Acta Mathematica Sinica, English Series
Jul., 2020, Vol. 36, No. 7, pp. 812–828
Published online: July 15, 2020
https://doi.org/10.1007/s10114-020-9047-4
http://www.ActaMath.com

Acta Mathematica Sinica, English Series

Adaptive Penalized Weighted Least Absolute Deviations Estimation for the Accelerated Failure Time Model

Ming Qiu WANG
School of Statistics, Qufu Normal University, Qufu 273165*, P. R. China*
E-mail: *wmq*0829*@gmail.com*

Yuan Shan WU **Qing Long YANG**[1)]
School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073*, P. R. China*
E-mail: *wu@zuel.edu.cn* *yangqinglong@zuel.edu.cn*

Abstract The accelerated failure time model always offers a valuable complement to the traditional Cox proportional hazards model due to its direct and meaningful interpretation. We propose a variable selection method in the context of the accelerated failure time model for survival data, which can simultaneously complete variable selection and parameter estimation. Meanwhile, the proposed method can deal with the potential outliers in survival times as well as heteroscedastic model errors, which are frequently encountered in practice. Specifically, utilizing the general nonconvex penalty, we propose the adaptive penalized weighted least absolute deviation estimator for the accelerated failure time model. Under some regularity conditions, we show that the proposed method yields consistent estimator and possesses the oracle property. In addition, we propose a new algorithm to compute the estimate in the high dimensional settings, and evaluate the practical utility of the proposed method through extensive simulation studies and two real examples.

Keywords Heteroscedastic errors, Kaplan–Meier estimator, least absolute deviation, nonconvex penalty, oracle property, outliers, robustness, survival analysis

MR(2010) Subject Classification 62N01, 62N02, 62F12

1 Introduction

More explanatory variables are usually needed to be introduced into the model to achieve more precise prediction. However, this method may take the risk of introducing unnecessary explanatory variables and increase the computational burdens. In recent two decades, lots of penalized variable selection procedures have been proposed to obtain better prediction with the aims of using as less explanatory variables as possible and meanwhile maintaining computational convenience. Examples include the bridge (Frank and Friedman [4]), the least absolute shrinkage and selection operator (Lasso, Tibshirani [19]), the smoothly clipped absolute deviation (Fan and

Received February 21, 2019, revised October 19, 2019, accepted December 5, 2019
Supported by the National Natural Science Foundation of China (Grant Nos. 11671311, 11771250 and 11971324), the Natural Science Foundation of Shandong Province (Grant No. ZR2019MA002) and the National Key Research and Development Program of China (Grant No. 2018YFC1314600)
1) Corresponding author

Li [3]), the gradient directed regularization method (Friedman and Popescu [5]), the adaptive Lasso (Zou [28]; Zhang and Lu [26]) and the minimax concave penalty (Zhang [25]).

Due to the direct interpretation of the log survival time using the covariates, the accelerated failure time (AFT) model provides a useful alternative to the Cox proportional hazards model (Wei [24]; Kalbfleisch and Prentice [11]). The AFT model has been frequently adopted to fit the survival data in practice. Recently, the penalized methods for variable selection and estimation in the AFT model has been attracted much attention. Huang et al. [8] studied the regularized estimation in the AFT model with high-dimensional covariates. Huang and Ma [7] studied variable selection in the AFT model via the bridge method. However, this work was built upon the penalized weighted least squares. It may suffer the sensitivity from the possible outliers in the response. When there exists heteroscedastic errors, it renders efficiency loss.

The least absolute deviation (LAD) method, however, can largely avoid the shortcomings from the least squares and enables to handle the heteroscedastic errors and outliers in the response and thus it is appealing in practice. Huang [9] studied the LAD estimation for the AFT model and obtained the consistency and asymptotic normality of the estimator. However, they did not consider the problem of variable selection. Cai et al. [2] studied the rank-based estimation procedure with Lasso-type penalty to develop parsimonious prediction models for the AFT model. Johnson et al. [10] proposed the Buckley–James estimation by using the SCAD penalty. Zhou et al. [27] proposed a novel method without conducting the Kaplan–Meier estimation to study the variable selection.

In this paper, we study the adaptive penalized weighted LAD (WLAD) estimator in the AFT model by using the Kaplan–Meier weights to account for censoring. The adaptive weighted penalty based on the nonconvex penalty is employed to complete the variable selection and parameter estimation. Under some mild conditions, we obtain the consistency of the estimator. Furthermore, with a proper choice of the tuning parameter, the resultant estimator enjoys the oracle property. In simulation studies, we consider three scenarios to demonstrate the finite sample performance of the penalized estimate. Furthermore, we use two real data examples to illustrate applications of the proposed method. The advantage of our method, based on the simulation studies, is that our method performs better than the adaptive Lasso. Unlike the adaptive Lasso, we adopt the derivative of the general nonconvex penalty as the weight, which contains the SCAD and MCP as special cases. In addition, we propose a new algorithm to compute the estimate in high dimensional settings, and study the finite sample performance of our method through simulations and real examples. When we analyze the high-dimensional microarray gene expression data, we do not screen the genes before modeling.

The main contributions of this work are two-fold. First, the proposed method inherits the advantages of the least absolute deviation method over the least squares method, which includes the robustness for potential outliers in survival response times and the ability of handling heteroscedastic model errors. Second, by utilizing the local asymptotic normality theory (Le Cam [15]; van der Vaart [20]), we make painstaking efforts to overcome the non-smooth loss function encoupled with a non-convex penalty to establish the asymptotic properties of the proposed method.

The rest of the paper is organized as follows. We propose the APWLAD estimator in Sec-

tion 2. We establish the asymptotic selection consistency of the proposed method in Section 3. Some computational issues for the APWLAD estimator and the choice of the tuning parameter are discussed in Section 4. We show the practical utility of the proposed method through both simulated and real data sets in Section 5. We conclude with some remarks in Section 6 and delineate the proofs of theorems in the Appendix.

2 Method

Let T denote the transformed failure time under a known monotone transformation, e.g., the logarithm function. Let C denote the censoring time under the same transformation. Consider the linear regression model

$$T = \boldsymbol{\beta}^\top \boldsymbol{X} + \varepsilon, \tag{2.1}$$

where $\boldsymbol{\beta}$ is an unknown d-vector regression coefficient, $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ is the associated d-dimensional predictor, and ε is the random error with an unknown distribution. Here we do not consider the intercept. Due to the right censoring, the observed survival time is denoted by $Y = \min(T, C)$ and the censoring indicator by $\delta = I(T \le C)$, where $I(\cdot)$ is an indicator function. For $i = 1, \ldots, n$, let $(Y_i, \delta_i, \boldsymbol{X}_i)$ be an independent and identically distributed sample distributed as $(Y, \delta, \boldsymbol{X})$.

Let $Y_{(1)} \le \cdots \le Y_{(n)}$ be the order statistics of Y_i's, $\delta_{(1)}, \ldots, \delta_{(n)}$ be the associated censoring indicators, and $\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(n)}$ be the associated covariates. Let F be the cumulative distribution function of T and $\hat{F}_n$ be one minus its Kaplan–Meier estimator. Following Stute and Wang [18],

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

where

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

are the so-called Kaplan–Meier weights. Consequently, the WLAD estimator is defined as the minimizer of

$$L_n(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} w_i |Y_{(i)} - \boldsymbol{\beta}^\top \boldsymbol{X}_{(i)}|. \tag{2.2}$$

We consider the penalized LAD objective function for estimating $\boldsymbol{\beta}$ as follows,

$$Q_n(\boldsymbol{\beta}) \equiv L_n(\boldsymbol{\beta}) + \sum_{j=1}^{d} p'_{\lambda_n}(|\beta_j^0|)|\beta_j|, \tag{2.3}$$

where $\boldsymbol{\beta}^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_d^0)^\top$ is the initial estimator which could be set as the unpenalized WLAD, and $p_{\lambda_n}(t)$ is a nonconvex penalty function. We consider a class of penalties that satisfy the following conditions: (a) Let $p'_\lambda(\cdot)$ be nonnegative, nonincreasing and continuous over $(0, \infty)$; (b) There exists a constant $a > 0$ such that $\lim_{t \to 0^+} p'_\lambda(t) = \lambda$, $p'_\lambda(t) \ge \lambda - t/a$ $(0 < t < a\lambda)$ and $p'_\lambda(t) = 0$ $(t \ge a\lambda)$. This class includes the smoothly clipped absolute deviation (Fan and Li [3]) and the minimax concave penalty (Zhang [25]) as special cases. The minimizer $\hat{\boldsymbol{\beta}}$ of (2.3) is called the adaptive penalized WLAD (APWLAD) estimator.

3 Asymptotic Properties

Let H be the cumulative distribution function of Y and let τ_Y, τ_T, and τ_C be the end points of the support of Y, T, and C, respectively. Let F^0 be the joint distribution function of $(\boldsymbol{X}, T)$. Denote

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

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

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

For $j = 1, \ldots, d$, let

$$\gamma_0(y) = \exp\left\{ \int_0^{y-} \frac{\tilde{H}^0(dw)}{1 - H(w)} \right\},$$
$$\gamma_{1j}(y; \boldsymbol{\beta}) = \frac{1}{1 - H(y)} \iint I(w > y) \mathrm{sgn}(w - \boldsymbol{\beta}^\top \boldsymbol{x}) x_j \gamma_0(w) \tilde{H}^{11}(d\boldsymbol{x}, dw),$$
$$\gamma_{2j}(y; \boldsymbol{\beta}) = \iint \frac{I(v < y, v < w) \mathrm{sgn}(w - \boldsymbol{\beta}^\top \boldsymbol{x}) x_j \gamma_0(w)}{\{1 - H(v)\}^2} \tilde{H}^0(dw) \tilde{H}^{11}(d\boldsymbol{x}, dw),$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ and the sign function $\mathrm{sgn}(x)$ is defined as -1, 0, and 1 if $x < 0$, $x = 0$, and $x > 0$, respectively. Let

$$\varphi_j = x_j \mathrm{sgn}(Y - \boldsymbol{\beta}_0^\top \boldsymbol{X}) \gamma_0(Y) \delta + \gamma_{1j}(y; \boldsymbol{\beta}_0)(1 - \delta) - \gamma_{2j}(y; \boldsymbol{\beta}_0).$$

Denote $\sigma_{ij} = \mathrm{Cov}(\varphi_i, \varphi_j)$ for $i, j = 1, \ldots, d$. Furthermore, let

$$\boldsymbol{\Sigma} = (\sigma_{ij})$$

and

$$\mathbf{A} = 2E[\boldsymbol{X}\boldsymbol{X}^\top f_\epsilon(0|\boldsymbol{X})].$$

Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0d})^\top$ be the true parameter value. We consider the sparsity model which means that the d covariates contain both important and trivial components. For simplicity, let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{20}^\top)^\top$, where $\boldsymbol{\beta}_{20} = \mathbf{0}$. Here $\boldsymbol{\beta}_{10} = (\beta_{01}, \ldots, \beta_{0d_0})^\top$ is a d_0-vector and $\mathbf{0}$ is a $(d - d_0)$-vector. Corresponding to the partition of $\boldsymbol{\beta}_0$, rewrite

$$\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$$

and

$$\boldsymbol{X}_i = (\boldsymbol{X}_{1i}^\top, \boldsymbol{X}_{2i}^\top)^\top.$$

Likewise, rewrite

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12} & \mathbf{A}_{22} \end{bmatrix}.$$

We impose the following conditions to establish the theoretic properties of the APWLAD estimator.

(C1) Let $F_\varepsilon(\cdot|\boldsymbol{x})$ be the conditional distribution function of ε given $\boldsymbol{X}=\boldsymbol{x}$ and $f_\varepsilon(\cdot|\boldsymbol{x})$ be its conditional density function. Then $F_\varepsilon(0|\boldsymbol{x})=0.5$ and $f_\varepsilon(e|\boldsymbol{x})$ is continuous in e in a neighborhood of 0 for almost all $\boldsymbol{x}$.

(C2) The censoring mechanism is completely random censoring in the sense that C is independent of T and $\boldsymbol{X}$.

(C3) $\tau_T<\tau_C$ or $\tau_T=\tau_C=\infty$.

(C4) The matrix $\mathbf{A}$ is finite and nonsingular.

(C5) (i) The covariate $\boldsymbol{X}$ is bounded and the right end point of the support of $\boldsymbol{\beta}_0^\top\boldsymbol{X}$ is strictly less than τ_Y; (ii) $E[\|\boldsymbol{X}\|^2\gamma_0^2(Y)\delta]<\infty$ and $\int|x_j|D^{1/2}(w)\tilde{F}^0(d\boldsymbol{x},dw)<\infty$ for $j=1,\ldots,d$, where

$$D(y)=\int_0^{y-}[\{1-H(w)\}\{1-G(w)\}]^{-1}G(dw)$$

and G is the distribution function of the censoring time C.

These conditions are common in survival analysis. The consistency of the resultant estimator is summarized in the following theorem.

Theorem 3.1 (Consistency) *Under Conditions* (C1)–(C5), *if* $\lambda_n\to 0$ *as* $n\to\infty$, *then we have*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_P\left(\frac{1}{\sqrt{n}}\right).$$

Theorem 3.1 states that the APWLAD estimator is a $\sqrt{n}$-consistent estimator under the assumption $\lambda_n\to 0$ and other mild conditions. The proof of Theorem 3.1 is presented in the Appendix.

Lemma 3.2 (Sparsity) *Let* $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_1^\top,\hat{\boldsymbol{\beta}}_2^\top)^\top$. *Under Conditions* (C1)–(C5), *if* $\lambda_n\to 0$ *and* $\sqrt{n}\lambda_n\to\infty$ *as* $n\to\infty$, *then*

$$\Pr(\hat{\boldsymbol{\beta}}_2=\mathbf{0})\to 1.$$

This lemma shows that the APWLAD estimator enjoys the sparsity, namely, some components of the estimator are zero. Note that the condition $\sqrt{n}\lambda_n\to\infty$ is important to achieve variable selection. The proof is provided in the Appendix. Furthermore, the oracle property of the proposed method is summarized in the following theorem.

Theorem 3.3 (Oracle property) *Under Conditions* (C1)–(C5), *if* $\lambda_n\to 0$ *and* $\sqrt{n}\lambda_n\to\infty$ *as* $n\to\infty$, *then*

(1) *Sparsity*:

$$\Pr(\hat{\boldsymbol{\beta}}_2=\mathbf{0})\to 1.$$

(2) *Asymptotic normality*:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\xrightarrow{D}N(\mathbf{0},\mathbf{A}_{11}^{-1}\boldsymbol{\Sigma}_{11}\mathbf{A}_{11}^{-1}).$$

Theorem 3.3 demonstrates that, with proper choice of the tuning parameter, the APWLAD estimator has an oracle property. As a result, we can conclude that variable selection and parameter estimation can be completed simultaneously with a suitable penalty for the accelerated failure time model. The proof is given in the Appendix.

4 Computation

In this section, we introduce the selection of the tuning parameter and the computation algorithm of the APWLAD estimator.

Selection of the tuning parameter. The tuning parameter λ_n plays an extremely important role in variable selection, and determines the sparsity of the final model. So it is critical to choose a proper λ_n. Wang et al. [23] and Wang et al. [21] found that Bayesian information criterion (BIC) is consistent in model selection. Following this idea, we propose the following BIC-type selector

$$\mathrm{BIC}_{\lambda_n} = \log\left(\sum_{i=1}^{n} w_i |Y_{(i)} - \hat{\boldsymbol{\beta}}_{\lambda_n}^{\top} \boldsymbol{X}_{(i)}|\right) + d_{\lambda_n} \frac{\log(n)}{n},$$

where $\hat{\boldsymbol{\beta}}_{\lambda_n}$ is the estimator with a given λ_n and d_{λ_n} is number of nonzero components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$.

Algorithm. It is easy to get the APWLAD estimator since it can be computed by the existing software. Specifically, let $Y_{(i)}^* = w_i Y_{(i)}$ and $\boldsymbol{X}_{(i)}^* = w_i \boldsymbol{X}_{(i)}$. We consider the augmented data $\{(\tilde{Y}_{(i)}, \tilde{\boldsymbol{X}}_{(i)}) : i = 1, \ldots, n+d\}$, where $\tilde{Y}_{(i)} = Y_{(i)}^*$, $\tilde{\boldsymbol{X}}_{(i)} = \boldsymbol{X}_{(i)}^*$ for $i = 1, \ldots, n$ and $\tilde{Y}_i = 0$, $\tilde{\boldsymbol{X}}_i = p'_{\lambda_n}(|\boldsymbol{\beta}^0|)\boldsymbol{e}_i$ for $i = n+1, \ldots, n+d$. Here,

$$p'_{\lambda_n}(|\boldsymbol{\beta}^0|) = (p'_{\lambda_n}(|\beta_j^0|), j = 1, \ldots, d)^{\top},$$

and $\boldsymbol{e}_i$ is a d-vector with the ith component being 1 and the remaining 0. The objective function (2.3) can be transformed into the following form

$$Q_n(\boldsymbol{\beta}) = \sum_{i=1}^{n+d} |\tilde{Y}_{(i)} - \boldsymbol{\beta}^{\top} \tilde{\boldsymbol{X}}_{(i)}|.$$

Consequently, we can apply some common optimization algorithms, such the Nelder–Meader method or the *rq*( ) function in the *R* package quantreg for quantile regression (Koenker [13]), to obtain the desired estimator when d is not large. However, when d is larger than n, we need a new method to compute the penalized estimate. Since the usual unpenalized WLAD $\boldsymbol{\beta}^0$ is unavailable in high dimensionality, we treat the marginal WLAD estimates as the initial estimates. To compute the penalized estimate, we approximate the objective function (2.2) by the square function, and then adopt the coordinate descent algorithm to carry out the minimization of (2.3) (Breheny and Huang [1]).

5 Numerical Examples

In this section, we evaluate the finite sample performance of the APWLAD estimator via several simulation experiments and two real examples. Since the SCAD and MCP are special cases of the nonconvex penalty considered in this paper, and they perform similarly, we only give the simulated results of the SCAD penalty, for simplicity. As suggested by Fan and Li [3], we fix $a = 3.7$.

5.1 Simulation Studies

For comparison, we also consider the other three methods including the Lasso, adaptive Lasso (ALasso) and Oracle. The weight of adaptive Lasso is chosen based on the unpenalized WLAD estimator. The Oracle estimator, which estimates the nonzero coefficients by excluding the

covariates of zero coefficients in advance, can not be obtained in practice. We treat it as the benchmark here.

For each simulation setting, 500 simulated data sets are generated. To examine the performances of estimators, we compute the median of the *mean absolute prediction error* (MAPE) evaluated based on another 500 independent testing samples for each iteration (Wang et al. [22]). The variable selection performance is assessed by (NT, NF, Corrfit, Overfit), where "NT" denotes the average number of zero coefficients truly set to zero, and "NF" gives the average number of nonzero coefficients incorrectly set to zero. "Corrfit" is average number the correct model is selected, and "Overfit" is average number including all the significant variables and some noise variables.

Example 5.1 This example considers the case that the number of covariates is fixed. In this example, n observations are generated from

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

where $\boldsymbol{\beta} = (2.0, 1.5, 1.0, 0.5, 0, 0, 0, 0)$, i.e., $d = 8$ and the covariate $\boldsymbol{X} = (X_1, \ldots, X_8)^\top$ are generated from independent $N(0, 1)$. We consider three different error distributions: the standard normal distribution, the standard extreme value distribution and the standard Cauchy distribution. The censoring time $C = \min(\widetilde{C}, \tau)$, where $\widetilde{C}$ is generated from a uniform distribution $U(0, \tau + 1)$ and the study duration τ is chosen to yield a censoring rate (Cen.) of 20% or 50%. Set $n = 50$ and 100.

Table 1 summarizes the mean absolute prediction errors and variable selection results. Standard errors are given in parentheses. We conclude conclusions from Table 1 as follows.

(1) All the three methods can reduce the model complexity in all considered cases. The prediction accuracy is similar for these methods in terms of median MAPE. However, the selection results and the accuracy of variable selection differ significantly. For example, when the error follows normal distribution and the censoring rate is 50%, the proportion of correctly fitted of the SCAD is 63.6% in the case of $n = 100$, but it is only 18.4% for the Lasso in that case and 54.8% for the adaptive Lasso. Overall, the SCAD penalty performs better than the Lasso and adaptive Lasso.

(2) Although the Lasso can also achieve a sparse model, while the proportion of correctly fitted is relatively lower than the other two penalties. A good method should not only exclude the superfluous variables but also give a relatively high accuracy. Therefore, it is unreasonable if we only evaluate a variable selection method by the indexes "NT" and "NF".

(3) We can also see that both the number of true zero coefficients identified and the proportion of the correct model selected increase as n increases for every fixed censoring rate. In addition, the censoring rate also affects estimation accuracy. When the censoring rate increases, both "NT" and "Correctly fitted" decrease for every fixed sample size.

(4) For the different error distributions, the proposed method always performs well. Since the results in terms of variable selection and prediction accuracy are reasonable, although the results of the Cauchy distribution are not better than that of the other two distributions.

Error	Cen.	n	Method	Corrfit	Overfit	NT	NF	MAPE
Normal	0.2	50	Lasso	0.178 (0.383)	0.644 (0.479)	2.704 (1.058)	0.188 (0.421)	1.026
			ALasso	0.408 (0.492)	0.278 (0.448)	3.596 (0.646)	0.326 (0.494)	1.036
			SCAD	0.464 (0.499)	0.254 (0.436)	3.604 (0.666)	0.288 (0.466)	1.039
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	1.009
		100	Lasso	0.264 (0.441)	0.692 (0.462)	2.822 (0.990)	0.044 (0.205)	0.968
			ALasso	0.668 (0.471)	0.226 (0.419)	3.722 (0.538)	0.106 (0.308)	0.975
			SCAD	0.688 (0.464)	0.184 (0.388)	3.766 (0.521)	0.128 (0.334)	0.983
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	0.974
	0.5	50	Lasso	0.098 (0.298)	0.696 (0.460)	2.118 (1.226)	0.208 (0.411)	1.537
			ALasso	0.282 (0.450)	0.426 (0.495)	3.068 (1.015)	0.294 (0.460)	1.565
			SCAD	0.362 (0.481)	0.338 (0.474)	3.238 (1.029)	0.302 (0.464)	1.586
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	1.546
		100	Lasso	0.184 (0.388)	0.756 (0.430)	2.538 (1.091)	0.060 (0.238)	1.476
			ALasso	0.548 (0.498)	0.316 (0.465)	3.542 (0.688)	0.138 (0.351)	1.505
			SCAD	0.636 (0.482)	0.228 (0.420)	3.670 (0.656)	0.136 (0.343)	1.527
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	1.514
Extreme	0.2	50	Lasso	0.190 (0.393)	0.512 (0.500)	2.956 (0.984)	0.340 (0.563)	1.299
			ALasso	0.356 (0.479)	0.190 (0.393)	3.668 (0.582)	0.502 (0.592)	1.306
			SCAD	0.462 (0.499)	0.182 (0.386)	3.686 (0.559)	0.374 (0.520)	1.317
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	1.288
		100	Lasso	0.318 (0.466)	0.584 (0.493)	3.028 (0.972)	0.098 (0.298)	1.241
			ALasso	0.622 (0.485)	0.166 (0.372)	3.782 (0.497)	0.214 (0.415)	1.255
			SCAD	0.652 (0.477)	0.108 (0.311)	3.862 (0.409)	0.242 (0.433)	1.264
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	1.249
	0.5	50	Lasso	0.134 (0.341)	0.538 (0.499)	2.540 (1.134)	0.372 (0.571)	2.044
			ALasso	0.238 (0.426)	0.316 (0.465)	3.194 (0.943)	0.498 (0.599)	2.126
			SCAD	0.316 (0.465)	0.252 (0.435)	3.354 (0.948)	0.472 (0.574)	2.150
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	2.101
		100	Lasso	0.230 (0.421)	0.600 (0.490)	2.836 (1.045)	0.174 (0.390)	1.969
			ALasso	0.446 (0.498)	0.270 (0.444)	3.570 (0.703)	0.288 (0.462)	2.032
			SCAD	0.536 (0.499)	0.154 (0.361)	3.754 (0.575)	0.314 (0.473)	2.047
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	2.042
Cauchy	0.2	50	Lasso	0.106 (0.308)	0.122 (0.328)	3.532 (0.776)	1.446 (1.146)	3.328
			ALasso	0.132 (0.339)	0.036 (0.186)	3.802 (0.493)	1.342 (0.940)	3.247
			SCAD	0.302 (0.460)	0.114 (0.318)	3.562 (0.689)	0.674 (0.639)	3.107
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	3.137
		100	Lasso	0.286 (0.452)	0.170 (0.376)	3.648 (0.627)	0.746 (0.853)	3.099
			ALasso	0.304 (0.460)	0.028 (0.165)	3.948 (0.256)	0.822 (0.695)	3.059
			SCAD	0.466 (0.499)	0.034 (0.181)	3.930 (0.278)	0.514 (0.528)	3.009
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	3.023

Error	Cen.	n	Method	Corrfit	Overfit	NT	NF	MAPE
Cauchy	0.5	50	Lasso	0.076 (0.265)	0.104 (0.306)	3.518 (0.874)	2.090 (1.453)	3.221
			ALasso	0.072 (0.259)	0.048 (0.214)	3.694 (0.711)	1.798 (1.113)	3.384
			SCAD	0.154 (0.361)	0.082 (0.275)	3.498 (0.812)	1.058 (0.785)	3.532
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	3.536
		100	Lasso	0.184 (0.388)	0.168 (0.374)	3.620 (0.670)	1.540 (1.561)	2.997
			ALasso	0.210 (0.408)	0.018 (0.133)	3.942 (0.251)	1.264 (1.034)	3.067
			SCAD	0.276 (0.447)	0.014 (0.118)	3.970 (0.171)	0.804 (0.595)	3.081
			Oracle	1.000 (0.000)	0.000 (0.000)	4.000 (0.000)	0.000 (0.000)	3.082

Table 1 Simulation results in Example 5.1

Error	Method	Corrfit	Overfit	NT	NF	MAPE
Normal	Lasso	0.136 (0.343)	0.772 (0.420)	9.006 (1.617)	0.092 (0.289)	0.979
	ALasso	0.530 (0.500)	0.346 (0.476)	10.490 (0.777)	0.124 (0.330)	0.985
	SCAD	0.552 (0.498)	0.310 (0.463)	10.546 (0.754)	0.138 (0.345)	0.997
	Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	0.978
	LS-Lasso	0.108 (0.311)	0.880 (0.325)	8.450 (1.719)	0.012 (0.109)	0.953
	LS-ALasso	0.792 (0.406)	0.078 (0.268)	10.910 (0.293)	0.130 (0.337)	0.960
	LS-SCAD	0.474 (0.500)	0.510 (0.500)	10.120 (1.070)	0.016 (0.126)	0.987
	LS-Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	0.965
Extreme	Lasso	0.160 (0.367)	0.644 (0.479)	9.326 (1.423)	0.200 (0.410)	1.273
	ALasso	0.490 (0.500)	0.258 (0.438)	10.612 (0.686)	0.252 (0.435)	1.278
	SCAD	0.564 (0.496)	0.180 (0.385)	10.730 (0.608)	0.256 (0.437)	1.290
	Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	1.274
	LS-Lasso	0.038 (0.191)	0.826 (0.379)	7.256 (1.929)	0.136 (0.343)	1.881
	LS-ALasso	0.272 (0.445)	0.360 (0.480)	10.194 (0.956)	0.376 (0.501)	1.903
	LS-SCAD	0.048 (0.214)	0.778 (0.416)	7.976 (1.662)	0.176 (0.386)	2.052
	LS-Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	2.064
Cauchy	Lasso	0.124 (0.330)	0.206 (0.405)	10.472 (0.786)	1.050 (0.997)	3.068
	ALasso	0.228 (0.420)	0.058 (0.234)	10.858 (0.467)	0.898 (0.702)	3.114
	SCAD	0.414 (0.493)	0.108 (0.311)	10.750 (0.540)	0.494 (0.532)	3.069
	Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	3.108
	LS-Lasso	0.016 (0.126)	0.212 (0.409)	7.384 (3.275)	1.408 (1.056)	3.923
	LS-ALasso	0.030 (0.171)	0.168 (0.374)	7.576 (3.345)	1.356 (0.963)	4.450
	LS-SCAD	0.002 (0.045)	0.674 (0.469)	3.468 (2.645)	0.366 (0.566)	4.972
	LS-Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	4.008

Table 2 Simulation results in Example 5.2

Example 5.2 This example considers the case that the number of covariates is high but

smaller than the sample size. Set $d = 15$ for $n = 100$. The censoring rate is 20%. The remaining settings are the same as that in Example 5.1. This example shows the comparisons between the LAD method and the least squares method. The least squares estimators with the Lasso, adaptive Lasso (ALasso), SCAD and Oracle are denoted by LS-Lasso, LS-ALasso, LS-SCAD, and LS-Oracle, respectively.

The simulation results are presented in Table 2, from which we can draw the similar conclusions as that in Example 5.1. As a result, when the parameter number increases, our method still has good performance. Furthermore, the least squares method performs better than the LAD method for the normal error, while it performs much worse than the LAD method for the standard extreme value distribution and the standard Cauchy distribution. This confirms the robust performance of the LAD method for censored data.

Example 5.3 This example considers the heteroscedastic errors. Predictors $X_j, j = 1, \ldots, d$ are generated in the following two steps. We first generate $X_j^*, j = 1, \ldots, d$ from independent $N(0, 1)$, and the next step is to set $X_j = \Phi(X_j^*)$ for $j = 1, 2, \ldots, d$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The response is generated according to the heteroscedastic location-scale model

$$T = \boldsymbol{\beta}^\top \boldsymbol{X} + 0.5X_5\epsilon.$$

The remaining settings are the same as that in Example 5.2.

The simulation results are presented in Table 3. From Table 3, it can be seen that the penalized penalized weighted LAD estimator still performs well in the case of the heteoscedastic error.

Error	Method	Corrfit	Overfit	NT	NF	MAPE
Normal	Lasso	0.092 (0.289)	0.908 (0.289)	8.898 (1.293)	0.000 (0.000)	0.440
	ALasso	0.912 (0.284)	0.078 (0.268)	10.918 (0.289)	0.010 (0.100)	0.437
	SCAD	0.920 (0.272)	0.068 (0.252)	10.928 (0.281)	0.012 (0.109)	0.442
	Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	0.442
Extreme	Lasso	0.164 (0.371)	0.834 (0.372)	9.116 (1.379)	0.002 (0.045)	0.479
	ALasso	0.726 (0.446)	0.216 (0.412)	10.748 (0.515)	0.058 (0.234)	0.479
	SCAD	0.842 (0.365)	0.114 (0.318)	10.852 (0.467)	0.044 (0.205)	0.478
	Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	0.482
Cauchy	Lasso	0.134 (0.341)	0.778 (0.416)	9.234 (1.150)	0.088 (0.284)	0.984
	ALasso	0.534 (0.499)	0.012 (0.109)	10.970 (0.182)	0.498 (0.582)	1.022
	SCAD	0.502 (0.500)	0.008 (0.089)	10.974 (0.171)	0.498 (0.516)	1.020
	Oracle	1.000 (0.000)	0.000 (0.000)	11.000 (0.000)	0.000 (0.000)	1.010

Table 3 Simulation results in Example 5.3

Example 5.4 In this example, we demonstrate the performance of the proposed estimator when the number of predictors is large compared with the number of observations. The true regression coefficients are $\beta_j = 2$ for $1 \le j \le 5$, and 0 otherwise. The sample size is chosen to

be 100, and $d = 100, 200$. The censoring rate is 20%. We keep the other setups as the same as Example 5.1.

From the simulation results summarized in Table 4, it can be seen that increasing the dimension of the predictors has a significant effect on the ratio of correct models identified and the number of predictors included in the final selected model. In general, the SCAD penalty performs best, and the Lasso performs worst.

Error	p	Method	Corrfit	Overfit	NT	NF	MAPE
Normal	100	Lasso	0.002 (0.045)	0.996 (0.063)	82.350 (6.379)	0.002 (0.045)	1.191
		ALasso	0.098 (0.298)	0.900 (0.300)	89.260 (4.149)	0.002 (0.045)	1.198
		SCAD	0.404 (0.491)	0.596 (0.491)	88.302 (7.574)	0.000 (0.000)	1.183
		Oracle	1.000 (0.000)	0.000 (0.000)	95.000 (0.000)	0.000 (0.000)	1.156
	200	Lasso	0.000 (0.000)	0.998 (0.045)	179.240 (6.598)	0.002 (0.045)	1.280
		ALasso	0.064 (0.245)	0.934 (0.249)	188.202 (4.242)	0.002 (0.045)	1.258
		SCAD	0.336 (0.473)	0.664 (0.473)	186.708 (8.549)	0.000 (0.000)	1.260
		Oracle	1.000 (0.000)	0.000 (0.000)	195.000 (0.000)	0.000 (0.000)	1.219
Extreme	100	Lasso	0.002 (0.045)	0.996 (0.063)	83.468 (5.907)	0.002 (0.045)	1.546
		ALasso	0.192 (0.394)	0.806 (0.396)	90.890 (3.756)	0.002 (0.045)	1.530
		SCAD	0.496 (0.500)	0.504 (0.500)	91.298 (5.729)	0.000 (0.000)	1.500
		Oracle	1.000 (0.000)	0.000 (0.000)	95.000 (0.000)	0.000 (0.000)	1.490
	200	Lasso	0.00 (0.000)	0.986 (0.118)	181.254 (6.272)	0.014 (0.118)	1.602
		ALasso	0.16 (0.367)	0.826 (0.379)	190.114 (3.959)	0.014 (0.118)	1.532
		SCAD	0.39 (0.488)	0.610 (0.488)	190.162 (6.547)	0.000 (0.000)	1.510
		Oracle	1.00 (0.000)	0.000 (0.000)	195.000 (0.000)	0.000 (0.000)	1.487
Cauchy	100	Lasso	0.036 (0.186)	0.808 (0.394)	87.884 (5.750)	0.436 (1.137)	3.673
		ALasso	0.326 (0.469)	0.516 (0.500)	92.656 (3.501)	0.444 (1.148)	3.554
		SCAD	0.242 (0.429)	0.646 (0.479)	91.890 (4.387)	0.376 (1.114)	3.643
		Oracle	1.000 (0.000)	0.000 (0.000)	95.000 (0.000)	0.000 (0.000)	3.386
	200	Lasso	0.010 (0.100)	0.724 (0.447)	186.126 (7.348)	0.708 (1.371)	5.788
		ALasso	0.216 (0.412)	0.512 (0.500)	191.592 (4.751)	0.724 (1.378)	5.965
		SCAD	0.134 (0.341)	0.656 (0.476)	189.806 (6.416)	0.628 (1.355)	4.930
		Oracle	1.000 (0.000)	0.000 (0.000)	195.000 (0.000)	0.000 (0.000)	5.230

Table 4 Simulation results in Example 5.4

5.2 Real Data Analysis

5.2.1 ACTG Data

We analysis the AIDS Clinical Trials Group Study 320 Data (ACTG Data) to verify the performance of the proposed method. The data come from a double-blind, placebo-controlled trial comparing nucleoside monotherapy with combination therapy in HIV-infected patients with CD4 cell counts from 200–500 per cubic millimeter provided by the AIDS Clinical Trials Group

Study 320 Study Team (Hammer et al. [6]). Randomization was stratified by CD4 cell counts at the time of screening. The primary outcome measure was time to AIDS defining event or death. There are 1151 observations available in this trial. The censoring rate is about 0.0834. We study the dependence of the patients' survival times on eleven covariates: continuous variables are cd4 (Baseline CD4 count cells/milliliter derived from multiple measurements), priorzdv (months of prior ZDV), age (age at baseline in years); categorical variables are tx (treatment indicator, 1 = including IDV, 0 = without IDV), txgrp (treatment group indicator, 1 = ZDV + 3TC, 2 = ZDV + 3TC + IDV, 3 = d4T + 3TC, 4 = d4T + 3TC + IDV), strat2 (CD4 stratum at screening, 0 if CD4 $\leq$ 50, 1 otherwise), sex (sex, 1 = male, 2 = female), raceth (race/ethnicity, 1 = White Non-Hispanic, 2 = Black Non-Hispanic, 3 = Hispanic (regardless of race), 4 = Asian, Pacific Islander, 5 = American Indian, Alaskan Native, 6 = other/unknown), ivdrug (IV drug use history, 1 = never, 2 = currently, 3 = previously), hemophil (Hemophiliac, 1 = Yes, 0 = No), karnof (Karnofsky performance scale, 100 = normal; no complaint; no evidence of disease, 90 = normal activity possible; minor signs/symptoms of disease, 80 = Normal activity with effort; some signs/symptoms of disease, 70 = cares for self; normal activity/active work not possible).

To estimate the standard errors, we apply the nonparametric 0.632 bootstrap method in which we sampled $0.632n$ from n observations without replacement. The bootstrap method is then repeated 100 times. Based on the bootstrapped samples, the corresponding bootstrapped estimator can be obtained following the same procedure as for the original sample and using the same tuning parameter λ. After proper scale adjustment, the sample standard deviation of the bootstrapped estimates provides an estimate for the standard error of $\hat{\boldsymbol{\beta}}$.

Variables selected via the WLAD-SCAD approach and their corresponding estimates are shown in Table 4. The estimated standard errors are given in parentheses. For comparison, we also provide the Lasso and adaptive Lasso estimates. From the results in Table 5 where " $-$ $(-)$ " indicates the method is not applicable, we can see that, although it removes the variable hemophil, the SCAD method is in line with the adaptive Lasso method. However, the Lasso only excludes three variables: strat2, sex and ivdrug.

Covariate	Lasso		ALasso		SCAD	
tx	-4.933	(0.012)	-5.234	(0.011)	-5.234	(0.013)
txgrp	4.704	(0.010)	4.917	(0.009)	4.926	(0.012)
strat2	–	(–)	–	(–)	–	(–)
sex	–	(–)	–	(–)	–	(–)
raceth	0.095	(0.004)	–	(–)	–	(–)
ivdrug	–	(–)	–	(–)	–	(–)
hemophil	0.213	(0.018)	–	(–)	0.382	(0.023)
karnof	0.031	(0.003)	–	(–)	–	(–)
cd4	0.191	(0.004)	0.219	(0.004)	0.254	(0.005)
priorzdv	0.003	(0.004)	–	(–)	–	(–)
age	0.122	(0.002)	0.090	(0.002)	0.077	(0.003)

Table 5 Estimation for the ACTG Data

5.2.2 Mantle Cell Lymphoma Data

Rosenwald et al. [16] studied the mantle cell lymphoma (MCL) data. The primary goal of this study is to identify genes that have good predictive power of patients' survival risk. Among 101 untreated patients with no history of previous lymphoma, 92 were classified as having MCL based on established morphologic and immunophenotypic criteria. During the followup, 64 patients died of MCL, and the other 28 patients were censored. The median survival time was 2.8 years (range 0.02 to 14.05 years). Lymphochip DNA microarrays were used to quantify mRNA expression in the lymphoma samples from the 92 patients. This data set is available at `http://llmpp.nih.gov/MCL/`, and contains expression values of 8810 cDNA elements. We exclude the genes containing missing values, then 6312 genes are used in analysis. We standardize these gene expressions to have zero mean and unit variance.

We fit this data by the AFT model, and use the proposed WLAD-SCAD approach for gene selection. We also give the selection results of the Lasso and adaptive Lasso as comparison. The tuning parameter is selected using the method introduced in Section 4. The estimated standard errors (in parentheses) are obtained using the method introduced in Section 5.2.1. From Table 6 where " $-$ $(-)$ " has the same meaning as above, we can see that the Lasso selects 16 genes, the adaptive Lasso identifies 9 genes, and the SCAD chooses only 5 genes which are all included by the Lasso and adaptive Lasso.

UNIQID	Lasso		ALasso		SCAD	
16312	−0.253	(0.049)	–	(–)	–	(–)
16443	−0.414	(0.031)	−0.862	(0.041)	–	(–)
16711	−0.045	(0.025)	–	(–)	–	(–)
16724	0.310	(0.031)	0.946	(0.058)	–	(–)
17911	−0.155	(0.028)	–	(–)	–	(–)
17924	−0.650	(0.051)	−1.204	(0.061)	−0.857	(0.066)
25226	−0.137	(0.033)	–	(–)	–	(–)
27315	0.195	(0.035)	0.405	(0.051)	–	(–)
27824	1.105	(0.062)	0.694	(0.054)	1.226	(0.068)
28148	−0.319	(0.034)	−0.601	(0.045)	–	(–)
29780	0.440	(0.034)	0.949	(0.063)	0.255	(0.061)
30917	−0.772	(0.049)	−1.474	(0.088)	−1.626	(0.097)
31318	−0.043	(0.025)	–	(–)	–	(–)
33781	0.175	(0.030)	–	(–)	–	(–)
33877	−0.036	(0.031)	–	(–)	–	(–)
34844	−0.837	(0.042)	−0.599	(0.045)	−1.450	(0.060)

Table 6 Estimation for the MCL data

6 Remark

We propose the APWLAD estimator for parameter estimation and variable selection in the AFT model. Under some mild conditions, the consistency and oracle properties of the AP-

WLAD estimator are established. Extensive numeric studies demonstrate that the APWLAD estimate performs well with fixed, diverging, and large number of covariates, respectively. As a conclusion, the APWLAD estimator is a plausible method in practice to conduct variable selection for the AFT model.

Acknowledgements We thank two referees for their constructive comments that have led to a substantial improvement of the paper.

References

[1] Breheny, P., Huang, J.: Coordinate descent algorithms for nonconvex penalized regression, with application to biological feature selection. *The Annals of Applied Statistics*, **5**, 232–253 (2011)

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

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

[4] Frank, I. E., Friedman, J. H.: A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148 (1993)

[5] Friedman, J. H., Popescu, B.: Gradient directed regularization. Technical Report, California: Stanford University, 2004

[6] Hammer, S. M., Katzenstein, D. A., Hughes, M. D., et al.: A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine*, **335**, 1081–1090 (1997)

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

[8] Huang, J., Ma, S., Xie, H.: Regularized estimation in the accelerated failure time model with high dimensional covariates. *Biometrics*, **62**, 813–820 (2006)

[9] Huang, J., Ma, S., Xie, H.: Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica*, **17**, 1533–1548 (2007)

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

[11] Kalbfleisch, J. D., Prentice, R. L.: The Statistical Analysis of Failure Time Data, Wiley, New York, 2002

[12] Knight, K.: Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, **26**, 755–770 (1998)

[13] Koenker, R.: Quantile Regression, Cambridge University Press, Cambridge, 2005

[14] Koenker, R., Zhao, Q.: Conditional quantile estimation and inference for ARCH models. *Econometric Theory*, **12**, 793–813 (1996)

[15] Le Cam, L.: Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *University of California Publications in Statistics*, **3**, 37–98 (1960)

[16] Rosenwald, A., Wright, G., Wiestner, A., et al.: The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, **3**, 185–197 (2003)

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

[18] Stute, W., Wang, J. L.: The strong law under random censorship. *The Annals of Statistics*, **14**, 1351–1365 (1993)

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

[20] Van der Vaart, A. W.: Asymptotic Statistics, Cambridge University Press, Cambridge, 1998

[21] Wang, H., Li, B., Leng, C.: Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B*, **71**, 671–683 (2009)

[22] Wang, H., Li, G., Jiang, G.: Robust regression shrinkage and consistent variable selection through the LAD-lasso. *Journal of Business & Economic Statistics*, **25**(3), 347–355 (2007)
[23] Wang, H., Li, R., Tsai, C. L.: Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568 (2007)
[24] Wei, L. J.: The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, **11**, 1871–1879 (1992)
[25] Zhang, C. H.: Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942 (2010)
[26] Zhang, H., Lu, W.: Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691–703 (2007)
[27] Zhou, Z., Jiang, R., Qian, W.: LAD variable selection for linear models with randomly censored data. *Metrika*, **76**, 287–300 (2013)
[28] Zou, H.: The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429 (2006)

Appendix

Proof of Theorem 3.1 Assume that the number of nonzero components in $\boldsymbol{\beta}_0$ is d_0. Without loss of generality, we rewrite $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{20}^\top)^\top$ and $\boldsymbol{\beta}_{20}^\top$ is a $(d-d_0)$-dimensional $\mathbf{0}$-vector. We intend to show that for any given $\epsilon > 0$, there exists a large constant C^* such that

$$\Pr\Big\{ \inf_{\|\mathbf{u}\|=C^*} Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > Q_n(\boldsymbol{\beta}_0) \Big\} \geq 1-\epsilon, \tag{A.1}$$

where $\mathbf{u} = (u_1, \ldots, u_d)^\top$ is a d-dimensional vector. Then we can claim that with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}: \|\mathbf{u}\| \leq C^*\}$, that is, there exists a local minimizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$. Because the objective function is a strictly convex function, the local minimizer is a global minimizer.

To show (A.1), let $D_n(\mathbf{u}) = Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\beta}_0)$. Simple calculations yield that

$$\begin{aligned}
nD_n(\mathbf{u}) &= n\sum_{i=1}^n w_i\{|Y_{(i)} - (\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u})^\top \boldsymbol{X}_{(i)}| - |Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)}|\} \\
&\quad + n\sum_{j=1}^d p'_{\lambda_n}(|\beta_j^0|)\{|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|\} \\
&\geq n\sum_{i=1}^n w_i\{|Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)} - n^{-1/2}\mathbf{u}^\top \boldsymbol{X}_{(i)}| - |Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)}|\} \\
&\quad + n\sum_{j=1}^{d_0} p'_{\lambda_n}(|\beta_j^0|)\{|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|\} \\
&= I_{1n} + I_{2n} + I_{3n},
\end{aligned}$$

where

$$\begin{aligned}
I_{1n} &= n\sum_{i=1}^n w_i\{|Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)} - n^{-1/2}\mathbf{u}^\top \boldsymbol{X}_{(i)}| - |Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)}|\} \\
&\quad + n^{1/2}\sum_{i=1}^n w_i \mathrm{sgn}(Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)})\boldsymbol{X}_{(i)}^\top \mathbf{u}, \\
I_{2n} &= -n^{1/2}\sum_{i=1}^n w_i \mathrm{sgn}(Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)})\boldsymbol{X}_{(i)}^\top \mathbf{u},
\end{aligned}$$

$$I_{3n} = n\sum_{j=1}^{d_0} p'_{\lambda_n}(|\beta_j^0|)\{|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|\}.$$

It follows from (6.5) in Huang et al. [9] that, under Conditions (C1)–(C5), for any fixed $\mathbf{u} \in \mathbb{R}^d$,

$$I_{1n} \xrightarrow{P} \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u},$$

as $n \to \infty$. Using Theorem 3.1 in Stute [17], we have that I_{2n} converges in distribution to $-\mathbf{u}^\top \mathbf{W}$, where $\mathbf{W}$ is a d-dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. In the other words, $I_{2n} = \mathbf{u}^\top O_P(1)$. For term I_{3n},

$$|I_{3n}| \le n^{1/2}\sum_{j=1}^{d_0} p'_{\lambda_n}(|\beta_j^0|)|u_j| \le (d_0 n)^{1/2}\max\{p'_{\lambda_n}(|\beta_j^0|): 1 \le j \le d_0\}\|\mathbf{u}\|.$$

Note that $p'_{\lambda_n}(|\beta_j^0|) = p'_{\lambda_n}(|\beta_j^0|)I(|\beta_j^0| \le a\lambda_n)$. By Condition $\lambda_n \to 0$, we have, for $j = 1, \ldots, d_0$, $|\beta_j^0| - a\lambda_n \xrightarrow{P} |\beta_j^0| > 0$. Thus for every $\eta > 0$,

$$\Pr\{\sqrt{n}p'_{\lambda_n}(|\beta_j^0|) > \eta\} \le \Pr(|\beta_j^0| \le a\lambda_n) \to 0,$$

which implies that $\sqrt{n}p'_{\lambda_n}(|\beta_j^0|) = o_P(1)$. Hence, by choosing a sufficiently large C^*, I_{1n} dominates I_{2n} and I_{3n}. Thus we complete the proof of Theorem 3.1 under Condition (C4). □

Proof of Lemma 3.2 Taking the first derivative of $Q_n(\boldsymbol{\beta})$ at any differentiable point with respect to β_j for $j = d_0 + 1, \ldots, d$, we can obtain that

$$n^{1/2}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} = -n^{1/2}\sum_{i=1}^{n} w_i \mathrm{sgn}(Y_{(i)} - \boldsymbol{\beta}^\top \boldsymbol{X}_{(i)})X_{(i)j} + n^{1/2}p'_{\lambda_n}(|\beta_j^0|)\mathrm{sgn}(\beta_j).$$

For any $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^\top, \tilde{\boldsymbol{\beta}}_2^\top)^\top$ satisfying that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) = O_P(1)$ and $\|\tilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20}\| \le \epsilon_n = Mn^{-1/2}$, where M is a constant, we firstly aim to show that

$$-n^{1/2}\sum_{i=1}^{n} w_i \mathrm{sgn}(Y_{(i)} - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_{(i)})\boldsymbol{X}_{(i)} = O_P(1). \tag{A.2}$$

Denote

$$\mathbf{V}(\mathbf{\Delta}) = n^{1/2}\sum_{i=1}^{n} w_i \mathrm{sgn}(Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)} - n^{-1/2}\boldsymbol{X}_{(i)}^\top \mathbf{\Delta})\boldsymbol{X}_{(i)}$$

for any $\mathbf{\Delta} \in \mathbb{R}^d$.

Note that we have obtained that $\mathbf{V}(\mathbf{0})$ converges weakly to the mean zero Gaussian random vector $\boldsymbol{W}$. On the other hand, under Conditions (C1) and (C4), it follows from Lemma A.2 in Koenker and Zhao [14] that

$$\sup_{\|\mathbf{\Delta}\| \le M} \|\mathbf{V}(\mathbf{\Delta}) - \mathbf{V}(\mathbf{0}) + f_\epsilon(0|\boldsymbol{X})\mathrm{cov}(\boldsymbol{W})\mathbf{\Delta}\| = o_P(1).$$

Taking $\mathbf{\Delta} = n^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ immediately yields that

$$\begin{aligned} &n^{1/2}\sum_{i=1}^{n} w_i \mathrm{sgn}(Y_{(i)} - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_{(i)})\boldsymbol{X}_{(i)} - n^{1/2}\sum_{i=1}^{n} w_i \mathrm{sgn}(Y_{(i)} - \boldsymbol{\beta}_0^\top \boldsymbol{X}_{(i)})\boldsymbol{X}_{(i)} \\ &\quad + f_\epsilon(0|\mathbf{x})\mathrm{cov}(\boldsymbol{W})n^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \end{aligned}$$

$$= o_P(1),$$

which implies (A.2).

Note that $p'_{\lambda_n}(t)/\lambda_n \to 1$ as $t \to 0$. For $j = d_0 + 1, \ldots, d$, we have $\sqrt{n}|\beta_j^0| = O_P(1)$. So $p'_{\lambda_n}(|\beta_j^0|)/\lambda_n \xrightarrow{P} 1$. Furthermore,

$$\sqrt{n}p'_{\lambda_n}(|\beta_j^0|) = \sqrt{n}\lambda_n \frac{p'_{\lambda_n}(|\beta_j^0|)}{\lambda_n} \to \infty.$$

Therefore, for $j = d_0 + 1, \ldots, d$,

$$n^{1/2}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = O_P(1) + n^{1/2}p'_{\lambda_n}(|\beta_j^0|)\mathrm{sgn}(\beta_j)\begin{cases} > 0, & \text{for } 0 < \tilde{\beta}_j < \epsilon_n, \\ < 0, & \text{for } \epsilon_n < \tilde{\beta}_j < 0. \end{cases}$$

This has the proof done. □

Proof of Theorem 3.3 For any $\mathbf{v} \in \mathbb{R}^{d_0}$, define $s_n(\mathbf{v}) = Q_n(\boldsymbol{\beta}_{10} + n^{-1/2}\mathbf{v}, \mathbf{0}) - Q_n(\boldsymbol{\beta}_{10}, \mathbf{0})$. Then

$$\begin{aligned} ns_n(\mathbf{v}) &= n\sum_{i=1}^{n} w_i\{|Y_{(i)} - \boldsymbol{\beta}_{10}^\top \boldsymbol{X}_{1(i)} - n^{-1/2}\mathbf{v}^\top \boldsymbol{X}_{1(i)}| - |Y_{(i)} - \boldsymbol{\beta}_{10}^\top \boldsymbol{X}_{1(i)}|\} \\ &\quad + n\sum_{j=1}^{d_0} p'_{\lambda_n}(|\beta_j^0|)\{|\beta_{0j} + n^{-1/2}v_j| - |\beta_{0j}|\} \\ &= I_{4n} + I_{5n} + I_{6n}, \end{aligned}$$

where

$$\begin{aligned} I_{4n} &= n\sum_{i=1}^{n} w_i\{|Y_{(i)} - \boldsymbol{\beta}_{10}^\top \boldsymbol{X}_{1(i)} - n^{-1/2}\mathbf{v}^\top \boldsymbol{X}_{1(i)}| - |Y_{(i)} - \boldsymbol{\beta}_{10}^\top \boldsymbol{X}_{1(i)}|\} \\ &\quad + n^{1/2}\sum_{i=1}^{n} w_i \mathrm{sgn}(Y_{(i)} - \boldsymbol{\beta}_{10}^\top \boldsymbol{X}_{1(i)})\mathbf{v}^\top \boldsymbol{X}_{1(i)}, \\ I_{5n} &= -n^{1/2}\sum_{i=1}^{n} w_i \mathrm{sgn}(Y_{(i)} - \boldsymbol{\beta}_{10}^\top \boldsymbol{X}_{1(i)})\mathbf{v}^\top \boldsymbol{X}_{1(i)}, \\ I_{6n} &= n\sum_{j=1}^{d_0} p'_{\lambda_n}(|\beta_j^0|)\{|\beta_{0j} + n^{-1/2}v_j| - |\beta_{0j}|\}. \end{aligned}$$

Using the similar arguments used in proof of consistency, under Conditions (C1)–(C5), we have that $I_{4n} \xrightarrow{P} \frac{1}{2}\mathbf{v}^\top \mathbf{A}_{11}\mathbf{v}$, where $\mathbf{A}_{11} = 2E[\boldsymbol{X}_1\boldsymbol{X}_1^\top f_\epsilon(0|\boldsymbol{X})]$, and that I_{5n} converges in distribution to $-\mathbf{v}^\top \mathbf{W}_1$, where $\mathbf{W}_1$ is a d_0-dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{11}$. Additionally, $I_{6n} \le \sqrt{n}\max\{p'_{\lambda_n}(|\beta_j^0|): 1 \le j \le d_0\}\sum_{j=1}^{d_0}|v_j|$, which converges to zero provided that $\lambda_n \to 0$ as $n \to \infty$. Hence, we can conclude that $ns_n(\mathbf{v})$ converges in distribution to $-\mathbf{v}^\top \mathbf{W}_1 + \frac{1}{2}\mathbf{v}^\top \mathbf{A}_{11}\mathbf{v}$ using the Slutsky theorem. Therefore, by Corollary 2 of Knight [12], it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{D} \mathbf{A}_{11}^{-1}\mathbf{W_1}$. Thus, we complete the proof. □