Statistica Sinica Preprint No: SS-2018-0089	
Title	High-Dimensional Variable Selection with Right Censored Length-biased Data
Manuscript ID	SS-2018-0089
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0089
Complete List of Authors	Di He Yong Zhou and Hui Zou
Corresponding Author	Hui Zou
E-mail	zouxx019@umn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

High-Dimensional Variable Selection with Right Censored Length-biased Data

Di He[1], Yong Zhou[1,2] and Hui Zou[3]

[1] *Shanghai University of Finance and Economics*

[2] *Chinese Academy of Sciences*

[3] *University of Minnesota*

Abstract: Length-biased data are inevitably encountered in various fields ranging from epidemiological cohort studies to studies of labor economics, attracting much attention in the survival literature. A crucial goal of survival analysis is to identify a subset of risk factors and their risk contributions among massive clinical covariates. However, there has been no work on variable selection for length-biased data due to the complex nature of such data and the lack of a convenient loss function. In this article, we propose an estimation method based on the penalized estimating equations to obtain both sparse and consistent estimator for length-biased data under the accelerated failure time (AFT) model. The proposed estimator possesses selection and estimation consistency property. In particular, we implement the method with SCAD penalty and local linear approximation algorithm. We suggest selecting the tuning parameter by extended BIC criterion in high-dimensional setting. Furthermore, a novel multi-stage SCAD

penalized estimating equations procedure is developed to achieve enhanced estimation accuracy and sparsity in variable selection. Simulation studies show that the proposed procedure has high accuracy and almost perfect sparsity. The Oscar Awards data is analyzed as an application of the proposed method.

Key words and phrases: Accelerated failure time model; High-dimensional variable selection; Length-biased data; Multi-stage penalization.

1. Introduction

Length-biased sampling, a special case of left truncation, is a frequently used, convenient and economical sampling technique for the collection of data in various fields, such as epidemiological cohort studies and studies of labor economics. Length-biased data assumes that the incidence of event onset follows a Poisson process (Zelen and Feinleib, 1969; Simon, 1980) known as the stationarity assumption which is often suitable in practice, or equivalently the truncation time follows a uniform distribution, and hence occurs when the probability that an item is sampled is proportional to its length. As a result, the observed time intervals from initiation to failure tend to be longer than those in the target population in a prevalent cohort study. An example of such data can be found in the study of dementia among elderly people by the Canadian study of Health and Aging (CSHA) (Asgharian et al., 2002; Addona and Wolfson, 2006; Shen et al., 2009; Qin

and Shen, 2010). There were more than 10,000 Canadians over the age of 65 recruited and screened for prevalence of dementia. The approximate initial date of dementia and the subsequent time of death and censoring were recorded, for the individuals who were found to have dementia in the study population. Those individuals who had dementia and did not survive to the examination time were excluded from the investigation, and only those individuals who had dementia and were still alive during the CSHA could be observed, which could lead to length-biased sampling.

Extensive methodology development has focused on estimating the unbiased target distribution in the presence of length-bias. One approach is based on the conditional distribution of the observations given the sampling process (Lagakos et al., 1988; Wang, 1991). Another approach is based on the unconditional distribution (Vardi, 1982, 1985; Gill et al., 1988; Asgharian et al., 2002, 2005), which requires the stationarity assumption. Recently, the analysis of right-censored and length biased data has attracted attentions from many researchers. A great challenge in analyzing such data is informative censoring, that is, the dependence between the right censoring time and the failure time. Another significant difficulty is that the observed length-biased data change the model structure assumed for the target population. Shen et al. (2009) developed estimating equation meth-

ods for semiparametric transformation and accelerated failure time (AFT) models to obtain the consistent estimators of the regression coefficients. Qin and Shen (2010) proposed two estimating equation approaches for the Cox model to analyze covariate effects. Ning et al. (2011) presented a generalized Buckley-James-type estimator under AFT model.

A crucial goal of survival analysis is to identify the risk factors and their risk contributions. By the advent of modern data collection technologies, a huge amount of clinical covariates, such as patients' personal characteristics, biomarkers and genotypes, are increasingly accessible from various sources by researchers. A necessary but challenging task is to select a subset of important variables upon which the hazard function or the survival time depends, since it helps medical researchers build comprehensible models to predict outcomes without information loss and leads to better disease diagnosis and treatment in the long run. This process, called variable selection or feature selection, has been widely studied for linear models with uncensored outcomes, including subset selection, least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), bridge regression (Fu, 1998), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006) and minimax concave penalty (MCP) (Zhang et al., 2010a). In the context of survival

data analysis, some of the techniques aforementioned have been extended to variable selection with censored outcome. For Cox's proportional hazards model, Tibshirani et al. (1997) applied LASSO to partial likelihood function, Fan and Li (2002) employed SCAD penalty and derived oracle property for its estimator, Zhang and Lu (2007) utilized adaptive LASSO and obtained its theoretical property. For other models, Lu and Zhang (2007) studied the proportional odds model by maximizing the penalized marginal likelihood of ranks, Zhang et al. (2010b) investigated semiparametric linear transformation models by penalizing a profiled score from the martingale estimating equation, Huang and Ma (2010) modeled the relationship between covariates and survival using the AFT models with bridge penalization for variable selection and parameter estimation, Liu and Zeng (2013) presented an estimation method for semiparametric transformation models that involves minimizing a weighted negative partial loglikelihood function plus an adaptive LASSO penalty. However, one may not select variables for length-biased data using the techniques above, because the estimation and inference treating the censored length-bias data as regular censored data leads to substantial bias and inaccuracy (Shen et al., 2009). Hence, it is necessary to develop a new method for such particular data.

To the best of our knowledge, there has been no work on variable selec-

tion for length-biased data especially, when the dimension of covariates is high. This is partially due to the information censoring length-biased data involve and biased sampling changing the model structure assumed for the target population. Another reason is that most estimation procedures for length-biased data are based on estimating equations, which is very different from the likelihood based methods such as the estimator for Cox's proportional hazards model. The complex nature of length-biased data and the lack of a convenient loss function hinder the existing variable selection methods being directly applied to such data.

In this article, we propose a simple yet powerful method to obtain both sparse and consistent estimator for length-biased data under AFT model. Our first contribution is to construct a working loss function based on the complex estimating equations for length-biased data and then minimize the working loss function with a sparse penalty. Due to the complex structure of length-biased data, we found that the typical penalization method does not produce a very good estimator with finite sample size, although the asymptotic theory supports it. Our second contribution is that we further develop a novel multi-stage sparse penalization procedure (such as SCAD) to achieve more efficient estimation and better sparsity in variable selection.

The remainder of the paper is organized as follows. A description of

length-biased data and the derivation of asymptotically unbiased estimating equation are given in Section 2. The estimator for length-biased data under AFT model is proposed in Section 3. Section 4 describes the implementation, in which the local linear approximation algorithm is introduced and tuning parameter selection problem is discussed. Section 5 derives the theoretical properties for the proposed estimator. Simulation studies and a real data analysis is presented in Section 6. Proof and detailed simulation results are given in supplementary materials.

2. Notation and Model

2.1 Length-biased data

Let $\tilde{T}$ be the uncensored survival time measured from the initiating event to failure without length-bias, A be the time from the initiating event to examination, V be the duration measured from examination to failure, and C be the censoring time from examination. $\tilde{T}$ is left truncated by A, which means, one can only observe T among those $\tilde{T} > A$ in a length-biased sampling, where $T = A + V$ is the observed survival time. Here, A is also known as the truncation variable (or backward recurrence time) and V as the residual survival time (or forward recurrence time).

With right censoring, we have a random sample $(Y_i, A_i, \delta_i, X_i), i = 1, 2, \cdots, n$, where $Y_i = \min(T_i, A_i + C_i), T_i = A_i + V_i, \delta_i = I(V_i \leq C_i)$,

X_i is a $(p+1) \times 1$ vector of covariates for the ith subject usually with intercept and n is the sample size. In addition, we assume C_i is independent of (A_i, V_i) given X_i following literature. We further assume that the right-censoring variable C is independent of covariates X.

Denote f_U as the unbiased density function of $\tilde{T}$, the density function for the length-biased data T conditional on $\tilde{T} > A$ given the covariates $X = x$ has the following form (Shen et al., 2009):

$$g(t|x) = \frac{t f_U(t|x)}{\mu(x)}, \quad \mu(x) = \int_0^\infty s f_U(s|x) ds$$

where $f_U(t|x)$ denotes the unbiased density given the covariates x, and $\mu(x) < \infty$.

2.2 Accelerated Failure Time Models

Consider the AFT model (Kalbfleisch and Prentice, 1980; Cox and Oakes, 1984), which is assumed that the logarithm of the survival time is linearly related to the covariates of interest as,

$$\log \tilde{T} = X^T \boldsymbol{\beta} + \epsilon, \tag{2.1}$$

where X is a covariate vector with intercept and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ parameter vector to estimate and ϵ has an unknown distribution with mean zero. According to Shen et al. (2009), estimating equations for estimation of parameters $\boldsymbol{\beta}$ can be derived using inverse probability of censoring weighting

techniques. Let $S_C(t) = P(C > t)$ be the survival function of C. Under the stationarity assumption, the joint distribution of (A, V) and (A, T) given covariates X has the following form

$$f_{A,V}(a, v|X = x) = f_U(a + v|x)I(a > 0.v > 0)/\mu(x),$$

which can be found in the literature (Zelen, 2006; Asgharian et al., 2005). The probability of observing the failure data is

$$\begin{aligned} P(A = a, Y = y, C \geq y - a|X = x) &= P(A = a, V = y - a, C \geq y - a|X = x) \\ &= f_U(y|x)S_C(y - a)/\mu(x). \end{aligned}$$

Based on the joint distribution of (A, Y) and C conditional on covariates X, we have

$$\begin{aligned} & E\Big[\frac{\delta}{\pi(Y)}(\log Y - X^T\boldsymbol{\beta})\Big] \\ =& E\Big\{E\Big[\frac{\delta}{\pi(Y)}(\log Y - X^T\boldsymbol{\beta})\Big|X = x\Big]\Big\} \\ =& E\Big\{\frac{1}{\mu(x)}\int_0^\infty \Big[\frac{1}{\pi(y)}\int_0^y S_C(y - a)da\Big] f_U(y|x)(\log y - x^T\boldsymbol{\beta})dy\Big\} \\ =& E\Big\{\frac{1}{\mu(x)}E\Big[(\log \tilde{T} - X^T\boldsymbol{\beta})\Big|X\Big]\Big\} = 0, \end{aligned}$$

where $\pi(t) = \int_0^t S_C(u)du$. Then the estimating equation can be constructed as:

$$\tilde{\mathbf{U}}(\boldsymbol{\beta}) = \sum_{i=1}^n X_i\frac{\delta_i}{\pi(Y_i)}(\log Y_i - X_i^T\boldsymbol{\beta}) = 0.$$

Since the censoring distribution is often unknown in practice, it is commonly used to replace the unknown censoring distribution by its consistent Kaplan-Meier estimator

$$\hat{S}_C(t) = \prod_{s \le t} \Big(1 - \frac{\Delta N_C(s)}{\overline{Y}(s)}\Big),$$

where $N_C(t) = \sum_{i=1}^n N_i^C(t)$, $N_i^C(t) = I(Y_i - A_i \le t, \delta_i = 0)$, $\overline{Y}(t) = \sum_{i=1}^n Y_i(t)$, $Y_i(t) = I(Y_i - A_i \ge t)$. Thus, an asymptotic unbiased estimating equation follows:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^n X_i \frac{\delta_i}{\hat{\pi}(Y_i)} (\log Y_i - X_i^T \boldsymbol{\beta}) = 0,$$

where $\hat{\pi}(t) = \int_0^t \hat{S}_C(u)du$ is a consistent plug-in estimator for $\pi(t)$.

Denote

$$\tilde{y}_{(p+1)\times 1} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i X_i \log Y_i}{\hat{\pi}(Y_i)} = \frac{1}{n} \mathbf{X}^T \mathbf{D} \mathbf{y},$$
$$\widetilde{\mathbf{X}}_{(p+1)\times(p+1)} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i X_i X_i^T}{\hat{\pi}(Y_i)} = \frac{1}{n} \mathbf{X}^T \mathbf{D} \mathbf{X} \tag{2.2}$$

as working data, where $\mathbf{D} = diag(\frac{\delta_1}{\hat{\pi}(Y_1)}, \cdots, \frac{\delta_n}{\hat{\pi}(Y_n)})$, $\mathbf{X} = (X_1, \cdots, X_n)^T$, $\mathbf{y} = (\log Y_1, \cdots, \log Y_n)^T$, then the asymptotic unbiased estimating equation can be written as

$$\mathbf{U}(\boldsymbol{\beta}) = n \cdot (\tilde{y} - \widetilde{\mathbf{X}} \boldsymbol{\beta}) = 0. \tag{2.3}$$

Consequently, a closed-form solution for $\boldsymbol{\beta}$ is

$$\widetilde{\mathbf{X}}^{-1} \tilde{y} = (\mathbf{X}^T \mathbf{D} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D} \mathbf{y}. \tag{2.4}$$

It is important to note that (2.4) only holds in low dimensions, because $\widetilde{\mathbf{X}}$ is not invertible when its dimension is greater than the rank of $\mathbf{D}$.

3. Methodology

3.1 Penalized estimating equations

In order to apply the modern penalization estimation method for variable selection in high-dimensions, we need to have a loss function because the common formulation of those methods is a loss plus a sparse penalty. For survival analysis with the Cox's proportional hazard model, the loss function is the negative log partial likelihood. For our study, due to the lack of a convenient loss function, variable selection is more challenging. To overcome the obstacle, we turn (2.3) into a working loss function. Note that finding the root to (2.3) is equivalent to solving the following minimization problem

$$\min_{\boldsymbol{\beta}}(\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T W(\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) \tag{3.1}$$

where W is a positive definite matrix free of $\boldsymbol{\beta}$. For example, a natural choice for W is the identity matrix. Next, we treat the quadratic function in (3.1) as a working loss function and consider minimizing the loss with a sparse penalty to encourage sparsity:

$$\min_{\boldsymbol{\beta}}(\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T W(\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda \sum_{j=2}^{p+1} P_\lambda(|\beta_j|). \tag{3.2}$$

In this work we consider a general folded concave penalty $P_\lambda(|t|)$ and its

definitions can be found in Section 5. It is important to note that the intercept is not penalized in (3.2).

The working loss function idea is related to the recent penalized generalized method of moments estimation studied by Caner (2009) and Fan and Liao (2011) in the econometrics literature, which is seldom seen in statistics field. Another related but different scheme is the penalized generalized estimating equations studied by Johnson et al. (2008) in semiparametric regression models and by Wang et al. (2012) for longitudinal data. These authors have reported very encouraging results. We tried the first approach and provided theoretical support. However, we also found in the numeric study that the resulting estimator, although reduces the dimension greatly, is still not very satisfactory. This issue has not been observed in the aforementioned related work (Caner, 2009; Fan and Liao, 2011; Johnson et al., 2008; Wang et al., 2012), which is due to the more complex structure of the censored length-biased data. This difficulty motivates us to further develop a new procedure which is presented in the next section.

3.2 Multi-stage penalized estimating equations

We propose an iterative multi-stage penalized estimating equation method. The multi-stage variable selection is discussed in Bühlmann and Meier (2008); Zou and Li (2008b) for penalized likelihood to reduce the num-

ber of false positives, which can be very urgent in biological applications since follow-up experiments can be costly and laborious.

Let the initial estimator be

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda^{(1)} \sum_{j=2}^{p+1} P_\lambda(|\beta_j|). \tag{3.3}$$

Suppose that at the k-th iteration we have a current estimator $\hat{\boldsymbol{\beta}}^{(k)}$. Denote active set $\mathcal{A}^k = \{j : \hat{\beta}_j^{(k)} \neq 0\}$, so $\hat{\boldsymbol{\beta}}_{\mathcal{A}^k}^{(k)}$ is the vector constituted by the nonzero components of $\hat{\boldsymbol{\beta}}^{(k)}$, and $\mathbf{X}_{\mathcal{A}^k}$ is the dimension-reduced design matrix with columns selected by $\mathcal{A}^k$. For computing the next iteration estimator $\hat{\boldsymbol{\beta}}^{(k+1)}$, we first compute dimension-reduced working data $(\tilde{y}_{\mathcal{A}^k}, \widetilde{\mathbf{X}}_{\mathcal{A}^k})$ by $\mathbf{X}_{\mathcal{A}^k}$ through (2.2). Then we consider the following optimization problem:

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}^k}^{(k+1)} = \min_{\boldsymbol{\beta}} (\tilde{y}_{\mathcal{A}^k} - \widetilde{\mathbf{X}}_{\mathcal{A}^k}\boldsymbol{\beta})^T W_{\mathcal{A}^k} (\tilde{y}_{\mathcal{A}^k} - \widetilde{\mathbf{X}}_{\mathcal{A}^k}\boldsymbol{\beta}) + \lambda^{(k+1)} \sum_{j \in \mathcal{A}^k; j \neq 1} P_\lambda(|\beta_j|), \tag{3.4}$$

where $W_{\mathcal{A}^k}$ is a computed working matrix computed based on $\hat{\boldsymbol{\beta}}^{(k)}$ and $\mathcal{A}^k$. Specifically, given $\hat{\boldsymbol{\beta}}^{(k)}$ and the data $\mathbf{y}, \mathbf{X}$,

$$\begin{aligned} W_{\mathcal{A}^k} &= \Big[\frac{1}{n}\sum_{i=1}^{n} X_i \Big(\frac{\delta_i}{\hat{\pi}(Y_i)}(\log Y_i - X_i^T \hat{\boldsymbol{\beta}}^{(k)})\Big)^2 X_i^T\Big]^{-1} \\ &= \Big[\frac{1}{n}\mathbf{X}_{\mathcal{A}^k}^T diag\big(\ (\mathbf{D}(\mathbf{y} - \mathbf{X}_{\mathcal{A}^k}\hat{\boldsymbol{\beta}}_{\mathcal{A}^k}^{(k)}))^2\ \big)\mathbf{X}_{\mathcal{A}^k}\Big]^{-1}, \end{aligned}$$

The interpretation of $W_{\mathcal{A}^k}$ is that it is an estimate of the inverse of the covariance matrix of the estimation equation. Note that we use identity

matrix as preliminary weighting matrix in the first step because we have no information about the covariance matrix of the estimation equation.

The penalization parameters $\lambda^{(k)}$ are not required to be the same. No matter how we choose $\lambda^{(k)}$, active sets sequences $\mathcal{A}^k$ are always nested, that is,

$$\mathcal{A}^k \supseteq \mathcal{A}^{k+1}$$

Therefore, we stop the iteration when we observe the convergence of the current active set, that is,

$$\text{if } \mathcal{A}^k = \mathcal{A}^{k+1} \quad \text{stop the iteration.}$$

By the nested property, the convergence is guaranteed.

After convergence, the active set is the selected subset of important variables. We also try to refit the coefficient by solving the unpenalized estimation equation with the selected subset. This final step is for reducing the estimation bias generated in the iterative penalization stage.

4. Implementation

4.1 LLA Algorithm and 2-step LLA solution

Our estimation method can work with all sparse penalties. In this work we focused on the folded concave penalties that include SCAD and MCP as two special cases. Since the penalty function is folded concave and non-differentiable at point 0, optimization objective function can be difficult

and sometimes has multiple local minimizers. We adopt the local linear approximation algorithm (LLA) proposed in Zou and Li (2008a) to compute the proposed estimator. Fan et al. (2014) proved that the computed local solution by LLA is the desired theoretical local solution, which resolved a final missing puzzle in the folded concave penalization picture. Here we directly present the LLA algorithm for solving (3.2). The same algorithm is applied repeatedly in the iterative multi-stage penalized estimating equations procedure. For its derivation and explanations, readers are referred to Zou and Li (2008a).

First, we compute the initial estimator as the LASSO penalized estimator

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T W (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda_{\text{lasso}} \sum_{j=2}^{p+1} |\beta_j|. \tag{4.1}$$

Given the LASSO estimator, we compute

$$\hat{\boldsymbol{\beta}}^{\text{lla1}} = \arg\min_{\boldsymbol{\beta}} (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T W (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \sum_{j=2}^{p+1} P'_\lambda(|\hat{\beta}_j^{\text{lasso}}|)|\beta_j|.$$

Given $\hat{\boldsymbol{\beta}}^{\text{lla1}}$, we compute

$$\hat{\boldsymbol{\beta}}^{\text{lla2}} = \arg\min_{\boldsymbol{\beta}} (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T W (\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) + \sum_{j=2}^{p+1} P'_\lambda(|\hat{\beta}_j^{\text{lla1}}|)|\beta_j|.$$

Following Fan et al. (2014), we stop with $\hat{\boldsymbol{\beta}}^{\text{lla2}}$ as the solution.

4.2 Tuning parameter selection

In a penalized estimation method, the choice of penalization parameter is very important. The tuning parameter selection method in Caner (2009) is based on subset selection which is only feasible in very low dimension. Fan and Liao (2011) did not consider the tuning parameter selection problem. Johnson et al. (2008) applied generalized cross-validation statistic and Wang et al. (2012) conducted cross validation to tune the parameter which are not applicable for length-biased data since the prediction error is hard to define.

In order to tune the regularization parameter λ, we borrow the extended BIC criterion from Chen and Chen (2008) for the linear regression model to the estimation equation setting. In the context of linear regression, the extended BIC is defined as

$$\frac{RSS}{\hat{\sigma}^2} + d\log n + 2\gamma \log \binom{p}{d}, \qquad 0 \leq \gamma \leq 1,$$

where n denotes the sample size, d denotes the number of free parameters, RSS is the residual sum of squares from the OLS fit, and $\hat{\sigma}^2$ is an estimator of error variance computed by the full model. Moreover, $\gamma = \frac{1}{2}$ for $p = n$ case as suggested by Chen and Chen (2008).

For the estimator from (3.4), we define the extended BIC criterion as

$$n \cdot (\tilde{y}_{\mathcal{A}^k} - \widetilde{\mathbf{X}}_{\mathcal{A}^k}\boldsymbol{\beta}_\lambda)^T W_{\mathcal{A}^k}(\tilde{y}_{\mathcal{A}^k} - \widetilde{\mathbf{X}}_{\mathcal{A}^k}\boldsymbol{\beta}_\lambda) + \|\boldsymbol{\beta}_\lambda\|_0 \cdot \log n + \log \binom{|\mathcal{A}^k|}{\|\boldsymbol{\beta}_\lambda\|_0}, \quad (4.2)$$

where $\|\cdot\|_0$ is the L0-norm. The idea is to treat $nW_{\mathcal{A}^k}$ as the role of $\frac{1}{\hat{\sigma}^2}$ in the original extended BIC for linear regression model.

For the estimator from (3.3), the working inverse covariance matrix is the identity matrix. If we consider the full model in order to get an analogue of $\hat{\sigma}^2$ in linear regression, note that "sample size" is $p+1$, "model size" is $\| \boldsymbol{\beta}_\lambda \|_0$, hence the "residuals" become zero because the number of parameters is equal to the sample size of working data. To avoid dividing zero, we define a similar extended BIC criterion as follows

$$(\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}_\lambda)^T(\tilde{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}_\lambda)\Big(1 + \frac{\| \boldsymbol{\beta}_\lambda\|_0 \log(p+1) + \log\binom{p+1}{\|\boldsymbol{\beta}_\lambda\|_0}}{p+1-\|\boldsymbol{\beta}_\lambda\|_0}\Big). \tag{4.3}$$

We use (4.3) in the first stage of the multi-stage penalized estimating equation procedure to get the first estimator. Then in the subsequent multi-stage procedure, we have $nW_{\mathcal{A}^k}$ and can apply (4.2) to tune the estimator. This practice has been tested in our simulation studies and worked well.

5. Theoretical properties

In this section we present the asymptotic results of our estimators for high-dimensional variable selection and estimation. Denote the true parameter in (2.1) as $\boldsymbol{\beta}^*$, the support set as $\mathcal{A} = \{j : \beta^* \neq 0\}$ and its cardinality as $s = |\mathcal{A}|$. The sparse estimation problem often assumes that s is much smaller than the dimension of $\boldsymbol{\beta}^*$.

Denote the problem we consider in (3.2) as

$$\min_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) + P_\lambda(|\boldsymbol{\beta}|)$$

with $\ell_n(\boldsymbol{\beta}) = \| W^{\frac{1}{2}}\tilde{y} - W^{\frac{1}{2}}\widetilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$ a convex loss and $P_\lambda(|\boldsymbol{\beta}|) = \sum_{j=2}^{p+1} P_\lambda(|\beta_j|)$. A true oracle estimator knows the true support set and is obtained by (2.4) using the true support set, that is

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}} = \left(\mathbf{X}_{\mathcal{A}}^T \mathbf{D} \mathbf{X}_{\mathcal{A}}\right)^{-1} \mathbf{X}_{\mathcal{A}}^T \mathbf{D} \mathbf{y}, \qquad \hat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{\text{oracle}} = \mathbf{0}.$$

Before presenting our theorems, we first state some conditions:

(A) $\|\boldsymbol{\beta}_{\mathcal{A}}^*\|_{\min} > (a+1)\lambda$, where $\|\cdot\|_{\min}$ is the minimum entrywise absolute value and a is a constant defined in Condition (E);

(B) $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$ are i.i.d. sub-Gaussian(σ) for some fixed constant $\sigma > 0$, that is, $E[\exp(t\epsilon_i)] \leq \exp(\sigma^2 t^2/2)$;

(C) There exists constant $M > m > 0$, such that $\frac{1}{M} < |\pi(Y)| < \frac{1}{m}$;

(D) $\kappa = \min_{\boldsymbol{\delta} \in \mathbb{R}^{p+1}; \boldsymbol{\delta} \neq 0: \|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1 \leq 3\|\boldsymbol{\delta}_{\mathcal{A}}\|_1} \frac{\|W^{\frac{1}{2}}\widetilde{\mathbf{X}}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_2^2} \in (0, +\infty)$;

(E) Assume the folded concave penalty $P_\lambda(|t|)$ defined on $t \in (-\infty, \infty)$ satisfying following assumptions:

(i) $P_\lambda(t)$ is increasing and concave in $t \in [0, \infty)$ with $P_\lambda(0) = 0$;

(ii) $P_\lambda(t)$ is differentiable in $t \in (0, \infty)$ with $P'_\lambda(0) := P'_\lambda(0+) \geq a_1\lambda$;

(iii) $P'_\lambda \geq a_1\lambda$ for $t \in (0, a_2\lambda]$;

(iv) $P'_\lambda = 0$ for $t \in [a\lambda, \infty)$ with pre-specified constant $a > a_2$.

Where a_1 and a_2 are two fixed positive constants.

(F) X is a $(p+1) \times 1$ vector of bounded covariates, not contained in a p-dimensional hyperplane;

(G) $\sup[t : \Pr(V > t) > 0] \geq \sup[t : \Pr(C > t) > 0] = t_0$ and $\Pr(\delta = 1) > 0$;

(H) $\int_0^{t_0} \left\{[(\int_t^{t_0} S_C(u)du)^2]/[S_C^2(t)S_V(t)]\right\} dS_C(t) < \infty$, where $S_V(t)$ is the survival function for residual failure time;

(I) $\det\left(E[\delta X_{\mathcal{A}}(\log Y - X_{\mathcal{A}}^T\boldsymbol{\beta}^*_{\mathcal{A}})/\pi(Y)]^{\otimes 2}\right) < \infty$, where for a vector v, $v^{\otimes 2} = vv^T$;

(J) $\det\left(\int_0^{t_0} \left\{H^{\otimes 2}(s)/[S_C^2(s)S_V(s)]\right\} dS_C(s)\right) < \infty$,
where $H(t) = E\left\{[\delta X_{\mathcal{A}} I(Y \geq s) \int_t^Y S_C(u)du(\log Y - X_{\mathcal{A}}^T\boldsymbol{\beta}^*_{\mathcal{A}})]/[\pi^2(Y)]\right\}$;

(K) $\boldsymbol{\Gamma}_{\mathcal{A}} \equiv \lim_{n\to\infty} \frac{1}{n}\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}}$ is nonsingular.

Conditions (A)-(B) can be found in Fan et al. (2014) in order to calculate the probability bound to ensure the convergence of LLA solution. Condition (D) is similar to the restricted eigenvalue condition considered by Bickel et al. (2009) in sparse linear regression. The assumptions in condition (E) can be found in Fan et al. (2014), which summarize the previous

works on the SCAD and the MCP. The derivatives of the SCAD penalty and MCP penalty are

$$P'_\lambda(t) = \lambda I_{\{t\le\lambda\}} + \frac{(a\lambda - t)_+}{a-1} I_{\{t>\lambda\}} \qquad \text{for some } a > 2,$$

$$P'_\lambda(t) = (\lambda - \frac{t}{a})_+ \qquad \text{for some } a > 1,$$

respectively. It is obvious that $a_1 = a_2 = 1$ for the SCAD, and $a_1 = 1 - a^{-1}, a_2 = 1$ for the MCP.

Conditions (F)-(K) are exactly the same as those in Shen et al. (2009). Under regularity conditions (F)-(K), they proved that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}} - \boldsymbol{\beta}^*_{\mathcal{A}})$ converges weakly to a normal distribution with mean zero and covariance matrix $\boldsymbol{\Gamma}_{\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A}}\boldsymbol{\Gamma}_{\mathcal{A}}^{-1}$, in which $\boldsymbol{\Sigma}_{\mathcal{A}}$ is the asymptotic covariance matrix of $n^{-\frac{1}{2}}\mathbf{U}_{\mathcal{A}}(\boldsymbol{\beta}^*_{\mathcal{A}}) = n^{-\frac{1}{2}}(\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{y} - \mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}^*_{\mathcal{A}})$. Further more, $\boldsymbol{\Gamma}_{\mathcal{A}}$ and $\boldsymbol{\Sigma}_{\mathcal{A}}$ can be consistently estimated by

$$\hat{\boldsymbol{\Gamma}}_{\mathcal{A}} = \frac{1}{n}\mathbf{X}_{\mathcal{A}}^T\mathbf{D}\mathbf{X}_{\mathcal{A}}, \tag{5.1}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{A}} = \frac{1}{n}\sum_{i=1}^n \left\{\delta_i X_{\mathcal{A}i}\frac{(\log Y_i - X_{\mathcal{A}i}^T\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}})}{\hat{\pi}(Y_i)} + \int_0^{t_0}\frac{\hat{H}(t)d\hat{M}_i(t)}{\eta(t)}\right\}^{\otimes 2}, \tag{5.2}$$

respectively, where

$$\hat{H}(t) = \frac{1}{n}\sum_{i=1}^n I(t\le Y_i)\delta_i X_{\mathcal{A}i}\int_t^{Y_i}\hat{S}_C(u)du\frac{(\log Y_i - X_{\mathcal{A}i}^T\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}})}{\hat{\pi}^2(Y_i)},$$

$$\hat{M}_i(t) = I(Y_i - A_i \le t, \delta_i = 0) - \int_0^t I(Y_i - A_i \ge u)d\hat{\Lambda}_C(u),$$

$$\eta(t) = \frac{1}{n}\sum_{i=1}^n I(Y_i - A_i \ge t),$$

and $\hat{\Lambda}_C(u)$ is the Nelson-Aalen estimator for the cumulative hazard function of C.

To connect the true oracle estimator with our LLA estimator, we define a so-called "working data oracle estimator" as

$$\tilde{\boldsymbol{\beta}}^{\text{oracle}} = (\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}, \mathbf{0}) = \underset{\boldsymbol{\beta}:\boldsymbol{\beta}_{\mathcal{A}^c}=\mathbf{0}}{\arg\min}\ \ell_n(\boldsymbol{\beta}).$$

Since $\ell_n(\boldsymbol{\beta})$ is convex, the solution above is unique, namely,

$$\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}} = \big(\widetilde{\mathbf{X}}_{\mathcal{A}}^T W \widetilde{\mathbf{X}}_{\mathcal{A}}\big)^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}^T W \tilde{y},$$

where $\widetilde{\mathbf{X}}_{\mathcal{A}}$ stands for the columns of $\widetilde{\mathbf{X}}$ corresponding to the support set and

$$\nabla_j \ell_n(\tilde{\boldsymbol{\beta}}^{\text{oracle}}) = 0, \qquad \forall j \in \mathcal{A},$$

where ∇_j denoted the subgradient with respect to the jth component of $\boldsymbol{\beta}$.

Denote $\mathbf{X}^o, \mathbf{X}^o_{\mathcal{A}}, \mathbf{X}^o_{\mathcal{A}^c}$ as the sub-matrixes formed by the rows in $\mathbf{X}, \mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}^c}$ where T_i is being observed, i.e. $\delta_i = 1$. For simplicity, write

$$\lambda_{\max}^{\mathcal{A}\mathcal{A}} = \lambda_{\max}\big(\frac{1}{n}\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{X}^o{}_{\mathcal{A}}\big), \quad \lambda_{\min}^{\mathcal{A}\mathcal{A}} = \lambda_{\min}\big(\frac{1}{n}\mathbf{X}^{oT}_{\mathcal{A}}\mathbf{X}^o{}_{\mathcal{A}}\big), \quad \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c} = \lambda_{\max}\big(\frac{1}{n}\mathbf{X}^{oT}_{\mathcal{A}^c}\mathbf{X}^o{}_{\mathcal{A}^c}\big),$$

where $\lambda_{\max}(\cdot), \lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of a matrix correspondingly.

We state the following asymptotic results for the estimator obtained by (3.2).

Theorem 1. *Consider the folded concave penalized problem (3.2) for any given positive definite matrix W with SCAD or MCP penalty. Denote $\lambda_{\max}^W, \lambda_{\min}^W$ as the maximum and minimum eigenvalues of W. Initialize the LLA algorithm by $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ which is obtained from (4.1). Given conditions (A)-(E) and let $a_0 = \min\{1, a_2\}$, if we pick $\lambda \geq \frac{3\sqrt{s}\lambda_{\text{lasso}}}{a_0\kappa}$, the solution of LLA algorithm $\hat{\boldsymbol{\beta}}$ converges to $\tilde{\boldsymbol{\beta}}^{\text{oracle}}$ after two iterations, with probability at least $1 - \delta_0^{\text{lasso}} - \delta_1 - \delta_2$, where*

$$\delta_1 = 2(p+1-s)\exp\Big(-\frac{na_1^2\lambda^2}{8\sigma^2M^2(\lambda_{\max}^W)^2\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c}(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\Big),$$

$$\delta_2 = 2s\exp\Big(-\frac{n\cdot m^4(\|\boldsymbol{\beta}_{\mathcal{A}}^*\|_{\min}-a\lambda)^2}{2\sigma^2M^2}\frac{\lambda_{\min}^{\mathcal{A}\mathcal{A}\,4}}{\lambda_{\max}^{\mathcal{A}\mathcal{A}}(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\big(\frac{\lambda_{\min}^W}{\lambda_{\max}^W}\big)^2\Big),$$

$$\delta_0^{\text{lasso}} = 2(p+1)\exp\Big(-\frac{n\lambda_{\text{lasso}}^2}{32\sigma^2(\lambda_{\max}^W)^2M^2(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^3}\Big).$$

Thus, we have $\Pr(supp(\hat{\boldsymbol{\beta}}) = \mathcal{A}) \to 1$ as n goes to infinity with $supp(\hat{\boldsymbol{\beta}})$ denoting the support set of $\hat{\boldsymbol{\beta}}$. Moreover, for any $\xi > 0$, $\theta \in (0, \frac{1}{2})$, we have

$$\Pr\Big(\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}\|_{\max} \leq \xi n^{-\theta}\Big) \geq 1 - \delta_0^{\text{lasso}} - \delta_1 - \delta_2 - \delta_3,$$

where

$$\delta_3 \leq 2s\exp\Big(-\frac{n^{1-2\theta}\xi^2}{16\sigma^2}\frac{1}{\lambda_{\max}^{\mathcal{A}\mathcal{A}}}\big[m^2\lambda_{\min}^{\mathcal{A}\mathcal{A}\,2}+\frac{M^4}{m^2}\frac{\lambda_{\min}^{\mathcal{A}\mathcal{A}\,4}}{(\lambda_{\max}^{\mathcal{A}\mathcal{A}}+\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\Big(\frac{\lambda_{\min}^W}{\lambda_{\max}^W}\Big)^2\big]\Big).$$

Remark 1. As discussed in Fan et al. (2014), it is also a good choice to use zero to initialize the LLA algorithm. If $\hat{\boldsymbol{\beta}}^{\text{initial}} = \mathbf{0}$, the first LLA iteration

gives a LASSO estimator with $\lambda_{\text{lasso}} = P'_\lambda(0)$. For both SCAD and MCP, $P'_\lambda(0) = \lambda$. If $\lambda_{\text{lasso}} = \lambda$ and $a_0\kappa \geq 3\sqrt{s}$, then after two more LLA iterations, or equivalently after three iterations when initialized by zero, the solution of LLA algorithm $\hat{\boldsymbol{\beta}}$ has the same asymptotic results in Theorem 1, as long as we replace δ_0^{lasso} there with

$$\delta_0^{\mathbf{0}} = 2(p+1)\exp\Big(-\frac{n\lambda^2}{32\sigma^2(\lambda_{\max}^{W})^2 M^2(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^3}\Big).$$

6. Numerical Studies

6.1 Simulations

In this section we assess the performance of our proposed methods by several numerical experiments and a real data analysis. We report average numbers of correct and incorrect non-zero coefficients, along with the average of mean squared errors based on 1000 simulated data sets, 200 sample size and $p = 20, 100, 400$ variables for the three penalized estimating equations estimators introduced above with LASSO penalty and SCAD penalty in (3.3), multi-stage SCAD penalty in (3.4). Here, mean squared errors are calculated by $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, where Σ is the population covariance matrix. To examine the inference results of non-zero coefficients in the final estimate, we report the biases (Bias), standard errors (SE), mean of asymptotic standard errors (ASE), and coverage probabilities (CP) of nominal 95% confidence intervals for the multi-stage SCAD penalized es-

timating equations estimator. Notice that the asymptotic standard errors are calculated by the sandwich formula with (5.1) and (5.2) using non-zero coefficients in the final estimate, and coverage probabilities are computed based on these asymptotic standard errors.

The length-biased and right censored data are generated according to the method in Shen et al. (2009). First we generate independent pairs $(A_i, \tilde{T}_i)$ and keep the pairs that satisfy $\tilde{T}_i > A_i$, where A_i is from a uniform distribution $U(0, \tau)$ and $\tilde{T}_i$ are generated from the models below. Here τ is chosen to be larger than the upper bound of the support of $\tilde{T}$ to satisfy the stationarity assumption. The censoring time C is generated from a uniform distribution $U(0, \omega_0)$, where ω_0 is chosen to achieve the desired censoring ratio. 10 %, 30% and 60% censoring rates are considered in our simulation. Tables in this paper present part of the simulation results. For full detailed results, see supplementary material online.

Example 1. The first example is adopted from Shen et al. (2009). Consider the AFT model

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

where $X = (1, X_1, X_2, \cdots, X_p)^T$ and $\boldsymbol{\beta} = (1, 1, 1, 0_{p-2})$. X_{2k-1} are i.i.d. Bernoulli variables with $P(X_{2k-1} = 1) = 0.5$, X_{2k} are i.i.d. uniform variables on $(0, 1)$, $k = 1, 2, \cdots$. Random error ϵ is generated from: (1)

$U(-0.5, 0.5)$, (2) $Exp(5) - 0.2$, (3) $N(0, 0.3^2)$.

Table 1 summarizes the average numbers of correct and incorrect non-zero coefficients, along with the average of mean squared errors. It can be inferred that all the true variables are selected by the three methods with almost 100% frequency. It can also be observed that LASSO and SCAD estimator will select far more incorrect non-zero coefficients than multi-stage SCAD, so the multi-stage is needed to achieve an almost perfect selection accuracy. In addition, we examine the inference results for the multi-stage SCAD penalized estimating equations estimator in Table 2. The empirical biases are mostly less than 3% and 5.62% in the worst case, indicating the proposed estimator achieves outstanding accuracy. Note that the exponential random error setting violates the sub-Gaussian error assumption, the simulation results presented in supplementary materials are still quite good. However, it is interesting to observe that the asymptotic standard errors calculated by the sandwich formula (5.1) and (5.2) are always slightly smaller than the Monte Carlo standard error, leading to a more or less decrease in the empirical coverage probabilities than 95% nominal level, especially dramatic when the dimension is high. The underestimation of the estimated standard errors than the sample standard errors can also be observed in variable selection literature for survival data (Lu and Zhang,

2007; Zhang and Lu, 2007; Johnson et al., 2008; Zhang et al., 2010b; Li and Gu, 2012). It is observed that the discrepancy between ASE and SE will decrease when sample size becomes large (Zhang and Lu, 2007; Li and Gu, 2012).

Example 2. Consider the underlying population distribution of $\tilde{T}$ follow

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

where $X = (1, X_1, \cdots, X_p)^T$ and X_is are marginally standard normal random variables with pairwise correlations $Cor(X_i, X_j) = \rho^{|i-j|}$, i.e. autoregressive correlation structure AR(ρ). $\rho = 0.5, 0.8$ are taken into account. We set $\boldsymbol{\beta} = (2, 0.3, 0.3, 0, 0, 0.3, 0_{p-5})$. ϵ is generated from $N(0, 0.2^2)$.

From Table 3 we see that the multi-stage SCAD penalized estimating equations estimator is still encouraging, though it has a little chance to miss some true variables when censoring rate is 60% and dimension is high. This false negative rate, however, can be acceptable when comparing a great amount decrease of false positives with LASSO and SCAD. The asymptotic standard errors in Table 4 are still underestimated as well as the coverage probabilities.

6.2 Real data

The proposed approach is applied to the Oscar Awards data analyzed

and compiled by Redelmeier and Singh (2001). The dataset could be found in Han et al. (2011), where a detailed description is given. It is a list of all 766 nominees for Oscar awards from 1929 to 2000, and only 327 died before the study ended. This means that the censoring ratio is about 57.3%.

Several authors (Redelmeier and Singh, 2001; Han et al., 2011; Chen et al., 2015; Ma et al., 2016) are interested in finding out whether winning an Oscar Award causes the actor or actress' expected lifetime to increase. Redelmeier and Singh (2001) fitted a Cox's proportional hazards model and claimed that life expectancy was 3.9 years longer for Oscar Award winners than for other less recognized performers. Han et al. (2011) stated that previous studies have suffered from healthy performer survivor bias, which is, candidates who are healthier will be able to act in more films and have more chance to win Oscar Awards. They adapted Robins' rank preserving structural accelerated failure time model and g-estimation method, and concluded there is no strong evidence that winning an Oscar increases life expectancy. Both Chen et al. (2015) and Ma et al. (2016) treated the survival time of performers as length-biased right-censored data, and they conducted monotone rank estimation method for transformation models and estimation method for semiparametric transformation models respectively to analyze this data. They all had the conclusion that a performer winning

Oscar may not have longer lifetime span than those without winning.

However, it is also of our interest to study the association between survival time and other nine variables of performers' information in the dataset as well as wining an Oscar Award. They are indicators including gender (male=1, female=0), born in USA (yes=1, no=0), white (yes=1, no=0), change name (yes=1, no=0), genre is drama (yes=1, no=0), and count variables with number of total films in career, number of four-star films, number of times the performer won an Oscar, number of times the performer was nominated for an Oscar.

Denote T as the time from birth to death, A as the truncation variable, that is the time from the performer's birth year to the first Oscar nomination year. Based on the formal test proposed by Addona and Wolfson (2006), the p-value of this test is 0.3, suggesting the dataset satisfies the stationarity assumption and can be treated as censored length-biased data.

We standardize the count variables and apply our proposed the multi-stage SCAD penalized estimating equations estimator to the dataset. The results of non-zero coefficient variables are shown in Table 5, along with their standard errors and 95% confidence intervals. The indicator of whether the performer has won an Oscar is not selected, implying winning an Oscar has nothing to do with life expectancy increase. Other significant variables

shows that female nominees tend to live longer than male nominees, USA performers are likely to live shorter than others, the number of total films and the number of four-star films in career have a positive effect on performers' life expectancy. Part of these results are consistent with those in Ma et al. (2016). The refitted model is

$$\log \tilde{T} = 4.2004 - 0.1106 * Gender - 0.1232 * USA + 0.0019 * NOTF + 0.0058 * NOFF.$$

To further explore the data and reduce possible modeling biases, we add all the possible interaction of variables as well as the quadratic terms of count variables to the initial model, totally 59 predictors. Table 6 presents the scaled variables selected by the multi-stage SCAD penalized estimating equations estimator. The refitted model is

$$\log \tilde{T} = 4.2029 + 0.0065 * NOFF - 0.1519 * Gender * USA + 0.0021 * USA * NOTF.$$

The binary variable indicating winning an Oscar is still outside of the active set. Again, the number of four-star films is selected, suggesting that it is a crucial predictor for the lifetime of movie stars and there is a positive association between a good physical condition and plenty of high-quality films.

7. Discussion

In this paper, we proposed an estimation method based on the penalized

estimating equations to achieve a sparse estimation with high-dimensional covariates for length-biased data under the AFT model. Theoretical results guarantee the selection and estimation consistency property of the proposed estimator. Moreover, a multi-stage penalized estimating equations procedure is developed to achieve enhanced estimation accuracy and sparsity. Numerical results also demonstrate the excellent performance of our estimator for both variable selection and model estimation.

Although we assumed C is independent of X in our paper because we may not know in advance that which covariates C depends on. However, generalizing derivations to the setting with a covariate-dependent censoring distribution is not conceptually difficult, such as fitting a semiparametric or parametric model and plug covariate-specific censoring distribution $S_c(\cdot|x)$ into the estimating equations (Shen et al., 2009; Chen and Zhou, 2012), as long as we know the dependent covariates in advance.

As suggested by a referee, we may consider an augmented-based estimator (Gorfine et al., 2017), treating censoring indicator as a special case of missing indicator. This estimator has a doubly-robust advantage, that is, the estimator is consistent when either the censoring distribution does not depend on the covariates, or the posited model for the conditional expectation is correct. This is a welcomed feature since we assumed censoring

distribution does not depend on the covariates for variable selection. However, the corresponding computation can be much more intensive and how to well choose a posited model for the conditional expectation term is worth study. This is an interesting problem which deserves further investigation.

Acknowledgements We thank the edtor, associate editor and referees for their helpful comments and suggestions. Zou's work is supported in part by NSF grant DMS-1505111. Zhou's work was supported by the State Key Program of National Natural Science Foundation of China (71331006), the State Key Program in the Major Research Plan of National Natural Science Foundation of China (91546202), National Center for Mathematics and Interdisciplinary Sciences (NCMIS), Key Laboratory of RCSDS, AMSS, CAS (2008DP173182) and Innovative Research Team of Shanghai University of Finance and Economics (IRTSHUFE13122402).

Supplementary Materials

The supplementary file contains proof of the theorem and full detailed tables for simulation studies.

References

Addona, V. and Wolfson, D. B. (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Analysis*,

Table 1: Average numbers of correct and incorrect non-zero coefficients and average of mean squared errors from 1000 simulated datasets for Example 1, with their standard errors shown in the parenthesis

			LASSO			SCAD			MS-SCAD		
error	p	censoring	C	I	MSE	C	I	MSE	C	I	MSE
unif	100	10%	2.00	28.68	0.036	2.00	36.48	0.041	2.00	0.78	0.005
			(0.03)	(8.89)	(0.015)	(0)	(8.96)	(0.009)	(0)	(1.6)	(0.007)
		30%	2.00	28.37	0.038	2.00	36.35	0.045	2.00	0.76	0.006
			(0.05)	(9.53)	(0.016)	(0)	(9.41)	(0.01)	(0)	(1.48)	(0.007)
		60%	2.00	30.66	0.047	2.00	38.61	0.061	2.00	1.28	0.013
			(0.03)	(11.03)	(0.02)	(0)	(9.91)	(0.014)	(0)	(2.29)	(0.017)
	400	10%	2.00	54.94	0.055	2.00	72.06	0.056	2.00	2.07	0.011
			(0)	(17.96)	(0.018)	(0)	(15.72)	(0.007)	(0)	(2.97)	(0.012)
		30%	2.00	57.94	0.053	2.00	77.25	0.061	2.00	2.14	0.014
			(0)	(19.63)	(0.018)	(0)	(17.22)	(0.008)	(0)	(3.16)	(0.015)
		60%	2.00	106.30	0.063	2.00	117.13	0.072	2.00	4.00	0.031
			(0)	(40.09)	(0.023)	(0)	(33.27)	(0.01)	(0)	(7.24)	(0.054)
normal	100	10%	2.00	29.56	0.038	2.00	37.41	0.043	2.00	0.40	0.004
			(0.03)	(9.31)	(0.016)	(0)	(9.25)	(0.011)	(0)	(0.94)	(0.005)
		30%	2.00	29.41	0.040	2.00	37.22	0.048	2.00	0.56	0.005
			(0)	(9.49)	(0.016)	(0)	(9.21)	(0.012)	(0)	(1.32)	(0.007)
		60%	2.00	31.50	0.049	2.00	39.28	0.065	2.00	0.84	0.010
			(0)	(10.75)	(0.021)	(0)	(9.93)	(0.016)	(0)	(1.85)	(0.016)
	400	10%	2.00	56.87	0.058	2.00	73.03	0.058	2.00	1.02	0.007
			(0.03)	(18.16)	(0.019)	(0)	(15.66)	(0.008)	(0)	(2.2)	(0.01)
		30%	2.00	59.95	0.058	2.00	78.39	0.063	2.00	1.43	0.010
			(0.03)	(20.14)	(0.019)	(0)	(17.35)	(0.009)	(0)	(2.93)	(0.015)
		60%	2.00	103.52	0.069	2.00	116.18	0.075	2.00	3.61	0.026
			(0)	(38.9)	(0.025)	(0)	(32.38)	(0.012)	(0.03)	(8.56)	(0.037)

Table 2: Estimates of coefficients for Multi-Stage SCAD, their biases, standard errors, mean of asymptotic standard errors, and coverage probabilities for nominal 95% confidence intervals from 1000 simulated datasets for Example 1

			unif				normal			
p	censoring		Bias	SE	ASE	CP	Bias	SE	ASE	CP
100	10%	b1	-0.0022	0.0534	0.0496	92.3	-0.0050	0.0556	0.0531	92.8
		b2	-0.0068	0.0947	0.0856	90.4	-0.0007	0.0992	0.0911	91.8
	30%	b1	-0.0040	0.0565	0.0532	92.5	-0.0047	0.0603	0.0560	91.6
		b2	-0.0135	0.0997	0.0913	92.5	-0.0111	0.1047	0.0958	92.1
	60%	b1	-0.0059	0.0698	0.0626	90.5	-0.0061	0.0738	0.0664	91.8
		b2	-0.0197	0.1277	0.1083	88.7	-0.0137	0.1254	0.1142	91.3
400	10%	b1	-0.0104	0.0542	0.0470	89.5	-0.0075	0.0558	0.0514	91.4
		b2	-0.0250	0.0967	0.0808	87.4	-0.0136	0.1016	0.0883	89.5
	30%	b1	-0.0124	0.0600	0.0503	87.5	-0.0085	0.0576	0.0542	91.8
		b2	-0.0226	0.1072	0.0867	85.8	-0.0278	0.1055	0.0929	88.8
	60%	b1	-0.0235	0.0762	0.0567	81.0	-0.0189	0.0780	0.0609	83.2
		b2	-0.0529	0.1575	0.0982	76.1	-0.0562	0.1551	0.1054	80.1

12(3):267–284.

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association*, 97(457):201–209.

Asgharian, M., Wolfson, D. B., et al. (2005). Asymptotic behavior of the unconditional npmle of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics*, 33(5):2109–2131.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4).

Table 3: Average numbers of correct and incorrect non-zero coefficients and average of mean squared errors from 1000 simulated datasets for Example 2, with their standard errors shown in the parenthesis; AR(ρ) is the autoregressive correlation structure for predictors

			LASSO			SCAD			MS-SCAD		
	p	censoring	C	I	MSE	C	I	MSE	C	I	MSE
AR(0.5)	100	10%	3.00	68.70	0.018	3.00	70.39	0.033	3.00	2.99	0.005
			(0)	(10.74)	(0.005)	(0)	(7.51)	(0.009)	(0)	(6.95)	(0.008)
		30%	3.00	66.87	0.022	3.00	68.68	0.044	3.00	5.27	0.009
			(0)	(10.16)	(0.007)	(0)	(7.82)	(0.015)	(0)	(8.82)	(0.013)
		60%	3.00	65.43	0.036	3.00	68.89	0.131	2.99	7.36	0.023
			(0)	(8.77)	(0.011)	(0)	(7.69)	(0.065)	(0.11)	(9.66)	(0.029)
	400	10%	3.00	230.92	0.030	3.00	263.35	0.437	3.00	3.98	0.006
			(0)	(29.47)	(0.005)	(0)	(24.64)	(0.096)	(0.08)	(8.22)	(0.011)
		30%	3.00	243.69	0.034	3.00	251.83	0.507	2.99	2.65	0.006
			(0)	(25.98)	(0.006)	(0)	(24.77)	(0.1)	(0.12)	(5.53)	(0.011)
		60%	3.00	213.96	0.036	3.00	218.36	0.551	2.91	1.98	0.015
			(0)	(25.11)	(0.01)	(0)	(25.66)	(0.091)	(0.34)	(4.59)	(0.051)
AR(0.8)	100	10%	3.00	40.19	0.010	3.00	44.45	0.019	3.00	1.80	0.004
			(0)	(11.61)	(0.003)	(0)	(9.43)	(0.005)	(0.03)	(4.06)	(0.005)
		30%	3.00	39.78	0.012	3.00	44.58	0.023	3.00	2.15	0.005
			(0)	(11.59)	(0.004)	(0)	(9.63)	(0.006)	(0.03)	(4.24)	(0.006)
		60%	3.00	43.01	0.020	3.00	47.42	0.042	2.99	4.36	0.014
			(0)	(10.89)	(0.007)	(0)	(9.68)	(0.016)	(0.12)	(5.84)	(0.017)
	400	10%	3.00	154.67	0.022	3.00	190.94	0.14	2.99	6.57	0.010
			(0)	(30.32)	(0.004)	(0)	(27.58)	(0.047)	(0.08)	(11.15)	(0.014)
		30%	3.00	169.42	0.026	3.00	211.08	0.218	2.98	3.00	0.007
			(0)	(32.17)	(0.006)	(0)	(27.18)	(0.072)	(0.13)	(6.26)	(0.011)
		60%	3.00	187.90	0.037	3.00	194.92	0.383	2.86	1.80	0.012
			(0)	(27.73)	(0.01)	(0)	(27.81)	(0.134)	(0.39)	(2.95)	(0.017)

Table 4: Estimates of coefficients for Multi-Stage SCAD, their biases, standard errors, mean of asymptotic standard errors, and coverage probabilities for nominal 95% confidence intervals from 1000 simulated datasets for Example 2; AR(ρ) is the autoregressive correlation structure for predictors

			AR(0.5)				AR(0.8)			
p	censoring		Bias	SE	ASE	CP	Bias	SE	ASE	CP
100	10%	b1	-0.0020	0.0211	0.0188	90.7	-0.0019	0.0338	0.0290	90.4
		b2	-0.0003	0.0228	0.0191	87.4	0.0003	0.0370	0.0307	90.6
		b5	-0.0017	0.0189	0.0166	89.4	-0.0023	0.0248	0.0203	87.3
	30%	b1	0.0008	0.0236	0.0196	88.6	0.0001	0.0350	0.0301	90.3
		b2	-0.0030	0.0244	0.0197	86.6	-0.0035	0.0373	0.0319	89.9
		b5	-0.0022	0.0220	0.0172	86.9	-0.0026	0.0276	0.0216	87.6
	60%	b1	-0.0025	0.0358	0.0222	78.4	0.0012	0.0496	0.0333	84.0
		b2	-0.0056	0.0384	0.0224	78.2	-0.0078	0.0574	0.0357	81.8
		b5	-0.0062	0.0376	0.0196	75.0	-0.0103	0.0426	0.0249	79.6
400	10%	b1	0.0028	0.0242	0.0186	88.5	0.0024	0.0362	0.0268	84.7
		b2	-0.0033	0.0278	0.0186	88.6	-0.0028	0.0406	0.0285	85.0
		b5	-0.0018	0.0218	0.0163	87.5	-0.0060	0.0301	0.0193	82.7
	30%	b1	-0.0002	0.0247	0.0201	90.0	0.0005	0.0387	0.0293	87.5
		b2	-0.0029	0.0303	0.0201	88.9	-0.0004	0.0436	0.0312	87.4
		b5	-0.0030	0.0316	0.0174	86.4	-0.0055	0.0409	0.0204	85.5
	60%	b1	-0.0066	0.0580	0.0246	83.7	0.0012	0.0675	0.0353	83.0
		b2	-0.0041	0.0568	0.0250	83.3	-0.0064	0.0832	0.0370	82.6
		b5	-0.0158	0.0657	0.0208	82.9	-0.0271	0.0861	0.0233	79.5

Table 5: Variable selection results for Oscar data

	Coef	SE	95% CI	
Gender	-0.1106	0.0328	(-0.1749 ,	-0.0463)
USA	-0.1232	0.0263	(-0.1747 ,	-0.0716)
NOTF	0.0666	0.0155	(0.0362 ,	0.0970)
NOFF	0.0359	0.0112	(0.0140 ,	0.0578)

† Note: Gender: male=1, female=0; USA: whether born in USA,yes=1, no=0; NOTF: number of total films; NOFF: number of four-star films.

Table 6: Variable selection results for Oscar data with quadratic and interaction terms

	Coef	SE	95% CI	
NOFF	0.0406	0.0092	(0.0226 ,	0.0585)
Gender*USA	-0.1519	0.0375	(-0.2253 ,	-0.0785)
USA*NOTF	0.0721	0.0166	(0.0396 ,	0.1046)

† Note: Gender: male=1, female=0; USA: whether born in USA,yes=1, no=0; NOTF: number of total films; NOFF: umber of four-star films.

Bühlmann, P. and Meier, L. (2008). Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1534–1541.

Caner, M. (2009). Lasso-type gmm estimator. *Econometric Theory*, 25(01):270–290.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, pages 759–771.

Chen, X., Shi, J., and Zhou, Y. (2015). Monotone rank estimation of transformation models with length-biased and right-censored data. *Science China Mathematics*, 58(10):1–14.

Chen, X. R. and Zhou, Y. (2012). Quantile regression for right-censored and length-biased data. *Acta Mathematicae Applicatae Sinica English*, 28(3):443–462.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, volume 21. CRC Press.

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

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, pages 74–99.

Fan, J. and Liao, Y. (2011). Ultra high dimensional variable selection with endogenous covariates. *Manuscript. Princeton University*.

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of statistics*, 42(3):819.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, pages 1069–1112.

Gorfine, M., Goldberg, Y., and Ritov, Y. (2017). A quantile regression model for failure-time data with time-dependent covariates. *Biostatistics*, 18(1):132–146.

Han, X., Small, D. S., Foster, D. P., Patel, V., et al. (2011). The effect of winning an oscar award on survival: correcting for healthy performer survivor bias with a rank preserving structural accelerated failure time model. *The Annals of Applied Statistics*, 5(2A):746–772.

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

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103(482):672–680. PMID: 20376193.

Kalbfleisch, J. and Prentice, R. (1980). *The statistical analysis of time failure data*. John Wiley

and Sons New York.

Lagakos, S. W., Barraj, L. M., and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to aids. *Biometrika*, 75(3):515–523.

Li, J. and Gu, M. (2012). Adaptive lasso for general transformation models with right censored data. *Computational Statistics & Data Analysis*, 56(8):2583–2597.

Liu, X. and Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika*, 100(4).

Lu, W. and Zhang, H. H. (2007). Variable selection for proportional odds model. *Statistics in Medicine*, 26(20):3771–3781.

Ma, H., Qiu, Z., and Zhou, Y. (2016). Semiparametric transformation models with length-biased and right-censored data under the case-cohort design. *Statistics and Its Interface*, 9(2):213–222.

Ning, J., Qin, J., and Shen, Y. (2011). Buckley–james-type estimator with right-censored and length-biased data. *Biometrics*, 67(4):1369–1378.

Qin, J. and Shen, Y. (2010). Statistical methods for analyzing right-censored length-biased data under cox model. *Biometrics*, 66(2):382–392.

Redelmeier, D. A. and Singh, S. M. (2001). Survival in academy award–winning actors and actresses. *Annals of Internal Medicine*, 134(10):955–962.

Shen, Y., Ning, J., and Qin, J. (2009). Analyzing length-biased data with semiparametric

transformation and accelerated failure time models. *Journal of the American Statistical Association*, 104(487):1192–1202.

Simon, R. (1980). Length biased sampling in etiologic studies. *American Journal of Epidemiology*, 111(4):444–452.

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

Tibshirani, R. et al. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, pages 616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, pages 178–203.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86(413):130–143.

Zelen, M. (2006). Forward and backward recurrence times and length biased sampling: age specific models. In *Probability, Statistics and Modelling in Public Health*, pages 1–11. Springer.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, 56(3):601–614.

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

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhang, H. H., Lu, W., and Wang, H. (2010b). On sparse estimation for semiparametric linear transformation models. *Journal of Multivariate Analysis*, 101(7):1594–1606.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Li, R. (2008a). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of statistics*, 36(4):1509.

Zou, H. and Li, R. (2008b). Rejoinder: One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36.

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

E-mail: (hedi8910@163.com)

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China; Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

E-mail: (yzhou@amss.ac.cn)

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

E-mail: (zouxx019@umn.edu)