Computational Statistics and Data Analysis 198 (2024) 107991

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A conditional approach for regression analysis of case K interval-censored failure time data with informative censoring

Mingyue Du [a,*], Xingqiu Zhao [b]

[a] School of Mathematics, Jilin University, Changchun 130012, China
[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Keywords:
Conditional approach, informative censoring
Penalized method
Variable selection

ABSTRACT

This paper discusses regression analysis of case K interval-censored failure time data, a general type of failure time data, in the presence of informative censoring with the focus on simultaneous variable selection and estimation. Although many authors have considered the challenging variable selection problem for interval-censored data, most of the existing methods assume independent or non-informative censoring. More importantly, the existing methods that allow for informative censoring are frailty model-based approaches and cannot directly assess the degree of informative censoring among other shortcomings. To address these, we propose a conditional approach and develop a penalized sieve maximum likelihood procedure for the simultaneous variable selection and estimation of covariate effects. Furthermore, we establish the oracle property of the proposed method and illustrate the appropriateness and usefulness of the approach using a simulation study. Finally we apply the proposed method to a set of real data on Alzheimer's disease and provide some new insights.

1. Introduction

This paper discusses regression analysis of case K interval-censored failure time data, a general type of failure time data, in the presence of informative censoring with the focus on simultaneous variable selection and estimation. Interval-censored data occur when the failure time of interest is known only to belong to an interval rather than being observed exactly and their analysis has recently attracted a great deal of attention (Sun, 2006). It is easy to see that such data can occur in many situations or fields and among others, one area that usually yields interval-censored data is periodic follow-up studies such as clinical trials. By informative censoring, we usually mean that the failure time of interest and the censoring mechanism or observation process may be correlated or the latter carries some information about the former (Sun, 2006).

One specific example of interval-censored failure time data that motivated this study is given by the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal follow-up study that started in 2004 and was designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of the Alzheimer's disease (AD). Due to the periodic follow-up nature, only interval-censored data are available on many variables such as the AD conversion time. Among other, one major goal of the initiative is to determine or identify the relevant biomarkers that can be used to predict the AD conversion. More details on the study will be given below and more examples of interval-censored data can be found in Sun (2006).

* Corresponding author.
E-mail address: mingydu@jlu.edu.cn (M. Du).

https://doi.org/10.1016/j.csda.2024.107991
Received 3 March 2024; Received in revised form 25 April 2024; Accepted 13 May 2024

Available online 23 May 2024

A great deal of literature has been established for variable selection under various contexts, and in particular, many methods have been proposed for right-censored failure time data situations (Fan and Li, 2002; Shi et al., 2014; Tibshirani, 1997; Zhang and Lu, 2007). Among the commonly used methods, the penalized estimation procedure, which optimizes an objective function with a penalty function, has recently become increasingly popular and in particular, many different penalty functions have been proposed. They include the L_1 penalty, the least absolute shrinkage and selection operator (LASSO) penalty (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the adaptive LASSO (ALASSO) penalty (Zou, 2006), the smooth integration of counting and absolute deviation (SICA) penalty (Lv and Fan, 2009), the seamless-L0 (SELO) penalty (Dicker et al., 2013), and the broken adaptive ridge (BAR) penalty (Liu and Li, 2016).

Some methods have been proposed for variable selection when one faces interval-censored failure time data (Du et al., 2021; Li et al., 2020; Wu and Cook, 2015; Zhao et al., 2020). For example, Wu and Cook (2015) and Zhao et al. (2020) proposed some penalized variable selection procedures under the weakly parametric proportional hazards (PH) model and the standard PH model, respectively. In particular, the latter provided a BAR penalty-based method and established the oracle property of the approach. In addition, Li et al. (2020) discussed the same problem under a class of semiparametric transformation models and Du et al. (2021) provided a unified approach under the PH model. However, all of methods above except Du et al. (2021) assume independent or non-informative censoring and it is well-known that in the presence of informative censoring, the analysis that ignores it would lead to biased estimation. More comments on the approach given in Du et al. (2021) are given below. Also it is worth noting that the variable selection with interval-censored data is much more challenging than with right-censored data under the PH model both numerically and theoretically. One main reason is that with the latter, a simple partial likelihood function that involves only regression parameters is available and usually used as the objective function. In contrast, with the former, the same is not true and one has to work with a much more complicated objective function as seen below.

Many authors have discussed regression analysis of interval-censored failure time data with informative censoring and this is especially the case for case I or current status data, a special case of interval-censored data where each subject is observed only once (Du et al., 2019; Du and Yu, 2023; Du et al., 2022; Li et al., 2017a, 2017b; Ma et al., 2015; Sun, 2006; Wang et al., 2016; Wang et al., 2018, 2020). For the situation, most of the existing methods such as that proposed in Du et al. (2021) are frailty-based procedures in which some frailty or latent variables are used to characterize the relationship between the failure time of interest and the observation process. A drawback of this approach is that some restrictions or distribution assumptions, which cannot usually be verified, have to be used. Another common drawback of the existing methods is that they need to assume that the observation process is a Poisson process. More importantly, such an approach cannot provide a direct estimation about the degree of informative censoring. To address these, we will propose a conditional approach that can overcome these shortcomings with the focus on simultaneous variable selection and estimation.

More specifically, we will present a conditional PH model and develop a penalized sieve maximum likelihood approach. In the method, B-splines functions will be used and the oracle property of the proposed estimators will be established. The idea behind the proposed model is similar to that discussed in Sun et al. (2005, 2007) for the analysis of longitudinal data with dependent observation processes, and the proposed method can also apply for joint analysis of interval-censored data and panel count data with the focus on the failure event (Xu et al., 2018). An example of the latter situations is given by the analysis of a disease onset and the hospitalization record of a patient with the disease onset being the focus. For the case, it is apparent that the record or the hospitalizations process may contain relevant information about the disease onset and thus a joint analysis needs to be conducted. More discussion on this will be given below.

In the following, we will first present the proposed conditional PH model along with some notation and assumptions that will be used throughout the paper in Section 2. Then in Section 3, we will propose a penalized sieve maximum likelihood procedure and establish the oracle property of the resulting estimators. For the implementation of the proposed method, a cyclic coordinate-wise optimization algorithm will be developed in Section 4. In Section 5, we will conduct a simulation study to assess the performance of the proposed approach and it indicates that the method works well for practical situations. In Section 6, we will apply the proposed methodology to the AD data discussed above and give some concluding remarks in Section 7.

2. A conditional PH model

Consider a failure time study that involves n independent subjects with T_i denoting the failure time of interest. Suppose that for subject i, there exists a p -dimensional vector of covariates denoted by Z_i and the subject is observed only at a sequence of time points denoted by $s_{i,1} < s_{i,2} < \cdots < s_{i,m_i}$, where m_i denotes the number of observations on subject i. That is, on the T_i's, only case K interval-censored data are available (Sun, 2006) and given by $O = \{O_i = \left\{m_i, Z_i, s_{i,j}, \delta_{ik} = I\left(T_i \in \left(s_{i,k-1}, s_{i,k}\right]\right); j = 1, 2, \ldots, m_i, k = 1, 2, \ldots, m_i + 1\right\}; i = 1, ..., n\}$, where $s_{i,0} = 0$ and $s_{i,m_i+1} = \infty$. Also suppose that one is mainly interested in identifying relevant or significant covariates and estimate their effects on the T_i's.

Define $N_i(t) = \sum_{j=1}^{m_i} I(s_{i,j} \le t)$, the observation process, and assume that T_i and $N_i(t)$ are correlated. That is, we have informative interval censoring. To describe the covariate effects on T_i as well as the possible correlation between T_i and $N_i(t)$, define $\mathcal{F}_{it} = \left\{N_i(s), 0 \le s \le t\right\}$, the history about the process N_i up to time t. In the following, we will assume that given Z_i and $\mathcal{F}_{it}$, the hazard function of T_i has the form

$$\lambda\left(t \mid Z_i, \mathcal{F}_{it}\right) = \lambda_0(t) \exp\left\{\boldsymbol{\beta}^{*T} Z_i + \alpha^T H\left(\mathcal{F}_{it}\right)\right\} = \lambda_0(t) \exp\left\{\boldsymbol{\beta}^T X_i(t)\right\}. \tag{1}$$

In the above, $\lambda_0(t)$ denotes an unknown baseline hazard function, $\boldsymbol{\beta}^* = (\beta_1, ..., \beta_p)^T$ is a p-dimensional vector of regression parameters, α is a p_1-dimensional vector of regression coefficients, $\boldsymbol{\beta} = (\alpha^T, \beta_1, \dots, \beta_p)^T$, $H(\cdot)$ is a vector of known functions of the counting process $N_i(t)$ up to time t, and $X_i(t) = (H\left(\mathcal{F}_{it}\right)^T, Z_i^T)^T$.

Model (1) is motivated by similar models on the longitudinal process discussed in Sun et al. (2005, 2007) for regression analysis of longitudinal data with a dependent observation process. It specifies that the failure time T_i may be related to the censoring or observation process $N_i(t)$ through the function H, which can be chosen in various forms. A natural and simple choice for H may be $H\left(\mathcal{F}_{it}\right) = N_i(t-)$, meaning that all information about T_i in $\mathcal{F}_{it}$ is given by the total number of observations. An alternative is that T_i depends on $\mathcal{F}_{it}$ only through a recent number of observations, say, in u time units, and this corresponds to $H\left(\mathcal{F}_{it}\right) = N_i(t-) - N_i(t-u)$. One could define H as a vector given by the foregoing two choices if both the total and recent numbers of observations may contain information about T_i. In the following, we will leave the distribution of $N_i(t)$ arbitrary but assume that it does not contain any unknown parameters in model (1).

3. Simultaneous variable selection and estimation

Now we consider the variable or covariate selection and for this, we will develop a penalized procedure with the use of the likelihood function as the objective function. Under the assumptions above, it is easy to see that the likelihood function of the observation data is proportional to

$$L_n\left(\alpha, \beta^*, \lambda_0\right) = \prod_{i=1}^{n}\prod_{k=1}^{m_i+1}\left[\exp\left\{-\int_0^{s_{i,k-1}} \lambda_0(t)\exp\left\{\boldsymbol{\beta}^{*T}Z_i + \boldsymbol{\alpha}^T H\left(\mathcal{F}_{it}\right)\right\}dt\right\} - \exp\left\{-\int_0^{s_{i,k}} \lambda_0(t)\exp\left\{\boldsymbol{\beta}^{*T}Z_i + \boldsymbol{\alpha}^T H\left(\mathcal{F}_{it}\right)\right\}dt\right\}\right]^{\delta_{ik}}. \tag{2}$$

If the estimation of unknown parameters was of main interest, it would be natural to maximize the likelihood function above. On the other hand, it is easy to see that the maximization would be difficult since it involves the unknown function $\lambda_0(t)$. To deal with this, we propose first to approximate $\lambda_0(t)$ by using B-splines functions.

Let τ denote the longest follow-up time and for the closed interval $[0, \tau]$, let $\mathcal{I} = \left\{t_i\right\}_1^{m_n+2l}$ with

$$0 = t_1 = \cdots = t_l < t_{l+1} < \cdots < t_{m_n+l} < t_{m_n+l+1} = \cdots = t_{m_n+2l} = \tau$$

being a sequence of knots that partition $[0, \tau]$ into $m_n + 1$ subintervals and $m_n = O(n^v)$, for $0 < v < 1/2$. Also let $\Psi_{l,\mathcal{I}}$ be the class linearly spanned by the B-splines basis functions $\left\{B_j, 1 \le j \le q_n\left(q_n = m_n + l\right)\right\}$ with order l and knots $\mathcal{I}$. That is,

$$\Psi_{l,\mathcal{I}} = \left\{\sum_{j=1}^{q_n}\gamma_j B_j : \gamma_j \in \mathbb{R}, j = 1, \cdots, q_n\right\}.$$

We now define a subclass of $\Psi_{l,\mathcal{I}}$ as $\Phi_{l,\mathcal{I}} = \left\{\sum_{j=1}^{q_n}\gamma_j B_j\right\}$ with $\gamma_j \ge 0$. According to the variation-diminishing properties of B-splines (Schumaker, 1981), $\Phi_{l,\mathcal{I}}$ is a class of nondecreasing splines on $[0, \tau]$. Then we can approximate the smooth nonnegative function $\lambda_0(t)$ by $\sum_{j=1}^{q_n}\gamma_j B_j(t)$ with the constraints $\gamma_j \ge 0$, $j = 1, \cdots, q_n$. It follows that for estimation of $(\alpha, \beta^*, \lambda_0)$, it would be natural to consider the log-likelihood function

$$l_n\left(\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\gamma}\right) = \sum_{i=1}^{n}\sum_{k=1}^{m_i+1}\delta_{ik}\log\left[\exp\left\{-\int_0^{s_{i,k-1}} \boldsymbol{\gamma}^T B(t)\exp\left\{\boldsymbol{\beta}^{*T}Z_i + \boldsymbol{\alpha}^T H\left(\mathcal{F}_{it}\right)\right\}dt\right\} - \exp\left\{-\int_0^{s_{i,k}} \boldsymbol{\gamma}^T B(t)\exp\left\{\boldsymbol{\beta}^{*T}Z_i + \boldsymbol{\alpha}^T H\left(\mathcal{F}_{it}\right)\right\}dt\right\}\right],$$

where $\boldsymbol{\gamma} = \left(\gamma_1, \cdots, \gamma_{q_n}\right)^T$ and $B(t) = \left(B_1(t), \cdots, B_{q_n}(t)\right)^T$. Let $\left(U_i, V_i\right]$ denote the smallest interval that brackets T_i. Then the likelihood function above can be rewritten as

$$l_n\left(\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\gamma}\right) = \sum_{i=1}^{n}\log\left[\exp\left\{-\int_0^{U_i} \boldsymbol{\gamma}^T B(t)\exp\left\{\boldsymbol{\beta}^{*T}Z_i + \boldsymbol{\alpha}^T H\left(\mathcal{F}_{it}\right)\right\}dt\right\} - \exp\left\{-\int_0^{V_i} \boldsymbol{\gamma}^T B(t)\exp\left\{\boldsymbol{\beta}^{*T}Z_i + \boldsymbol{\alpha}^T H\left(\mathcal{F}_{it}\right)\right\}dt\right\}\right].$$

For simultaneous variable selection and estimation, we propose to maximize the penalized log likelihood function

$$\ell_p(\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\gamma}) = l_n(\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\gamma} \mid X_i) - \sum_{j=1}^{p} P_\lambda\left(\left|\beta_j\right|\right),$$

where $P_\lambda\left(\left|\beta_j\right|\right)$ denotes a penalty function characterized by the tuning parameter λ. In the following, we will focus on the BAR penalty function (Dai et al., 2018; Zhao et al., 2020) although the proposed method is valid with the use of other penalty functions too. Some comments on this will be given below. Let $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\gamma}}$ denote the BAR estimators of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}$ given by the maximization above. In the following, we will establish the oracle property of $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^{*T})^T$.

Let $\boldsymbol{\beta}_0 = (\alpha_0^T, \beta_{0,1}, \ldots, \beta_{0,p})^T$ denote the true value of $\boldsymbol{\beta}$. Without loss of generality, assume that we can write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, \boldsymbol{\beta}_{02}^T)^T$, where $\boldsymbol{\beta}_{01}$ is a $p_1 + q$ vector consisting of α_0 and all q ($q \ll p$) nonzero components and $\boldsymbol{\beta}_{02}$ the remaining zero components. Correspondingly, we will divide the BAR estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ in the same way. Also for a vector of $\boldsymbol{\theta}_1$ and given $\boldsymbol{\beta}_1$, define $Q_{n1}(\boldsymbol{\theta}_1) = Q_{n1}(\boldsymbol{\theta}_1|\boldsymbol{\beta}_1) = l_{n1}(\boldsymbol{\theta}_1) - \lambda_n \boldsymbol{\theta}_1^T D_1(\boldsymbol{\beta}_1)\boldsymbol{\theta}_1$, where $l_{n1}(\boldsymbol{\theta}_1) = l_n(\boldsymbol{\theta}_1, 0|X)$ and $D_1(\boldsymbol{\beta}_1) = \mathrm{diag}\{\mathbf{0}_{p_1\times 1}, \beta_1^{-2}, \ldots, \beta_q^{-2}\}$. In the following, we assume that $p < n$ but p and q can diverge or increase with the sample size n. For the oracle property, we need the following regularity conditions.

(C1). The maximum spacing of the knots satisfies $\Delta = \max_{l+1\le j\le m_n+l+1}\left|t_j - t_{j-1}\right| = O(n^{-\nu})$.

(C2). (i) The parameter space of $\left(\alpha^T, \beta^{*T}\right)^T$, $\mathcal{R}$, is bounded and convex on $\mathbb{R}^{p_1+p}$, and the true parameter $\left(\alpha_0, \beta_0^*\right) \in \mathcal{R}^\circ$, where $\mathcal{R}^\circ$ is the interior of $\mathcal{R}$. There exists a constant $C_0 > 0$ such that $\lambda_0(t) \ge C_0$ for $t \in [0, \tau]$. In addition, the true failure rate λ_0 is differentiable up to order r and all derivatives are bounded in $[0, \tau]$, where $r \ge 1$. (ii) $N_i(\tau)$ ($i = 1, 2, ..., n$) are bounded by a constant. There exists $Z_0 > 0$ such that $P(\|Z\| \le Z_0) = 1$ and $E(ZZ^T)$ is nonsingular. (iii) For some $\eta \in (0, 1)$, we have that $a^T \mathrm{Var}(X \mid U)a \ge \eta a^T E\left(XX^T \mid U\right)a$ and $a^T \mathrm{Var}(X \mid V)a \ge \eta a^T E\left(XX^T \mid V\right)a$ almost surely for all $a \in \mathcal{R}^{p_1+p}$.

(C3). There exists a compact neighborhood $\mathcal{B}_0$ of the true value $\boldsymbol{\beta}_0$ and a positive definite $(p + p_1) \times (p + p_1)$ matrix $I(\boldsymbol{\beta}_0)$ such that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}_0} \| - n^{-1}\ddot{l}_n(\boldsymbol{\beta}) - I(\boldsymbol{\beta}_0)\| \xrightarrow{a.s.} 0,$$

where $\ddot{l}_n(\boldsymbol{\beta})$ denotes the second derivative of $l_n(\boldsymbol{\beta}|X)$.

(C4). There exists some constant $C > 1$ such that

$$C^{-1} < \lambda_{min}(-n^{-1}\ddot{l}_n(\boldsymbol{\beta})) \le \lambda_{max}(-n^{-1}\ddot{l}_n(\boldsymbol{\beta})) < C$$

for sufficiently large n, where $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ denote the smallest and largest eigenvalues of the matrix, respectively.

(C5). There exist positive constants a_0 and a_1 such that $a_0 \le |\beta_{0,j}| \le a_1$, $1 \le j \le q$.

(C6). As $n \to \infty$, we have that $p^2 q/\sqrt{n} \to 0$, $\lambda_n/\sqrt{n} \to 0$, $\xi_n/\sqrt{n} \to 0$, $\lambda_n\sqrt{q/n} \to 0$, and $\lambda_n^2/(p\sqrt{n}) \to \infty$.

Theorem. *Assume that the regularity conditions (C1) - (C6) described above hold. Then as $n \to \infty$ and with probability tending to 1, the BAR estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ exists and has the following properties:*

(i) $\hat{\boldsymbol{\beta}}_2 = 0$.

(ii) $\hat{\boldsymbol{\beta}}_1$ is the unique fixed point of $f(\boldsymbol{\beta}_1)$, where $f(\boldsymbol{\beta}_1)$ is a solution to $\dot{Q}_{n1}(\boldsymbol{\theta}_1) = 0$ with $\dot{Q}_{n1}(\boldsymbol{\theta}_1)$ denoting the first derivative of $Q_{n1}(\boldsymbol{\theta}_1)$.

(iii) $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})$ converges in distribution to the multivariate normal distribution $N_{p_1+q}(0, \Sigma_1(\boldsymbol{\beta}_0)^{-1})$, where $\Sigma_1(\boldsymbol{\beta}_0)$ is given in the Appendix.

The proof of the results above is sketched in the Appendix. For the determination of the proposed estimator, we will present a cyclic coordinate-wise optimization algorithm in the next section.

4. Cyclic coordinate-wise optimization algorithm

First, we will consider the determination of $\hat{\boldsymbol{\beta}}^*$ and for this, we will take turn to update each element β_i of $\boldsymbol{\beta}^*$ while keeping all other elements of $\boldsymbol{\beta}^*$ as well as $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ fixed at their current estimates. More specifically, define

$$g\left(\beta_j\right) = \sum_{i=1}^{n} \log\left[\exp\left\{-\int_0^{U_i} \hat{\boldsymbol{\gamma}}^T B(t) \exp\left\{\sum_{l\ne j} \hat{\beta}_l^T Z_{il} + \beta_j Z_{ij} + \hat{\boldsymbol{\alpha}}^T H\left(\mathcal{F}_{it}\right)\right\} dt\right\} - \exp\left\{-\int_0^{V_i} \hat{\boldsymbol{\gamma}}^T B(t) \exp\left\{\sum_{l\ne j} \hat{\beta}_l^T Z_{il} + \beta_j Z_{ij} + \hat{\boldsymbol{\alpha}}^T H\left(\mathcal{F}_{it}\right)\right\} dt\right\}\right].$$

Then at the kth iteration, we need to determine $\hat{\beta}_j^{(k)}$, the value of β_j that maximizes $h\left(\beta_j\right) = g\left(\beta_j\right) - P_\lambda\left(\left|\beta_j\right|\right)$. Note that by borrowing the LQA idea discussed in Fan and Li (2001), $g\left(\beta_j\right)$ can be approximated by the second-order Taylor expansion

$$g\left(\beta_j\right) \approx g\left(\hat{\beta}_j^{(k-1)}\right) + g'\left(\hat{\beta}_j^{(k-1)}\right)\left(\beta_j - \hat{\beta}_j^{(k-1)}\right) + \frac{1}{2} g''\left(\hat{\beta}_j^{(k-1)}\right)\left(\beta_j - \hat{\beta}_j^{(k-1)}\right)^2,$$

where g' and g'' denote the first and second derivatives of g, respectively. In consequence, we can obtain the close form iterative solution as

$$\hat{\beta}_j^{(k)} = \hat{\beta}_j^{(k-1)} - \frac{h'\left(\hat{\beta}_j^{(k-1)}\right)}{h''\left(\hat{\beta}_j^{(k-1)}\right)},$$

where $h'\left(\hat{\beta}_j^{(k-1)}\right)$ and $h''\left(\hat{\beta}_j^{(k-1)}\right)$ are the first and second derivatives of $h\left(\beta_j\right) = g\left(\beta_j\right) - \lambda\beta_j^2 / \left(\hat{\beta}_j^{(k-1)}\right)^2$ with respect to β_j evaluated at $\hat{\beta}_j^{(k-1)}$, respectively.

Note that our experience indicates that in the iteration above for each element of $\boldsymbol{\beta}^*$, one only needs to update the estimate once. This is because the algorithm will update the estimates of $\boldsymbol{\beta}^*$, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}$ alternately and there is little reason to find the estimates of $\boldsymbol{\beta}^*$ with a high precision in one iteration based on the current estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. For the determination of the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ in the iteration, we suggest to employ the quasi-Newton algorithm since it is more convenient for the situation. The following gives the summary of the algorithm discussed above.

Step 1: Set $k = 0$ and choose the initial estimates $\hat{\boldsymbol{\gamma}}^{(0)}$, $\hat{\boldsymbol{\alpha}}^{(0)}$, and $\hat{\boldsymbol{\beta}}^{*(0)}$.

Step 2: At the kth iteration, obtain $\hat{\boldsymbol{\alpha}}^{(k)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$ by using the R function optim with $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}^{*(k-1)}$.

Step 3: With $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}^{(k)}$ and $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{(k)}$, use the coordinate descent algorithm to determine

$$\hat{\boldsymbol{\beta}}^{*(k)} = \underset{\boldsymbol{\beta}^*}{\operatorname{argmax}} \left\{ l_n\left(\hat{\boldsymbol{\alpha}}^{(k)}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}}^{(k)}\right) - \sum_{j=1}^{p} P_\lambda\left(\left|\beta_j\right|\right) \right\}.$$

Step 4: Repeat Steps 2 to 3 until the convergence or k exceeding a given large number.

Note that for the better performance of the algorithm above, as with most algorithms, it is important to choose good initial estimates. For this, we suggest to use the ridge estimate or the estimate with the ridge penalty given by

$$\hat{\boldsymbol{\beta}}^{*(0)} = \hat{\boldsymbol{\beta}}^*_{\text{Ridge}} = \underset{\boldsymbol{\beta}^*}{\operatorname{argmax}} \left\{ l_n(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \hat{\boldsymbol{\gamma}}) - \xi \sum_{j=1}^{p} \beta_j^2 \right\}$$

with the application of the algorithm above, where ξ is another tuning parameter to be discussed below. To check the convergence in Step 4 above, one may apply various criteria. In the numerical studies below, we used the mean absolute difference between the consecutive estimates of all parameters defined as $N^{-1}\left\|\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k-1)}\right\|_1 = N^{-1}\sum_{l=1}^{N} \mid \hat{\theta}_l^{(k)} - \hat{\theta}_l^{(k-1)} \mid < \epsilon$ with setting $\epsilon = 10^{-4}$. Here $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^{*T}, \boldsymbol{\gamma}^T\right)^T$, N denotes the dimension of $\boldsymbol{\theta}$, and $\hat{\theta}_l^{(k)}$ represents the lth component of $\hat{\boldsymbol{\theta}}^{(k)}$.

To implement the algorithm above, also one needs to choose both tuning parameters ξ and λ and for this, the simulation study below suggests that the estimation results seem to be robust with ξ and one only needs to choose λ. For this, we propose to employ or minimize the BIC, which is data-dependent and defined as

$$\text{BIC}_\lambda = -2l_n(\hat{\boldsymbol{\theta}}) + df_\lambda \cdot \log(n).$$

In the above, $\hat{\boldsymbol{\theta}}$ is the final estimator of $\boldsymbol{\theta}$, $l_n(\hat{\boldsymbol{\theta}})$ denotes the logarithm of the observed data likelihood function, and df_λ represents the total number of nonzero estimates in $\hat{\boldsymbol{\theta}}$ in the ultimate model, which serves as the degrees of freedom. The numerical results in the simulation study below suggest that BIC works well in practical situations. Of course, one could employ other methods such as Akaike information criterion or the cross-validation.

5. A simulation study

In this section, we present some results obtained from a simulation study conducted to assess the performance of the penalized variable selection procedure proposed in the previous sections. To generate the simulated data, the covariate vector Z was first generated from the multivariate normal distribution with mean zero, variance one, and the correlation between Z_j and Z_k being $\rho^{|j-k|}$ with $\rho = 0.5$, $j, k = 1, ..., p$. Then the total numbers of observation times m_i's and the observation times were generated. For the former, m_i was assumed to follow the uniform distribution over $\{1, 2, 3, 4, 5, 6\}$ and for the latter, given m_i, the observation times $s_{i,j}$'s were taken to be the order statistics of the m_i random variables from the uniform distribution over $(0.02, 1)$. Given the Z_i's and $s_{i,j}$'s or $N_i(t)$'s, the true failure times T_i s were generated under model (1) with $\lambda_0(t) = 1$ or $\lambda_0(t) = 1/(t+1)$ and $H\left(\mathcal{F}_{it}\right) = N_i(t-)$.

To assess the performance of the proposed method, we calculated three evaluation metrics, the mean weighted squared error (MSE), the true positive rate (TPR) and the false positive rate (FPR). Here MSE was defined to be $\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*\right)^T E\left(ZZ'\right)\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*\right)$, and TPR and FPR were defined as

$$TPR = \frac{TP}{R} \text{ and } FPR = \frac{FP}{TO - R}.$$

In the above, TP, true positive, denotes the number of the variables with non-zero coefficients that were correctly identified by the approach, FP, false positive, the number of the variables with zero coefficients that were incorrectly identified, R is the number of non-zero predictors, and TO is the total number of the predictors in the simulated data. The results given below are based on $n = 200$ or 400 and $p = 8$ or 30 with 100 replications.

Table 1
Simulation results for informatively interval-censored data.

	Sample Size	MMSE(SD)	TPR	FPR
		$p = 8$		
$\Lambda_0(t) = t$	$n = 200$	0.077(0.189)	1	0.014
	$n = 400$	0.033(0.059)	1	0.012
$\Lambda_0(t) = \log(t+1)$	$n = 200$	0.080(0.143)	1	0.014
	$n = 400$	0.033(0.054)	1	0.014
		$p = 30$		
$\Lambda_0(t) = t$	$n = 200$	0.261(0.483)	1	0.017
	$n = 400$	0.060(0.073)	1	0.010
$\Lambda_0(t) = \log(t+1)$	$n = 200$	0.289(0.450)	1	0.018
	$n = 400$	0.076(0.076)	1	0.010

Table 2
Simulation results for joint analysis with $N_i(t)$ being Poisson process.

	Sample Size	MMSE(SD)	TPR	FPR
		$p = 8$		
$\Lambda_0(t) = t$	$n = 200$	0.055(0.111)	1	0.018
	$n = 400$	0.027(0.041)	1	0.014
$\Lambda_0(t) = \log(t+1)$	$n = 200$	0.055(0.138)	1	0.018
	$n = 400$	0.026(0.034)	1	0.006
		$p = 30$		
$\Lambda_0(t) = t$	$n = 200$	0.132(0.352)	1	0.016
	$n = 400$	0.061(0.087)	1	0.015
$\Lambda_0(t) = \log(t+1)$	$n = 200$	0.158(0.493)	1	0.018
	$n = 400$	0.068(0.072)	1	0.012

Table 1 presents the results on the covariate selection given by the approach proposed in the previous sections. Here we set $\beta^{*T} = (1, 1, 0, 0, 0, 0, 0, 1)$ or $(1, 1, \mathbf{0}_{26}, 1, 1)$, and $\alpha = 1$. Also we used the cubic B-splines and took $m_n = n^{\upsilon}$ with $\upsilon = 1/4$. For a given number of the interior knots m_n, the equally spaced knots were chosen. For the selection of the tuning parameter λ_n, the BIC criterion based on the grid search was used, and for the tuning parameter ξ, we set $\xi = 100$ since as mentioned above, the results are not sensitive to the choice of ξ. One can see from Table 1 that the proposed procedure seems to work well for the situations considered here. As expected, the performance got better when the sample size increased or for smaller p.

As discussed above, the proposed method also applies to joint analysis of interval-censored data and panel count data with $N_i(t)$ representing a recurrent event process that may be correlated to the failure process of interest and on which only panel count data are available (Xu et al., 2018). To evaluate the performance of the proposed approach for the situation, for the $N_i(t)$'s, we generated the panel count data from the Poisson process or mixed Poisson process with the mean function $\Lambda_1(t)\exp(\zeta Z_i)$ and set

$$N_i(s_{i,j}) = N_i(s_{i,1}) + \{N_i(s_{i,2}) - N_i(s_{i,1})\} + \ldots + \{N_i(s_{i,j}) - N_i(s_{i,j-1})\}.$$

In the above, it was assumed that

$$N_i(s_{i,1}) \sim \text{Poisson}\left\{\Lambda_1(s_{i,1})\exp(\zeta Z_i)\right\},$$
$$N_i(s_{i,j}) - N_i(s_{i,j-1}) \sim \text{Poisson}\left\{\left[\Lambda_1(s_{ij}) - \Lambda_1(s_{i,j-1})\right]\exp(\zeta Z_i)\right\},$$

or

$$N_i(s_{i,1}) \sim \text{Poisson}\left\{\eta\Lambda_1(s_{i,1})\exp(\zeta Z_i)\right\},$$
$$N_i(s_{i,j}) - N_i(s_{i,j-1}) \sim \text{Poisson}\left\{\eta\left[\Lambda_1(s_{ij}) - \Lambda_1(s_{i,j-1})\right]\exp(\zeta Z_i)\right\}$$

for $j = 2, \ldots, m_i, i = 1, \ldots, n$. Here ζ is a vector of regression parameters as β^*, η follows the gamma distribution with mean one and variance 0.25, and $\Lambda_1(t) = t$. The other information was generated in the same way as above.

Tables 2 and 3 give the results on the covariate selection given by the proposed approach with $\zeta^T = (0.2, 0.2, 0, 0, 0, 0, 0, 0.2)$ or $(0.2, 0.2, \mathbf{0}_{26}, 0.2, 0.2)$ and the $N_i(t)$'s being the Poisson process or the mixed Poisson processes, respectively. It is apparent that they are similar to those given in Table 1 and indicate that the proposed method works well. To further see the performance of the proposed approach in terms of estimating the baseline cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$, Figs. 1 and 2 present the averages of the estimates given by the proposed method corresponding to the situation considered in Table 3 with $n = 400$,

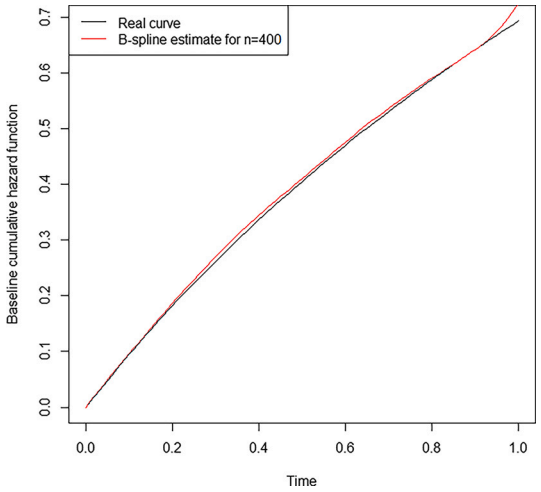

Fig. 1. Estimates of $\Lambda_0(t) = \log(t+1)$ for the mixed Poisson process with p = 8. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

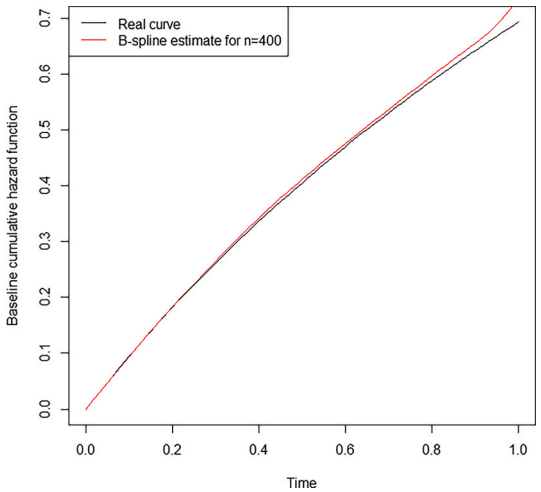

Fig. 2. Estimates of $\Lambda_0(t) = \log(t+1)$ for the mixed Poisson process with p = 30.

$\Lambda_0(t) = \log(t+1)$, and $p = 8$ or $p = 30$, respectively. For comparison, the true curve is included in the figures too. They suggest that the proposed method seems to perform well. We also considered other situations or set-ups and obtained similar results.

6. Analysis of the Alzheimer's disease data

Now we apply the methodology proposed in the previous sections to the data arising from the Alzheimer's Disease Neuroimaging Initiative (ADNI) described above. As mentioned before, in the study, the information on many biomarkers were collected in order to identify these that can be used for the early detection and tracking of the Alzheimer's disease (AD), and the participants were examined periodically for their AD status. Based on their cognitive conditions, the participants were initially grouped into three groups, cognitively normal, mild cognitive impairment (MCI) and Alzheimer's disease. Among others, one variable of interest is the time from the baseline visit date to the AD conversion, and as expected, many patients dropped out of the study early and some missed their scheduled visits. Thus the participants have different observation times and only interval-censored data are available on the AD conversion time.

For the analysis here, we will define T_i and N_i to represent the AD conversion time and the observation process, respectively. By following Li et al. (2017a,2017b) and others, we will focus on the 319 participants in the MCI group for whom the information on 24 covariates is complete in order to identify important prognostic factors for the AD conversion. These 24 demographic and clinical covariates were identified as possible important factors associated with the AD conversion by Li et al. (2017a,2017b), who considered a similar problem by performing a simple or individual analysis. In addition to the information on the covariates, the observed data for each participant include the number of observations as well as the observation times and the event indicators.

Table 3
Simulation results for joint analysis with $N_i(t)$ being mixed-Poisson process.

	Sample Size	MMSE(SD)	TPR	FPR
		$p = 8$		
$\Lambda_0(t) = t$	$n = 200$	0.054(0.137)	1	0.022
	$n = 400$	0.026(0.038)	1	0.006
$\Lambda_0(t) = \log(t+1)$	$n = 200$	0.056(0.145)	1	0.018
	$n = 400$	0.026(0.042)	1	0.014
		$p = 30$		
$\Lambda_0(t) = t$	$n = 200$	0.146(0.299)	1	0.016
	$n = 400$	0.062(0.077)	1	0.010
$\Lambda_0(t) = \log(t+1)$	$n = 200$	0.125(0.344)	1	0.016
	$n = 400$	0.064(0.095)	1	0.013

Table 4
Analysis results for the ADNI with $H(t) = N(t-)$.

	Proposed Method		Du et al. (2021)		Li et al. (2020)	
Factor	Estimate	SSE	Estimate	SSE	Estimate	SSE
PTEDUCAT	-	-	0.077	0.110	-	-
PTMARRY	-	-	-	-	-	-
CDRSB	-	-	0.211	0.177	-	-
ADAS11	-	-	-	-	-	-
ADAS13	-	-	0.276	0.173	0.245	0.220
ADASQ4	-	-	-	-	-	-
MMSE	-	-	-0.126	0.124	-	-
RAVLT.i	-0.597	0.156	-0.577	0.212	-0.582	0.220
RAVLT.l	-	-	0.231	0.269	-	-
RAVLT.f	-	-	-	-	-	-
RAVLT.perc.f	-	-	0.142	0.328	-	-
DIGITSCOR	-	-	-	-	-	-
TRABSCOR	-	-	-	-	-	-
FAQ	0.353	0.159	0.358	0.198	0.297	0.176
Ventricles	-	-	-	-	-	-
Hippocampus	-	-	-0.070	0.182	-	-
WholeBrain	-	-	-	-	-	-
Entorhinal	-0.433	0.227	-0.306	0.165	-0.184	0.218
Fusiform	-	-	-	-	-	-
MidTemp	-0.328	0.179	-0.652	0.228	-0.434	0.219
ICV	-	-	0.311	0.254	-	-
Age	-0.265	0.182	-0.322	0.152	-0.208	0.199
APOEϵ4	-	-	0.471	0.170	0.137	0.151
Gender	-	-	-	-	-	-

Furthermore, for the analysis, as in the simulation study, we will employ the BIC to select the optimal tuning parameter λ while setting $\xi = 100$.

The factor selection and estimation results are given in Table 4 with the use of $H\left(\mathcal{F}_{it}\right) = N_i(t-)$. Here as in the simulation study, we set the equally spaced knots for the given number of the interior knots m_n. We tried other ways for the knot selection and obtained similar results. For each selected factor, in addition to the estimated factor effect, we also provide the estimated standard error obtained by using the bootstrap procedure with 100 bootstrap samples randomly drawn with replacement from the observed data. One can see from Table 4 that five factors, RAVLT.i, FAQ, Entorhinal, MidTemp, and Age, were selected. Among them, RAVLT.i and FAQ seem to have significant effects on the AD conversion, while Entorhinal and MidTemp also appear to be marginally correlated with the AD conversion.

For comparison, we also include in Table 4 the results given by Li et al. (2020) and Du et al. (2021), which selected 7 and 14 factors, respectively. Note that the former assumes non-informative interval censoring, while the latter employs some latent variables to model informative interval censoring. It is clear that all of the three methods gave similar results for the selected factors and most of the extra factors selected by the two other methods did not seem to have much effects on the AD conversion. In other words, these results indicate that the censoring indeed seems to be informative and the two other methods tend to over-select factors. We also considered other choices for $H\left(\mathcal{F}_{it}\right)$ and obtained similar results.

7. Concluding remarks

This paper considered regression analysis of informatively interval-censored failure time data. Corresponding to the existing joint modeling approaches, a conditional approach was proposed. More specifically, a conditional PH model was presented and a penalized sieve maximum likelihood estimation approach was proposed. For the implementation of the proposed method, a cyclic coordinate-wise optimization algorithm was developed and the oracle property of the resulting estimators was established. As discussed above, the proposed procedure also applies to joint analysis of interval-censored data and panel count data, and one advantage of the proposed method is that it does not impose any assumption on the censoring or observation processes. The numerical results indicated that the procedure works well for practical situations.

Note that as mentioned above, unlike the existing methods that allow for informative interval censoring, the proposed method took a conditional approach that avoids the use of frailty variables and their related distribution assumption as well as the commonly used Poisson process assumption for the situation. Also unlike the existing methods for variable selection based on interval-censored data such as that given in Zhao et al. (2020), the proposed approach allows for informative censoring. In other words, the proposed method is more general and robust, and furthermore, it allows one to directly estimate the degree of the dependent censoring.

In the proposed method, we have focused on the use of the BAR penalty function and as mentioned above, the method is still valid with other penalty functions replacing the BAR penalty function. However, some minor modifications to the algorithm may be needed depending on the penalty function and also the proof of the oracle property could be different. Also in the proposed method, B-spline functions have been used to approximate the baseline hazard function. As an alternative, one could apply other spline functions and develop similar variable selection procedures.

Acknowledgement

We wish to thank two reviewers for their helpful and useful comments and suggestions that greatly improved the paper. This research was supported in part by the National Natural Science Foundation of China grant (12101522) to the first author and the National Natural Science Foundation of China grant (12271459) and The Hong Kong Polytechnic University grant to the second author.

Appendix A. Proof of the oracle property

In this appendix, we will sketch the proof of the oracle property described in the theorem and the main idea behind the proof is similar to that of Zhao et al. (2020). To prove the theorem, we need the following three lemmas. For simplicity, we assume that $\boldsymbol{\alpha}$ is a one-dimensional variable coefficient.

Lemma 1 *(Consistency of the ridge estimator). Let $\boldsymbol{\beta}_{ridge}$ denote the ridge estimator defined as $\boldsymbol{\beta}_{ridge} = \arg\max_{\boldsymbol{\beta}}\{l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}|X) - \xi_n \sum_{j=1}^{p} \beta_j^2\}$, and suppose that the conditions (C1) - (C6) hold. Then we have that $\|\boldsymbol{\beta}_{ridge} - \boldsymbol{\beta}_0\| = O_p(\sqrt{p/n})$.*

Proof. Denote $\mathcal{L}(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}|X) - \xi_n \sum_{j=1}^{p} \beta_j^2 = l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}|X) - n\sum_{j=1}^{p} p_{\xi_n}(\beta_{0j})$, $a_n = \max_{1\le j\le q}\{|\dot{p}_{\xi_n}(\beta_{0j})| : \beta_{0j} \neq 0\}$, and $b_n = \max_{1\le j\le q}\{|\ddot{p}_{\xi_n}(\beta_{0j})| : \beta_{0j} \neq 0\}$. For ridge regression we can see that $p_{\xi_n}(\beta_{0j}) = \beta_{0j}^2 \xi_n/n$ for $j = 1, \ldots, p$. Thus the first and second derivatives of $p_{\xi_n}(\beta_{0j})$ are $\dot{p}_{\xi_n}(\beta_{0j}) = 2\beta_{0j}\xi_n/n$ and $\ddot{p}_{\xi_n}(\beta_{0j}) = 2\xi_n/n$ respectively. From Conditions (C5) and (C6) we have that $a_n \le 2a_1\xi_n/n = o(n^{-1/2})$ and $b_n \le 2\xi_n/n = o(n^{-1/2})$. Therefore, $a_n \to 0$, and $b_n \to 0$.

Let $\alpha_n = \sqrt{p}(n^{-1/2} + a_n)$, then using the similar manipulation as those in Cai et al. (2005), we can prove that, for any given $\epsilon > 0$, there exists a large constant C_0 such that

$$P\{\sup_{\|\boldsymbol{\nu}\|=C_0} \mathcal{L}(\boldsymbol{\beta}_0 + \alpha_n\boldsymbol{\nu}) < \mathcal{L}(\boldsymbol{\beta}_0)\} \ge 1 - \epsilon,$$

which implies that there exists a local maximiser, $\boldsymbol{\beta}_{ridge}$, such that $\|\boldsymbol{\beta}_{ridge} - \boldsymbol{\beta}_0\| = O_p(\alpha_n) = O_p(\sqrt{p/n})$.

To describe Lemma 2, for a vector of $\boldsymbol{\theta}$ and given $\boldsymbol{\beta}$, define $Q_n(\boldsymbol{\theta}) \equiv Q_n(\boldsymbol{\theta}; \boldsymbol{\beta}, X) = l_n(\boldsymbol{\theta}|X) - \lambda_n\boldsymbol{\theta}^T D(\boldsymbol{\beta})\boldsymbol{\theta}$, where $D(\boldsymbol{\beta}) = \mathrm{diag}\{0, \beta_1^{-2}, \ldots, \beta_p^{-2}\}$. Then the first and second derivatives of $Q_n(\boldsymbol{\theta})$ are $\dot{Q}_n(\boldsymbol{\theta}) = \dot{l}_n(\boldsymbol{\theta}|X) - 2\lambda_n D(\boldsymbol{\beta})\boldsymbol{\theta}$, and $\ddot{Q}_n(\boldsymbol{\theta}) = \ddot{l}_n(\boldsymbol{\theta}|X) - 2\lambda_n D(\boldsymbol{\beta})$.

Lemma 2. *Suppose $g(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta})^T, g_2(\boldsymbol{\beta})^T)^T$ is a solution to $\dot{Q}_n(\boldsymbol{\theta}) = 0$ and let $\{\delta_n\}$ be a sequence of positive real numbers satisfying $\delta_n \to \infty$ and $\delta_n^2 p/\lambda_n \to 0$. Furthermore, define $\mathcal{H}_n \equiv \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T : |\boldsymbol{\beta}_1| = (|\alpha|, |\beta_1|, \ldots, |\beta_q|)^T \in [1/K_0, K_0]^{q+1}, \|\boldsymbol{\beta}_2\| \le \delta_n\sqrt{p/n}\}$, where $K_0 > 1$ is a constant such that $|\boldsymbol{\beta}_{01}| \in [1/K_0, K_0]^{q+1}$. Then under the regularity conditions (C1)−(C6) and with probability tending to 1, we have that*

(i) $\sup_{\boldsymbol{\beta}\in H_n} \dfrac{\|g_2(\boldsymbol{\beta})\|}{\|\boldsymbol{\beta}_2\|} < \dfrac{1}{C_0}$ for some constant $C_0 > 1$;

(ii) $g(\cdot)$ is a mapping from $\mathcal{H}_n$ to itself.

Proof. Taking the first-order Taylor expansion for $\dot{Q}_n(\theta)$ at β_0 in a neighborhood of $g(\beta)$, we have that, $\dot{Q}_n(\beta_0) = \dot{Q}_n(g(\beta)) + \ddot{Q}_n(\beta^*)(\beta_0 - g(\beta))$, where β_0 is the true parameter vector, and β^* lies between β_0 and $g(\beta)$. Then, $\ddot{Q}_n(\beta^*)g(\beta) = -\dot{Q}_n(\beta_0) + \ddot{Q}_n(\beta^*)\beta_0$ since $\dot{Q}_n(g(\beta)) = 0$. Substituting $\dot{Q}_n(\theta)$ and $\ddot{Q}_n(\theta)$ to the above equation, we have

$$\left[\frac{1}{n}\ddot{l}_n(\beta^*|X) - \frac{2\lambda_n}{n}D(\beta)\right]g(\beta) = \frac{1}{n}\ddot{l}_n(\beta^*|X)\beta_0 - \frac{1}{n}\dot{l}_n(\beta_0|X). \tag{A1}$$

Denote $H_n(\beta^*) = -\frac{1}{n}\ddot{l}_n(\beta^*|X)$ and from (C3), $H_n(\beta^*)^{-1}$ exists. Then multiplying both sides of (A1) by $H_n(\beta^*)^{-1}$,

$$g(\beta) - \beta_0 + \frac{2\lambda_n}{n}H_n(\beta^*)^{-1}D(\beta)g(\beta) = \frac{1}{n}H_n(\beta^*)^{-1}\dot{l}_n(\beta_0|X). \tag{A2}$$

Partition $H_n(\beta^*)^{-1}$ and $D(\beta)$ into

$$H_n(\beta^*)^{-1} = \begin{pmatrix} A & B \\ B^T & G \end{pmatrix} \quad \text{and} \quad D(\beta) = \begin{pmatrix} D_1(\beta_1) & 0 \\ 0 & D_2(\beta_2) \end{pmatrix},$$

where A is a $(q+1)\times(q+1)$ matrix, $D_1(\beta_1) = \mathrm{diag}\{0, \beta_1^{-2}, \ldots, \beta_q^{-2}\}$ and $D_2(\beta_2) = \mathrm{diag}\{\beta_{q+1}^{-2}, \ldots, \beta_p^{-2}\}$. Then (A2) can be rewritten as

$$\begin{pmatrix} g_1(\beta) - \beta_{01} \\ g_2(\beta) \end{pmatrix} + \frac{2\lambda_n}{n}\begin{pmatrix} AD_1(\beta_1)g_1(\beta) + BD_2(\beta_2)g_2(\beta) \\ B^T D_1(\beta_1)g_1(\beta) + GD_2(\beta_2)g_2(\beta) \end{pmatrix} = \frac{1}{n}H_n(\beta^*)^{-1}\dot{l}_n(\beta_0|X). \tag{A3}$$

By arguments similar to those in Theorem 1 of Cai et al. (2005), Conditions (C1)−(C5) guarantee that $\|\frac{1}{n}H_n(\beta^*)^{-1}\dot{l}_n(\beta_0|X\| = O_p(\sqrt{p/n})$, therefore,

$$\sup_{\beta\in H_n}\|g_2(\beta) + \frac{2\lambda_n}{n}B^T D_1(\beta_1)g_1(\beta) + \frac{2\lambda_n}{n}GD_2(\beta_2)g_2(\beta)\| = O_p(\sqrt{p/n}). \tag{A4}$$

Note that $|\beta_1| \in [1/K_0, K_0]^{q+1}$, $\|g_1(\beta)\| \le \|g(\beta)\| \le \|\hat{\beta}\| = O_p(\sqrt{p})$, where $\hat{\beta}$ is equal to $g(\beta)$ with $\xi = 0$, and furthermore, from $\|BB^T\| - \|A^2\| \le \|BB^T + A^2\| \le \|H_n(\beta^*)^{-2}\| < C^2$, we can derive $\|B\| \le \sqrt{2}C$ and

$$\sup_{\beta\in H_n}\left\|\frac{2\lambda_n}{n}B^T \boldsymbol{D}_1(\beta_1)g_1(\beta)\right\| \le \frac{2\lambda_n}{n}\sup_{\beta\in H_n}\left\|B^T\right\|\left\|D_1(\beta_1)\right\|\left\|g_1(\beta)\right\| = o_p(\sqrt{p/n}), \tag{A5}$$

then (A4) can be rewritten as $\sup_{\beta\in H_n}\|g_2(\beta) + \frac{2\lambda_n}{n}GD_2(\beta_2)g_2(\beta)\| = O_p(\sqrt{p/n})$. At the same time, $\frac{2\lambda_n}{n}\|GD_2(\beta_2)g_2(\beta)\| \ge \frac{2\lambda_n}{n}\frac{1}{C}\|D_2(\beta_2)g_2(\beta)\|$, and thus

$$\frac{1}{C}\|\frac{2\lambda_n}{n}D_2(\beta_2)g_2(\beta)\| - \|g_2(\beta)\| \le \sup_{\beta\in H_n}\|g_2(\beta) + \frac{2\lambda_n}{n}GD_2(\beta_2)g_2(\beta)\| \le \delta_n(\sqrt{p/n}). \tag{A6}$$

Let $m_{g_2(\beta)/\beta_2} = (g_2(\beta_{q+1})/\beta_{q+1}, g_2(\beta_{q+2})/\beta_{q+2}, \ldots, g_2(\beta_p)/\beta_p)^T$, then $g_2(\beta) = D_2(\beta_2)^{-1/2}m_{g_2(\beta)/\beta_2}$. Furthermore, it follows from the Cauchy-Schwarz inequality and the assumption $\|\beta_2\| \le \delta_n\sqrt{p/n}$ that

$$\frac{1}{C}\|\frac{2\lambda_n}{n}D_2(\beta_2)g_2(\beta)\| \ge \frac{2\lambda_n}{nC}\frac{\sqrt{n}}{\delta_n\sqrt{p}}\|m_{g_2(\beta)/\beta_2}\|, \quad \text{and} \tag{A7}$$

$$\|g_2(\beta)\| = \|(D_2(\beta_2))^{-1/2}m_{g_2(\beta)/\beta_2}\| \le \|m_{g_2(\beta)/\beta_2}\| \cdot \|\beta_2\| \le \|m_{g_2(\beta)/\beta_2}\|\delta_n\sqrt{p/n}. \tag{A8}$$

By (A6), (A7) and (A8), we have the following inequality

$$\frac{2\lambda_n}{nC}\frac{\sqrt{n}}{\delta_n\sqrt{p}}\|m_{g_2(\beta)/\beta_2}\| - \frac{\delta_n\sqrt{p}}{\sqrt{n}}\|m_{g_2(\beta)/\beta_2}\| \le \frac{\delta_n\sqrt{p}}{\sqrt{n}}.$$

Immediately from $p\delta_n^2/\lambda_n \to 0$, we have $\|m_{g_2(\beta)/\beta_2}\| \le \frac{1}{\frac{2\lambda_n}{p\delta_n^2 C} - 1} < \frac{1}{C_0}$, $(C_0 > 1)$, with probability tending to one. Hence with probability tending to one, $\|g_2(\beta)\| \le \|\beta_2\|\|m_{g_2(\beta)/\beta_2}\| \le \frac{1}{C_0}\|\beta_2\|$ as $n \to \infty$, which implies that conclusion (i) holds and $\|g_2(\beta)\| \le \delta_n\sqrt{p/n}$ with probability tending to 1.

To prove (ii), we only need to verify that $\|g_1(\beta) - \beta_{01}\| \le \delta_n\sqrt{p/n}$ with probability tending to 1. Analogously, from (A3), we have $\sup_{\beta\in H_n}\left\|\frac{2\lambda_n}{n}AD_1(\beta_1)g_1(\beta)\right\| = o_p(\sqrt{p/n})$, and $\sup_{\beta\in H_n}\left\|g_1(\beta) - \beta_{01} + \frac{2\lambda_n}{n}BD_2(\beta_2)g_2(\beta)\right\| = O_p(\sqrt{p/n}) \le \delta_n\sqrt{p/n}$. Again by (A6), we know that as $n \to \infty$ and with probability tending to one,

$$\sup_{\beta\in H_n}\left\|\frac{2\lambda_n}{n}BD_2(\beta_2)g_2(\beta)\right\|\le C(\|g_2(\beta)\|+\delta_n\sqrt{p/n})\|B\|\le 2\sqrt{2}C^2\delta_n\sqrt{p/n}.$$

Since $\left\|g_1(\beta)-\beta_{01}\right\|-\frac{2\lambda_n}{n}\left\|BD_2(\beta_2)g_2(\beta)\right\|\le\sup_{\beta\in H_n}\left\|g_1(\beta)-\beta_{01}+\frac{2\lambda_n}{n}BD_2(\beta_2)g_2(\beta)\right\|\le\delta_n\sqrt{p/n}$, then $\sup_{\beta\in H_n}\left\|g_1(\beta)-\beta_{01}\right\|\le(2\sqrt{2}C^2+1)\delta_n\sqrt{p/n}\to 0$ with probability tending to one, which implies that for any $\epsilon>0$, $P(\|g_1(\beta)-\beta_{01}\|\le\epsilon)\to 1$. Thus it follows from $\beta_{01}\in[1/K_0,K_0]^{q+1}$ that $g_1(\beta)\in[1/K_0,K_0]^{q+1}$ holds for large n, which implies that conclusion (ii) holds. This completes the proof. □

Since $\beta_{02}=0$, we can express the objective function of this reduced model as $Q_{n1}(\theta_1)=l_{n1}(\theta_1)-\lambda_n\theta_1^T D_1(\beta_1)\theta_1$.

Lemma 3. *Let $f(\beta_1)$ be a solution to $\dot{Q}_{n1}(\theta_1)=0$, then under regularity conditions (C1)−(C6) and with probability tending to 1,*
(i) $f(\beta_1)$ is a contraction mapping from $[1/K_0,K_0]^{q+1}$ to itself;
(ii) $\sqrt{n}(\hat{\beta}_1^o-\beta_{01})\xrightarrow{D}N(0,\Sigma_1)$, where $\hat{\beta}_1^o$ is the unique fixed point of $f(\beta_1)$ and $\Sigma_1(\beta_0)$ is shown below.

Proof. (i) Similar as the derivation of (A2), through the first order Taylor expansion, we have that

$$f(\beta_1)-\beta_{01}+\frac{2\lambda_n}{n}H_{n1}(\beta_1^*)^{-1}D_1(\beta_1)f(\beta_1)=\frac{1}{n}H_{n1}(\beta_1^*)^{-1}\dot{l}_{n1}(\beta_{01}),\tag{A9}$$

where $H_{n1}(\beta_1^*)=-n^{-1}\ddot{l}_{n1}(\beta_1^*)$ and β_1^* lies between β_{01} and $f(\beta_1)$. From $n^{-1}\dot{l}_{n1}(\beta_{01})=O_p(q/n)$ we know that,

$$\sup_{|\beta_1|\in[1/K_0,K_0]^{q+1}}\|f(\beta_1)-\beta_{01}+\frac{2\lambda_n}{n}H_{n1}(\beta_1^*)^{-1}D_1(\beta_1)f(\beta_1)\|=O_p(\sqrt{q/n}).$$

By (C3) and similar as the proof process of Lemma 2, we have that

$$\sup_{|\beta_1|\in[1/K_0,K_0]^{q+1}}\|\frac{2\lambda_n}{n}H_{n1}(\beta_1^*)^{-1}D_1(\beta_1)f(\beta_1)\|=o_p(\sqrt{q/n}).\tag{A10}$$

Thus with the probability tending to one, $\sup_{|\beta_1|\in[1/K_0,K_0]^{q+1}}\|f(\beta_1)-\beta_{01}\|\le\delta_n\sqrt{q/n}\to 0$, which implies that $P\{f(\beta_1)\in[1/K_0,K_0]^{q+1}\}\to 1$ as $n\to\infty$. That is, $f(\beta_1)$ is a mapping from $[1/K_0,K_0]^{q+1}$ to itself. Next to prove $f(\beta_1)$ is a contraction mapping, we need show that $\sup_{|\beta_1|\in[1/K_0,K_0]^{q+1}}\|\dot{f}(\beta_1)\|=o_p(1)$.

From $\dot{Q}_{n1}(f(\beta_1))=0$ we have $\dot{l}_{n1}(f(\beta_1))=2\lambda_n D_1(\beta_1)f(\beta_1)$. Taking the derivative with respect to β_1^T on both sides of the above equation and rearranging terms, we obtain that

$$\left[\frac{2\lambda_n}{n}D_1(\beta_1)+H_{n1}(f(\beta_1))\right]\dot{f}(\beta_1)=\frac{4\lambda_n}{n}f(\beta_1)\mathrm{diag}(0,\beta_1^{-3},\ldots,\beta_q^{-3}),\tag{A11}$$

where $\dot{f}(\beta_1)=\partial f(\beta_1)/\partial\beta_1{}^T$. From the fact that $\lambda_n/\sqrt{n}\to 0$, $\|f(\beta_1)\|$ and $\|\beta_1\|$ are bounded, we have $\sup_{|\beta_1|\in[1/K_0,K_0]^{q+1}}\times\frac{4\lambda_n}{n}\left\|f(\beta_1)\mathrm{diag}(0,\beta_1^{-3},\ldots,\beta_q^{-3})\right\|=o_p(1)$. Again, since $1/C\|\dot{f}(\beta_1)\|\le\|H_{n1}(f(\beta_1))\dot{f}(\beta_1)\|\le C\|\dot{f}(\beta_1)\|$ and $1/K_0^2\|\dot{f}(\beta_1)\|\le\|D_1(\beta_1)\dot{f}(\beta_1)\|\le K_0^2\|\dot{f}(\beta_1)\|$, and from (A11), we can reach the conclusion that $\sup_{|\beta_1|\in[1/K_0,K_0]^{q+1}}\|\dot{f}(\beta_1)\|=o_p(1)$, which implies that $f(\cdot)$ is a contraction mapping from $[1/K_0,K_0]^{q+1}$ to itself with probability tending to one. Hence according to the contraction mapping theorem, there exists one unique fixed-point $\hat{\beta}_1^o\in[1/K_0,K_0]^{q+1}$ such that $f(\hat{\beta}_1^o)=\hat{\beta}_1^o$.

(ii) From (A9), we have $f(\beta_1)=\left[H_{n1}(\beta_1^*)+\frac{2\lambda_n}{n}D_1(\beta_1)\right]^{-1}\left[H_{n1}(\beta_1^*)\beta_{01}+\frac{1}{n}\dot{l}_{n1}(\beta_{01})\right]$. Denote $\Phi(\hat{\beta}_1^o)=\left[H_{n1}(\beta_1^*)+\frac{2\lambda_n}{n}D_1(\hat{\beta}_1^o)\right]^{-1}$, then we have $\hat{\beta}_1^o=f(\hat{\beta}_1^o)=\Phi(\hat{\beta}_1^o)H_{n1}(\beta_1^*)\beta_{01}+\frac{1}{n}\Phi(\hat{\beta}_1^o)\dot{l}_{n1}(\beta_{01})\right]$, and

$$\sqrt{n}(\hat{\beta}_1^o-\beta_{01})=\sqrt{n}\left\{\Phi(\hat{\beta}_1^o)H_{n1}(\beta_1^*)-I_{q+1}\right\}\beta_{01}+\frac{1}{\sqrt{n}}\Phi(\hat{\beta}_1^o)\dot{l}_{n1}(\beta_{01})=\Pi_1+\Pi_2,$$

with $\Pi_1=\sqrt{n}\left\{\Phi(\hat{\beta}_1^o)H_{n1}(\beta_1^*)-I_{q+1}\right\}\beta_{01}$ and $\Pi_2=\frac{1}{\sqrt{n}}\Phi(\hat{\beta}_1^o)\dot{l}_{n1}(\beta_{01})$.

Furthermore, it follows from Woodbury matrix identity and Condition (C5) that

$$\|\Pi_1\|=\frac{2\lambda_n}{\sqrt{n}}\|H_{n1}(\beta_1^*)^{-1}D_1(\hat{\beta}_1^o)\Phi(\hat{\beta}_1^o)H_{n1}(\beta_1^*)\beta_{01}\|\le\frac{2\lambda_n}{\sqrt{n}}K_0^2\|\beta_{01}\|=o_p(1).$$

Similarly using Woodbury matrix identity to Π_2, we have

$$\begin{aligned}\|\Pi_2\| &= \frac{1}{\sqrt{n}} H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2}\Big\{ I_{q+1} - \frac{2\lambda_n}{n} D_1(\hat{\boldsymbol{\beta}}_1^o)\Phi(\hat{\boldsymbol{\beta}}_1^o)\Big\}\dot{l}_{n1}(\boldsymbol{\beta}_{01}) \\ &= \frac{1}{\sqrt{n}} H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2}\dot{l}_{n1}(\boldsymbol{\beta}_{01}) - \frac{2\lambda_n}{\sqrt{n}} H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2} D_1(\hat{\boldsymbol{\beta}}_1^o)\Phi(\hat{\boldsymbol{\beta}}_1^o)\frac{1}{n}\dot{l}_{n1}(\boldsymbol{\beta}_{01}) \\ &= \frac{1}{\sqrt{n}} H_{n1}(\boldsymbol{\beta}_1^*)^{-1/2}\dot{l}_{n1}(\boldsymbol{\beta}_{01}) + o_p(1) \to N_{q+1}(0, I_{q+1}).\end{aligned}$$

Therefore, $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_{01}) \to N_{q+1}(0, H_{n1}(\boldsymbol{\beta}_1^*)^{-1})$. □

Proof of the theorem. According to the definition of $g(\boldsymbol{\beta})$, $g(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta})^T, g_2(\boldsymbol{\beta})^T)^T$ is the solution of $\dot{Q}_n(\boldsymbol{\theta}) = 0$, that is, $g_1(\boldsymbol{\beta})$ is the solution of $\dot{Q}_{n1}(\boldsymbol{\theta}_1) = 0$ and $g_2(\boldsymbol{\beta})$ is the solution of $\dot{Q}_{n2}(\boldsymbol{\theta}_2) = 0$.

(i) According to the definitions of the BAR estimator $\hat{\boldsymbol{\beta}}$ and Lemma 1 and Lemma 2(i), we have that $\hat{\boldsymbol{\beta}}_2 = \lim_{k\to\infty} g_2(\hat{\boldsymbol{\beta}}^{(k)}) = 0$ holds with the probability tending to 1.

(ii) Since $\hat{\boldsymbol{\beta}}_1 = \lim_{k\to\infty} g_1(\hat{\boldsymbol{\beta}}^{(k)})$, next we should show that $P(\lim_{k\to\infty} \|g_1(\hat{\boldsymbol{\beta}}^{(k)}) - \hat{\boldsymbol{\beta}}_1^o\| = 0) \to 1$, where $\hat{\boldsymbol{\beta}}_1^o$ is the unique fixed point of $f(\boldsymbol{\beta}_1)$ defined in Lemma 3.

From (i) we can see that $\lim_{\boldsymbol{\beta}_2\to 0} g_2(\boldsymbol{\beta};\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = 0$, and thus $\lim_{\boldsymbol{\beta}_2\to 0} g_1(\boldsymbol{\beta};\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = f(\boldsymbol{\beta}_1)$ holds. Also, for any $\hat{\boldsymbol{\beta}}_2^{(k)}$, $g(\boldsymbol{\beta};\boldsymbol{\beta}_1,\hat{\boldsymbol{\beta}}_2^{(k)})$ is a mapping of $\boldsymbol{\beta}_1$, and with $k\to\infty$ and probability tending to one, we have that

$$\eta_k \equiv \sup_{g_1(\boldsymbol{\beta})\in[1/K_0,K_0]^{q+1}} \Big\| f(\boldsymbol{\beta}_1) - g_1(\boldsymbol{\beta};\boldsymbol{\beta}_1,\hat{\boldsymbol{\beta}}_2^{(k)})\Big\| \to 0. \tag{A12}$$

On the other hand, since $f(\cdot)$ is a contraction mapping, there exists a constant $C_1 > 1$ such that

$$\|f(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\boldsymbol{\beta}}_1^o\| = \|f(\hat{\boldsymbol{\beta}}_1^{(k)}) - f(\hat{\boldsymbol{\beta}}_1^o)\| \le \frac{1}{C_1}\|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^o\|. \tag{A13}$$

Let $h_k = \|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^o\|$, then it follows from (A12) and (A13) that

$$\begin{aligned}h_{k+1} = \|\hat{\boldsymbol{\beta}}_1^{(k+1)} - \hat{\boldsymbol{\beta}}_1^o\| &\le \|g_1(\hat{\boldsymbol{\beta}}^{(k)}) - f(\hat{\boldsymbol{\beta}}_1^{(k)})\| + \|f(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\boldsymbol{\beta}}_1^o\| \\ &\le \eta_k + \frac{1}{C_1}h_k.\end{aligned}$$

From (A12), for any $\epsilon \ge 0$, there exists $N > 0$ such that when $k > N$, $0 \le \eta_k < \epsilon$.

Employing some recursive calculation, we have $h_k \to 0$ as $k\to\infty$. Hence, with probability tending to one, we have $\|\hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\boldsymbol{\beta}}_1^o\| \to 0$ as $k\to\infty$. Since $\hat{\boldsymbol{\beta}}_1 \equiv \lim_{k\to\infty}\hat{\boldsymbol{\beta}}_1^{(k)}$, it follows from the uniqueness of the fixed-point that $P(\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1^o) \to 1, \quad k\to\infty$.

(iii) The asymptotic normality of $\hat{\boldsymbol{\beta}}_1$ follows from part (ii) of Lemma 3. □

References

Cai, J., Fan, J., Li, R., Zhou, H., 2005. Variable selection for multivariate failure time data. Biometrika 92, 303–316.

Dai, L., Chen, K., Sun, Z., Liu, Z., Li, G., 2018. Broken adaptive ridge regression and its asymptotic properties. J. Multivar. Anal. 168, 334–351.

Dicker, L., Huang, B., Lin, X., 2013. Variable selection and estimation with seamless-L_0 penalty. Stat. Sin. 23, 929–962.

Du, M., Hu, T., Sun, J., 2019. Semiparametric probit model for informative current status data. Stat. Med. 38, 2219–2227.

Du, M., Yu, M., 2023. Regression analysis of multivariate interval-censored failure time data with a cured subgroup and informative censoring. J. Nonparametr. Stat. https://doi.org/10.1080/10485252.2023.2280016.

Du, M., Zhao, H., Sun, J., 2021. A unified approach to variable selection for Cox proportional hazards model with interval-censored failure time data. Stat. Methods Med. Res. 30, 1833–1849.

Du, M., Zhao, X., Sun, J., 2022. Variable selection for case-cohort studies with informatively interval-censored outcomes. Comput. Stat. Data Anal. 172, 107484.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle property. J. Am. Stat. Assoc. 96, 1348–1360.

Fan, J., Li, R., 2002. Variable selection for Cox's proportional hazards model and frailty model. Ann. Stat. 30, 74–99.

Li, K., Chan, W., Doody, R.S., Quinn, J., Luo, S., Initiative, A.D.N., 2017a. Prediction of conversion to alzheimer's disease with longitudinal measures and time-to-event data. J. Alzheimer's Dis. 58, 361–371.

Li, S., Hu, T., Wang, P., Sun, J., 2017b. Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity. Comput. Stat. Data Anal. 110, 75–86.

Li, S., Wu, Q., Sun, J., 2020. Penalized estimation of semiparametric transformation models with interval-censored data and application to Alzheimer's disease. Stat. Methods Med. Res. 29 (8), 2151–2166.

Liu, Z., Li, G., 2016. Efficient regularized regression with L_0 penalty for variable selection and network construction. Comput. Math. Methods Med., 3456153.

Lv, J., Fan, Y., 2009. A unified approach to model selection and sparse recovery using regularized least squares. Ann. Stat. 37, 3498–3528.

Ma, L., Hu, T., Sun, J., 2015. Sieve maximum likelihood regression analysis of dependent current status data. Biometrika 102, 731–738.

Schumaker, L., 1981. Spline functions: basic theory. Wiley, New York, NY.

Shi, Y., Cao, Y., Jiao, Y., Liu, Y., 2014. SICA for Cox's proportional hazards model with a diverging number of parameters. Acta Math. Appl. Sin. Engl. Ser. 30, 887–902.

Sun, J., 2006. The statistical analysis of interval-censored failure time data. Springer, New York.

Sun, J., Park, D-H., Sun, L., Zhao, X., 2005. Semiparametric regression analysis of longitudinal data with informative observation times. J. Am. Stat. Assoc. 100, 882–889.

Sun, J., Sun, L., Liu, D., 2007. Regression analysis of longitudinal data in the presence of informative observation and censoring times. J. Am. Stat. Assoc. 102, 1397–1406.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Tibshirani, R., 1997. The Lasso method for variable selection in the Cox model. Stat. Med. 16, 385–395.
Wang, P., Zhao, H., Sun, J., 2016. Regression analysis of case K interval-censored failure time data in the presence if informative censoring. Biometrics 72, 1103–1112.
Wang, S., Wang, C., Wang, P., Sun, J., 2018. Semiparametric analysis of the additive hazards model with informatively interval-censored failure time data. Comput. Stat. Data Anal. 125, 1–9.
Wang, S., Wang, C., Wang, P., Sun, J., 2020. Estimation of the additive hazards model with case K interval-censored failure time data in the presence of informative censoring. Comput. Stat. Data Anal. 144, 106891.
Wu, Y., Cook, R., 2015. Penalized regression for interval-censored times of disease progression: selection of HLA markers in psoriatic arthritis. Biometrics 71, 782–791.
Xu, D., Zhao, H., Sun, J., 2018. Joint analysis of interval-censored failure time data and panel count data. Lifetime Data Anal. 24, 94–109.
Zhang, H., Lu, W.B., 2007. Adaptive Lasso for Cox's proportional hazards model. Biometrika 94, 1–13.
Zhao, H., Wu, Q., Li, G., Sun, J., 2020. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. J. Am. Stat. Assoc. 115 (529), 204–216.
Zou, H., 2006. The Adaptive Lasso and its oracle properties. J. Am. Stat. Assoc. 101, 1418–1429.