \to \infty$. For $k = 1, \ldots, K$, let $\omega_{\mathcal{I}_k, i}$, $Y_{\mathcal{I}_k,(i)}$ and $\boldsymbol{X}_{\mathcal{I}_k,(i)}$ represent ω_i, $Y_{(i)}$ and $\boldsymbol{X}_{(i)}$ for the kth dataset $\mathcal{D}_k$, respectively.

Then $\widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k} - \boldsymbol{\beta}_0$ can be written as

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k} - \boldsymbol{\beta}_0 = \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1} \sum_{i=1}^{n^*} \omega_{\mathcal{I}_k,i} \boldsymbol{X}_{\mathcal{I}_k,(i)} (Y_{\mathcal{I}_k,(i)} - \boldsymbol{X}_{\mathcal{I}_k,(i)}^\top \boldsymbol{\beta}_0) = \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1} \boldsymbol{M}_{\mathcal{I}_k}. \quad (1)$$

Stute (1996) stated that $E(\boldsymbol{M}_{\mathcal{I}_k})$ decreases to zero at any polynomial rate and

$$\lim_{n\to\infty} \mathrm{Var}(\sqrt{n^*} \boldsymbol{M}_{\mathcal{I}_k}) = \boldsymbol{\Sigma}_{\boldsymbol{W}}.$$

Further, we can obtain $E(\boldsymbol{M}_{\mathcal{I}_k}) = o(n^{-1/2})$. Since $\boldsymbol{M}_{\mathcal{I}_k} (k = 1, \ldots, K)$ are mutually independent, it follows from the Central Limit Theorem that

$$\sqrt{n} \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{M}_{\mathcal{I}_k} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_{\boldsymbol{W}} \boldsymbol{\Sigma}_0^{-1}).$$

Coupled with Lemma 19.24 in van der Vaart (1998) and the fact $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k} \xrightarrow{P} \boldsymbol{\Sigma}_0$ stated in Lemma A.1 we have

$$\sqrt{n} \frac{1}{K} \sum_{k=1}^{K} \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}_k}^{-1} \boldsymbol{M}_{\mathcal{I}_k} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_{\boldsymbol{W}} \boldsymbol{\Sigma}_0^{-1}). \quad (2)$$

Combining (1), (2), and $\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} = K^{-1} \sum_{k=1}^{K} \widetilde{\boldsymbol{\beta}}_{\mathcal{I}_k}$, we have

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Sigma}_{\boldsymbol{W}} \boldsymbol{\Sigma}_0^{-1}).$$

Thus, the proof of Lemma A.2 is completed. ■

Proof of Theorem 1. Note that $Q_n^\dagger(\boldsymbol{\beta})$ is a strictly convex function in $\boldsymbol{\beta}$, thus we only need to show that $Q_n^\dagger(\boldsymbol{\beta})$ has a $\sqrt{n}$-consistent local minimizer. By Fan & Li (2001), we should verify that for any $\varepsilon > 0$, there exists a sufficiently large constant C such that

$$P\left\{ \inf_{\boldsymbol{u}\in\mathbb{R}^p, \|\boldsymbol{u}\|=C} Q_n^\dagger(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{u}) > Q_n^\dagger(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon, \quad (3)$$

where $\|\cdot\|$ is the Euclidean norm. Let $\boldsymbol{u} = (u_0, u_1, \ldots, u_p)^\top$. For $\|\boldsymbol{u}\| = C$, Taylor's expansion entails

$$
\begin{aligned}
& n\{Q_n^\dagger(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{u}) - Q_n^\dagger(\boldsymbol{\beta}_0)\} \\
&= -2\boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\{\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0)\} + \boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\boldsymbol{u} + n\lambda \sum_{j=0}^{p} \frac{1}{|\widetilde{\beta}_{\mathrm{DAC},j}|}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \\
&\geq \boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\boldsymbol{u} - 2\|\boldsymbol{u}\| \left\| \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0) \right\| + n\lambda \sum_{j=0}^{d_0} \frac{1}{|\widetilde{\beta}_{\mathrm{DAC},j}|}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \\
&\geq \boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\boldsymbol{u} - 2\|\boldsymbol{u}\| \left\| \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0) \right\| - c_1\sqrt{n}\lambda\|\boldsymbol{u}\| \\
&\geq \boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\boldsymbol{u} - 2C \left\| \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0) \right\| - c_1\sqrt{n}\lambda C \\
&= \eta_1 - \eta_2 - \eta_3,
\end{aligned}
$$

where c_1 is a positive constant. According to Lemma A.1, we obtain that with probability converging to 1, $\eta_1 \geq \rho_{\min}(\boldsymbol{\Sigma}_0)\|\boldsymbol{u}\|_2^2 > \frac{1}{2}\rho_{\min}(\boldsymbol{\Sigma}_0)C^2$, where $\rho_{\min}(\boldsymbol{\Sigma}_0) > 0$ is the minimal eigenvalue of $\boldsymbol{\Sigma}_0$. On the other hand, Lemma A.2 implies that $\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0)\| = O_P(1)$. Thus we have $\eta_2 = O_P(1)C$. As $\sqrt{n}\lambda \to 0$, we obtain that $\eta_3 = o(1)C$. It is clear that for a sufficiently large C, η_1 dominates η_2 and η_3 with probability converging to 1, which implies that Equation (3) holds. Thus $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$.

Proof of Theorem 2. We first prove the selection consistency of the proposed estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$. If $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},j} \neq 0$ for some $d_0 < j \leq p$, then we have

$$
\sqrt{n}\,\frac{\partial Q_n^\dagger(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}} = 2\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{(j)}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} - \widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}\right) + \sqrt{n}\lambda\frac{\mathrm{sgn}(\widehat{\beta}_{\mathrm{DAC},j})}{|\widetilde{\beta}_{\mathrm{DAC},j}|},
$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{(j)}$ is the j-th row of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}$. By Theorem 1 and Lemma A.2, we have $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} - \widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} = O_P(n^{-1/2})$. Combined with Lemma A.1, it follows that $\widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}^{(j)}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}} - \widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}}\right) = O_P(1)$. On the other hand, from $n\lambda \to \infty$ and the fact $\widetilde{\boldsymbol{\beta}}_{DAC} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$ as stated in Lemma A.2, we have

$$
\sqrt{n}\lambda\frac{\mathrm{sgn}(\widehat{\beta}_{\mathrm{DAC},j})}{|\widetilde{\beta}_{\mathrm{DAC},j}|} = \frac{n\lambda}{|\sqrt{n}\widetilde{\beta}_{\mathrm{DAC},j}|}\mathrm{sgn}(\widehat{\beta}_{\mathrm{DAC},j}) \to \infty
$$

holds in probability. Therefore, with probability converging to 1, $\sqrt{n}\,\frac{\partial Q_n^\dagger(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}} \neq 0$, which implies $P\left(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},j} = 0\right) \to 1$ for all $d_0 < j \leq p$.

Thus, we complete the proof of the first part of Theorem 2.

Now we turn to show the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\mathrm{DAC}}$. Let $\boldsymbol{u} = (u_0, u_1, \ldots, u_p)^\top$ and $S_n(\boldsymbol{u}) = n\{Q_n^\dagger(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{u}) - Q_n^\dagger(\boldsymbol{\beta}_0)\}$. By Taylor's expansion, we have

$$\begin{aligned} S_n(\boldsymbol{u}) = &-2\boldsymbol{u}^\top \widehat{\Sigma}_{\mathrm{DAC}}\{\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0)\} + \boldsymbol{u}^\top \widehat{\Sigma}_{\mathrm{DAC}}\boldsymbol{u} \\ &+n\lambda \sum_{j=0}^{p} \frac{1}{|\widetilde{\beta}_{\mathrm{DAC},j}|}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|). \end{aligned}$$

It follows from Lemma A.1, Lemma A.2 and Slutsky's theorem that

$$\boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\boldsymbol{u} \xrightarrow{P} \boldsymbol{u}^\top \boldsymbol{\Sigma}_0 \boldsymbol{u} \tag{4}$$

and

$$2\boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{DAC}}\{\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{\mathrm{DAC}} - \boldsymbol{\beta}_0)\} \xrightarrow{d} 2\boldsymbol{u}^\top \boldsymbol{W}, \tag{5}$$

where $\boldsymbol{W}$ is defined in Theorem 2. If $0 \le j \le d_0$ (i.e., $\beta_{0j} \neq 0$), by Lemma A.2, we have

$$\frac{\sqrt{n}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|)}{|\widetilde{\beta}_{\mathrm{DAC},j}|} \xrightarrow{P} u_j \mathrm{sgn}(\beta_{0j}),$$

which implies

$$n\lambda \frac{|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|}{|\widetilde{\beta}_{\mathrm{DAC},j}|} = \sqrt{n}\lambda \frac{\sqrt{n}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|)}{|\widetilde{\beta}_{\mathrm{DAC},j}|} \xrightarrow{P} 0$$

using $\sqrt{n}\lambda \to 0$. Corresponding to $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,a}^\top, \boldsymbol{\beta}_{0,b}^\top)^\top$, we write $\boldsymbol{u} = (\boldsymbol{u}_a^\top, \boldsymbol{u}_b^\top)^\top$. If $d_0 < j \le p$, Lemma A.2 and the assumption $n\lambda \to \infty$ in the statement of Theorem 2 imply that

$$n\lambda \frac{|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|}{|\widetilde{\beta}_{\mathrm{DAC},j}|} = n\lambda \frac{|u_j|}{|\sqrt{n}\widetilde{\beta}_{\mathrm{DAC},j}|} = \begin{cases} 0, & u_j = 0, \\ +\infty, & u_j \neq 0. \end{cases}$$

Therefore we conclude that

$$n\lambda \sum_{j=0}^{p} \frac{1}{|\widetilde{\beta}_{\mathrm{DAC},j}|}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) = \begin{cases} 0, & \boldsymbol{u}_b = \boldsymbol{0}, \\ +\infty, & \text{otherwise}. \end{cases} \tag{6}$$

Define

$$S(\boldsymbol{u}) = \begin{cases} \boldsymbol{u}_a^\top \boldsymbol{\Sigma}_{11}\boldsymbol{u}_a - 2\boldsymbol{u}_a^\top \boldsymbol{W}_a, & \boldsymbol{u}_b = \boldsymbol{0}, \\ +\infty, & \text{otherwise}, \end{cases}$$

where $\boldsymbol{W}_a$ is defined in Theorem 2. Using Equations (4)–(6) and Slutsky's theorem, we have

$$S_n(\boldsymbol{u}) \xrightarrow{d} S(\boldsymbol{u}).$$

Note that $(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{W}_a, \mathbf{0})^\top$ is the unique minimizer of $S(\boldsymbol{u})$. By the argmax continuous mapping theorem (van der Vaart & Wellner, 1996), we conclude that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{DAC},a} - \boldsymbol{\beta}_{0,a}) \xrightarrow{d} \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{W}_a,$$

which completes the proof of the second part of Theorem 2.

Received 9 July 2009
Accepted 8 July 2010