AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(1): 1195–1207.
DOI: 10.3934/math.2023060
Received: 19 June 2022
Revised: 30 September 2022
Accepted: 10 October 2022
Published: 18 October 2022

Research article

Variable selection and estimation for accelerated failure time model via seamless-L_0 penalty

Yin Xu[1,*] **and Ning Wang**[2]

[1] Department of Statistics, School of Economics, Jinan University, Guangzhou 510632, China

[2] Department of Statistical Science, University College London, WC1E 6AE, UK

* **Correspondence:** Email: xuyin@stu2020.jnu.edu.cn.

Abstract: Survival data with high dimensional covariates have been collected in medical studies and other fields. In this work, we propose a seamless L_0 (SELO) penalized method for the accelerated failure time (AFT) model under the framework of high dimension. Specifically, we apply the SELO to do variable selection and estimation under this model. Under appropriate conditions, we show that the SELO selects a model whose dimension is comparable to the underlying model, and prove that the proposed procedure is asymptotically normal. Simulation results demonstrate that the SELO procedure outperforms other existing procedures. The real data analysis is considered as well which shows that SELO selects the variables more correctly.

Keywords: variable selection; AFT model; weighted least squares; seamless L_0
Mathematics Subject Classification: 14Q15, 62N02, 62E20

1. Introduction

Analyzing high-dimensional survival data has become an important topic in statistics, among which finding covariates with good predictive power of survival is a fundamental step. In the variable selection, penalized least squares procedures is an attractive approach. Penalized least squares procedures are used for variable selection and estimation which help predict estimators.

As a useful alternative to the Cox model [2], the AFT model [10] based on linear regression models has an intuitive form compared to Cox model. The AFT model with an unspecified error distribution has been studied commonly for right-censored data. Two approaches in this aspect have gained attractive attention. One uses the Kaplan-Meier estimator to obtain the ordinary least squares estimator. The other is the rank-based estimator, which is motivated by the score function of the partial likelihood. See for examples in [1, 13, 17].

Identifying significant factors with predictive power, many techniques for linear regression models have been extended to the Cox regression and the AFT model. Penalized methods have drawn extensive attentions, which are for imposing some penalties to the regression coefficients. By balancing the goodness of fit and model complexity, penalization approaches lead the complex models to a profile. There exists plenty of methods used in gene expression analysis with survival data; see for examples in [16, 19]. Moreover, various penalization methods of consistent selection have also been proposed. Examples include the adaptive Lasso [19], the smoothly clipped absolute deviations (SCAD) [6], the minimax concave penalty (MCP) [20] and the bridge penalty. It has been shown that the bridge penalty had the oracle estimation in the linear regression models having divergent number of covariates. For the AFT models, there also exists much literature (e.g., [7, 9, 22]). To name but a few, Huang et al. [7] considered the regularization approaches for estimation in the AFT model with high-dimension covariates based on Stute's weighted least squares method. Huang and Ma [8] considered variable selection for AFT model with bridge method. Wang and Song [18] applied adaptive lasso to the AFT models. In recent years, there are still a lot of studies on the AFT models. For example, Chai et al. [3] considered a set of low-dimensional covariates of main interest and a set of high-dimensional covariates that may also affect survival under the accelerated failure time model. Choi and Choi [4] proposed the logistic-kernel smoothing procedure for the semi-parametric AFT model with high-dimensional right-censored data. Li et al. [12] proposed a unified Expectation-Maximization approach combined with the L_1-norm penalty to perform variable selection and parameter estimation simultaneously in the accelerated failure time model with right-censored survival data of moderate sizes.

This article is motivated by the need for considering a seamless-L_0 (SELO) penalty [5] under the AFT model, which is a smooth function similar to the L_0 penalty. Under appropriate conditions, we show that the SELO selects a model whose dimension is comparable to the underlying model and prove that the proposed estimators is asymptotically normal. Monte Carlo simulations to evaluate the finite sample performance of the proposed procedure are computed. The proposed method is also demonstrated through an empirical analysis.

The rest of this paper is organized as follows. The AFT model is based on SELO penalization and computational algorithm are introduced in Section 2. In Section 3, we further propose an accurate variable selection for high dimensional sparse AFT model based on seamless L_0. The root n consistency and the asymptotic normality of the resulting estimate are established. We simulate Monte Carlo simulation study to examine the finite sample performance of the proposed estimate in Section 4. A real data example is used to illustrate the proposed methodology in Section 5.

2. SELO estimation in the AFT model

Let T_i be the logarithm of the failure time and X_i be the p-dimensional covariate vector. The AFT model assumes

$$T_i = \alpha + X_i\beta + \varepsilon_i,\ i = 1, \ldots, n, \tag{2.1}$$

where α is the intercept, $\beta \in \mathcal{R}^p$ is an unknown vector of interest, and ε_i is the random error. When T_i is subject to right censoring, we can only observe (Y_i, δ_i, X_i), where $Y_i = \min(T_i, C_i)$, X_i be the p-dimensional covariate vector for the ith row of the $n \times p$ matrix X which is the covariate matrix, C_i is the logarithm of the censoring time, and $\delta_i = I\{T_i \leq C_i\}$ is the censoring indicator. We assume that (Y_i, δ_i, X_i), $i = 1, \ldots, n$, come from the same distribution.

Let $\hat{F}_n$ be the Kaplan-Meier estimator of the distribution function F of T. $\hat{F}_n$ can be written as

$$\hat{F}_n(y) = \sum_{i=1}^{n} w_i I\{Y_{(i)} \leq y\},$$

where the w_i's are the jumps in the Kaplan-Meier estimator expressed as $w_1 = \frac{\delta_{(1)}}{n}$, $w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}$, $j = 2, \ldots, n$. w_i's are also called the Kaplan-Meier weights; see for examples in Stute and Wang [14]. Here $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics of Y_i's and $\delta_{(1)}, \ldots, \delta_{(n)}$ are the associated censoring indicators. Similarly, let $X_{(1)}, \ldots, X_{(n)}$ be the associated covariates of the ordered Y_i's. The weighted least square(WLS) loss function is

$$\frac{1}{2} \sum_{i=1}^{n} w_i (Y_{(i)} - \alpha - X_{(i)}\beta)^2. \tag{2.2}$$

Let $\bar{X}_w = \sum_{i=1}^{n} w_i X_{(i)} / \sum_{i=1}^{n} w_i$, and $\bar{Y}_w = \sum_{i=1}^{n} w_i Y_{(i)} / \sum_{i=1}^{n} w_i$, denoted by $X_{(i)}^* = (nw_i)^{1/2}(X_{(i)} - \bar{X}_w)$ and $Y_{(i)}^* = (nw_i)^{1/2}(Y_{(i)} - \bar{Y}_w)$. The weighted least square(WLS) objective function (2.2) can be written as

$$\ell_n(\beta) = \frac{1}{2} \sum_{i=1}^{n} (Y_{(i)}^* - X_{(i)}^T \beta)^2.$$

Penalized regression problem has been studied extensively. LASSO is one of the most popular and widely studied L_1 penalty. But it has been proved that its estimator may be inconsistent for model selection. The smoothly clipped absolute deviations (SCAD) and the minimax concave penalty (MCP) are another two popular penalties. SCAD is a continuous penalty and its estimator has oracle property. MCP also performs well in variable selection , whose estimator is consistent. However, L_0 penalty directly penalizes the non-zero parameters,whose drawback is the difficulty of computing because of its discontinuity. Seamless-L_0 (SELO) was proposed in Dicker [5], which was explicitly designed to minic L_0 penalty. It has been found that SELO possessed good theoretical properties.

We now describe the variable selection for AFT model via SELO. Coordinate descent is introduced to solve this problem. We propose tuning parameter λ using cross-validation. The SELO penalized objective function is,

$$Q(\beta) = \ell_n(\beta) + \sum_{j=1}^{p} p_{SELO}(\beta_j), \tag{2.3}$$

where SELO(β_j) is defined as,

$$p_{SELO}(\beta_j) = P_{SELO,\lambda,\tau}(\beta_j) = \frac{\lambda}{\log(2)} \log\left(\frac{|\beta_j|}{|\beta_j| + \tau} + 1\right),$$

and λ is tuning parameter. When λ is large, SELO may select small estimators. In the paragraph, λ is determined by Cross Validation. It is easy to see that when τ is enough small, $p_{SELO}(\beta_j) \approx \lambda I\{\beta_j \neq 0\}$, which is similar to L_0 penalty.

To minimize (2.3), we utilize coordinate descent algorithm. Coordinate descent algorithm [21] has been widely used in penalized regression problem, which optimizes an objective function by calculating a single parameter at a time until convergence is reached. Dicker [5] described this algorithm for obtaining SELO estimators. The algorithm is formulated in terms of the tuning parameter λ. For a fixed value of λ, it can be implemented in the following steps.

```
Algorithm of AFT model with SELO penalty
Step 1. Initialize $\beta_j^{(0)} = \mathbf{0}$, $j = 1, \cdots, p$.
Step 2. For the $k$-th iteration, we calculate the parameter from $\beta_1^k$ to $\beta_p^k$.
        $\tilde{\beta}_i^{(k)} = \mathrm{argmin} Q(\tilde{\beta}_1^k, \ldots, \tilde{\beta}_{i-1}^k, \beta_i, \beta_{i+1}^{k-1}, \ldots, \beta_p^{k-1})$.
Step 3. If $|\tilde{\beta}^{(k+1)} - \tilde{\beta}^{(k)}|$ is small or $k$ is very large, return $\beta^{(k+1)}$;
        otherwise increase $k$ to $k + 1$ and go to Step 2.
Step 4. Repeat Steps 2 and 3 until convergence.
```

3. Theoretical properties

In this section, we prove the consistency and asymptotic normality of WLS estimator via SELO under some conditions. Following the notation of Stute [14, 15], let H denote the distribution function of Y. Under the assumption of independence between T and C, $1 - H(y) = (1 - F(y))(1 - G(y))$, where F and G are the distribution functions of T and C. Let τ_Y, τ_T and τ_C be the endpoints of the support of Y, T and C. We put

$$\tilde{F}^0(x, y) = \begin{cases} F^0(x, y) & y < \tau_H, \\ F^0(x, \tau_H) + 1_{\{\tau_H \in A\}} F^0(x, \{\tau_H\}) & y \geq \tau_H. \end{cases}$$

Now, introduce the following sub-distribution functions:

$$\tilde{H}^1(x, y) = P(X \leq x, Z \leq y, \delta = 1) \;\; and \;\; \tilde{H}^0(y) = P(Z \leq y, \delta = 0).$$

Under random censoring the limit variance becomes much more complicated. Let

$$\gamma_0(y) = \exp\left\{\int_0^{y-} \frac{\tilde{H}^0(dz)}{1 - H(z)}\right\}, \;\; \gamma_1(y) = \frac{1}{1 - H(y)} \int 1_{y<w}(w - x^T\beta) x_j \gamma_0(w) \tilde{H}^1(dx, dw)$$

and

$$\gamma_2(y) = \int\int \frac{1_{\{v<y, v<w\}}(w - x^T\beta) x_j \gamma_0(w)}{[1 - H(v)]^2} \tilde{H}^0(dz) \tilde{H}^1(dx, dw).$$

We assume that

(A1) (a) $E[(Y - X^T\beta^*)^2 XX^T\delta] < \infty$, (b) $\int |(w - x^T\beta^*) x_j| D^{1/2}(w) \tilde{F}^0(dx, dw) < \infty$, for $j = 1, ..., p$ and $D(y) = \int_0^{y-} [(1 - H(w))(1 - G(w))]^{-1} G(dw)$.

(A2) $\lambda = O(1)$, $\tau = O(p^{1/2}/n^{3/2})$, $\lambda\sqrt{n/p} \to \infty$ and $p\sigma^2/n \to 0$ when $n \to \infty$.

(A3) $r \leq \lambda(E(XX^T)) \leq R$, where r and R are positive constant.

(A4) $\lim_{n\to\infty} n^{-1} \max_{1\leq i\leq n} \sum_{j=1}^p w_i^2 x_{ij}^2 = 0$.

(A5) $E\left(\left|\frac{\epsilon}{\sigma}\right|^{2+\frac{2\delta}{1+\delta}}\right) < M$ for some $M < 0$ and $\delta > 0$.

Condition (A1) is usually used for the proof of consistency in Stute [14]. Condition (A2) restricts the size of λ and τ. Condition (A3) gives the bound of the eigenvalues of $E(XX^T)$, which is used in Theorem 3.1. Conditions (A4) and (A5) are used for the proof the asymptotic normality of SELO estimators and are related to the Lindeberg condition of Lindeberg-Feller CLT.

Theorem 3.1. *Suppose that conditions (A1)–(A5), then*

(i) $\lim_{n\to\infty} P(\{j : \hat{\boldsymbol{\beta}}_j \neq 0\} = A) = 1$, $A = \{j;\beta_j \neq 0\}$.

(ii) $\sqrt{n}(n^{-1}X_A^T W X_A/\sigma^2)^{1/2}(\hat{\beta}_A - \beta_A^*) \xrightarrow{D} (\sigma^2(X_A^T W X_A))^{-\frac{1}{2}} G_A$ *where* $G \sim N(0,\Sigma)$ *and* $\Sigma = Var\{\delta\gamma_0(Y)(Y - X\beta^*)X + (1-\delta)\gamma_1(Y;\beta^*) - \gamma_2(Y;\beta^*)\}/n$ *and* G_A *is the part of* G *corresponding to* β_A^*.

The proof is put in the supplementary.

4. Numerical results

Let T be generated from $T = X\beta + \epsilon$, where $\epsilon \sim N(0,\sigma)$. The covariates $X = (X_1, ..., X_p)$ are standard normal. Here we set $\sigma = 0.1$. The censoring variables are generated as uniformly distributed $U(0, C_0)$ and independent of the events, where C_0 is chosen to obtain the censoring rate 25% and 40%. Set two sample sizes $n = 200$ and $n = 400$. The tuning parameter λ is chosen by cross validation. For each value of n, we simulated 1000 independent datasets $\{(y_1, x_1^T), ..., (y_n, x_n^T)\}$. For each dataset, we calculated estimates of β. For each estimator $\hat{\beta}$, we recorded: the model size, $\hat{A} = \{j;\hat{\beta}_j \neq 0\}$; an indicator of whether or not the true model was selected, $I\{\hat{A} = A\}$; the false positive rate, $|\hat{A} - A|/|\hat{A}|$; the false negative rate, $|A - \hat{A}|/(p - |\hat{A}|)$; and the model error, $(\hat{\beta} - \beta^*)^T(\hat{\beta} - \beta^*)$. The column labeled "size", "rate","F+","F-" and "MSE" represent the above indicators. Results for SELO, LASSO, SCAD and MCP are summarized in the tables. Furthermore, we use the V-fold cross-validation to determine the tuning parameter. The CV score is $\sum_{v=1}^{V}[\ell_n(\hat{\beta}^{(-v)}) - \ell_n^{(-v)}(\hat{\beta}^{(-v)})]$. In this article, we set $V = 5$.

4.1. Simulation I

The example was conducted with $p = 8$ and set $\beta = (3, 1.5, 0, 0, 1, 0, 0, 0) \in R^8$, where Table1 summarizes the variable selection results based on SELO, LASSO, SCAD and MCP when censoring rates are 25% and 40%. Overall, SELO performs better than other three methods, which selects the correct model more frequently. For instance, when the censoring rate is 25%, SELO selects the true model most accurately. The true model size is 3 and the average size from SELO is 3.43. LASSO performs worse than other three methods both in model size and correct rate. LASSO, SCAD and MCP select model with average size 4.12, 4.05 and 4.04 and select the correct model in 42%, 46.7% and 47%. Similar results perform when the censoring rate is 40%. But we can easily see that the situation when the censoring rate is 25% is better than that when the censoring rate is 40%. When n increases, we can see that the results are better. For instance, SELO selects 3.07 variables when $n = 400$, which is more accurate compared to 3.43 when $n = 200$. Other indicators can also prove it.

Table 1. Simulation results for $p = 8$.

		25% censoring					40% censoring				
n	Method	size	rate	F+	F-	MSE	size	rate	F+	F-	MSE
200	SELO	3.43	0.568	0.137	0.156	2.61	3.62	0.38	0.267	0.287	2.55
	LASSO	4.12	0.429	0.212	0.428	3.62	4.43	0.338	0.276	0.567	2.72
	SCAD	4.05	0.467	0.198	0.417	2.41	4.36	0.346	0.271	0.537	2.49
	MCP	4.04	0.47	0.196	0.412	2.46	4.33	0.359	0.267	0.531	2.45
400	SELO	3.07	0.894	0.035	0.029	2.27	3.27	0.589	0.147	0.139	2.43
	LASSO	3.80	0.620	0.136	0.277	3.44	3.96	0.519	0.178	0.373	2.66
	SCAD	3.74	0.641	0.131	0.275	2.20	3.93	0.534	0.171	0.363	2.42
	MCP	3.74	0.641	0.131	0.275	2.24	3.92	0.533	0.170	0.361	2.40

4.2. Simulation II

The example was conducted with $p = 50$ and set $\beta = (3, 1.5, 0, 0, 2, 0, 3, 0, 0, 2, 0, \dots, 0) \in R^{50}$, the rest of β is zero. Other settings were similar to the case in simulation I and the results are listed in Table 2. It is easy to see that SELO remains better performance compared to other three methods. The model size from SELO is 5.68, which is the closest to the true model when the censoring rate is 25% compared to 9.64, 9.27 and 8.86 for LASSO, SCAD and MCP. And it also performs better in correct rate and other indicators. For instance, SELO selects 52% correct variables which is better compared to 3%, 5% and 7% for LASSO, SCAD and MCP. The indicators "F+" and "F-" of SELO are the smallest among four methods. Similar results perform when $p = 50$ compared to $p = 8$. It can be concluded that the results are worse when the censoring rate is 40%. However, when n increases from 200 to 400, SELO selects more correct models.

Table 2. Simulation results for $p = 50$.

		25% censoring					40% censoring				
n	Method	size	rate	F+	F-	MSE	size	rate	F+	F-	MSE
200	SELO	5.68	0.29	0.194	0.028	3.05	6.27	0.27	0.267	0.287	2.55
	LASSO	9.64	0.03	0.416	0.127	5.32	9.27	0.08	0.276	0.567	2.72
	SCAD	9.27	0.05	0.396	0.116	5.19	8.91	0.12	0.271	0.537	2.49
	MCP	8.86	0.07	0.374	0.104	5.88	8.42	0.10	0.267	0.531	2.45
400	SELO	5.09	0.52	0.138	0.015	5.00	5.48	0.29	0.245	0.031	2.805
	LASSO	9.2	0.03	0.396	0.124	5.18	9.42	0.07	0.405	0.116	3.356
	SCAD	8.66	0.10	0.357	0.101	5.20	8.99	0.09	0.375	0.103	2.830
	MCP	8.56	0.13	0.347	0.099	5.53	8.70	0.11	0.357	0.095	2.662

4.3. Simulation III

The example is set under 25% and 40% censoring, and we also estimate the mean estimated variance across 1000 simulated datasets when n is 200 and 400. From the Tables 3 and 4, we see that SELO with tuning over $\tau \in \{0.001, 0.01, 0.1, 0.5\}$ seems to give better variance when compared to SELO with $\tau = 0.01$ fixed. We can see that the 1000 estimators remain stable whenever censoring rate are 25% and 40% and the simulation results perform better when n increases to 400.

Table 3. Variance of SELO estimator under 25% censoring.

n	τ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
200	0.001	0.322	0.254	0.103	0.104	0.270	0.103	0.09	0.11
	$\{0.001, 0.01, 0.1, 0.5\}$	0.319	0.204	0.011	0.004	0.940	0.01	0.02	0.001
400	0.001	0.152	0.125	0.045	0.043	0.135	0.043	0.03	0.04
	$\{0.001, 0.01, 0.1, 0.5\}$	0.145	0.120	0.009	0.026	0.136	0.005	0.02	0.02

Table 4. Variance of SELO estimator under 40% censoring.

n	τ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
200	0.001	0.319	0.376	0.041	0.032	0.326	0.037	0.03	0.32
	$\{0.001, 0.01, 0.1, 0.5\}$	0.316	0	0.010	0.007	0.007	0.002	0.003	0.012
400	0.001	0.157	0.322	0.003	0.003	0.162	0.003	0.015	0.021
	$\{0.001, 0.01, 0.1, 0.5\}$	0.154	0.077	0.001	0	0.098	0	0	0.002

4.4. PBC Data

PBC data was collected in the Mayo Clinic trial of primary biliary cirrhosis of liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data were participated in the randomized trial and contained complete data. The additional 112 cases did not participate in the clinical trial. After deleting the missing data, the remaining 276 datasets are used for the analysis. We consider 17 covariates: age, albumin, alk.phos, ascites, ast, bili, chol, copper, platelet, edema, hepato, protime, sex, spiders, stage, trt, trig.

Among 276 samples without losing data, we calculated Table 5. The optimal values of lambda with SELO is small, which is 0.008. And the optimal values of lambda that are chosen by CV for LASSO, SCAD and MCP are 0.011. LASSO selects 6 variables and SELO selects 3 variables. LASSO selects the genes including sex, heptato, bili, albumin, protime, stage. SCAD selects the same variables like MCP. However, SELO selects sex,albumin and protime. The variables selected by SELO are also contained by other three methods. Meanwhile, we also calculate the AIC (Akaike Information Criteria) $AIC = n\log(\hat{\sigma}^2) + 2(d+1)$ (d is the non-zero parameter and $\hat{\sigma}$ is the error between the estimator and true value) for four methods which show that SELO results better with smaller AIC. In Table 6, we also calculated p values for the coefficients of variables selected by SELO. We calculated p values for several steps. Firstly, we calculated t values called t_k where $t_k = \hat{\beta}/s$ and s is the sum of squares error of the estimators. Secondly, we found the value of $t_{\alpha/2}(n-K)$ where $\alpha = 0.05$ and $n-K$ is the degree. Finally, we calculated the values of $P(T{>}t_{\alpha/2}(n-K)) = P(T{<}t_{\alpha/2}(n-K))$.

Table 5. PBC data: Estimated coefficients and selected variables.

Method	Model size	R^2	Covariate	AIC
SELO	3	0.349	sex,albumin,protime	68.72
LASSO	6	0.249	sex,heptato,bili,albumin,protime,stage	85.76
SCAD	5	0.299	sex,heptato,bili,albumin,protime	75.48
MCP	5	0.299	sex,heptato,bili,albumin,protime	75.48

Table 6. PBC data: Significance test and p value.

variable	coefficient	p-value
sex	0.624	1.53×10^{-11}<0.05
albumin	0.341	1.2×10^{-4}
protime	0.210	1.004×10^{-4}
hepato	-0.045	0.217
bili	-0.068	0.116
stage	-0.036	0.255

Table 6 indicates the significance test results. From the results, we can see that the p value of sex, albumin and protime are all less than 0.05. The variables selected by SELO are all significant. Overall, SELO selected a simple model compared to another three methods.

5. Conclusions

Statistical analysis of failure time with high dimension covariates is an important topic. In this article, we investigate a new method (SELO) for the AFT model with high dimension covariates, for simultaneous variable selection and estimation. A real dataset (PBC) is analyzed and SELO selects some important covariates. Our numerical results indicate that SELO performs better than another three methods. In this article, we address the situation where $p \ll n$ and prove the oracle property under the condition $p/n \to 0$. We allow both n and p to diverge but p goes to infinity more slowly than n. The situation where p is much larger than n will be extended in the future research.

Acknowledgments

The authors would like to thanks the editors and three reviewers for their valuable and helpful comments. The authors also thanks Guangren Yang and Yiming Liu for their advice for revising the paper.

Conflict of interest

The authors declare no conflict of interest.

References

1. J. Buckley, I. James, Linear regression with censored data, *Biometrika*, **66** (1979), 429–436. https://doi.org/10.1093/biomet/66.3.429
2. D. R. Cox, Regression models and life-tables (with discussion), *J. Roy. Stat. Soc. Ser. B*, **34** (1972), 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
3. H. Chai, Q. Z. Zhang, J. Huang, S. G. Ma, Inference for low-dimensional covariates in a high-dimensional accelerated failure time model, *Stat. Sinica*, **29** (2019), 877–894. https://doi.org/10.5705/ss.202016.0449
4. T. Choi, S. Choi, A fast algorithm for the accelerated failure time model with high-dimensional time-to-event data, *J. Stat. Comput. Simul.*, **91** (2021), 3385–3403. https://doi.org/10.1080/00949655.2021.1927034
5. L. Dicker, B. S. Huang, X. H. Lin, Variable selection and estimation with the seamless-L 0 penalty, *Stat. Sinica*, **23** (2013), 929–962. https://dx.org/10.5705/ss.2011.074
6. J. Q. Fan, R. Z. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.*, **96** (2001), 1348–1360. https://doi.org/10.1198/016214501753382273
7. J. Huang, S. G. Ma, H. L. Xie, Regularized estimation in the accelerated failure time model with high-dimensional covariates, *Biometrics*, **62** (2006), 813–820. https://doi.org/10.1111/j.1541-0420.2006.00562.x
8. J. Huang, S. G. Ma, Variable selection in the accelerated failure time model via the bridge method, *Lifetime Data Anal.*, **16** (2010), 176–195. https://doi.org/10.1007/s10985-009-9144-2
9. S. M. Hu, J. S. Rao, *Sparse penalization with censoring constraints for estimating high dimensional AFT models with applications to microarray data analysis*, Technical reports, University of Miami, 2010.
10. J. D. Kalbfleisch, R. L. Prentice, *The statistical analysis of failure time data*, John Wiley & Sons. Inc., New Jersey, **2** (2011), 168–170. https://doi.org/10.1016/0197-2456(81)90009-X
11. Y. D. Kim, H. Choi, H. S. Oh, Smoothly clipped absolute deviation on high dimensions, *J. Am. Stat. Assoc.*, **103** (2008), 1665–1673. https://doi.org/10.1198/016214508000001066
12. Y. Li, M. X. Liang, L. Mao, S. J. Wang, Robust estimation and variable selection for the accelerated failure time model, *Stat. Med.*, **40** (2021), 4473–4491. https://doi.org/10.1002/sim.9042
13. Y. Ritov, Estimation in a linear regression model with censored data, *Ann. Stat.*, **18** (1990), 354–372. https://doi.org/10.1214/aos/1176347502
14. W. Stute, Consistent estimation under random censorship when covariables are present, *J. Multivariate Anal.*, **45** (1993), 89–103. https://doi.org/10.1006/jmva.1993.1028
15. W. Stute, Distributional convergence under random censorship when covariables are present, *Scand. J. Stat.*, **23** (1996), 461–471. https://doi.org/10.1016/s0167-7152(98)00069-8
16. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Roy. Stat. Soc. B*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

17. A. A. Tsiatis, Estimating regression parameters using linear rank tests for censored data, *Ann. Stat.*, **18** (1990), 354–372. https://doi.org/354-372.10.1214/aos/1176347504

18. X. G. Wang, L. X. Song, Adaptive Lasso variable selection for the accelerated failure models, *Commun. Stat.-Theor. M.*, **40** (2011), 4372–4386. https://doi.org/10.1080/03610926.2010.513785

19. H. Zou, The adaptive lasso and its oracle properties, *J. Am. Stat. Assoc.*, **101** (2006), 1418–1429. https://doi.org/10.1198/016214506000000735

20. H. Zou, Nearly unbiased variable selection under minimax concave penalty, *J. Am. Stat. Assoc.*, **38** (2010), 894–942. https://doi.org/894-942.10.1214/09-AOS729

21. W. J. Fu, Penalized regressions: The bridge versus the lasso, *J. Comput. Graph. Stat.*, **7** (1998). https://doi.org/397-416.10.1214/09-AOS729

22. M. H. R. Khan, J. E. H. Shaw, Variable selection for survival data with a class of adaptive elastic net techniques, *Stat. Comput.*, **26** (2016), 725–741. https://doi.org/10.1007/s11222-015-9555-8

Supplementary

In order to complete the proofs of Theorem 3.1 (i), we first show two lemmas.

Lemma 1. Recall that

$$Q(\beta) = \frac{1}{2}\sum_{i=1}^{n} w_i(Y_{(i)} - X_{(i)}^{\mathrm{T}}\beta)^2 + \sum_{j=1}^{p} P_{SELO}(\beta_j). \tag{A.1}$$

Then for every $r \in (0,1)$, there exists a constant C_0>0 such that

$$\liminf_{n\to\infty} P\left[\underset{\|\beta-\beta^*\|\le C\sqrt{p\sigma^2/n}}{\operatorname{argmin}} Q_n(\beta) \subset \left\{\beta \in R^p; \|\beta-\beta^*\|<C\sqrt{\frac{p\sigma^2}{n}}\right\}\right]>1-r$$

where $C \ge C_0$.

Proof. Let $\alpha_n = \sqrt{p\sigma^2/n}$ and fix $r \in (0,1)$. To prove the Lemma 1, it suffices to show that if C>0 is large enough, then

$$P\left\{\sup_{\|u\|=1} Q(\beta^* + C\alpha_n u)>Q(\beta^*)\right\} \ge 1-\epsilon.$$

Furthermore, define $Q_n(u) = Q(\beta^* + C\alpha_n u) - Q(\beta^*)$.Then,

$$\begin{aligned} Q_n(u) &= \frac{1}{2}C^2\alpha_n^2 u^T\left[\sum_{i=1}^{n} w_i X_{(i)}X_{(i)}^T\right]u - \sum_{i=1}^{n} w_i(Y_{(i)} - X_{(i)}^T\beta^*)X_{(i)}^T C\alpha_n u + \sum_{j=1}^{p}[P_{SELO}(\beta_j^* + C\alpha_n u_j) - P_{SELO}(\beta_j^*)] \\ &\ge \frac{1}{2}C^2\alpha_n^2 u^T\left[\sum_{i=1}^{n} w_i X_{(i)}X_{(i)}^T\right]u - \sum_{i=1}^{n} w_i(Y_{(i)} - X_{(i)}^T\beta^*)X_{(i)}^T C\alpha_n u + \sum_{j\in K(u)}[P_{SELO}(\beta_j^* + C\alpha_n u_j) - P_{SELO}(\beta_j^*)]. \end{aligned}$$

By the results of Stute (1993, 1996), we have

$$\sum_{i=1}^{n} w_i X_{(i)}X_{(i)}^T \xrightarrow{P} E(XX^T), \quad \textit{and} \quad \sqrt{n}\sum_{i=1}^{n} w_i(Y_{(i)} - X_{(i)}^T\beta^*)X_{(i)}^T \xrightarrow{D} W,$$

where $W \sim N(0, \Sigma)$, with Σ defined in the theorem. The last term in $Q_n(u)$ where $K(u) = \{j; P_{SELO}(\beta_j^* + C\alpha_n u_j) - P_{SELO}(\beta_j^*) < 0\}$. The fact that P_{SELO} is concave on $[0, \infty)$ imply that, for each β, $P_{SELO}(\beta_j^* + C\alpha_n u_j) - P_{SELO}(\beta_j^*) \geq -C\alpha_n |u_j| P'_{SELO}(\beta_j + C\alpha_n u_j)$

$$\begin{aligned} Q_n(u) &\geq \frac{1}{2}C^2\alpha_n^2 u^T \left[\sum_{i=1}^n w_i X_{(i)} X_{(i)}^T\right] u - \sum_{i=1}^n w_i (Y_{(i)} - X_{(i)}^T \beta^*) X_{(i)}^T C\alpha_n u - \sum_{j \in K(u)} C\alpha_n |u_j| P'_{SELO}(\beta_j^* + C\alpha_n u_j) \\ &= \frac{1}{2}C^2\alpha_n^2 u^T \left[\sum_{i=1}^n w_i X_{(i)} X_{(i)}^T\right] u - \sum_{i=1}^n w_i (Y_{(i)} - X_{(i)}^T \beta^*) X_{(i)}^T C\alpha_n u - C\alpha_n \sum_{j \in K(u)} \frac{\lambda}{\log(2)} \frac{\tau}{(2|\beta_j^*| + \tau)(|\beta_j^*| + \tau)} \\ &\triangleq I_1 + I_2 + I_3. \end{aligned}$$

Under conditions (A1) and (A3), for I_1,

$$\sum_{i=1}^n w_i X_{(i)} X_{(i)}^T \xrightarrow{P} E(XX^T), \quad and \quad I_1 \geq \frac{1}{2}C^2\alpha_n^2 R.$$

For I_2,

$$\sqrt{n} \sum_{i=1}^n w_i (Y_{(i)} - X_{(i)}^T \beta^*) X_{(i)}^T \xrightarrow{D} N(0, \Sigma), \quad and \quad I_2 = O_p\left(\frac{C\alpha_n}{\sqrt{n}}\right).$$

For I_3, by condition (A2), we have $I_3 = o_p(C\alpha_n)$. We conclude that if $C>0$ is large enough, then $\inf_{||u||=1} Q_n(u)>0$ holds for all n sufficiently large, with probability at least $1 - r$. This finishes the proof of the Lemma 1. □

Lemma 2. We assume that $C>0$, $Q(\beta)$ is similar to that in Lemma 1. Under the conditions (A1)–(A3),

$$\lim_{n\to\infty} P\left[\underset{||\beta-\beta^*|| \leq C\sqrt{p\sigma^2/n}}{\operatorname{argmin}} Q_n(\beta) \subset \{\beta \in R^p; \beta_{A^c} = 0\}\right] = 1,$$

where $A^c = \{1, ..., p\}/A$ is the complement of A in $\{1, ..., p\}$.

Proof. Suppose that $\beta \in R^p$ and that $||\beta - \beta^*||<C\alpha_n u$. Define $\tilde{\beta} \in R^p$ by $\tilde{\boldsymbol{\beta}}_{A^c} = 0$ and $\tilde{\beta}_A = \beta_A$. Similar to the proof of Lemma 1, if $D_n(\beta,\tilde{\beta}) = Q_n(\beta) - Q_n(\tilde{\beta})$, then

$$\begin{aligned} D_n(\beta,\tilde{\beta}) &= \frac{1}{2}\sum_{i=1}^n w_i (Y_{(i)} - X_{(i)}^T \beta)^2 - \frac{1}{2}\sum_{i=1}^n w_i (Y_{(i)} - X_{(i)}^T \tilde{\beta})^2 + \sum_{j \in A^c} P_{SELO}(\beta_j) \\ &= \frac{1}{2}\sum_{i=1}^n w_i (\beta - \tilde{\beta})^T X_{(i)} X_{(i)}^T (\beta - \tilde{\beta}) - \sum_{i=1}^n w_i (Y_{(i)} - X_{(i)}^T \tilde{\beta}) X_{(i)}^T (\beta - \tilde{\beta}) + \sum_{j \in A^c} P_{SELO}(\beta_j) \\ &\triangleq I_1 + I_2 + I_3. \end{aligned}$$

For I_1 and I_2, under the conditions (A1) and (A3),

$$I_1 + I_2 = O_p(||\beta - \tilde{\beta}|| \sqrt{\frac{p\sigma^2}{n}}).$$

For I_3, P_{SELO} is concave $||\beta||<C\sqrt{p\sigma^2/n} \rightarrow |\beta_j|<C\sqrt{p\sigma^2/n}$, and

$$\sum_{j\in A^c} P_{SELO}(\beta_j) > \frac{\tau}{(2|C\sqrt{p\sigma^2/n}|+\tau)(|C\sqrt{p\sigma^2/n}|+\tau)}||\beta-\tilde{\beta}|| > 0.$$

Under the condition (A2),we combine the results of I_1, I_2 and I_3. So, $D_n(\beta,\tilde{\beta})>0$. □

Combining the proof of Lemmas 1 and 2, we can have the conclusion of Theorem 3.1 (i).

Proof of Theorem 3.1 (ii). We consider the proof related to the Lindeberg condition of Lindeberg-Feller CLT. Under the conditions (A1)–(A5), let $\hat{\beta}_A$ be a estimator where $A = \{j; \hat{\beta}_j \neq 0\}$.

We can easily have the following form,

$$\hat{\beta}_A = \beta_A^* + \left(X_A^T W X_A\right)^{-1} X_A^T \epsilon - \left(X_A^T W X_A\right)^{-1} p'_A(\hat{\beta})$$

and

$$\sqrt{n}\left(\frac{(n^{-1}X_A^T W X_A)}{\sigma^2}\right)^{1/2}(\hat{\beta}_A - \beta_A^*) = \left(\sigma^2 X_A^T W X_A\right)^{-1/2} X_A^T W X_A \epsilon - \left(\sigma^2 \sum_{i=1}^{n} X_A^T W X_A\right)^{-1/2} p'_A(\hat{\beta}).$$

To prove,

$$\left(\sigma^2 X_A^T W X_A\right)^{-\frac{1}{2}} X_A^T W \epsilon \rightarrow N(0,G) \ \ and \ \ \left(\sigma^2 X_A^T W X_A\right)^{-\frac{1}{2}} X_A^T W \epsilon = \sum_{i=1}^{n} w_{i,n},$$

where $w_{i,n} = \left(\sigma^2 X_A^T W X_A\right)^{-\frac{1}{2}} w_i x'_{(i),A}\epsilon_i$. Let $\eta_{i,n} = w_i x^T_{(i),A}\left(X_A^T W X_A\right)^{-\frac{1}{2}} (X_A^T W X_A)^{-\frac{1}{2}} w_i x_{(i),A}$.

By the condition of Lindeberg-Feller CLT, we have

$$E[||w_{i,n}||^2; ||w_{i,n}||^2>\delta_0] = \eta_{i,n} E\left[\frac{\epsilon_i^2}{\sigma^2}; \eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0\right] = \eta_{i,n}\int_{\eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0} \frac{\epsilon_i^2}{\sigma^2} dF(x).$$

By Holder inequality, set $\frac{1}{p} = \frac{2}{2+\delta}$, $\frac{1}{q} = \frac{\delta}{\delta+2}$,

$$\begin{aligned} E[||w_{i,n}||^2; ||w_{i,n}||^2>\delta_0] &\leq \eta_{i,n}\left(\int_{\eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0}\left|\frac{\epsilon_i}{\sigma}\right|^{2p} dF(x)\right)^{\frac{1}{p}}\left(\int_{\eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0} 1^q dF(x)\right)^{\frac{1}{q}} \\ &= \eta_{i,n}\left(\int_{\eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0}\left|\frac{\epsilon_i}{\sigma}\right|^{2+\delta} dF(x)\right)^{\frac{2}{2+\delta}} P\left\{\eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0\right\}^{\frac{\delta}{2+\delta}} \\ &= \eta_{i,n} E\left(\left|\frac{\epsilon_i}{\sigma}\right|^{2+\delta}\right)^{\frac{2}{2+\delta}} P\left\{\eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0\right\}^{\frac{\delta}{2+\delta}}. \end{aligned}$$

By Markov inequality,

$$P\left\{\eta_{i,n}\frac{\epsilon_i^2}{\sigma^2}>\delta_0\right\}^{\frac{\delta}{2+\delta}} \leq \left(\frac{\eta_{i,n}}{\delta_0} E\left\{\frac{\epsilon_i^2}{\sigma^2}\right\}\right)^{\frac{\delta}{2+\delta}} = \delta_0^{\frac{-\delta}{2+\delta}} \eta^{\frac{\delta}{2+\delta}} E\left(\left|\frac{\epsilon_i}{\sigma}\right|^2\right)^{\frac{\delta}{2+\delta}}$$

and

$$E[\|w_{i,n}\|^2; \|w_{i,n}\|^2 > \delta_0] \quad \leq \quad \eta^{1+\frac{\delta}{2+\delta}} \delta_0^{-\frac{\delta}{\delta+2}} E\left(\left|\frac{\epsilon}{\sigma}\right|\right)^{\frac{4\delta+4}{2+\delta}} .$$

We showed that $\sum_{i=1}^n \eta_{i,n} = \sum_{i=1}^n w_i$ and $\eta_{i,n} \leq \|(n^{-1} X_A^T W X_A)^{-\frac{1}{2}}\|^2 \max\limits_{1\leq i\leq n} \sum_{j=1}^q \frac{1}{n} w_i^2 x_{ij}^2$

$$\sum_{i=1}^n E[\|w_{i,n}\|^2; \|w_{i,n}\|^2 > \delta_0] \quad \leq \quad \delta_0^{-\frac{\delta}{2+\delta}} E\left(\left|\frac{\epsilon}{\sigma}\right|^{2+\frac{2\delta}{1+\delta}}\right) \sum_{i=1}^n w_i^2 x_{ij}^2 \max_{1\leq i\leq n} \eta_{i,n}^{\frac{\delta}{2+\delta}} .$$

By conditions (A4) and (A5), $\sum_{i=1}^n E[\|w_{i,n}\|^2; \|w_{i,n}\|^2 > \delta_0] \to 0$.

By conditions (A1)–(A3), $\| \left(\sigma^2 \sum_{i=1}^n X_A^T W X_A\right)^{-1/2} p'_A(\hat{\beta})\| = o_p(1)$,
and

$$\sqrt{n} \sum_{i=1}^n w_i (Y_{(i)} - X_{(i)}^T \beta^*) X_{(i)}^T \xrightarrow{D} N(0, \Sigma),$$

where, $\Sigma = Var\{\delta\gamma_0(Y)(Y - X\beta^*)X + (1-\delta)\gamma_1(Y;\beta^*) - \gamma_2(Y;\beta^*)\}$.

So, $(\sigma^2 X_A^T W X_A)^{-\frac{1}{2}} X_A^T W \epsilon \xrightarrow{D} (\sigma^2 X_A^T W X_A)^{-\frac{1}{2}} G_A$,where $G \sim N(0, \Sigma)$ follows. □

AIMS Press