Statistica Sinica **30** (2020), 193-215
doi:https://doi.org/10.5705/ss.202018.0089

HIGH-DIMENSIONAL VARIABLE SELECTION WITH RIGHT-CENSORED LENGTH-BIASED DATA

Di He[1,2], Yong Zhou[3] and Hui Zou[4]

[1]*Shanghai University of Finance and Economics,* [2]*Nanjing University,*
[3]*East China Normal University and* [4]*University of Minnesota*

Abstract: Length-biased data are common in various fields, including epidemiology and labor economics, and they have attracted considerable attention in survival literature. A crucial goal of a survival analysis is to identify a subset of risk factors and their risk contributions from among a vast number of clinical covariates. However, there is no research on variable selection for length-biased data, owing to the complex nature of such data and the lack of a convenient loss function. Therefore, we propose an estimation method based on penalized estimating equations to obtain a sparse and consistent estimator for length-biased data under an accelerated failure time model. The proposed estimator possesses the selection and estimation consistency property. In particular, we implement our method using a SCAD penalty and a local linear approximation algorithm. We suggest selecting the tuning parameter using the extended BIC in high-dimensional settings. Furthermore, we develop a novel multistage SCAD penalized estimating equation procedure to achieve improved estimation accuracy and sparsity in the variable selection. Simulation studies show that the proposed procedure has high accuracy and almost perfect sparsity. Oscar Awards data are 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 used to collect data in fields such as epidemiology and labor economics. For length-biased data, we assume 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. Equivalently, we can assume that 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 Canadian Study of Health and Aging (CSHA) on dementia among elderly people (Asgharian, M'Lan and Wolfson (2002); Addona and Wolfson (2006); Shen, Ning and Qin (2009); Qin and Shen (2010)). The study recruited and screened more than 10,000 Canadians over the age of 65 for the prevalence of dementia. The approximate initial date of dementia and the subsequent time of death and censoring were recorded for individuals found to have dementia. Those individuals who had dementia and did not survive to the examination time were excluded from the investigation. Thus, 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, Barraj and De Gruttola (1988); Wang (1991)). Another approach is based on the unconditional distribution (Vardi (1982, 1985); Gill, Vardi and Wellner (1988); Asgharian, M'Lan and Wolfson (2002); Asgharian and Wolfson (2005)), which requires the stationarity assumption. Recently, the analysis of right-censored and length-biased data has attracted the attentions of many researchers. Informative censoring, that is, the dependence between the right-censoring time and the failure time, can make analyses of such data difficult. Another significant difficulty is that the observed length-biased data may change the model structure that has been assumed for the target population. Shen, Ning and Qin (2009) developed estimating equation methods for semiparametric transformation and accelerated failure time (AFT) models to obtain consistent estimators of the regression coefficients. Qin and Shen (2010) proposed two estimating equation approaches for using the Cox model to analyze covariate effects. Ning, Qin and Shen (2011) presented a generalized Buckley–James-type estimator under an AFT model.

A crucial goal of survival analyses is to identify the risk factors and their risk contributions. Modern data-collection technologies are making vast amounts of data on clinical covariates accessible, including patients' personal characteristics, biomarkers, and genotypes. A necessary, but challenging task is to select a subset of important variables upon which the hazard function or the survival time depends. This helps medical researchers build comprehensible models to predict outcomes without information loss, leading to better disease diagnoses and treatments in the long run. The process, called variable selection or feature selection, has been widely studied for linear models with uncensored outcomes, including subset selection, the least absolute shrinkage and selection oper-

ator (LASSO) (Tibshirani (1996)), bridge regressions (Fu (1998)), the smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)), elastic nets (Zou and Hastie (2005)), the adaptive LASSO (Zou (2006)), and the minimax concave penalty (MCP) (Zhang (2010)). In the context of survival data analyses, some of the aforementioned techniques have been extended to variable selection with censored outcomes. For Cox's proportional hazards model, Tibshirani (1997) applied the LASSO to a partial likelihood function, Fan and Li (2002) employed a SCAD penalty to derive the oracle property for its estimator, and Zhang and Lu (2007) utilized the adaptive LASSO to obtain its theoretical properties. For other models, Lu and Zhang (2007) studied the proportional odds model, where they maximize the penalized marginal likelihood of ranks. Zhang, Lu and Wang (2010) 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 AFT models, with bridge penalization for the variable selection and parameter estimation. Liu and Zeng (2013) presented an estimation method for semiparametric transformation models that minimizes a weighted negative partial loglikelihood function plus an adaptive LASSO penalty. However, we cannot select variables for length-biased data using the above techniques, because an estimation or inference that treats the censored length-bias data as regular censored data will lead to substantial bias and inaccuracy (Shen, Ning and Qin (2009)). Hence, it is necessary to develop a new method for such data.

To the best of our knowledge, there is no existing work on variable selection for length-biased data, especially when the dimension of the covariates is high. As mentioned earlier, this is due to the information censoring 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. Such procedures are quite 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 from being applied directly to such data.

We propose a simple, yet powerful method for obtaining sparse and consistent estimators for length-biased data under an AFT model. Our first contribution is to construct a working loss function based on complex estimating equations for length-biased data, after which, we minimize the working loss function using a sparse penalty. Owing to the complex structure of length-biased data, we find that the typical penalization method does not produce a very good estimator for

a finite sample size, although the asymptotic theory supports such an estimator. Our second contribution is a novel multistage sparse penalization procedure (e.g., the SCAD) that achieves a more efficient estimation and better sparsity during variable selection.

The remainder of the paper is organized as follows. In Section 2, we describe length-biased data and derive an asymptotically unbiased estimating equation. The estimator for length-biased data under an AFT model is proposed in Section 3. Section 4 describes the implementation. Here, we introduce the local linear approximation algorithm and discuss the tuning parameter selection problem. Section 5 derives the theoretical properties for the proposed estimator. Simulation studies and a real-data analysis are presented in Section 6. All proofs and detailed simulation results are given in the Supplementary Material.

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. Here, $\tilde{T}$ is left truncated by A, which means we can only observe T of $\tilde{T} > A$ in a length-biased sample, 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 $(Y_i, A_i, \delta_i, X_i)$, for $i = 1, 2, \ldots, n$, 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 an intercept, and n is the sample size. In addition, we assume C_i is independent of (A_i, V_i), given X_i, following the literature. We further assume that the right-censoring variable C is independent of the 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 that the covariates $X = x$ have the following form (Shen, Ning and Qin (2009)):

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

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

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

where X is a covariate vector with intercept, and $\boldsymbol{\beta}$ is a $(p+1)\times 1$ parameter vector to be estimated, and ϵ has an unknown distribution with mean zero. According to Shen, Ning and Qin (2009), the equations for estimating $\boldsymbol{\beta}$ can be derived using the 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 the covariates X, has the following form:

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

as discussed in the literature (Zelen (2006); Asgharian and Wolfson (2005)). The probability of observing the failure data is

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

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

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

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

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

Because the censoring distribution is often unknown in practice, researchers often to replace the unknown censoring distribution by its consistent Kaplan–Meier estimator

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

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

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

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 $\mathbf{D} = diag(\delta_1/(\hat{\pi}(Y_1)), \ldots, \delta_n/(\hat{\pi}(Y_n)))$, $\mathbf{X} = (X_1, \ldots, X_n)^T$, and $\mathbf{y} = (\log Y_1, \ldots, \log Y_n)^T$. 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

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

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 a modern penalization estimation method for variable selection in high-dimensions, we need to have a loss function because the common formulation of such methods is a loss plus a sparse penalty. For survival analyses using Cox's proportional hazard model, the loss function is the negative log partial likelihood. For our study, owing to the lack of a convenient loss function, variable selection is more challenging. To overcome this obstacle, we change (2.3) into a working loss function. Note that finding the roots of (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 minimize the loss, using 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 $P_\lambda(|t|)$, defined in Section 5. 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. A related scheme contains the penalized generalized estimating equations studied by Johnson, Lin and Zeng (2008) for semiparametric regression models, and by Wang, Zhou and Qu (2012) for longitudinal data. These studies reported encouraging results. We tried the first approach, and our results provided theoretical support. However, our numeric study showed that the resulting estimator, despite reducing the dimension, is still not satisfactory. This issue is not identified in the aforementioned works (Caner (2009); Fan and Liao (2011); Johnson, Lin and Zeng (2008); Wang, Zhou and Qu (2012)), owing to the more complex structure of censored length-biased data. This difficulty motivates us to develop a new procedure, which is presented in the next section.

3.2. Multistage penalized estimating equations

We propose an iterative multistage penalized estimating equation method. Multistage variable selection is discussed in Bühlmann and Meier (2008); Zou and Li (2008b) for the penalized likelihood to reduce the number of false positives, which can be serious in biological applications, because 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 kth iteration we have the estimator $\hat{\boldsymbol{\beta}}^{(k)}$. Denote the active set $\mathcal{A}^k=\{j:\hat{\beta}_j^{(k)}\neq 0\}$, where $\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{A}^k}$ 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 $\mathcal{A}^k$. To compute the next iteration estimator $\hat{\boldsymbol{\beta}}^{(k+1)}$, we first compute the dimension-reduced working data $(\tilde{y}_{\mathcal{A}^k}, \widetilde{\mathbf{X}}_{\mathcal{A}^k})$ using $\mathbf{X}_{\mathcal{A}^k}$ and (2.2). Then, we consider the following optimization problem:

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

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

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

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 the identity matrix as the 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. Regardless of our choice of $\lambda^{(k)}$, the active set 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 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, 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 reduces the estimation bias generated in the iterative penalization stage.

4. Implementation

4.1. LLA algorithm and two-step LLA solution

Our estimation method can work with all sparse penalties. In this work, we focused on folded concave penalties, which include the SCAD and MCP as special cases. Because the penalty function is folded concave and nondifferentiable at point 0, the optimization objective function can be difficult, and sometimes has multiple local minimizers. We adopt the local linear approximation (LLA) algo-

rithm proposed in Zou and Li (2008a) to compute the proposed estimator. Fan, Xue and Zou (2014) proved that the local solution computed by the LLA is the desired theoretical local solution. Here, we present the LLA algorithm for solving (3.2). The same algorithm is applied repeatedly in the iterative multistage penalized estimating equations procedure. For its derivation and explanations, refer 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, Xue and Zou (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 that is only feasible for a very low dimension. Fan and Liao (2011) did not consider the tuning parameter selection problem. Johnson, Lin and Zeng (2008) applied a generalized cross-validation statistic, and Wang, Zhou and Qu (2012) conducted cross validation to tune the parameter, neither of which is applicable for length-biased data because the prediction error is difficult to define.

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

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

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 $p = n$, as suggested

by Chen and Chen (2008).

For the estimator from (3.4), we define the extended BIC 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 $nW_{\mathcal{A}^k}$ in the same way as $\frac{1}{\hat{\sigma}^2}$ in the original extended BIC for a 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 the linear regression, "sample size" is $p+1$ and "model size" is $\| \boldsymbol{\beta}_\lambda \|_0$. Hence, the "residuals" become zero because the number of parameters is equal to the sample size of the working data. To avoid dividing by zero, we define a similar extended BIC, as follows:

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

We use (4.3) in the first stage of the multistage penalized estimating equation procedure to obtain the first estimator. Then, in the subsequent multistage procedure, we have $nW_{\mathcal{A}^k}$ and can apply (4.2) to tune the estimator. This practice is tested in our simulation studies, and works 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

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

where $\ell_n(\boldsymbol{\beta}) = \| W^{1/2}\tilde{y} - W^{1/2}\widetilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$ is 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 this set; that is,

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

Before presenting our theorems, we first state several 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) $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are independent and identically distributed (i.i.d.) sub-Gaussian(σ), for some fixed constant $\sigma > 0$; that is, $E[\exp(t\epsilon_i)] \le \exp(\sigma^2 t^2/2)$;

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

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

(E) Assume a 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 $P'_\lambda(0) := P'_\lambda(0+) \ge a_1\lambda$;

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

(iv) $P'_\lambda = 0$, for $t \in [a\lambda, \infty)$, with prespecified constant $a > a_2$;

where a_1 and a_2 are fixed positive constants.

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

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

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

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

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

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

Conditions (A)–(B) can be found in Fan, Xue and Zou (2014), who calculate the probability bound to ensure the convergence of the LLA solution. Condition (D) is similar to the restricted eigenvalue condition considered by Bickel, Ritov and Tsybakov (2009) in a sparse linear regression. The assumptions in condition (E) can be found in Fan, Xue and Zou (2014), who summarize previous works on the SCAD and MCP. The derivatives of the SCAD penalty and MCP penalty are

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

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

respectively. Clearly, $a_1 = a_2 = 1$ for the SCAD, and $a_1 = 1 - a^{-1}, a_2 = 1$ for the MCP.

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

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

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

respectively, where

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

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

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

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

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

Because $\ell_n(\boldsymbol{\beta})$ is convex, the above solution is unique; that is,

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

where $\widetilde{\mathbf{X}}_{\mathcal{A}}$ denotes 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 denotes the subgradient with respect to the jth component of $\boldsymbol{\beta}$.

Denote $\mathbf{X}^o$, $\mathbf{X}^o_{\mathcal{A}}$ and $\mathbf{X}^o_{\mathcal{A}^c}$ as the submatrixes formed by the rows in $\mathbf{X}$, $\mathbf{X}_{\mathcal{A}}$ and $\mathbf{X}_{\mathcal{A}^c}$, respectively, where T_i is being observed; that is, $\delta_i = 1$. For simplicity,

write

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

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

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 a SCAD or an MCP penalty. Denote $\lambda_{\max}^W, \lambda_{\min}^W$ as the maximum and minimum eigenvalues of W, respectively. Initialize the LLA algorithm using $\hat{\boldsymbol{\beta}}^{\text{lasso}}$, which is obtained from* (4.1). *Given conditions* (A)–(E) *and letting $a_0 = \min\{1, a_2\}$, if we pick $\lambda \geq (3\sqrt{s}\lambda_{\text{lasso}})/(a_0\kappa)$, the solution of the 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*

$$\delta_1 = 2(p+1-s)\exp\bigg(-\frac{na_1^2\lambda^2}{8\sigma^2 M^2(\lambda_{\max}^W)^2\lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c}(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\bigg),$$
$$\delta_2 = 2s\exp\bigg(-\frac{n\cdot m^4(\|\boldsymbol{\beta}^*_{\mathcal{A}}\|_{\min} - a\lambda)^2}{2\sigma^2 M^2}\frac{{\lambda_{\min}^{\mathcal{A}\mathcal{A}}}^4}{\lambda_{\max}^{\mathcal{A}\mathcal{A}}(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^2}\bigg(\frac{\lambda_{\min}^W}{\lambda_{\max}^W}\bigg)^2\bigg),$$
$$\delta_0^{\text{lasso}} = 2(p+1)\exp\bigg(-\frac{n\lambda_{\text{lasso}}^2}{32\sigma^2(\lambda_{\max}^W)^2 M^2(\lambda_{\max}^{\mathcal{A}\mathcal{A}} + \lambda_{\max}^{\mathcal{A}^c\mathcal{A}^c})^3}\bigg).$$

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 $\xi > 0$, $\theta \in (0, 1/2)$, we have*

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

where

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

Remark 1. Fan, Xue and Zou (2014) suggest using 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 the SCAD and the MCP, $P'_\lambda(0) = \lambda$. If $\lambda_{\text{lasso}} = \lambda$ and $a_0\kappa \geq 3\sqrt{s}$, then after two further LLA iterations, or equivalently, after three iterations when initialized by zero, the solution of the LLA algorithm $\hat{\boldsymbol{\beta}}$ has the same asymptotic results as those described in Theorem 1, as long as we replace δ_0^{lasso} with

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

6. Numerical Studies

6.1. Simulations

In this section, we assess the performance of our proposed methods using several numerical experiments and a real-data analysis. We report the average numbers of correct and incorrect nonzero coefficients, along with the average mean squared errors based on 1,000 simulated data sets, a sample size of 200 and $p = 20, 100, 400$ variables for the penalized estimators with a LASSO penalty and a SCAD penalty in (3.3), and a multistage SCAD penalty in (3.4). Here, the mean squared errors are calculated as $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, where Σ is the population covariance matrix. To examine the inference results of the nonzero coefficients in the final estimate, we report the biases (Bias), standard errors (SE), means of asymptotic standard errors (ASE), and coverage probabilities (CP) of the nominal 95% confidence intervals for the multistage SCAD penalized estimating equations estimator. Note that the asymptotic standard errors are calculated using the sandwich formula, with (5.1) and (5.2) using nonzero coefficients in the final estimate, and the 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, Ning and Qin (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. We consider censoring rates of 10%, 30%, and 60% in our simulation. The tables shown here present part of the simulation results. For the full detailed results, see the online Supplementary Material.

Example 1. The first example is adopted from Shen, Ning and Qin (2009). Consider the AFT model

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

where $X = (1, X_1, X_2, \ldots, X_p)^T$ and $\boldsymbol{\beta} = (1, 1, 1, 0_{p-2})$. Here, X_{2k-1} are i.i.d. Bernoulli variables, with $P(X_{2k-1} = 1) = 0.5$, and X_{2k} are i.i.d. uniform vari-

Table 1. Average numbers of correct and incorrect nonzero coefficients and average of mean squared errors from 1,000 simulated data sets for Example 1, with their standard errors shown in parentheses.

			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)	(1.60)	(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)	(0.010)	(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)	(0.020)	(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)	(0.010)	(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)	(2.20)	(0.010)
		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)

ables on (0, 1), for $k = 1, 2, \ldots$. The random error ϵ is generated from: (1) $U(-0.5, 0.5)$, (2) $Exp(5) - 0.2$, and (3) $N(0, 0.3^2)$.

Table 1 summarizes the average numbers of correct and incorrect nonzero coefficients, along with the average 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 the LASSO and SCAD estimators select far more incorrect nonzero coefficients than the multistage SCAD does. Thus, the multistage SCAD is needed to achieve an almost perfect selection accuracy. In addition, we examine the inference results for the multistage SCAD penalized estimator in Table 2. The empirical biases are mostly less than 3%, and 5.62% in the worst case, indicating that the proposed estimator achieves outstanding accuracy. Note that although the exponential random error setting violates the subGaussian error assumption, the simulation results presented in the Supple-

Table 2. Estimates of coefficients for multistage SCAD, their biases, standard errors, mean of asymptotic standard errors, and coverage probabilities for nominal 95% confidence intervals from 1,000 simulated data sets 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

mentary Material are still quite good. However, it is interesting to observe that the asymptotic standard errors calculated using the sandwich formula (5.1) and (5.2) are always slightly smaller than the Monte Carlo standard errors, leading to an approximate decrease in the empirical coverage probabilities from the 95% nominal level, especially when the dimension is high. The underestimation of estimated standard errors from sample standard errors can also be observed in the variable selection literature for survival data (Lu and Zhang (2007); Zhang and Lu (2007); Johnson, Lin and Zeng (2008); Zhang, Lu and Wang (2010); Li and Gu (2012)). Note that the discrepancy between the ASE and SE decreases when the sample size becomes large (Zhang and Lu (2007); Li and Gu (2012)).

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

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

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

From Table 3 we see that the multistage SCAD estimator is still encouraging, though there is a small chance of missing true variables when the censoring rate is 60% and the dimension is high. This false negative rate can be acceptable when comparing a decrease of false positives with the LASSO and SCAD. The

Table 3. Average numbers of correct and incorrect nonzero coefficients and average of mean squared errors from 1,000 simulated data sets for Example 2, with their standard errors shown in the parentheses; 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)

asymptotic standard errors in Table 4 are still underestimated, as the coverage probabilities.

6.2. Real data

The proposed approach is applied to Oscar Awards data analyzed and compiled by Redelmeier and Singh (2001). The data set can 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, of whom 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, Shi and Zhou (2015); Ma, Qiu and Zhou (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

Table 4. Estimates of coefficients for multistage SCAD, their biases, standard errors, mean of asymptotic standard errors, and coverage probabilities for nominal 95% confidence intervals from 1,000 simulated data sets 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

claimed that the 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 a healthy performer survivor bias. Thus, candidates who are healthier can act in more films and have a greater chance of winning an Oscar Award. 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, Shi and Zhou (2015) and Ma, Qiu and Zhou (2016) treated the survival time of performers as length-biased right-censored data, and they conducted a monotone rank estimation method for transformation models and an estimation method for semiparametric transformation models. They all concluded that a performer winning Oscar may not have longer lifetime span than those without winning.

However, we also wish to study the association between the survival time and nine other variables of performers' information in the data set, including wining an Oscar Award. These indicators include gender (male=1, female=0), born in USA (yes=1, no=0), white (yes=1, no=0), change name (yes=1, no=0),

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.

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, and 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 data set satisfies the stationarity assumption and can be treated as censored length-biased data.

We standardize the count variables and apply our proposed the multistage SCAD penalized estimator to the data set. The results of nonzero 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, US performers are likely to live shorter lives than others, and the number of films and the number of four-star films in career have a positive effect on performers' life expectancy. Parts of these results are consistent with those in Ma, Qiu and Zhou (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 the possible modeling bias, we add all possible interactions of variables and the quadratic terms of count variables to the initial model, yielding 59 predictors. Table 6 presents the scaled variables selected by the multistage 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 denoting winning an Oscar is still outside of the active set.

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.

Again, the number of four-star films is selected, suggesting that it is a crucial predictor for the lifetime of movie stars, and that there is a positive association between being in good physical condition and many 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 an AFT model. The theoretical results guarantee the selection and estimation consistency property of the proposed estimator. Moreover, a multistage penalized estimating equations procedure is developed to improve the estimation accuracy and sparsity. Numerical results demonstrate the excellent performance of our estimator for both variable selection and model estimation.

We assume that C is independent of X because we may not know in advance 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 substituting a covariate-specific censoring distribution $S_c(\cdot|x)$ into the estimating equations (Shen, Ning and Qin (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, Goldberg and Ritov (2017)), treating the censoring indicator as a special case of the missing indicator. This estimator has a doubly robust advantage, because the estimator is consistent regardless of whether the censoring distribution depends on the covariates, or whether the posited model for the conditional expectation is correct. This is a welcome feature because we assume that the censoring distribution does not depend on the covariates for the variable selection. However, the corresponding computation can be much more intensive, thus,

choosing a posited model for the conditional expectation term is worth studying. This is an interesting problem that deserves further investigation.

Supplementary Material

The Supplementary Material contains the proofs of the theorems and detailed tables for the simulation studies.

Acknowledgments

We thank the editor, associate editor, and referees for their helpful comments and suggestions. Zou's work was supported, in part, by NSF grant DMS-1505111. Zhou's work was supported by the State Key Program in the Major Research Plan of National Natural Science Foundation of China (91546202) and the Key Laboratory of Advanced Theory and Application in Statistics and Data Science, MOE.

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* **12**, 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**, 201–209.

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

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

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

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

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 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**, 1–14.

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

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Vol. 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**, 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty

model. *The Annals of Statistics* **30**, 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**, 819.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 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* **16**, 1069–1112.

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

Han, X., Small, D. S., Foster, D. P. and Patel, V. (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**, 746–772.

Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis* **16**, 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**, 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**, 515–523.

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

Liu, X. and Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika* **100**, 859–876.

Lu, W. and Zhang, H. H. (2007). Variable selection for proportional odds model. *Statistics in Medicine* **26**, 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**, 213–222.

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

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

Redelmeier, D. A. and Singh, S. M. (2001). Survival in academy award–winning actors and actresses. *Annals of Internal Medicine* **134**, 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**, 1192–1202.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*

Statistical Society. Series B (Statistical Methodological), 267–288.

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

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics* **10**, 616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13**, 178–203.

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

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* **86**, 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*, 1–11. Springer.

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

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

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

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 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**, 301–320.

Zou, H. and Li, R. (2008a). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533.

Zou, H. and Li, R. (2008b). Rejoinder: One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1561–1566.

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

School of Economics, Nanjing University, Nanjing, 210046, China.

E-mail: hedi8910@163.com

Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai 200062, China.

E-mail: yzhou@amss.ac.cn

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

E-mail: zouxx019@umn.edu

(Received June 2017; accepted March 2018)