Statistica Sinica Preprint No: SS-2025-0283	
Title	A Primal Dual Active Set with Continuation Algorithm for ℓ_0-penalized High-dimensional Accelerated Failure Time Model
Manuscript ID	SS-2025-0283
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0283
Complete List of Authors	Peili Li, Ruoying Hu, Yanyun Ding and Yunhai Xiao
Corresponding Authors	Yunhai Xiao
E-mails	yhxiao@henu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

A PRIMAL DUAL ACTIVE SET WITH CONTINUATION METHOD FOR ℓ_0-PENALIZED ACCELERATED FAILURE TIME MODEL

Peili Li, Ruoying Hu, Yanyun Ding and Yunhai Xiao

Henan University and Shenzhen Polytechnic University

Abstract: The accelerated failure time model has garnered attention due to its intuitive linear regression interpretation and has been successfully applied in fields such as biostatistics, clinical medicine, economics, and social sciences. This paper considers a weighted least squares estimation method with an ℓ_0-penalty based on right-censored data in a high-dimensional setting. For practical implementation, we adopt an efficient primal dual active set algorithm and utilize a continuous strategy to select the appropriate regularization parameter. By employing the mutual incoherence property and restricted isometry property of the covariate matrix, we perform an error analysis for the estimated variables in the active set during the iteration process. Furthermore, we identify a distinctive monotonicity in the active set and show that the algorithm terminates at the oracle solution in a finite number of steps. Finally, we perform extensive numerical experiments using both simulated data and real breast cancer datasets to assess the performance benefits of our method in comparison to other existing approaches.

Key words and phrases: High-dimensional accelerated failure time model, ℓ_0-penalty, oracle solution, primal dual active set with continuation algorithm, weighted least squares method.

1. Introduction

In research fields, such as mathematics, statistics, biology, and economics, survival analysis is commonly used to explore the mechanisms behind event occurrence or failure. In these fields, it often involves investigating various factors influencing the subjects to

identify the key ones, which allows us to assess and predict when events are likely to happen. In recent years, a variety of survival models have been introduced, including the Bayesian parametric survival model, Cox proportional hazards model, accelerated failure time (AFT) model, and support vector machine model. Among these models, the AFT model establishes a linear relationship between the logarithm or a known monotonic transformation of the failure time and the covariates. This linear regression approach contributes to its broader applicability compared to other models. Mathematically, the AFT model can be expressed as

$$\ln(T_i) = X_i^\top \beta^* + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where T_i is the failure time, $X_i \in \mathbb{R}^p$ is the covariate vector, $\beta^* \in \mathbb{R}^p$ is the regression coefficient, and ϵ_i's are the independent and identically distributed (i.i.d.) random errors. Here, we assume that T_i is right-censored, meaning the starting time of observation is known, but the event termination time is unknown, which prevents us from determining the exact survival time. But we only know that the survival time is greater than the observation time. At this point, we can obtain the observed data $\{(Y_i, \delta_i, X_i)\}_{i=1}^n$, where $Y_i = \min\{\ln(T_i), \ln(C_i)\}$, C_i is the censoring time and $\delta_i = \mathbf{1}_{\{T_i \le C_i\}}$ is the censoring indicator.

It is well known that there are several methods available to estimate the coefficient β^* appeared in (1.1), including the Buckley-James method (Buckley and James, 1979), rank-based method (Ying, 1993) and weighted least-squares method (Stute, 1996; Stute and Wang, 1993). In this paper, we primarily focus on the estimation and variable selection in high-dimensional settings, where the sample size n is assumed to be much smaller than the dimension p. In such cases, sparse penalty is often employed, which adds a

penalty term to the traditional loss function. For example, Johnson (2008) and Johnson, Lin and Zeng (2008) considered smoothly clipped absolute deviation penalty (Fan and Li, 2001) to the Buckley-James estimator. Cai, Huang and Tian (2009) introduced an adaptive lasso method (Zou, 2006) based on rank-based estimation. Hu and Chai (2013), Huang and Ma (2010), Huang, Ma and Xie (2006), and Khan and Shaw (2016) extended lasso (Tibshirani, 1996), adaptive weighted elastic net (Hong and Zhang, 2010; Zou and Zhang, 2009), bridge penalty (Frank and Friedman, 1993) and minimax concave penalty (Zhang, 2010) to weighted least-squares estimation, respectively. Specifically, Cheng et al. (2022) extended the non-convex, non-smooth ℓ_0-penalty to weighted least-squares estimation. The ℓ_0-penalty is well-suited for sparse estimation, as it effectively mitigates overfitting and enhances interpretability. Additionally, Cheng et al. (2022) extended the support detection and root finding algorithm (Huang et al., 2018) and derived ℓ_∞-error bounds for the solution sequence. Under specific conditions, they demonstrated that the estimated support accurately includes the true support within a finite number of steps.

In this paper, we focus on the high-dimensional AFT model with an ℓ_0-penalty. We observe that the challenges presented by high dimensionality, heavy censoring, and non-convex, non-smooth characteristics greatly complicate the model estimation and variable selection. For practical implementation, in this paper, we extend the efficient primal dual active set with continuation (PDASC) algorithm proposed by Jiao, Jin and Lu (2015), referred to as AFT-PDASC. In each iteration, the active and inactive sets are initially identified based on the primal and dual variables from the previous iteration. Both the primal variable over the active set and dual variable can be explicitly updated by utilizing the corresponding KKT conditions. Furthermore, a continuation strategy for the regularization parameter is incorporated to achieve algorithm's convergence. We observe that the active set exhibits a special monotonicity in each iteration, where it gradually grows

and eventually matches the true active set as the regularization parameter decreases. We also analyze the ℓ_2- and ℓ_∞- error bounds of the solution sequence on the active set, assuming the mutual incoherence property and restricted isometry property of the covariate matrix. In addition, we prove that the iterative process of this algorithm terminates in a finite number of steps at the oracle solution. Finally, we evaluate the estimation performance of the AFT-PDASC algorithm under the high-dimensional AFT model using both simulated and real data. In extensive comparisons with other existing algorithms, the AFT-PDASC algorithm consistently yields more stable and efficient results.

Organization: The remainder of this paper is organized as follows. In Section 2, we present the ℓ_0-penalized AFT model, then transform it into a standard least squares estimation model through some simple algebraic manipulations, and finally list the iterative framework of AFT-PDASC. In Section 3, we present some relevant theoretical results, which includes error analysis, the monotonicity analysis of the active set, and the finite-step termination property of AFT-PDASC. In Section 4, we design some simulation and real data experiments to evaluate the numerical performance of AFT-PDASC. Finally, in Section 5, we conclude this paper.

Notation: For a vector $\beta \in \mathbb{R}^p$, we denote $\|\beta\| := \sqrt{\sum_{i=1}^p \beta_i^2}$ and $\|\beta\|_\infty := \max_{1\leq i\leq p}\{|\beta_i|\}$. The symbol $\|\beta\|_{T,\infty}$ is the T-th largest elements (in absolute value), $\|\beta\|_0$ denotes the number of non-zero elements in β and $\text{supp}(\beta) := \{i \mid \beta_i \neq 0, i = 1, 2, \ldots, p\}$. For an index set $\mathcal{A}$, we denote $|\mathcal{A}|$ as its length and $\mathcal{A}^c$ is the complement. In addition, we also denote $\beta_\mathcal{A} := \{\beta_i, i \in \mathcal{A}\} \in \mathbb{R}^{|\mathcal{A}|}$. For a matrix $X \in \mathbb{R}^{n\times p}$, we use $\|X\|_\infty$ to represent its maximum value (in absolute value) and use the notation $X_\mathcal{A} \in \mathbb{R}^{n\times|\mathcal{A}|}$. We use β^* and $\hat{\beta}$ to denote the true and the estimated regression coefficient, respectively. We assume the true coefficient β^* has K non-zero components along with its support set denoted as $\mathcal{A}^*$, i.e., $|\mathcal{A}^*| = K$. Finally, we denote $\mathcal{S} =: \{1, 2, \ldots, p\}$ and let $\mathcal{I}^* := \mathcal{S} \setminus \mathcal{A}^*$.

2. AFT-PDASC algorithm for ℓ_0-penalized AFT model

2.1 ℓ_0-penalized AFT model

Let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ be the order statistics of Y_i's, $\delta_{(1)}, \ldots, \delta_{(n)}$ be the associated censoring indicators, and $X_{(1)}, \ldots, X_{(n)}$ be the associated covariates. The weighted least-squares estimation method with an ℓ_0-penalty for AFT model (1.1) is given by:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^{n} w_{(i)} \big(Y_{(i)} - X_{(i)}^\top \beta\big)^2 + \lambda \|\beta\|_0, \tag{2.1}$$

where $w_{(i)}$'s are the jumps in Kaplan-Meier estimator in the form of $w_{(1)} = \delta_{(1)}/n$ and that

$$w_{(i)} = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \big(\frac{n-j}{n-j+1}\big)^{\delta_{(j)}}, \quad i = 2, \ldots, n,$$

and $\lambda > 0$ is a regularization parameter to control the sparsity level of β. For convenience, we make a standardization on (2.1). Define

$$\tilde{Y} := \Big(\sqrt{w_{(1)}} Y_{(1)}, \ldots, \sqrt{w_{(n)}} Y_{(n)}\Big)^\top \quad \text{and} \quad \tilde{X} := \Big(\sqrt{w_{(1)}} X_{(1)}, \ldots, \sqrt{w_{(n)}} X_{(n)}\Big)^\top.$$

Assume that $\|\tilde{X}_i\|_2 > 0$, where $\tilde{X}_i \in \mathbb{R}^n$ is the i-th column of $\tilde{X}$. Define

$$D := \operatorname{diag}\Big(\frac{1}{\|\tilde{X}_1\|_2}, \ldots, \frac{1}{\|\tilde{X}_p\|_2}\Big),$$

and denote $\eta := D^{-1}\beta$, $\bar{X} := \tilde{X} D$, $\bar{Y} := \tilde{Y}$. Furthermore, let $\bar{Y} := \bar{X}\eta^* + \gamma$, where γ is an error satisfying $\|\gamma\| \leq \bar{\epsilon}$, and $\bar{\epsilon} \geq 0$ is a noise level. Using these notations, we can rewrite (2.1) as the following ℓ_0-penalized least-squares problem:

$$\min_{\eta \in \mathbb{R}^p} L(\eta) := \frac{1}{2} \|\bar{Y} - \bar{X}\eta\|^2 + \lambda \|\eta\|_0. \tag{2.2}$$

In the following subsection, we will develop an algorithm to solve (2.2) to obtain the solution $\hat{\eta}$. Subsequently, the estimated coefficient $\hat{\beta}$ of (1.1) can be computed by using $\hat{\beta} = D\hat{\eta}$.

2.2 AFT-PDASC algorithm

Due to the non-convex and non-smooth nature of the objective function in (2.2), we focus on a coordinate-wise minimization approach. A vector $\eta = (\eta_1, \ldots, \eta_p)^\top \in \mathbb{R}^p$ is said to be a coordinate-wise minimizer of $L(\eta)$ if it is a local minimizer along each coordinate direction, that is,

$$\eta_i \in \arg\min_{t\in\mathbb{R}} L\big(\eta_1, ..., \eta_{i-1}, t, \eta_{i+1}, ..., \eta_p\big).$$

The following lemma presents the KKT conditions for the coordinate-wise minimizers of $L(\eta)$, which plays a crucial role in the development of our subsequent algorithm. For the proof, one can refer to Huang et al. (2021) and Li et al. (2022).

Lemma 1. *If $\bar{\eta}$ is a minimizer of (2.2), then there exists a $\bar{d}$ such that the following KKT system holds:*

$$\begin{cases} \bar{d} = \bar{X}^\top(\bar{Y} - \bar{X}\bar{\eta}), \\ \bar{\eta} = S_\lambda(\bar{\eta} + \bar{d}), \end{cases} \tag{2.3}$$

where the i-th element of $S_\lambda(x)$ is defined as

$$\Big(S_\lambda(x)\Big)_i \begin{cases} = 0, & if\ |x_i| < \sqrt{2\lambda}, \\ \in \{0, x_i\}, & if\ |x_i| = \sqrt{2\lambda}, \\ = x_i, & if\ |x_i| > \sqrt{2\lambda}. \end{cases} \tag{2.4}$$

Conversely, if $\bar{\eta}$ and $\bar{d}$ satisfy (2.3), then $\bar{\eta}$ is a local minimizer of model (2.2).

Let $\bar{\mathcal{A}} := \text{supp}(\bar{\eta})$ and $\bar{\mathcal{I}} := \bar{\mathcal{A}}^c$. Thus, it can be seen from (2.3) and (2.4) that

$$\bar{\mathcal{A}} := \left\{ i : |\bar{\eta}_i + \bar{d}_i| > \sqrt{2\lambda} \right\}, \quad \bar{\mathcal{I}} := \left\{ i : |\bar{\eta}_i + \bar{d}_i| \leq \sqrt{2\lambda} \right\}, \tag{2.5}$$

and that

$$\bar{\eta}_{\bar{\mathcal{I}}} = \mathbf{0}, \quad \bar{d}_{\bar{\mathcal{A}}} = \mathbf{0}, \quad \bar{\eta}_{\bar{\mathcal{A}}} = \left(\bar{X}_{\bar{\mathcal{A}}}^\top \bar{X}_{\bar{\mathcal{A}}} \right)^{-1} \bar{X}_{\bar{\mathcal{A}}}^\top \bar{Y}, \quad \bar{d}_{\bar{\mathcal{I}}} = \bar{X}_{\bar{\mathcal{I}}}^\top \left(\bar{Y} - \bar{X}_{\bar{\mathcal{A}}} \bar{\eta}_{\bar{\mathcal{A}}} \right). \tag{2.6}$$

Based on (2.5) and (2.6), we can iteratively derive $\bar{\eta}$ and $\bar{d}$ by applying a specific rule. Typically, for the active set $\mathcal{A}^k$, we can compute the primal variable η^k and dual variable d^k using the KKT conditions. The latest β^k can also be derived using the relation $\beta^k = D\eta^k$. This process is repeated until the active sets in successive steps become consistent, or the maximum number of iterations is reached.

It is important to note that the primal-dual active set method is essentially equivalent to the well-known semismooth Newton method, which demonstrates local superlinear convergence, see Hintermüller, Ito and Kunisch (2002) for more details. However, to fully leverage this characteristic, a good initial guess is often required. In this paper, we employ a continuation technique for the regularization parameter λ to address this issue. Specifically, given an initial value λ_0, we set $\lambda_k = \lambda_0 \rho^k$ with $\rho \in (0, 1)$. Based on λ_k, we iterate the above inner loop to obtain $\eta(\lambda_k)$ and $d(\lambda_k)$. Next, we use $\eta(\lambda_k)$ and $d(\lambda_k)$ as the initial values for the corresponding problem with λ_{k+1} and continue the iterative process until a discrete analogue of the discrepancy principle is satisfied.

Based on the above analysis, we outline the iterative framework of the AFT-PDASC below.

AFT-PDASC Algorithm

1. Choose $\lambda_0 \geq \frac{1}{2}\|\bar{X}^\top \bar{Y}\|_\infty^2$ and define $\mathcal{A}(\lambda_0) := \varnothing$ and $\eta(\lambda_0) := \mathbf{0}$. Let $d(\lambda_0) := \bar{X}^\top \bar{Y}$ and choose $\rho \in (0,1)$ and $\bar{\epsilon} > 0$. Choose positive integers $K_{\max} \in \mathcal{N}$ and $J_{\max} \in \mathcal{N}$.

2. for $k = 1, 2, \ldots, K_{\max}$ **do**

3. Let $\lambda_k := \rho\lambda_{k-1}$, $\mathcal{A}_0 := \mathcal{A}(\lambda_{k-1})$, and $(\eta^0, d^0) := (\eta(\lambda_{k-1}), d(\lambda_{k-1}))$.

4. **for** $j = 1, 2, \ldots, J_{\max}$ **do**

5. Compute $\mathcal{A}_j = \{i : |\eta_i^{j-1} + d_i^{j-1}| > \sqrt{2\lambda_k}\}$ and let $\mathcal{I}_j = \mathcal{A}_j^c$;

6. Break if $\mathcal{A}_j = \mathcal{A}_{j-1}$;

7. Otherwise, compute

$$\eta_{\mathcal{I}_j}^j := \mathbf{0}, \quad d_{\mathcal{A}_j}^j := \mathbf{0},$$

$$\eta_{\mathcal{A}_j}^j := (\bar{X}_{\mathcal{A}_j}^\top \bar{X}_{\mathcal{A}_j})^{-1} \bar{X}_{\mathcal{A}_j}^\top \bar{Y}, \quad d_{\mathcal{I}_j}^j := \bar{X}_{\mathcal{I}_j}^\top (\bar{Y} - \bar{X}_{\mathcal{A}_j} \eta_{\mathcal{A}_j}^j).$$

8. Compute $\beta^j := D\eta^j$.

9. **end for**

10. Choose $\tilde{j} = \min(J_{\max}, j)$. Compute

$$\mathcal{A}(\lambda_k) := \left\{i : |\eta_i^{\tilde{j}} + d_i^{\tilde{j}}| > \sqrt{2\lambda_k}\right\} \text{ and } \left(\eta(\lambda_k), d(\lambda_k)\right) := (\eta^{\tilde{j}}, d^{\tilde{j}}).$$

11. Compute $\beta(\lambda_k) := D\eta(\lambda_k)$.

12. Stop if $\|\bar{Y} - \bar{X}\eta(\lambda_k)\| \leq \bar{\epsilon}$. Otherwise, continue.

13. end for

It is important to highlight that Cheng et al. (2022) also proposed an algorithm based on the KKT system for the coordinate-wise minimizers of model (2.2), but there

are several key differences compared to AFT-PDASC. Theoretically, they derived high-probability results under the assumption that the errors are subgaussian. In contrast, we obtain deterministic results under the assumption that the errors are deterministic perturbations. In computing the active set, Cheng et al. used $\|\eta+\tau d\|_{T,\infty}$ to eliminate the parameter λ, but this approach relies on a new parameter T. Additionally, a step size τ is also required for computing the active set, which makes both the theoretical analysis and numerical performance highly dependent on this step size. In summary, the performance of the algorithm proposed by Cheng et al. (2022) heavily depends on the selection of the parameters T and τ, which cannot be adaptively selected and can be a challenging task in practical implementation. In contrast, our approach directly uses the regularization parameter λ and employs a continuation technique to achieve adaptive selection. This not only avoids the tedious process of parameter tuning, but also eliminates dependence on sensitive parameters. Most importantly, the numerical comparisons with Cheng et al. (2022) demonstrate that the algorithm presented in this paper can consistently achieve high-precision solutions generally.

3. Theoretical analysis

In this section, we derive an error analysis for the estimated coefficients on the active set during the iterative process, leveraging the mutual incoherence property (MIP) and the restricted isometry property (RIP) of the covariate matrix $\bar{X}$. Additionally, we provide a finite-step termination analysis of AFT-PDASC at the oracle solution by using the special monotonicity of the active set. For simplicity, we let $\lambda > 0$ and $s > 0$, and define $\mathcal{T}_{\lambda,s} := \{i : |\eta_i^*| \geq \sqrt{2\lambda}s\}$ to identify the indices of η^* where the values are relatively large. In addition, we also define the least-squares solution on the true support set $\mathcal{A}^*$ as the oracle solution $\eta^o \in \mathbb{R}^{|\mathcal{A}^*|}$, which obeys the form $\eta^o := (\bar{X}_{\mathcal{A}^*}^\top \bar{X}_{\mathcal{A}^*})^{-1}\bar{X}_{\mathcal{A}^*}^\top \bar{Y}$.

For the purpose of the subsequent theoretical analysis, we need a couple of assumptions on the covariate matrix $\bar{X}$:

Assumption 1. *The following conditions hold:*

(1) The mutual coherence $\nu = \max_{1\le i\neq j\le p}\left\{|\bar{X}_i^\top \bar{X}_j|\right\}$ *of the covariate matrix* $\bar{X}$ *is small, where* $\bar{X}_i$ *is the* i*-th column of* $\bar{X}$*.*

(2) There exists a constant $\delta \in (0,1)$ *such that* $(1-\delta)\|\eta\|^2 \le \|\bar{X}\eta\|^2 \le (1+\delta)\|\eta\|^2$ *for any* $\eta \in \mathbb{R}^p$ *satisfying* $\|\eta\|_0 \le s$*. Here, we define* δ_s *as the infimum of all parameters* δ *for which the RIP holds.*

The following lemma provides some basic estimates under the MIP and RIP conditions. One may refer to Jiao, Jin and Lu (2015), Dai and Milenkovic (2009) and Needell and Tropp (2009) for its proof.

Lemma 2. *Assume* $\mathcal{A}$ *and* $\mathcal{B}$ *are disjoint subsets of* $\mathcal{S}$*. Then*

1.

$$\|\bar{X}_{\mathcal{A}}^\top \bar{Y}\|_\infty \le \|\bar{Y}\|, \quad \|\bar{X}_{\mathcal{B}}^\top \bar{X}_{\mathcal{A}}\eta_{\mathcal{A}}\|_\infty \le \nu|\mathcal{A}|\|\eta_{\mathcal{A}}\|_\infty,$$
$$\|(\bar{X}_{\mathcal{A}}^\top \bar{X}_{\mathcal{A}})^{-1}\eta_{\mathcal{A}}\|_\infty \le \frac{\|\eta_{\mathcal{A}}\|_\infty}{1-(|\mathcal{A}|-1)\nu}, \ if \ (|\mathcal{A}|-1)\nu < 1.$$

2.

$$\|\bar{X}_{\mathcal{A}}^\top \bar{X}_{\mathcal{A}}\eta_{\mathcal{A}}\| \gtrless (1\mp\delta_{|\mathcal{A}|})\|\eta_{\mathcal{A}}\|, \quad \|(\bar{X}_{\mathcal{A}}^\top \bar{X}_{\mathcal{A}})^{-1}\eta_{\mathcal{A}}\| \gtrless \frac{\|\eta_{\mathcal{A}}\|}{1\pm\delta_{|\mathcal{A}|}},$$
$$\|\bar{X}_{\mathcal{A}}^\top \bar{X}_{\mathcal{B}}\| \le \delta_{|\mathcal{A}|+|\mathcal{B}|}, \quad \|(\bar{X}_{\mathcal{A}}^\top \bar{X}_{\mathcal{A}})^{-1}\bar{X}_{\mathcal{A}}^\top \bar{Y}\| \le \frac{\|\bar{Y}\|}{\sqrt{1-\delta_{|\mathcal{A}|}}}, \quad \delta_s \le \delta_{s'}, \ if \ s < s'.$$

For the subsequent theoretical analysis, we also require the following assumption on the noise level $\bar{\epsilon}$.

Assumption 2. *The noise level $\bar{\epsilon}$ is small in sense that $\bar{\epsilon} \leq \alpha \min_{i\in\mathcal{A}^*}\{|\eta_i^*|\}$ for a certain one $0 \leq \alpha < 1/2$.*

We now provide the main theoretical results of the estimation method (2.2) based on the MIP condition in Assumption 1 (1) on the covariate matrix $\bar{X}$.

Theorem 1. *Suppose that Assumption 2 holds and $\mathcal{A}(\lambda_{-1}) = \mathcal{A}(\lambda_0) := \emptyset$. If $\nu < \frac{1-2\alpha}{2K-1}$, then for any $\rho \in \big(((2K-1)\nu + 2\alpha)^2, 1\big)$, it holds that:*

(1) In the λ_k subproblem, for $j = 1, 2, \ldots, J_{\max}$, we set $\mathcal{B} := \mathcal{A}^*\backslash\mathcal{A}_j$ and $\mathcal{I}_j := \mathcal{S}\backslash\mathcal{A}_j$. If $|\mathcal{A}_j| \leq K$, then we have

$$\begin{aligned}
\big\|\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j}\big\|_\infty &\leq \frac{1}{1-(|\mathcal{A}_j|-1)\nu}\Big(\nu|\mathcal{B}|\cdot\|\eta^*_{\mathcal{B}}\|_\infty + \bar{\epsilon}\Big),\\
\big\|\beta^j_{\mathcal{A}_j} - \beta^*_{\mathcal{A}_j}\big\|_\infty &\leq \frac{\|D\|_\infty}{1-(|\mathcal{A}_j|-1)\nu}\Big(\nu|\mathcal{B}|\cdot\|D^{-1}\|_\infty\cdot\|\beta^*_{\mathcal{B}}\|_\infty + \bar{\epsilon}\Big),\\
\big\|\beta^j_{\mathcal{A}_j} - \beta^*_{\mathcal{A}_j}\big\| &\leq \frac{\sqrt{K}\|D\|_\infty}{1-(|\mathcal{A}_j|-1)\nu}\Big(\nu|\mathcal{B}|\cdot\|D^{-1}\|_\infty\cdot\|\beta^*_{\mathcal{B}}\|_\infty + \bar{\epsilon}\Big).
\end{aligned}$$

(2) For $k = 0, 1, 2, \ldots, K_{\max}$, there exist $s_1, s_2 \in \big(1/(1-K\nu+\nu-\alpha), 1/(K\nu+\alpha)\big)$ with $s_1 > s_2$, such that the active sets $\mathcal{A}(\lambda_k)$ have the special monotonicity, that is,

$$\text{If } \mathcal{T}_{\lambda_k,s_1} \subseteq \mathcal{A}(\lambda_{k-1}) \subseteq \mathcal{A}^*, \quad \text{then } \mathcal{T}_{\lambda_k,s_2} \subseteq \mathcal{A}(\lambda_k) \subseteq \mathcal{A}^*. \tag{3.1}$$

Furthermore, AFT-PDASC terminates in a finite number of steps.

(3) Let $\alpha \leq \frac{1-2(K-1)\nu}{K+3}$ and let $\xi := \frac{1-2(K-1)\nu-2\alpha-\alpha^2}{2K}\min_{i\in\mathcal{A}^*}\{|\eta_i^*|^2\}$. Then, for any $\lambda \in \big(\frac{\bar{\epsilon}^2}{2}, \xi\big)$, AFT-PDASC terminates at the oracle solution η^o.

Proof. (1) It can be seen from the iterative process of algorithm AFT-PDASC that $\eta^j_{\mathcal{A}_j} =$

$(\bar{X}_{\mathcal{A}_j}^\top \bar{X}_{\mathcal{A}_j})^{-1}\bar{X}_{\mathcal{A}_j}^\top \bar{Y}$. Recalling that $\bar{Y} = \bar{X}_{\mathcal{A}^*}\eta^*_{\mathcal{A}^*} + \gamma$, we have

$$\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j} = \Big(\bar{X}_{\mathcal{A}_j}^\top \bar{X}_{\mathcal{A}_j}\Big)^{-1}\bar{X}_{\mathcal{A}_j}^\top\Big(\bar{X}_{\mathcal{A}^*}\eta^*_{\mathcal{A}^*} + \gamma - \bar{X}_{\mathcal{A}_j}\eta^*_{\mathcal{A}_j}\Big) = \Big(\bar{X}_{\mathcal{A}_j}^\top \bar{X}_{\mathcal{A}_j}\Big)^{-1}\bar{X}_{\mathcal{A}_j}^\top\Big(\bar{X}_{\mathcal{B}}\eta^*_{\mathcal{B}} + \gamma\Big).$$

From the fact that $\nu < \frac{1-2\alpha}{2K-1}$, $|\mathcal{A}_j| \le K$, and $0 \le \alpha < \frac{1}{2}$, we can easily get that $(|\mathcal{A}_j| - 1)\nu < 1$. Then, using the inequalities in Lemma 2, we can further obtain that

$$\begin{aligned}\big\|\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j}\big\|_\infty &\le \frac{1}{1-(|\mathcal{A}_j|-1)\nu}\Big(\|\bar{X}_{\mathcal{A}_j}^\top \bar{X}_{\mathcal{B}}\eta^*_{\mathcal{B}}\|_\infty + \|\bar{X}_{\mathcal{A}_j}^\top\gamma\|_\infty\Big)\\ &\le \frac{1}{1-(|\mathcal{A}_j|-1)\nu}\Big(\nu|\mathcal{B}|\cdot\|\eta^*_{\mathcal{B}}\|_\infty + \bar{\epsilon}\Big).\end{aligned}$$

Additionally, based on the algebraic relationship between β and η, we have

$$\big\|\beta^j_{\mathcal{A}_j} - \beta^*_{\mathcal{A}_j}\big\|_\infty \le \|D\|_\infty\big\|\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j}\big\|_\infty \le \frac{\|D\|_\infty}{1-(|\mathcal{A}_j|-1)\nu}\Big(\nu|\mathcal{B}|\cdot\|D^{-1}\|_\infty\cdot\|\beta^*_{\mathcal{B}}\|_\infty + \bar{\epsilon}\Big),$$

and

$$\big\|\beta^j_{\mathcal{A}_j} - \beta^*_{\mathcal{A}_j}\big\| \le \sqrt{|\mathcal{A}_j|}\big\|\beta^j_{\mathcal{A}_j} - \beta^*_{\mathcal{A}_j}\big\|_\infty \le \frac{\sqrt{K}\|D\|_\infty}{1-(|\mathcal{A}_j|-1)\nu}\Big(\nu|\mathcal{B}|\cdot\|D^{-1}\|_\infty\cdot\|\beta^*_{\mathcal{B}}\|_\infty + \bar{\epsilon}\Big).$$

(2) From the fact $\nu < \frac{1-2\alpha}{2K-1}$, we know that $(2K-1)\nu + 2\alpha < 1$, which leads to $K\nu + \alpha < 1 - K\nu + \nu - \alpha$. Then, for any $s_1 \in \big(\frac{1}{1-K\nu+\nu-\alpha}, \frac{1}{K\nu+\alpha}\big)$, we have

$$s_1 > 1 + (K\nu - \nu + \alpha)s_1, \text{ and } 1 + (K\nu - \nu + \alpha)s_1 > \frac{1}{1-K\nu+\nu-\alpha},$$

i.e., $\frac{1}{1-K\nu+\nu-\alpha} < 1 + (K\nu - \nu + \alpha)s_1 < s_1 < \frac{1}{K\nu+\alpha}$. Let $s_2 := 1 + (K\nu - \nu + \alpha)s_1$, then we have that $s_1, s_2 \in (1/(1-K\nu+\nu-\alpha), 1/(K\nu+\alpha))$ and that $s_1 > s_2$. Define the function $f(s_1) := s_2/s_1 = \big(1 + (K\nu - \nu + \alpha)s_1\big)/s_1$, which has a monotonic decreasing

property within the interval $\big(\frac{1}{1-K\nu+\nu-\alpha}, \frac{1}{K\nu+\alpha}\big)$, then we have

$$f\Big(\frac{1}{K\nu+\alpha}\Big) < f(s_1) < f\Big(\frac{1}{1-K\nu+\nu-\alpha}\Big),$$

that is, $(2K-1)\nu+2\alpha < f(s_1) < 1$. This means that for any $\rho \in \big(((2K-1)\nu+2\alpha)^2, 1\big)$, there exists a s_1 within the interval $\big(\frac{1}{1-K\nu+\nu-\alpha}, \frac{1}{K\nu+\alpha}\big)$ such that $s_2/s_1 = \sqrt{\rho}$.

In light of these analysis, for $k = 0, 1, 2, \ldots, K_{\max}$, we now prove the following special monotonicity:

$$\text{If } \mathcal{T}_{\lambda_k, s_1} \subseteq \mathcal{A}(\lambda_{k-1}) \subseteq \mathcal{A}^*, \quad \text{then } \mathcal{T}_{\lambda_k, s_2} \subseteq \mathcal{A}(\lambda_k) \subseteq \mathcal{A}^*.$$

First, we consider the case of $k = 0$. From Lemma 2, it holds that

$$\|\eta^*\|_\infty = \|\eta^*_{\mathcal{A}^*}\|_\infty = \|(\bar{X}^\top_{\mathcal{A}^*}\bar{X}_{\mathcal{A}^*})^{-1}(\bar{X}^\top_{\mathcal{A}^*}\bar{X}_{\mathcal{A}^*})\eta^*_{\mathcal{A}^*}\|_\infty \le \frac{\|(\bar{X}^\top_{\mathcal{A}^*}\bar{X}_{\mathcal{A}^*})\eta^*_{\mathcal{A}^*}\|_\infty}{1-(K-1)\nu}.$$

Furthermore, it also holds that

$$\|\bar{X}^\top\bar{Y}\|_\infty \ge \|(\bar{X}^\top_{\mathcal{A}^*}\bar{X}_{\mathcal{A}^*})\eta^*_{\mathcal{A}^*}\|_\infty - \|\bar{X}^\top\gamma\|_\infty \ge \big(1-(K-1)\nu\big)\|\eta^*\|_\infty - \bar{\epsilon}.$$

Given $s_2 > \frac{1}{1-K\nu+\nu-\alpha}$, we have $1-(K-1)\nu > 1/s_2 + \alpha$. Additionally, using Assumption 2, we can obtain that $\|\bar{X}^\top\bar{Y}\|_\infty > \frac{1}{s_2}\|\eta^*\|_\infty$. This indicates that $\|\eta^*\|_\infty < s_2\|\bar{X}^\top\bar{Y}\|_\infty$. Using the fact that $\lambda_0 \ge \frac{1}{2}\|\bar{X}^\top\bar{Y}\|^2_\infty$ and the definition of $\mathcal{T}_{\lambda,s}$, we can get $\mathcal{T}_{\lambda_0,s_1} = \emptyset$ and $\mathcal{T}_{\lambda_0,s_2} = \emptyset$. Therefore, it holds that $\mathcal{T}_{\lambda_0,s_1} \subseteq \mathcal{A}(\lambda_{-1}) \subseteq \mathcal{A}^*$ and that $\mathcal{T}_{\lambda_0,s_2} \subseteq \mathcal{A}(\lambda_0) \subseteq \mathcal{A}^*$.

Assume that the monotonicity (3.1) holds when $k = l$, that is, "if $\mathcal{T}_{\lambda_l,s_1} \subseteq \mathcal{A}(\lambda_{l-1}) \subseteq \mathcal{A}^*$, then $\mathcal{T}_{\lambda_l,s_2} \subseteq \mathcal{A}(\lambda_l) \subseteq \mathcal{A}^*$". Then, we will show that this assertion also holds when $k = l+1$. From the relations $\lambda_{l+1} = \rho\lambda_l$ and $\rho = s_2^2/s_1^2$, we know that $\mathcal{T}_{\lambda_{l+1},s_1} = \mathcal{T}_{\lambda_l,s_2}$.

Based on the result for $k = l$, we have $\mathcal{T}_{\lambda_{l+1},s_1} \subseteq \mathcal{A}(\lambda_l) \subseteq \mathcal{A}^*$. Therefore, it remains to prove $\mathcal{T}_{\lambda_{l+1},s_2} \subseteq \mathcal{A}(\lambda_{l+1}) \subseteq \mathcal{A}^*$. Note that $\mathcal{A}(\lambda_l)$ and $\mathcal{A}(\lambda_{l+1})$ are essentially the initial guess and final output of the active set in the λ_{l+1}-problem, respectively. Therefore, we only need to demonstrate that the following assertion holds during the inner iterations of the λ_{l+1}-problem, i.e.,

$$\text{If } \mathcal{T}_{\lambda_{l+1},s_1} \subseteq \mathcal{A}_j \subseteq \mathcal{A}^*, \quad \text{then } \mathcal{T}_{\lambda_{l+1},s_2} \subseteq \mathcal{A}_{j+1} \subseteq \mathcal{A}^*. \tag{3.2}$$

From the relation $\eta^j_{\mathcal{A}_j} = \eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j} + \eta^*_{\mathcal{A}_j}$ for any $i \in \mathcal{A}_j$, we get

$$|\eta^j_i| \geq |\eta^*_i| - \|\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j}\|_\infty \geq |\eta^*_i| - \frac{\nu|\mathcal{B}| \cdot \|\eta^*_{\mathcal{B}}\|_\infty + \bar{\epsilon}}{1 - (|\mathcal{A}_j| - 1)\nu}.$$

Additionally, from the update formula of the dual variable d, it can be seen that

$$\begin{aligned} d^j_i &= \bar{X}_i^\top \Big(\bar{Y} - \bar{X}_{\mathcal{A}_j}\eta^j_{\mathcal{A}_j}\Big) = \bar{X}_i^\top \Big(\bar{X}_{\mathcal{A}^*}\eta^*_{\mathcal{A}^*} + \gamma - \bar{X}_{\mathcal{A}_j}\eta^j_{\mathcal{A}_j}\Big) \\ &= \bar{X}_i^\top \Big(\bar{X}_{\mathcal{A}^*}\eta^*_{\mathcal{A}^*} - \bar{X}_{\mathcal{A}_j}\eta^*_{\mathcal{A}_j} + \bar{X}_{\mathcal{A}_j}\eta^*_{\mathcal{A}_j} + \gamma - \bar{X}_{\mathcal{A}_j}\eta^j_{\mathcal{A}_j}\Big) \\ &= \bar{X}_i^\top \Big(\bar{X}_{\mathcal{B}}\eta^*_{\mathcal{B}} + \gamma - \bar{X}_{\mathcal{A}_j}(\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j})\Big). \end{aligned}$$

Then, for any $i \in \mathcal{B}$, using the result in (1), Lemma 2, and $\bar{X}_i^\top \bar{X}_i = 1$, we know that

$$\begin{aligned} |d^j_i| &= \Big|\bar{X}_i^\top \bar{X}_i \eta^*_i + \bar{X}_i^\top \Big(\bar{X}_{\mathcal{B}\setminus\{i\}}\eta^*_{\mathcal{B}\setminus\{i\}} + \gamma - \bar{X}_{\mathcal{A}_j}(\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j})\Big)\Big| \\ &\geq |\eta^*_i| - |\bar{X}_i^\top \bar{X}_{\mathcal{B}\setminus\{i\}}\eta^*_{\mathcal{B}\setminus\{i\}}| - |\bar{X}_i^\top \gamma| - |\bar{X}_i^\top \bar{X}_{\mathcal{A}_j}(\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j})| \\ &\geq |\eta^*_i| - (|\mathcal{B}| - 1)\nu\|\eta^*_{\mathcal{B}}\|_\infty - \bar{\epsilon} - |\mathcal{A}_j|\nu\|\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j}\|_\infty \\ &\geq |\eta^*_i| + \nu\|\eta^*_{\mathcal{B}}\|_\infty - |\mathcal{B}|\nu\Big(1 + \frac{|\mathcal{A}_j|\nu}{1 - (|\mathcal{A}_j| - 1)\nu}\Big)\|\eta^*_{\mathcal{B}}\|_\infty - \bar{\epsilon}\Big(1 + \frac{|\mathcal{A}_j|\nu}{1 - (|\mathcal{A}_j| - 1)\nu}\Big). \end{aligned} \tag{3.3}$$

Similarly, for any $i \in \mathcal{I}^*$, we have

$$\begin{aligned}|d_i^j| &\le |\bar{X}_i^\top \bar{X}_{\mathcal{B}} \eta_{\mathcal{B}}^*| + |\bar{X}_i^\top \gamma| + |\bar{X}_i^\top \bar{X}_{\mathcal{A}_j}(\eta_{\mathcal{A}_j}^j - \eta_{\mathcal{A}_j}^*)| \\ &\le |\mathcal{B}|\nu \|\eta_{\mathcal{B}}^*\|_\infty + \bar{\epsilon} + |\mathcal{A}_j|\nu \|\eta_{\mathcal{A}_j}^j - \eta_{\mathcal{A}_j}^*\|_\infty \\ &\le |\mathcal{B}|\nu\Big(1 + \frac{|\mathcal{A}_j|\nu}{1-(|\mathcal{A}_j|-1)\nu}\Big)\|\eta_{\mathcal{B}}^*\|_\infty + \bar{\epsilon}\Big(1 + \frac{|\mathcal{A}_j|\nu}{1-(|\mathcal{A}_j|-1)\nu}\Big). \end{aligned} \tag{3.4}$$

From $|\mathcal{A}_j| = K - |\mathcal{B}|$, $\frac{|\mathcal{B}|\nu+\alpha}{1-K\nu+\nu+|B|\nu} \le \frac{K\nu+\alpha}{1+\nu}$, and $\bar{\epsilon} \le \alpha \min_{i\in\mathcal{A}^*} |\eta_i^*| \le \alpha \|\eta_{\mathcal{B}}^*\|_\infty$, we get

$$\begin{aligned}&|\mathcal{B}|\nu\Big(1 + \frac{|\mathcal{A}_j|\nu}{1-(|\mathcal{A}_j|-1)\nu}\Big)\|\eta_{\mathcal{B}}^*\|_\infty + \bar{\epsilon}\Big(1 + \frac{|\mathcal{A}_j|\nu}{1-(|\mathcal{A}_j|-1)\nu}\Big) \\ \le &\frac{|\mathcal{B}|\nu+\alpha}{1-K\nu+\nu+|\mathcal{B}|\nu}(1+\nu)\|\eta_{\mathcal{B}}^*\|_\infty \le (K\nu+\alpha)\|\eta_{\mathcal{B}}^*\|_\infty.\end{aligned}$$

Therefore, it can be further concluded from (3.3) and (3.4) that

$$|d_i^j| \ge |\eta_i^*| - (K\nu - \nu + \alpha)\|\eta_{\mathcal{B}}^*\|_\infty, \ \forall i \in \mathcal{B}, \quad \text{and} \quad |d_i^j| \le (K\nu+\alpha)\|\eta_{\mathcal{B}}^*\|_\infty, \ \forall i \in \mathcal{I}^*.$$

Using $\mathcal{T}_{\lambda_{l+1},s_1} \subseteq \mathcal{A}_j$, we get $\|\eta_{\mathcal{B}}^*\|_\infty < s_1\sqrt{2\lambda_{l+1}}$. Next, we will prove $\mathcal{T}_{\lambda_{l+1},s_2} \subseteq \mathcal{A}_{j+1}$. For any $i \in \mathcal{I}_j \cap \mathcal{T}_{\lambda_{l+1},s_2}$, it holds that

$$|d_i^j| > s_2\sqrt{2\lambda_{l+1}} - (K\nu - \nu + \alpha)s_1\sqrt{2\lambda_{l+1}} = [s_2 - (K\nu - \nu + \alpha)s_1]\sqrt{2\lambda_{l+1}} = \sqrt{2\lambda_{l+1}},$$

which means that $i \in \mathcal{A}_{j+1}$. For any $i \in \mathcal{A}_j \cap \mathcal{T}_{\lambda_{l+1},s_2}$, from the relation $\frac{|\mathcal{B}|\nu+\alpha}{1-(|\mathcal{A}_j|-1)\nu} \le \frac{|\mathcal{B}|\nu+\alpha+(|\mathcal{A}_j|-1)\nu}{1-(|\mathcal{A}_j|-1)\nu+(|\mathcal{A}_j|-1)\nu} = K\nu - \nu + \alpha$, we can deduce that

$$|\eta_i^j| \ge |\eta_i^*| - \frac{|\mathcal{B}|\nu+\alpha}{1-(|\mathcal{A}_j|-1)\nu}\|\eta_{\mathcal{B}}^*\|_\infty > s_2\sqrt{2\lambda_{l+1}} - (K\nu - \nu + \alpha)s_1\sqrt{2\lambda_{l+1}} = \sqrt{2\lambda_{l+1}},$$

which means $i \in \mathcal{A}_{j+1}$. Therefore, $\mathcal{T}_{\lambda_{l+1},s_2} \subseteq \mathcal{A}_{j+1}$. Then, we get conclusion $\mathcal{A}_{j+1} \subseteq \mathcal{A}^*$

by proving $\mathcal{I}^* \subseteq \mathcal{I}_{j+1}$. For all $i \in \mathcal{I}^*$, we have $|d_i^j| < s_1(K\nu+\alpha)\sqrt{2\lambda_{l+1}} < \sqrt{2\lambda_{l+1}}$, which means $i \in \mathcal{I}_{j+1}$. By this point, we have proven the assertion (3.2).

Furthermore, we have proven that, for $k = 0, 1, 2, \ldots, K_{\max}$, the active sets $\mathcal{A}(\lambda_k)$ have the following special monotonicity:

$$\text{If } \mathcal{T}_{\lambda_k, s_1} \subseteq \mathcal{A}(\lambda_{k-1}) \subseteq \mathcal{A}^*, \quad \text{then } \mathcal{T}_{\lambda_k, s_2} \subseteq \mathcal{A}(\lambda_k) \subseteq \mathcal{A}^*.$$

From the above conclusion, it is clear that the active set $\mathcal{A}(\lambda_k)$ is always contained within $\mathcal{A}^*$. Additionally, for (2.2), we define η_{λ_1} and η_{λ_2} as the optimal solutions corresponding to λ_1 and λ_2, respectively. Here, we assume that $\lambda_1 > \lambda_2$. Then, we have

$$\begin{aligned}
\frac{1}{2}\|\bar{Y} - \bar{X}\eta_{\lambda_1}\|^2 + \lambda_1\|\eta_{\lambda_1}\|_0 &\leq \frac{1}{2}\|\bar{Y} - \bar{X}\eta_{\lambda_2}\|^2 + \lambda_1\|\eta_{\lambda_2}\|_0, \\
\frac{1}{2}\|\bar{Y} - \bar{X}\eta_{\lambda_2}\|^2 + \lambda_2\|\eta_{\lambda_2}\|_0 &\leq \frac{1}{2}\|\bar{Y} - \bar{X}\eta_{\lambda_1}\|^2 + \lambda_2\|\eta_{\lambda_1}\|_0.
\end{aligned}$$

Adding both sides of the two inequalities above, we can obtain $\lambda_1\|\eta_{\lambda_1}\|_0 + \lambda_2\|\eta_{\lambda_2}\|_0 \leq \lambda_1\|\eta_{\lambda_2}\|_0 + \lambda_2\|\eta_{\lambda_1}\|_0$, i.e., $(\lambda_1 - \lambda_2)(\|\eta_{\lambda_1}\|_0 - \|\eta_{\lambda_2}\|_0) \leq 0$. Therefore, we get $\|\eta_{\lambda_1}\|_0 \leq \|\eta_{\lambda_2}\|_0$. This indicates that as k increases, λ_k continuously decreases, and $\mathcal{A}(\lambda_k)$ will eventually either be $\mathcal{A}^*$ itself or a proper subset of $\mathcal{A}^*$. In a word, both cases lead to the termination criterion in the 12-th line of AFT-PDASC is satisfied, so that the algorithm is terminated in finite steps.

(3) From the proof process we know that, when AFT-PDASC terminates, the estimated active set $\hat{\mathcal{A}}$ is either $\mathcal{A}^*$ itself or a proper subset of $\mathcal{A}^*$. Next, we will proceed by contradiction to prove that the final active set must be exactly $\mathcal{A}^*$ itself but cannot be a proper subset. We assume that $\hat{\mathcal{A}} \subsetneqq \mathcal{A}^*$, and then $\mathcal{B} := \mathcal{A}^* \backslash \hat{\mathcal{A}}$ is non-empty. Let

$\hat{\mathcal{I}} := \mathcal{S} \setminus \hat{\mathcal{A}}$. Noting that $\hat{\eta}$ represents the estimated regression coefficient. Then, we have

$$L(\hat{\eta}) = \frac{1}{2}\|\bar{Y} - \bar{X}\hat{\eta}\|^2 + \lambda\|\hat{\eta}\|_0 = \frac{1}{2}\|\bar{X}_{\mathcal{B}}\eta^*_{\mathcal{B}} + \gamma - \bar{X}_{\hat{\mathcal{A}}}(\hat{\eta}_{\hat{\mathcal{A}}} - \eta^*_{\hat{\mathcal{A}}})\|^2 + \lambda|\hat{\mathcal{A}}|.$$

Let $i_{\mathcal{A}^*} \in \{i \in \hat{\mathcal{I}} : |\eta^*_i| = \|\eta^*_{\mathcal{B}}\|_\infty\}$, then it is easy to see that $i_{\mathcal{A}^*} \in \mathcal{B}$ and $|\eta^*_{i_{\mathcal{A}^*}}| = \|\eta^*_{\mathcal{B}}\|_\infty$. Furthermore, we have

$$\begin{aligned}
L(\hat{\eta}) =& \frac{1}{2}\|\bar{X}_{i_{\mathcal{A}^*}}\eta^*_{i_{\mathcal{A}^*}} + \bar{X}_{\mathcal{B}\setminus\{i_{\mathcal{A}^*}\}}\eta^*_{\mathcal{B}\setminus\{i_{\mathcal{A}^*}\}} + \gamma - \bar{X}_{\hat{\mathcal{A}}}(\hat{\eta}_{\hat{\mathcal{A}}} - \eta^*_{\hat{\mathcal{A}}})\|^2 + \lambda|\hat{\mathcal{A}}| \\
\geq& \frac{1}{2}|\eta^*_{i_{\mathcal{A}^*}}|^2 - |\eta^*_{i_{\mathcal{A}^*}}|\Big(|\langle \bar{X}_{i_{\mathcal{A}^*}}, \bar{X}_{\mathcal{B}\setminus\{i_{\mathcal{A}^*}\}}\eta^*_{\mathcal{B}\setminus\{i_{\mathcal{A}^*}\}}\rangle| + |\langle \bar{X}_{i_{\mathcal{A}^*}}, \gamma\rangle| \\
&+ |\langle \bar{X}_{i_{\mathcal{A}^*}}, \bar{X}_{\hat{\mathcal{A}}}(\hat{\eta}_{\hat{\mathcal{A}}} - \eta^*_{\hat{\mathcal{A}}})\rangle|\Big) + \lambda|\hat{\mathcal{A}}| \\
\geq& \frac{1}{2}|\eta^*_{i_{\mathcal{A}^*}}|^2 - |\eta^*_{i_{\mathcal{A}^*}}|\Big((|\mathcal{B}| - 1)\nu|\eta^*_{i_{\mathcal{A}^*}}| + \bar{\epsilon} + |\hat{\mathcal{A}}|\nu\|\hat{\eta}_{\hat{\mathcal{A}}} - \eta^*_{\hat{\mathcal{A}}}\|_\infty\Big) + \lambda|\hat{\mathcal{A}}| \\
\geq& \frac{1}{2}|\eta^*_{i_{\mathcal{A}^*}}|^2 - |\eta^*_{i_{\mathcal{A}^*}}|\Big((|\mathcal{B}| - 1)\nu|\eta^*_{i_{\mathcal{A}^*}}| + \bar{\epsilon} + \frac{|\hat{\mathcal{A}}|\nu}{1 - (|\hat{\mathcal{A}}| - 1)\nu}(|\mathcal{B}|\nu|\eta^*_{i_{\mathcal{A}^*}}| + \bar{\epsilon})\Big) + \lambda|\hat{\mathcal{A}}|.
\end{aligned}$$

In addition, using Assumption 2, we have $\bar{\epsilon} \leq \alpha \min_{i\in\mathcal{A}^*}\{|\eta^*_i|\} \leq \alpha|\eta^*_{i_{\mathcal{A}^*}}|$. Then, we can deduce that

$$\begin{aligned}
L(\hat{\eta}) \geq& \frac{1}{2}|\eta^*_{i_{\mathcal{A}^*}}|^2 - |\eta^*_{i_{\mathcal{A}^*}}|\Big((|\mathcal{B}| - 1)\nu|\eta^*_{i_{\mathcal{A}^*}}| + \alpha|\eta^*_{i_{\mathcal{A}^*}}| \\
&+ \frac{|\hat{\mathcal{A}}|\nu}{1 - (|\hat{\mathcal{A}}| - 1)\nu}(|\mathcal{B}|\nu|\eta^*_{i_{\mathcal{A}^*}}| + \alpha|\eta^*_{i_{\mathcal{A}^*}}|)\Big) + \lambda|\hat{\mathcal{A}}| \\
=& |\eta^*_{i_{\mathcal{A}^*}}|^2\Big(\frac{1}{2} - (|\mathcal{B}| - 1)\nu - \alpha - \frac{|\hat{\mathcal{A}}|\nu(|\mathcal{B}|\nu + \alpha)}{1 - (|\hat{\mathcal{A}}| - 1)\nu}\Big) + \lambda|\hat{\mathcal{A}}| \\
=& |\eta^*_{i_{\mathcal{A}^*}}|^2\Big(\frac{1}{2} - (K - 1)\nu - \alpha\Big) + |\eta^*_{i_{\mathcal{A}^*}}|^2|\hat{\mathcal{A}}|\nu\Big(1 - \frac{|\mathcal{B}|\nu + \alpha}{1 - (|\hat{\mathcal{A}}| - 1)\nu}\Big) + \lambda|\hat{\mathcal{A}}|.
\end{aligned}$$

Here, $1 - \frac{|\mathcal{B}|\nu+\alpha}{1-(|\hat{\mathcal{A}}|-1)\nu} = \frac{1-\big((|\hat{\mathcal{A}}|+|\mathcal{B}|-1)\nu+\alpha\big)}{1-(|\hat{\mathcal{A}}|-1)\nu}$. It has been analyzed earlier that $(|\hat{\mathcal{A}}| - 1)\nu < 1$, and $(|\hat{\mathcal{A}}| + |\mathcal{B}| - 1)\nu + \alpha < (2K - 1)\nu + 2\alpha < 1$, so we get $1 - \frac{|\mathcal{B}|\nu+\alpha}{1-(|\hat{\mathcal{A}}|-1)\nu} > 0$. Therefore, we can conclude that $L(\hat{\eta}) \geq |\eta^*_{i_{\mathcal{A}^*}}|^2\Big(\frac{1}{2} - (K - 1)\nu - \alpha\Big)$.

Next, we will first explain the reasonableness of the setting for the value range of

λ. From $0 \leq \alpha < \frac{1}{2}$, we have $\alpha^2 < \alpha$. With the relation of $\alpha \leq \frac{1-2(K-1)\nu}{K+3}$, we can easily obtain $(K+1)\alpha^2 + 2\alpha < (K+3)\alpha \leq 1 - 2(K-1)\nu$. Furthermore, it has $K\alpha^2 < 1 - 2(K-1)\nu - 2\alpha - \alpha^2$. It is clear from the definition of ξ and Assumption 2 that $\xi > \frac{\alpha^2}{2} \min_{i\in\mathcal{A}^*} |\eta_i^*|^2 \geq \bar{\epsilon}^2/2$. Therefore, the setting of the value range of λ is reasonable. Based on the value of λ, we have

$$\begin{aligned} L(\hat{\eta}) - L(\eta^o) &\geq |\eta^*_{i_{\mathcal{A}^*}}|^2 \Big(\frac{1}{2} - (K-1)\nu - \alpha\Big) - \frac{1}{2}\bar{\epsilon}^2 - \lambda K \\ &\geq \Big(\frac{1}{2} - (K-1)\nu - \alpha - \frac{\alpha^2}{2}\Big) \min_{i\in\mathcal{A}^*}\{|\eta_i^*|^2\} - \lambda K \\ &= \xi K - \lambda K > 0. \end{aligned}$$

In other words, it gets $\frac{1}{2}\|\bar{Y} - \bar{X}\hat{\eta}\|^2 + \lambda|\hat{\mathcal{A}}| > \frac{1}{2}\bar{\epsilon}^2 + \lambda K$. Since $|\hat{\mathcal{A}}| < K$, then it yields $\|\bar{Y} - \bar{X}\hat{\eta}\| > \bar{\epsilon}$. This contradicts the stopping condition in 12-th line of AFT-PDASC, so the previous assumption is incorrect. In other words, $\hat{\mathcal{A}}$ cannot be a proper subset of $\mathcal{A}^*$, it must be $\mathcal{A}^*$ itself. This also indicates that AFT-PDASC terminates at the oracle solution η^o. □

At the end of this section, using the Assumption 1 (2) that the covariate matrix $\bar{X}$ satisfies the RIP condition, we also present the error analysis, the monotonicity of the active set, and the finite-step termination of AFT-PDAS at the oracle solution η^o. Since the proof process is similar to that of Theorem 1, we omit it here.

Theorem 2. *Suppose that Assumption 2 holds. Let* $\delta := \delta_{K+1} < \frac{1-2\alpha}{2\sqrt{K}+1}$, *then for any* $\rho \in \Big((\frac{2\delta\sqrt{K}+2\alpha}{1-\delta})^2, 1\Big)$, *it holds that:*

(1) In the λ_k *subproblem, for* $j = 1, 2, \ldots, J_{\max}$, *we set* $\mathcal{B} := \mathcal{A}^* \backslash \mathcal{A}_j$ *and* $\mathcal{I}_j = \mathcal{S} \backslash \mathcal{A}_j$. *If*

$|\mathcal{A}_j| \leq K$, *then we have*

$$\left\|\eta^j_{\mathcal{A}_j} - \eta^*_{\mathcal{A}_j}\right\| \leq \frac{\delta_{|\mathcal{A}_j|+|\mathcal{B}|}}{1-\delta_{|\mathcal{A}_j|}} \|\eta^*_{\mathcal{B}}\| + \frac{1}{\sqrt{1-\delta_{|\mathcal{A}_j|}}}\bar{\epsilon},$$
$$\left\|\beta^j_{\mathcal{A}_j} - \beta^*_{\mathcal{A}_j}\right\| \leq \frac{\|D\|\delta_{|\mathcal{A}_j|+|\mathcal{B}|}}{1-\delta_{|\mathcal{A}_j|}} \left\|D^{-1}\beta^*_{\mathcal{B}}\right\| + \frac{\|D\|}{\sqrt{1-\delta_{|\mathcal{A}_j|}}}\bar{\epsilon}.$$

(2) *For* $k = 0, 1, 2, \ldots, K_{\max}$, *there exist* $s_1, s_2 \in (\frac{1-\delta}{1-\delta-\delta\sqrt{K}-\alpha}, \frac{1-\delta}{\delta\sqrt{K}+\alpha})$ *with* $s_1 > s_2$, *such that the active sets* $\mathcal{A}(\lambda_k)$ *have the following special monotonicity:*

$$\text{If } \mathcal{T}_{\lambda_k,s_1} \subseteq \mathcal{A}(\lambda_{k-1}) \subseteq \mathcal{A}^*, \quad \text{then } \mathcal{T}_{\lambda_k,s_2} \subseteq \mathcal{A}(\lambda_k) \subseteq \mathcal{A}^*.$$

Furthermore, AFT-PDASC terminates in a finite number of steps.

(3) *Let* $\alpha \leq (1-2\delta-\delta^2)/4$ *and* $\xi = (\frac{1-\delta}{2} - \frac{\delta^2}{1-\delta} - \frac{\alpha}{\sqrt{1-\delta}} - \frac{1}{2}\alpha^2)\min_{i\in\mathcal{A}^*}\{|\eta^*_i|^2\}$. *Then, for any* $\lambda \in (\frac{\bar{\epsilon}^2}{2}, \xi)$, *AFT-PDASC terminates at the oracle solution* η^o.

4. Numerical experiments

In this section, we demonstrate the progressiveness of the estimation method in (2.1) and highlight the effectiveness of the AFT-PDASC algorithm for handling censored data from both numerical simulations and real data applications. We also conduct performance comparisons with several state-of-the-art estimation methods, focusing on the impact of model parameters and the accuracy of the estimations. The algorithms selected for comparison include the support detection and root finding algorithm (SDAR) (Cheng et al., 2022), Orthogonal Matching Pursuit (OMP) (Pati, Rezaiifar and Krishnaprasad, 1993), Greedy Gradient Pursuit (GreedyGP) (Blumensath and Davies, 2008), Accelerated Iterative Hard Thresholding (AIHT) (Blumensath, 2012), and Hard Thresholding Pursuit (HTP) (Foucart, 2011). All the experiments are performed with Microsoft Windows 11

and MATLAB R2022a, and run on a PC with an Intel(R) Xeon(R) W-2295 CPU at 3GHz and 128 GB of memory.

We now outline the data generation process used in the simulation study. We first generate a $n \times p$ random Gaussian matrix $\check{X}$ whose entries are i.i.d. $\mathcal{N}(0,1)$. Then X is generated with $X_1 := \check{X}_1$, $X_p := \check{X}_p$ and $X_i := \check{X}_i + \kappa(\check{X}_{i+1} + \check{X}_{i-1})$ for $i = 2, \ldots, p-1$. Here, κ measures the strength of the correlation between the covariates. To generate the true regression coefficient β^*, we first randomly select a subset of $\mathcal{S}$ to form the true active set $\mathcal{A}^*$. Let $R := m_2/m_1$, where $m_2 = \max\{|\beta_i^*| : i \in \mathcal{A}^*\}$ and $m_1 = \min\{|\beta_i^*| : i \in \mathcal{A}^*\}$. Subsequently, the K nonzero coefficients in β^* are distributed uniformly within the interval (m_1, m_2). Then we set $\ln(T_i) := X_i^\top \beta^* + \epsilon_i$ for $i = 1, \ldots, n$, where ϵ_i is generated independently from $\mathcal{N}(0, \sigma^2)$. The censoring time C_i follows a uniform distribution $U(0, \zeta)$, where ζ controls the censoring rate. Then for $i = 1, \ldots, n$, the response variable is generated by $Y_i := \min\{\ln(T_i), \ln(C_i)\}$. In AFT-PDASC, we use a grid search method to select the appropriate regularization parameter. Specifically, we choose $\lambda_0 = \frac{1}{2}\|\bar{X}^\top \bar{Y}\|_\infty^2$ and set $\lambda_{\min} := 10^{-15}\lambda_0$, and then divide the interval $[\lambda_{\min}, \lambda_0]$ into N equally spaced subintervals. It is easy to observe that as N increases, the decay factor ρ also becomes larger. The values of the other parameters will be given as they occur.

4.1 Performance evaluation on a simple simulated data

In this part, we analyze the numerical performance of algorithm AFT-PDASC using 100 independent experiments. Here, we choose $n = 500$, $p = 1000$, $K = 10$, $\kappa = 0.3$, $N = 100$, and $J_{\max} = 2$. Let $c.r$ be the censoring rate, and in this case, we consider $c.r = 0.3$. At the same time, we fix the 10 non-zero elements of the true regression coefficients β^* to be $\beta_{34}^* = 5$, $\beta_{166}^* = -1$, $\beta_{278}^* = 2$, $\beta_{354}^* = -3$, $\beta_{409}^* = 4$, $\beta_{520}^* = -5$, $\beta_{666}^* = 1$, $\beta_{708}^* = -4$,

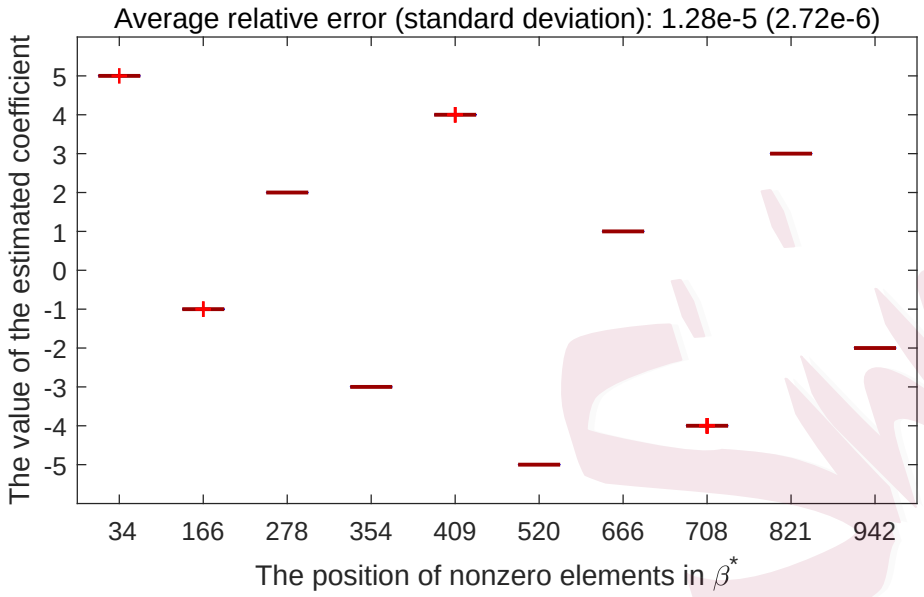

Figure 1: The boxplot of AFT-PDASC under 100 independent experiments.

$\beta^*_{821} = 3$, $\beta^*_{942} = -2$. The boxplot in Figure 1 illustrates the estimation performance of AFT-PDASC in this test.

From Figure 1, it is evident that the AFT-PDASC algorithm not only accurately locates each non-zero element but also consistently estimates their values correctly in nearly every trial. Additionally, the average relative error ($ReErr = \frac{1}{100}\sum \frac{\|\hat{\beta}-\beta^*\|}{\|\beta^*\|}$) and standard deviation between the true regression coefficient and the estimated coefficient are very small, further highlighting the excellent estimation performance of the method in (2.1) and the effectiveness of the AFT-PDASC algorithm.

4.2 Performance comparisons with other state-of-the-art algorithms

In this part, we conduct the performance comparisons of the AFT-PDASC algorithm with SDAR, OMP, GreedyGP, AIHT, and HTP, focusing on the impact of model parameters and estimation performance. To demonstrate the numerical stability of each algorithm, we perform a simulation analysis based on the results of 100 independent repetitions.

Let $\widehat{\mathcal{A}}^{(i)}$ be the active set obtained in the i-th experiment. In addition to recording the computing time (Time) and ReErr, we introduce an indicator defined as "Probability $= \frac{1}{100}\sum_{i=1}^{100}\mathbf{1}_{\{\widehat{\mathcal{A}}^{(i)}=\mathcal{A}^*\}}$", which measures the probability of accurately recovering the true active set. Furthermore, we also consider the Mattews correlation coefficient (MCC), which is commonly used in statistics to evaluate the performance of variable selection, and its specific definition is given as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively. It is easy to see from its definition that the value range of MCC is $[-1, 1]$. Superior variable selection is achieved when the result is closer to 1.

(a). Parameters' values influence evaluation:

In this test, we examine the impact of certain parameter values, namely $\{n, p, K, \kappa\}$, on the active set recovery performance of each algorithm. The specific values of all parameters for each experimental setting are as follows:

- $n = \{200 : 50 : 500\}$, $p = 1000$, $K = 15$, $R = 10^3$, $\kappa = 0.3$, $c.r = 0.3$, $\sigma = 1e-3$, $N = 150$, $J_{\max} = 6$.
- $n = 500$, $p = \{1200 : 100 : 2000\}$, $K = 15$, $R = 10^3$, $\kappa = 0.3$, $c.r = 0.3$, $\sigma = 1e-3$, $N = 150$, $J_{\max} = 6$.
- $n = 500$, $p = 1000$, $K = \{10 : 10 : 200\}$, $R = 10^3$, $\kappa = 0.3$, $c.r = 0.3$, $\sigma = 1e-3$, $N = 150$, $J_{\max} = 6$.
- $n = 500$, $p = 1000$, $K = 15$, $\kappa = \{0.1 : 0.1 : 0.8\}$, $R = 10^3$, $c.r = 0.3$, $\sigma = 1e-3$,

4.2 Performance comparisons with other state-of-the-art algorithms

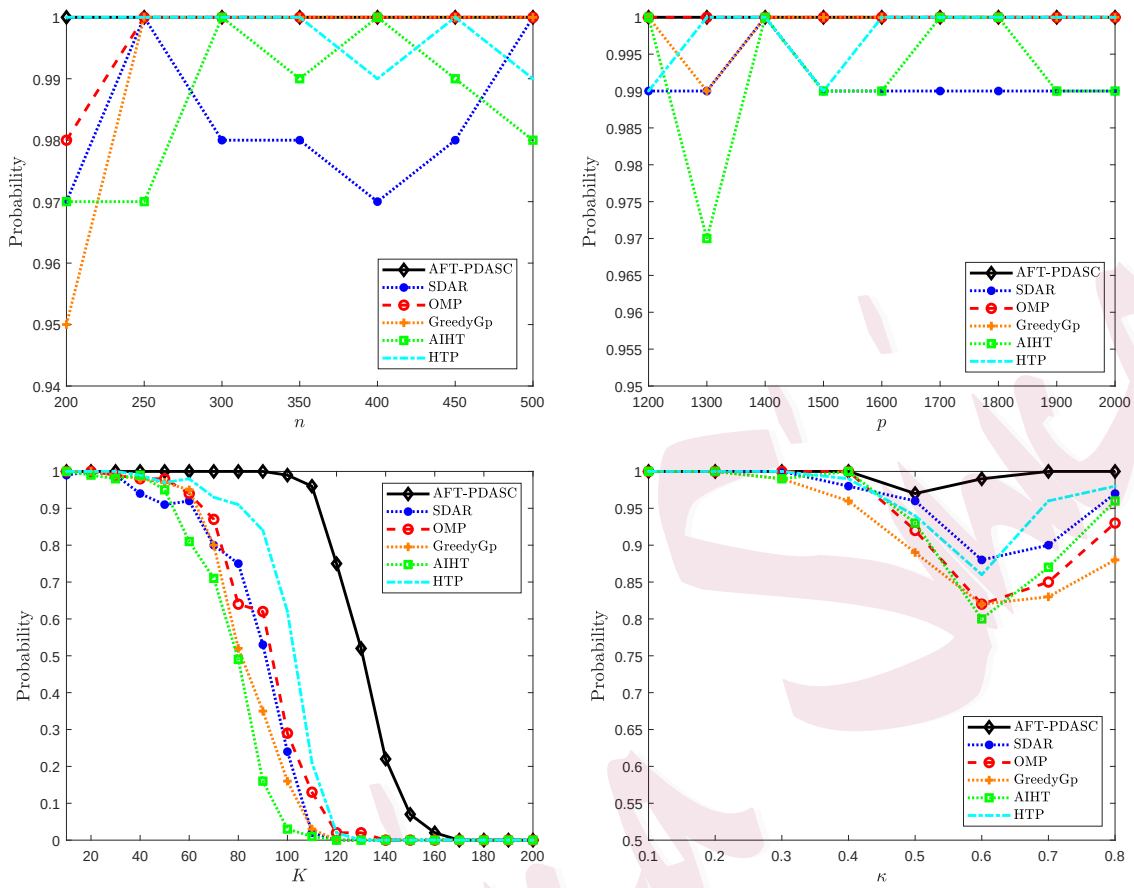

Figure 2: The "Probability" values of each algorithm under different parameters' settings

$N = 150,\ J_{\max} = 6.$

Figure 2 shows the results of "Probability" for the six algorithms under different values of the parameters $\{n, p, K, \kappa\}$. It is evident from the figure that the regression probability of the AFT-PDASC algorithm is consistently higher than that of the other algorithms, with the value of "Probability" approaching 1 in most cases. This phenomenon not only demonstrates that the AFT-PDASC algorithm is more stable compared to other methods, but also indicates that it can achieve accurate recovery across a wide range of parameter settings.

(b). Performance comparisons on simulated data:

To further highlight the accuracy and efficiency of the AFT-PDASC algorithm, we compare it with the other algorithms in terms of four key metrics: Time, ReErr, Proba-

bility and MCC. Additionally, for each result, we also provide the corresponding standard deviation obtained from 100 independent trials to quantify the stability of the considered algorithms. In this test, we consider three different levels of correlation for the simulation matrix by setting $\kappa = 0.1$, 0.4 and 0.7. Additionally, we consider three different dimensions by setting $p = 1000$, 2000 and 3000. Other parameters's values are fixed as $n = 500$, $K = 20$, $R = 10^3$, $c.r = 0.3$, $\sigma = 1e-3$, $N = 100$, and $J_{\max} = 2$. The results of the different algorithms in terms of Time, ReErr, Probability and MCC are listed in Table 1.

From the results in Table 1, it is evident that when $\kappa = 0.1$, i.e., when the correlation between covariates is relatively low, all six algorithms can quickly achieve good estimation results and perform accurate recovery across the three dimensions. However, when $\kappa = 0.4$ and $\kappa = 0.7$, i.e., when the correlation between covariates is high, AFT-PDASC outperforms the other algorithms in terms of relative error, recovery probability and MCC, consistently achieving precise estimation. It is worth noting that the AFT-PDASC algorithm employs a grid search method for its regularization parameter, which results in higher computing time compared to other algorithms. However, the overall time required remains relatively low and within an acceptable range. In summary, the AFT-PDASC algorithm exhibits both high computational efficiency and numerical robustness in the simulation tests.

(c). Performance comparisons on real data:

In this section, we perform numerical tests using the NKI70 breast cancer dataset, which is available in the R package. This dataset includes data from 144 breast cancer patients with lymph node-positive metastasis-free survival, along with gene expression measurements for 70 genes that were identified as prognostic for metastasis-free survival in a previous study. The censoring rate for this dataset is 66.67%. For the AFT-PDASC

4.2 Performance comparisons with other state-of-the-art algorithms

Table 1: Numerical results of each algorithm on simulated data

κ	p	Methods	Time(s)	ReErr	Probability	MCC
0.1	1000	AFT-PDASC	2.19e-2(3.77e-3)	2.02e-7(9.05e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	2.35e-3(4.22e-4)	2.02e-7(9.05e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		OMP	1.53e-3(2.23e-4)	2.02e-7(9.05e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		GreedyGP	1.64e-3(4.62e-4)	1.12e-6(3.77e-7)	1.00(0.00e-0)	1.00(0.00e-0)
		AIHT	3.99e-3(3.95e-4)	2.08e-7(9.29e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		HTP	8.72e-3(1.03e-3)	2.02e-7(9.05e-8)	1.00(0.00e-0)	1.00(0.00e-0)
	2000	AFT-PDASC	2.96e-2(2.82e-3)	2.15e-7(6.55e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	3.63e-3(5.43e-4)	2.15e-7(6.55e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		OMP	2.44e-3(5.75e-4)	2.15e-7(6.55e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		GreedyGP	1.94e-3(5.04e-4)	1.20e-6(4.58e-7)	1.00(0.00e-0)	1.00(0.00e-0)
		AIHT	7.90e-3(8.64e-4)	2.25e-7(6.93e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		HTP	1.76e-2(2.28e-3)	2.15e-7(6.55e-8)	1.00(0.00e-0)	1.00(0.00e-0)
	3000	AFT-PDASC	4.41e-2(3.11e-3)	1.99e-7(6.81e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	3.99e-3(5.06e-4)	1.99e-7(6.81e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		OMP	3.09e-3(6.54e-4)	1.99e-7(6.81e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		GreedyGP	3.93e-3(8.67e-4)	1.10e-6(4.22e-7)	1.00(0.00e-0)	1.00(0.00e-0)
		AIHT	1.28e-2(1.21e-3)	2.09e-7(7.15e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		HTP	2.55e-2(2.84e-3)	1.99e-7(6.81e-8)	1.00(0.00e-0)	1.00(0.00e-0)
0.4	1000	AFT-PDASC	2.05e-2(3.08e-3)	1.86e-7(9.23e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	2.43e-3(6.27e-4)	5.82e-5(2.66e-4)	0.95(2.19e-1)	0.99(1.11e-2)
		OMP	1.56e-3(2.50e-4)	1.87e-7(9.24e-8)	0.96(1.96e-1)	0.99(8.95e-3)
		GreedyGP	1.71e-3(4.72e-4)	2.92e-4(2.00e-3)	0.96(1.96e-1)	0.99(2.36e-2)
		AIHT	9.18e-3(1.75e-3)	5.66e-4(3.76e-3)	0.94(2.38e-1)	0.98(4.88e-2)
		HTP	1.04e-2(1.54e-3)	2.88e-5(2.03e-4)	0.98(1.40e-1)	0.99(7.17e-3)
	2000	AFT-PDASC	2.98e-2(2.88e-3)	1.90e-7(6.37e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	3.94e-3(4.83e-4)	4.58e-5(3.23e-4)	0.98(1.40e-1)	0.99(1.12e-2)
		OMP	2.61e-3(5.37e-4)	1.90e-7(6.37e-8)	0.99(1.00e-1)	0.99(8.80e-3)
		GreedyGP	1.95e-3(5.32e-4)	1.41e-3(1.41e-2)	0.99(1.00e-1)	0.99(4.54e-2)
		AIHT	1.72e-2(2.66e-3)	4.97e-5(3.62e-4)	0.98(1.40e-1)	0.99(1.59e-2)
		HTP	1.98e-2(2.48e-3)	1.90e-7(6.37e-8)	1.00(0.00e-0)	1.00(0.00e-0)
	3000	AFT-PDASC	4.36e-2(3.20e-3)	1.76e-7(6.98e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	4.55e-3(5.35e-4)	2.40e-5(1.69e-4)	0.98(1.40e-1)	0.99(7.08e-3)
		OMP	3.15e-3(6.13e-4)	1.76e-7(6.98e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		GreedyGP	3.95e-3(8.43e-4)	7.42e-6(6.28e-5)	1.00(0.00e-0)	1.00(0.00e-0)
		AIHT	2.63e-2(4.43e-3)	2.72e-4(2.48e-3)	0.98(1.40e-1)	0.99(2.24e-2)
		HTP	3.00e-2(3.22e-3)	1.19e-5(1.17e-4)	0.99(1.00e-1)	0.99(1.00e-2)
0.7	1000	AFT-PDASC	2.16e-2(2.95e-3)	1.57e-7(9.21e-8)	0.99(1.00e-1)	0.99(1.07e-2)
		SDAR	3.05e-3(4.85e-4)	1.36e-3(7.45e-3)	0.80(4.02e-1)	0.98(4.44e-2)
		OMP	1.79e-3(4.06e-4)	1.66e-7(9.69e-8)	0.74(4.40e-1)	0.98(2.50e-2)
		GreedyGP	1.63e-3(4.37e-4)	1.02e-2(4.37e-2)	0.69(4.64e-1)	0.94(1.05e-1)
		AIHT	1.25e-2(2.63e-3)	2.73e-3(1.18e-2)	0.76(4.29e-1)	0.96(8.41e-2)
		HTP	1.16e-2(1.36e-3)	1.85e-4(7.67e-4)	0.89(3.14e-1)	0.99(2.00e-2)
	2000	AFT-PDASC	3.05e-2(3.22e-3)	1.57e-7(6.53e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	4.25e-3(6.18e-4)	1.73e-4(7.12e-4)	0.93(2.56e-1)	0.99(2.14e-2)
		OMP	2.44e-3(7.07e-4)	1.60e-7(6.84e-8)	0.85(3.58e-1)	0.99(1.25e-2)
		GreedyGP	1.82e-3(4.11e-4)	4.95e-3(1.87e-2)	0.84(3.68e-1)	0.97(8.35e-2)
		AIHT	2.21e-2(3.90e-3)	2.15e-3(6.52e-3)	0.78(4.16e-1)	0.96(7.83e-2)
		HTP	2.18e-2(2.41e-3)	2.17e-4(7.92e-4)	0.91(2.87e-1)	0.99(2.48e-2)
	3000	AFT-PDASC	4.44e-2(3.43e-3)	1.46e-7(7.02e-8)	1.00(0.00e-0)	1.00(0.00e-0)
		SDAR	4.87e-3(8.30e-4)	1.79e-4(7.97e-4)	0.93(2.56e-1)	0.99(2.64e-2)
		OMP	3.23e-3(6.83e-4)	1.47e-7(7.01e-8)	0.91(2.87e-1)	0.99(7.99e-3)
		GreedyGP	3.87e-3(8.38e-4)	1.27e-3(7.04e-3)	0.92(2.72e-1)	0.98(4.18e-2)
		AIHT	3.24e-2(5.65e-3)	1.71e-3(8.32e-3)	0.88(3.26e-1)	0.98(5.41e-2)
		HTP	3.23e-2(3.84e-3)	1.52e-4(5.97e-4)	0.92(2.72e-1)	0.99(2.44e-2)

algorithm, we set parameters as $N = 100$ and $J_{\max} = 1$. Additionally, we compare the performance of AFT-PDASC with other algorithms mentioned earlier, including SDAR, OMP, GreedyGP, AIHT, and HTP. The computational results are summarized in Table 2.

From Table 2, we observe that the AFT-PDASC algorithm selects the fewest number

of genes in most cases, while the OMP method seemly selects the most. Moreover, the estimated coefficients for the genes chosen by all six algorithms share the same mathematics sign. Specifically, for the 6th gene, 'ALDH4A1', the regression coefficients estimated by AFT-PDASC and AIHT are quite similar. For the 20th gene, 'MMP9', the coefficient estimated by AFT-PDASC is close to those obtained by OMP and GreedyGP. In the case of the 27th gene, 'CDC42BPA', AFT-PDASC's coefficient is comparable to that of GreedyGP. Lastly, for the 66th gene, 'CENPA', the estimated coefficient from AFT-PDASC is similar to those from SDAR and GreedyGP. These observations suggest that the AFT-PDASC algorithm not only selects fewer genes for the NKI70 dataset but also exhibits superior estimation performance.

Table 2: Numerical results of each algorithm on real data 'nki70'

Gene name	Number	AFT-PDASC	SDAR	OMP	GreedyGP	AIHT	HTP
TSPYL5	1	-	-	-	-	-0.38	-
Contig63649_RC	2	-	-	-1.66	-0.53	-	-
AA555029_RC	5	-	-	-1.28	-	-	-
ALDH4A1	6	-2.61	-	-2.20	-1.70	-2.64	-2.00
Contig32125_RC	10	-	-	-0.97	-	-	-
SCUBE2	15	-0.41	-	-	-	-	-
EXT1	16	4.18	1.67	6.55	3.95	3.74	3.36
GNAZ	18	-	0.55	-	-	-	0.81
MMP9	20	-3.91	-2.52	-3.99	-3.90	-2.71	-3.59
RUNDC1	21	-	-1.09	-1.97	-	-	-
GMPS	24	-	-	-2.02	-	-	-
KNTC2	25	-	-	0.09	1.89	-	-
WISP1	26	-	-	-2.72	-	-	-
CDC42BPA	27	2.25	-	4.15	2.26	-	1.62
GSTM3	30	-	-1.19	-1.50	-	-1.26	-0.81
GPR180	31	-	1.02	-	-	-	-
RAB6B	32	-	-	-0.82	-	-	-
MTDH	37	-1.46	-1.23	-	-	-	-1.21
DCK	44	-	-	-2.72	-1.67	-	-
SLC2A3	47	1.49	1.66	2.69	2.41	-	1.71
CDCA7	51	-	-0.64	-	-	-0.86	-
MS4A7	53	-	-	0.79	-	-	-
MCM6	54	-	1.81	-	-	-	-
AP2B1	55	-	0.96	-	-	-	-
PALM2.AKAP2	62	-	-	-	-	1.84	-
LGP2	63	-	1.30	1.85	0.97	-	-
CENPA	66	-2.05	-2.09	-1.45	-1.97	-0.85	-1.66
NM_004702	68	-	-	-	-	-0.41	-
ESM1	69	-	-	1.21	-	-	-
C20orf46	70	-	-0.95	-	-	-0.81	-0.78

5. Conclusions

In this paper, we addressed the AFT problem with right-censored survival data by employing a weighted least-squares method with an ℓ_0-penalty for parameter estimation and

variable selection. For practical implementations, we developed an efficient primal dual active set algorithm and utilize a continuous strategy to select the appropriate regularization parameter. From a theoretical perspective, we provided an error analysis for the estimated coefficients, grounded in certain assumptions regarding the covariate matrix and noise level. Additionally, we proved that the algorithm converges to the oracle solution in a finite number of steps by demonstrating the special monotonicity of the active set throughout the iterative process. We also conducted extensive tests of the AFT-PDASC algorithm on both simulated and real data, showing its superior performance in comparison to other leading algorithms. Thus, we conclude that the AFT-PDASC algorithm is an effective tool for analyzing high-dimensional censored survival data.

Acknowledgements

We would like to thank professor Xiliang Lu from Wuhan university for his guidance and significant assistance in the theoretical analysis of this paper.

Declarations

The authors report there are no competing interests to declare. All authors contributed to the study conception and design. All authors read and approved the final manuscript.

Funding

The work of P. Li is supported by the National Natural Science Foundation of China (Grant No. 12301420), the Key Scientific Research Project of Universities in Henan Province (Grant No. 25A110002). The work of Y. Ding is supported by the Shenzhen Polytechnic University Research Fund (Grant No. 6024310021K). The work of Y. Xiao is supported by the National Natural Science Foundation of China (Grant No. 12471307 and

12271217), the Natural Science Foundation of Henan Province (Grant No. 232300421018).

References

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

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics* **21**(1), 76–99.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* **23**(4), 461–471.

Stute, W. and J.-L. Wang (1993). The strong law under random censorship. *The Annals of Statistics* **21**(3), 1591–1607.

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **70**(2), 351–370.

Johnson, B. A., D. Y. Lin and D. Zeng (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**(482), 672–680.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456), 1348–1360.

Cai, T., J. Huang and L. Tian (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65**(2), 394–404.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429.

Hu, J. and H. Chai (2013). Adjusted regularized estimation in the accelerated failure time model with high dimensional covariates. *Journal of Multivariate Analysis* **122**, 96–114.

Huang, J. and S. Ma (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis* **16**(2), 176–195.

Huang, J., S. Ma and H. Xie (2006). Regularized estimation in the accelerated failure time model with high-

dimensional covariates. *Biometrics* **62**(3), 813–820.

Khan, M. H. R. and J. E. H. Shaw (2016). Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing* **26**(3), 725–741.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1), 267–288.

Hong, D. and F. Zhang (2010). Weighted elastic net model for mass spectrometry imaging processing. *Mathematical Modelling of Natural Phenomena* **5**(3), 115–133.

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37**(4), 1733–1751.

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**(2), 109–135.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**(2), 894–942.

Cheng, C., X. Feng, J. Huang, Y. Jiao and S. Zhang (2022). ℓ_0-regularized high-dimensional accelerated failure time model. *Computational Statistics & Data Analysis* **170**, 107430.

Huang, J., Y. Jiao, Y. Liu and X. Lu (2018). A constructive approach to ℓ_0 penalized regression. *Journal of Machine Learning Research* **19**(10), 1–37.

Jiao, Y., B. Jin and X. Lu (2015). A primal dual active set with continuation algorithm for the ℓ_0-regularized optimization problem. *Applied and Computational Harmonic Analysis* **39**(3), 400–426.

Huang, J., Y. Jiao, B. Jin, J. Liu, X. Lu and C. Yang (2021). A unified primal dual active set algorithm for nonconvex sparse recovery. *Statistical Science* **36**(2), 215–238.

Li, P., Y. Jiao, X. Lu and L. Kang (2022). A data-driven line search rule for support recovery in high-dimensional data analysis. *Computational Statistics & Data Analysis* **174**, 107524.

Hintermüller, M., K. Ito and K. Kunisch (2002). The primal-dual active set strategy as a semismooth Newton

method. *SIAM Journal on Optimization* **13**(3), 865–888.

Dai, W. and O. Milenkovic (2009). Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory* **55**(5), 2230–2249.

Needell, D. and J. A. Tropp (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis* **26**(3), 301–321.

Pati, Y. C., R. Rezaiifar and P. S. Krishnaprasad (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, 40–44.

Blumensath, T. and M. E. Davies (2008). Gradient pursuits. *IEEE Transactions on Signal Processing* **56**(6), 2370–2382.

Blumensath, T. (2012). Accelerated iterative hard thresholding. *Signal Processing* **92**(3), 752–756.

Foucart, S. (2011). Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis* **49**(6), 2543–2563.

Henan University

E-mail: (lipeili@henu.edu.cn)

Henan University

E-mail: (ruoyinghu@163.com)

Shenzhen Polytechnic University

E-mail: (dingyanyun@szpu.edu.cn)

Henan University

E-mail: (yhxiao@henu.edu.cn)