Statistical Papers
https://doi.org/10.1007/s00362-023-01486-1

REGULAR ARTICLE

Variable selection in proportional odds model with informatively interval-censored data

Bo Zhao[1] · **Shuying Wang**[1] · **Chunjie Wang**[1]

Received: 12 February 2023 / Revised: 8 August 2023

Abstract

The proportional odds (PO) model is one of the most commonly used models for regression analysis of failure time data in survival analysis. It assumes that the odds of the failure is proportional to the baseline odds at any point in time given the covariate. The model focus on the situation that the ratio of the hazards converges to unity as time goes to infinity, while the proportional hazards (PH) model has a constant ratio of hazards over time. In the paper, we consider a general type of failure time data, case K interval-censored data, that include case I or case II interval-censored data as special cases. We propose a PO model-based unified penalized variable selection procedure that involves minimizing a negative sieve log-likelihood function plus a broken adaptive ridge penalty, with the initial values obtained from the ridge regression estimator. The proposed approach allows dependent censoring, which occurs quite often and could lead to biased or misleading estimates without considering it. We show that the proposed selection method has oracle properties and the estimator is semiparametrically efficient. The numerical studies suggest that the proposed approach works well for practical situations. In addition, the method is applied to an ADNI study that motivates this investigation.

Keywords Informatively interval-censored data · Variable selection · Proportional odds model · Broken adaptive ridge penalty · Sieve maximum likelihood

Mathematics Subject Classification 62G05 · 62J07 · 62N01 · 62N02

✉ Shuying Wang
wangshuying0601@163.com

Bo Zhao
jgtjx0313@163.com

Chunjie Wang
wangchunjie@ccut.edu.cn

[1] School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China

Published online: 29 September 2023

1 Introduction

Interval-censored failure time data often occur in follow-up studies and clinical trials. A general type for interval-censored data is the case K interval-censored data (Wang et al. 2016, 2018, 2020a, b, 2023; Zhao et al. 2021), for which there exists a sequence of observation points for each subject and the failure time of interest is known only to belong to a window or an interval. It includes case I or case II interval-censored data as special cases. For case I interval-censored data(or current status data), each subject is observed only once and the observed data have the monitoring time and the current status of the event of interest(occurred or not) (Huang 1996; Ma et al. 2015; Hu et al. 2017). Another type is case II interval-censored data meaning that there exist two observation points for each subject, the observed information is described by the left and right end points of intervals, respectively. In addition, the right-censored data arise when the right end points of an interval is infinite (Kalbfleisch and Prentice 2002).

Many methods and models have been proposed for the censored data in the literature (Sun 2006). The PH model and the PO model are two popular frameworks in investigating the association between risk factors and disease occurrence or death (Cox 1972; Rossini and Tsiatis 1996; Huang and Rossini 1997; Shen 1998; Yang and Prentice 1999; Wang and Wang 2021). The PH model assumes a constant ratio of hazards over time, however, this assumption may not be appropriate in real applications. As an important alternative model, the PO model specifies that the odds of the failure given any covariate is proportional to the baseline odds at any time point. And the regression parameters in the PO model have a nice interpretation in terms of the log odds ratio of the failure. In addition, the PO model is considered more appropriate than the PH model when there exists an effective cure or the morbidity rates converge with time (Murphy et al. 1997).

A topic of widespread interest is variable selection and numerous methods have been developed in statistical analysis. In particular, for linear models with outcomes that are not censored, certain conventional methods are employed like backward selection, forward selection, and best subset selection. Lately, there has been productive investigation focused on the penalized estimation method which is that maximizes an objective function with a penalty function. These methods consist of the least absolute shrinkage and selection operator(LASSO) procedure (Tibshirani 1996), the smoothly clipped absolute deviation(SCAD) procedure (Fan and Li 2001), the adaptive LASSO(ALASSO) procedure (Zou 2006), the smooth integration of counting and absolute deviation(SICA) procedure (Lv and Fan 2009), the seamless-L_0(SELO) procedure(Dicke et al. 2013), and the broken adaptive ridge (BAR) regression (Dai et al. 2018; Zhao et al. 2020; Sun et al. 2022a, b). A number of authors have investigated the variable selection for the right-censored failure time data, and especially, Tibshirani (1997), Fan and Li (2002) and Zhang and Lu (2007) generalized the LASSO, SCAD, and ALASSO penalty-based procedures, respectively, to the Cox proportional hazards model situation. Furthermore, Lu and Zhang (2007) studied the proportional odds model for the right-censored data by utilizing a penalized marginal likelihood based on ranks.

For the variable selection based on the interval-censored data, many procedures have been investigated (Scolas et al. 2016; Wu and Cook 2015; Zhao et al. 2020; Li et

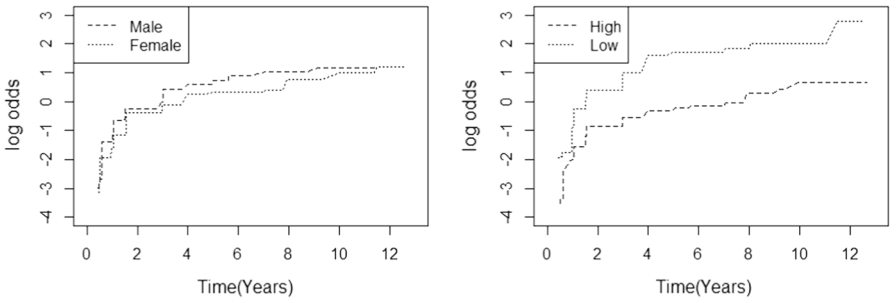

Fig. 1 Estimate of the log odds on AD conversion by any given time for participants with different PTGENDER (left) and FAQ (right)

al. 2021; Sun et al. 2022a, b; Du and Sun 2022). Specifically, two parametric procedures were developed in Scolas et al. (2016) and Wu and Cook (2015), respectively. Zhao et al. (2020) considered a semiparametric procedure and proposed the broken adaptive ridge (BAR) regression on the proportional hazards model. And Sun et al. (2022a, b) developed a variable selection technique for multivariate interval-censored data. Du and Sun (2022) reviewed variable selection procedures for noninformative or independent interval-censored failure time data. One drawback for majority of the methods mentioned above is that they all assumed that the failure time and the observation process are independent. Nevertheless, the assumption may not be true in many real situations. Corresponding to this, Du et al. (2021) developed an approach to variable selection for informative interval-censored data, and proposed a two-step method relying on the proportional hazards model and assuming the Poisson process for the observation process. However, note that the proportional hazards model may not be an appropriate choice when homogeneity between different groups increases over time and it is more preferable to use the proportional odds model to analyze the data.

A motivating example in this article is an Alzheimer's Disease Neuroimaging Initiative (ADNI) study which can be found on the website http://adni.loni.usc.edu. This research includes medical, imaging measures, biomarkers and others covariates. An interesting aspect in this investigation is to monitor the progress of participants and also to identify the covariates or risk factors for AD conversion time. To establish a suitable model, we draw the empirical plot of the log odds function for participants based on the variables PTGENDER (Male and Female) and FAQ (high and low), respectively. Figure 1 presents the nonparametric maximum likelihood estimates (NPMLE). Note that the vertical difference between the two curves remains relatively constant which suggests that the proportional odds model maybe reasonable. Moreover, the initial visit or censoring mechanism may be related to the time of AD conversion. Hence, in this article, we will introduce a method for selecting variables using the proportional odds model for informatively case K interval-censored failure time data. This method does not rely on the assumption of a Poisson distribution for the observation process.

The remainder of the paper is organized as follows. We begin with introducing data structure and model assumptions in Sect. 2. In Sect. 3, we propose an sieve penalized variable selection procedure that combines the two-step approach and the sieve method. The number of covariates is allowed to diverge with the sample size and a

recursive algorithm for the determination of the BAR estimators is developed. Results obtained from an extensive simulation study are presented in Sect. 4 and indicate that the proposed method seems to work well for practical situations. In Sect. 5, the proposed procedure is employed to a set of real data and Sect. 6 contains some discussion and concluding remarks.

2 Data and models

Consider a failure time study that consists of n independent subjects. For each subject i, let T_i denote the failure time of interest and suppose that there exists a p-dimensional vector of covariates denoted by $x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$, $i = 1, 2, \ldots, n$. In the practical applications, the failure time T_i may not be observed exactly instead that we obtain a sequence of observation time points denoted by $U_{i0} = 0 < U_{i1} < U_{i2} < \ldots < U_{iK_i} < \infty$ and the indicator $\delta_{ij} = I(U_{ij-1} < T_i \le U_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, K_i$, where K_i denotes the total number of observation points. Then, we have a point process $\widetilde{N}_i(t) = \sum_{j=1}^{K_i} I(U_{ij} \le t)$, which characterizes the observation process on subject i and jumps only at the observation times. Let τ_i denote the follow-up time which is independent of failure time T_i on subject i, we have $K_i = \widetilde{N}_i(\tau_i)$ and one observes case K interval-censored data which have the form

$$O = \Big\{ O_i = (x_i, \tau_i, U_{ij}, \delta_{ij}, K_i, j = 1, \ldots, K_i), i = 1, 2, \ldots, n \Big\}.$$

As mentioned above, it is apparent that case I interval-censored data or current status data occur if observation points $K_i = 1$ for $i = 1, \ldots, n$. That is, each subject is observed only once and the only observed information for the event of interest is whether the event has occurred no later than the observation time. If observation points $K_i = 2$, the data are usually referred to case II interval-censored data. It is meaning that each subject is observed twice and the observed information is described by two variables that represent the left and right end points of an interval (Sun 2006), respectively. In addition, they reduce to the right-censored data, if the failure time $T_i > U_{iK_i}$ or $\sum_{j=1}^{K_i} \delta_{ij} = 0$ (Kalbfleisch and Prentice 2002).

Note that there exist two processes in the informatively interval-censored data, which are the failure time process and the observation process. In many situations, the two processes may be correlated. A typical example of the informative censoring can be observed in health or medical follow-up studies, such as clinical trials where patients may pay less or more visits depending on their conditions than the pre-specified schedule. To describe the covariate effects and the relationship between the failure time of interest and the censoring mechanism, assuming that there exists a latent variable b_i and given the covariates x_i and b_i, the variable T_i follows the proportional odds frailty model

$$\frac{F(t \mid x_i, b_i)}{1 - F(t \mid x_i, b_i)} = \Lambda_0(t) \exp\left(x_i^\top \boldsymbol{\beta}_a + b_i \beta_b\right), \tag{1}$$

where $\Lambda_0(t) = \frac{F_0(t)}{1-F_0(t)}$ denotes a completely unknown baseline odds function, $F_0(t)$ is the baseline cumulative distribution function, and $\boldsymbol{\beta}_a$, β_b are unknown regression parameters. The survival function corresponding to the proposed model is

$$S(t \mid x_i, b_i) = \left[1 + \Lambda_0(t) \exp\left(x_i^\top \boldsymbol{\beta}_a + b_i \beta_b\right)\right]^{-1}.$$

For the observation process, it will be assumed that given x_i and b_i, $\widetilde{N}_i(t)$ has the rate function

$$E(d\widetilde{N}_i(t) \mid x_i, b_i) = \lambda_{0h}(t) \exp(x_i^\top \boldsymbol{\gamma} + b_i) dt\,, \tag{2}$$

where $\lambda_{0h}(t)$ is a completely unknown continuous baseline rate function and $\boldsymbol{\gamma}$ is a vector of regression parameters as $\boldsymbol{\beta}_a$. Define $\Lambda_{0h}(t) = \int_0^t \lambda_{0h}(s)ds$, and assume that $\Lambda_{0h}(\tau_0) = 1$, where τ_0 denotes the longest follow-up time. Also it will be assumed that given x_i and b_i, T_i and $\widetilde{N}_i(t)$ are conditional independent. Note that models (1) and (2) with $b_i = 0$ have been commonly used in the analysis of failure time data (Klein and Moeschberger 2003) and event history data (Cook and Lawless 2007), respectively. It is apparent that the parameter β_b represents the extent of the association between the failure time and the observation process. The positive value of β_b represents that the observation process and the failure time process are positively correlated, while the negative value of β_b indicates the negative association between them. In addition, if $\beta_b = 0$, these processes are independent. Moreover, we simply suppose that the observation process has the rate function, rather than a nonhomogeneous Poisson process, and it is much more flexible than the others (Wang et al. 2016, 2018).

For inference, define $\boldsymbol{\beta} = (\boldsymbol{\beta}_a^\top, \beta_b)^\top$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \Lambda_0)^\top$. Under the assumptions above, if the distribution of the b_i's was known, one can write the likelihood function as

$$L(\boldsymbol{\beta}, \Lambda_0 \mid b_i) = \prod_{i=1}^{n} \prod_{j=1}^{K_i} \left[\left[S(U_{i,j-1} \mid x_i, b_i) - S(U_{ij} \mid x_i, b_i)\right]^{\delta_{ij}} \left[S(U_{iK_i} \mid x_i, b_i)\right]^{1-\sum_{j=1}^{n} \delta_{ij}} \right]. \tag{3}$$

Note that the conditional likelihood mentioned above includes latent frailty variable b_i's, the baseline odds function $\Lambda_0(t)$, and regression parameter $\boldsymbol{\beta}$. If the distribution of the latent frailty variable b_i's was known, an EM algorithm can be employed to maximize the log-likelihood function to estimate the parameters of interest (Wang et al. 2020a, b). However, it is easy to find that the distribution of the b_i's may be unknown in many real applications. To solve the problem, we first propose to estimate the b_i's and then conduct the sieve penalized variable selection in the following section.

3 Sieve penalized variable selection procedure

Now we will focus on the likelihood-based borrow-strength approach for a latent variable in the models above. It should be noted that the latent effects b_i's are not

known in (3) and therefore, a natural approach is to estimate latent effects b_i's by employing a borrow-strength estimation procedure (Wang et al. 2018; Zhao et al. 2021) and then directly maximize the working likelihood function to estimate the unknown parameters $\Lambda_0(t)$, and $\boldsymbol{\beta}$ by substituting in the estimator of b_i into likelihood function (3).

3.1 Borrow-strength Sieve approach

Note that the shared frailty variables b_i's are not observed and their distribution is unknown. To address this, we propose the two-step method, as described by Huang and Wang (2004) and Zhao et al. (2021). In the subsequent sections, we will first outline the estimation and inference process for model (2), following the similar methodology employed by Huang and Wang (2004) and Zhao et al. (2021). Specifically, define $N_i(t) = \widetilde{N}_i(t \wedge \tau_i)$ for the i-th subject, where $t \wedge \tau_i = \min(t, \tau_i)$. Then we obtain that

$$N_i(t) = \int_0^{t\wedge\tau_i} d\widetilde{N}_i(u) = \int_0^t I(\tau_i \geq u) d\widetilde{N}_i(u).$$

Under the assumptions of the observation process, we derive that

$$d\Lambda_{0h}(t) = \frac{d[EN_i(t)]}{E(\exp(x_i^\top \boldsymbol{\gamma} + b_i) I(\tau_i \geq t))}$$

by the conditional expectation of $N_i(t)$.

According to the estimation process and results in Zhao et al. (2021), we have

$$\log \Lambda_{0h}(\tau_0) - \log \Lambda_{0h}(t) = \int_t^{\tau_0} \frac{d[EN_i(s)]}{E(N_i(s) I(\tau_i \geq s))}, \tag{4}$$

which does not depend on frailty variable b_i and covariates x_i. Combining the assumption $\Lambda_{0h}(\tau_0) = 1$, the straightforward calculation yields

$$\Lambda_{0h}(t) \approx \prod_{t \leq s \leq \tau_i} \left[1 - \frac{d[EN_i(s)]}{E[N_i(s) I(\tau_i \geq s)]}\right].$$

It is thus natural to suggest that one can estimate $\Lambda_{0h}(t)$ by

$$\widehat{\Lambda}_{0h}(t) = \prod_{s_{(l)} > t} \left(1 - \frac{d_{(l)}}{R_{(l)}}\right). \tag{5}$$

In the above, the $s_{(l)}$'s are the ordered and distinct values of observation times $\{U_{ij}\}$, $d_{(l)}$ is the number of the observation times equal to $s_{(l)}$, and $R_{(l)}$ the total number of observation events with observation times and observation terminating time satisfying $U_{ij} \leq s_{(l)} \leq \tau_i$.

To achieve the regression estimation of parameter $\boldsymbol{\gamma}$, based on the similar idea in Huang and Wang (2004) and Zhao et al. (2021), we employ the following weighted estimating equations

$$\sum_{i=1}^{n} w_i \tilde{x}_i \left(K_i \widehat{\Lambda}_{0h}^{-1}(\tau_i) - \gamma_0 \exp(x_i^\top \boldsymbol{\gamma}) \right) = 0 , \tag{6}$$

where $\tilde{x}_i^\top = (1, x_i^\top)$, $\gamma_0 = E(e^{b_i})$, and the w_i's are some weights that could depend on the x_i's, τ_i's and Λ_{0h} (Huang and Wang 2004). The estimators of $\boldsymbol{\gamma}$ and γ_0, $\hat{\boldsymbol{\gamma}}$ and $\hat{\gamma}_0$ can be obtained by solving the estimation equations (6) in the above. Then, the estimation of b_i is given by

$$\hat{b}_i = \log \left\{ \frac{K_i}{\widehat{\Lambda}_{0h}(\tau_i) \exp(x_i^\top \hat{\boldsymbol{\gamma}})} \right\}.$$

For the inference about model (1), as previously stated, given the b_is, model (1) becomes the usual proportional odds model. Therefore, it is natural to replace b_i by the estimator $\hat{b}_i$ and the likelihood function (3) is rewritten by

$$\begin{aligned} &L(\boldsymbol{\beta}, \Lambda_0 \mid \hat{b}_i) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{K_i} \left[\left(S(U_{i,j-1} \mid x_i, \hat{b}_i) - S(U_{ij} \mid x_i, \hat{b}_i) \right)^{\delta_{ij}} \left(S(U_{iK_i} \mid x_i, \hat{b}_i) \right)^{1-\sum_{i=1}^{n} \delta_{ij}} \right]. \end{aligned} \tag{7}$$

Now we focus on the estimator $\boldsymbol{\theta}$ in general and it is apparent that a natural approach would be to maximize the log-likelihood function $l(\boldsymbol{\beta}, \Lambda_0 \mid \hat{b}_i) = \log(L(\boldsymbol{\beta}, \Lambda_0 \mid \hat{b}_i))$. However, it is obvious that the computation becomes challenging due to the presence of infinite-dimensional parameters Λ_0. To address this issue and maintain the modeling flexibility, we employ the Bernstein polynomials method to approximate Λ_0(t) (Wang and Ghosh 2012). Specifically, let Θ denote the parameter space of $\boldsymbol{\theta}$, and define the sieve space

$$\Theta_n = \left\{ \boldsymbol{\theta}_n = (\boldsymbol{\beta}^\top, \Lambda_n)^\top \right\} = \mathcal{B} \otimes \Phi_n ,$$

where

$$\mathcal{B} = \left\{ \boldsymbol{\beta} \in \mathcal{R}^{p+1}, \parallel \boldsymbol{\beta} \parallel \leq M \right\},$$

$$\Phi_n = \left\{ \Lambda_n(t) = \sum_{l=0}^{m} \alpha_l B_l(t, m, a_1, a_2) : \sum_{0 \leq l \leq m} \mid \alpha_l \mid \leq M_n, 0 \leq \alpha_0 \leq \alpha_1 \leq \ldots \leq \alpha_m \right\}.$$

M is a constant, $M_n = o(n^{a_0})$, $0 < a_0 < 1/2$. $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)^\top$ is the unknown sieve parameter vector to be estimated and

$$B_l(t, m, a_1, a_2) = C_m^l \left(\frac{t - a_1}{a_2 - a_1} \right)^l \left(1 - \frac{t - a_1}{a_2 - a_1} \right)^{m-l},$$

where m denotes the degree of the Bernstein polynomials which is usually taken to be $m = O(n^{\nu})$ for some $0 < \nu < 1/2$. And $0 \le a_1 < a_2 < \infty$ with (a_1, a_2) usually taken as the range of observed data.

Given the observation points U_{ij}, $\hat{b}_i$ and $\Lambda_n(t)$, the working likelihood function has the following form

$$L_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \prod_{j=1}^{K_i} \left[\left(S_n(U_{i,j-1} \mid x_i, \hat{b}_i) - S_n(U_{ij} \mid x_i, \hat{b}_i) \right)^{\delta_{ij}} \left(S_n(U_{iK_i} \mid x_i, \hat{b}_i) \right)^{1-\sum_{i=1}^{n} \delta_{ij}} \right],$$

where $S_n(t \mid x_i, \hat{b}_i) = [1 + \Lambda_n(t) \exp(x_i^{\top} \boldsymbol{\beta}_a + \hat{b}_i \beta_b)]^{-1}$. If someone is solely concerned with approximating $\boldsymbol{\beta}$, it is natural to concentrate on the sieve profile log-likelihood function $l_p(\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}} l_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$, where $l_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \log L_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$. In the following, we will propose a penalized or regularized procedure for simultaneous estimation and covariate selection based on $l_p(\boldsymbol{\beta})$.

3.2 Penalized variable selection procedure

For the simultaneous estimation and covariate selection, we will consider a broken adaptive ridge (BAR) penalized operator proposed by Zhao et al. (2020), which combines the strengths of the quadratic regularization and the adaptive weighted bridge shrinkage in interval-censored data. Specifically, define $\breve{\boldsymbol{\beta}} = (\breve{\beta}_1, \ldots, \breve{\beta}_{p-1}, \breve{\beta}_p)^{\top}$ denotes a consistent estimator of $\boldsymbol{\beta}$, the corresponding penalized objective function is

$$l_{pp}(\boldsymbol{\beta} \mid \breve{\boldsymbol{\beta}}) = -2l_p(\boldsymbol{\beta}) + \sum_{s=1}^{p} P(\mid \beta_s \mid; \lambda_n) = -2l_p(\boldsymbol{\beta}) + \lambda_n \sum_{s=1}^{p} \frac{\beta_s^2}{\breve{\beta}_s^2}, \tag{8}$$

where parameter λ_n denotes a non-negative penalization tuning parameter. As mentioned by Dai et al. (2018), Zhao et al. (2020) and Sun et al. (2022a, b), one main advantage behind the proposed procedure above is that $\frac{\beta_s^2}{\breve{\beta}_s^2}$ is expected to converge to $I(\mid \beta_s \mid \neq 0)$ in probability as n goes to infinity if given the consistency of $\breve{\boldsymbol{\beta}}$.

To obtain the broken adaptive ridge estimates, we need to minimize objective function $l_{pp}(\boldsymbol{\beta} \mid \breve{\boldsymbol{\beta}})$ in (7). Specifically, for given an initial value $\hat{\boldsymbol{\beta}}^{(0)}$, which is a consistent estimator of $\boldsymbol{\beta}$, one can update $\hat{\boldsymbol{\beta}}^{(k)}$ iteratively by the following reweighed L_2-penalized proportional odds regression estimator

$$\hat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_p(\boldsymbol{\beta}) + \lambda_n \sum_{s=1}^{p} \frac{\beta_s^2}{(\hat{\beta}_s^{(k-1)})^2} \right\}, \tag{9}$$

In the algorithm, according to the ideas of Wang and Leng (2007) and Zhao et al. (2020), the log-likelihood function $l_p(\boldsymbol{\beta})$ can be approximated as an iterative least

square procedure by the Newton–Raphson update. Therefore, by a second-order Taylor expansion, minimizing (7) is asymptotically equivalent to minimizing

$$\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{s=1}^{p} \frac{\beta_s^2}{(\hat{\beta}_s^{(k-1)})^2}. \tag{10}$$

In the above, $\boldsymbol{y} = (\boldsymbol{Z}^\top)^{-1}[\dot{l}_n(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) - \ddot{l}_n(\boldsymbol{\beta} \mid \boldsymbol{\alpha})\boldsymbol{\beta}]$ and matrix $\boldsymbol{Z}$ be the Cholesky decomposition of $-\ddot{l}_n(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$ or $\boldsymbol{Z}^\top\boldsymbol{Z} = -\ddot{l}_n(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$, where $\dot{l}_n(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) = \frac{\partial l_n(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}}$, $\ddot{l}_n(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) = \frac{\partial l_n^2(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}$ represent the first and second partial derivatives of $l_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ about $\boldsymbol{\beta}$, respectively. In the following section, we can obtain the broken adaptive ridge regression estimator by minimizing objective function (9) for a fixed λ_n by the subsequent iterative algorithm. The specific iterative algorithm is outlined below.

- Step 1. Set $k = 0$ and choose an initial estimator $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\boldsymbol{\beta}}^{(0)\top}, \hat{\boldsymbol{\alpha}}^{(0)\top})^\top$. One can take $\hat{\boldsymbol{\alpha}}^{(0)\top} = \mathbf{0}$ and take the ridge regression estimator

$$\hat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\beta} \left\{ -2l_p(\boldsymbol{\beta}) + \varrho_n \sum_{s=1}^{p} \beta_s^2 \right\},$$

  where ϱ_n is a non-negative tuning parameter.
- Step 2. At the $(k+1)$-th iteration, for the current estimate compute $\hat{\boldsymbol{\alpha}}^{(k)}$, compute $\boldsymbol{\Omega}_n = \ddot{l}_n(\boldsymbol{\beta}^{(k)} \mid \boldsymbol{\alpha}^{(k)})$, $\boldsymbol{D}(\boldsymbol{\beta}^{(k)}) = diag(\beta_1^{(k)}, \ldots, \beta_{p-1}^{(k)}, \beta_p^{(k)}, 0)$, and $\boldsymbol{\xi}_n = \dot{l}_n(\boldsymbol{\beta}^{(k)} \mid \boldsymbol{\alpha}^{(k)}) - \ddot{l}_n(\boldsymbol{\beta}^{(k)} \mid \boldsymbol{\alpha}^{(k)})\boldsymbol{\beta}^{(k)}$. And obtain updated estimates

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left\{ \boldsymbol{\Omega}_n + 2\lambda_n \boldsymbol{D}(\hat{\boldsymbol{\beta}}^{(k)}) \right\} \boldsymbol{\xi}_n.$$

- Step 3. At the $(k+1)$th iteration, for the current estimate $\hat{\boldsymbol{\beta}}^{(k+1)}$, obtain the updated estimate $\hat{\boldsymbol{\alpha}}^{(k+1)}$ by solving equation $\frac{\partial l_n(\hat{\boldsymbol{\beta}}^{(k+1)},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0$.
- Step 4. Repeat Steps 2–3 until the convergence is achieved.

Note that in the iterative process above, one needs the Cholesky decomposition of $-\ddot{l}_n(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$ when calculating $\boldsymbol{\Omega}_n$ and $\boldsymbol{\xi}_n$. And on the covariate selection, let the estimates of the components of $\boldsymbol{\beta}$ as zero if the estimate values are less than a pre-specified threshold. By following Wang et al. (2007), we used the threshold of 10^{-6}. To implement the algorithm above, one needs to select two tuning parameters ϱ_n and λ_n. An optimal tuning parameter can result in a parsimonious model with good prediction performance. Wang et al. (2009) has showed that Bayesian information criterion (BIC) is consistent in model selection, and we employ the BIC-type criterion to choose the tuning parameter. Also as pointed out by others and shown in the numerical study, the BAR-based approach is not sensitive to ϱ_n and thus it can be taken to be a constant(e.g., 50 or 100). For a given λ_n, we can obtain BAR estimators $\hat{\boldsymbol{\beta}}_{\lambda_n}$ and $\hat{\boldsymbol{\alpha}}_{\lambda_n}$ and the BIC-type criterion is defined by

$$BIC_{\lambda_n} = -2l_n(\hat{\boldsymbol{\beta}}_{\lambda_n}, \hat{\boldsymbol{\alpha}}_{\lambda_n}) + \log(n) df_{\lambda_n},$$

where df_{λ_n} is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_{\lambda_n}$.

In the following, we establish the asymptotic properties of the proposed BAR estimator $\hat{\boldsymbol{\beta}}^*$ and denote $\boldsymbol{\beta}_0 = (\beta_{0,1}, \beta_{0,2}, \ldots, \beta_{0,p}, \beta_{0,p+1})^\top$ as the true of $\boldsymbol{\beta}$. Without losing generality, assume parameter $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{(01)}, \boldsymbol{\beta}_{(02)})$, where $\boldsymbol{\beta}_{(01)}$ consists of all $(q+1)(q << p)$ nonzero components. Corresponding to this, we will divide BAR estimator $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1^{*\top}, \hat{\boldsymbol{\beta}}_2^{*\top})^\top$ in some ways. The following Theorem describes the oracle property of BAR estimator with limits being $n \to \infty$ and the proof was given in the Appendix.

Theorem 1 *Suppose that the regularity conditions (C1)–(C8) described in the Appendix hold. Then as* $n \to \infty$ *and with probability tending to* 1*, the BAR estimator* $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1^*, \hat{\boldsymbol{\beta}}_2^*)$ *exists and has the following properties:*

(I) $\hat{\boldsymbol{\beta}}_2^* = 0$*, with probability tending to 1.*

(II) $\hat{\boldsymbol{\beta}}_1^*$ *is the unique fixed-point of the equation* $\boldsymbol{\beta}_1 = (\boldsymbol{\Omega}_n^{(1)} + 2\lambda_n \boldsymbol{D}_1(\boldsymbol{\beta}_1))^{-1} \boldsymbol{\xi}_n^{(1)}$. *Here* $\boldsymbol{D}_1(\boldsymbol{\beta}_1) = diag(\beta_1^{-2}, \beta_2^{-2}, \ldots, \beta_q^{-2}, 0)$*,* $\boldsymbol{\Omega}_n^{(1)}$ *denotes the* $(q+1) \times (q+1)$ *leading submatrix of* $\boldsymbol{\Omega}_n$ *and* $\boldsymbol{\xi}_n^{(1)}$ *denotes the vector that consists of the first* $q+1$ *component of* $\boldsymbol{\xi}_n$*.*

(III) $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_{(01)})$ *converges in distribution to a multivariate normal distribution* $N_{q+1}(0, \Sigma)$*, where the variance-covariance matrix* Σ *is defined in the Appendix.*

4 Simulation study

In this section, to assess the performance of the finite sample of the penalized variable selection procedure, we present some results obtained from an extensive simulation study. In the study, we first generated the covariates x_i's from the multivariate normal distribution with mean zero, variance one, and the correlation between x_j and x_k being $\rho^{|j-k|}$ with $\rho = 0.2$ or $\rho = 0.5$, $j, k = 1, \ldots, p$. The latent variables b_i's were generated from the normal distribution with mean 0 and variance 1. The failure time of interest was generated from model (1) with $\Lambda_0(t) = t^2$ or $\Lambda_0(t) = \log(t+1) + t^{1.5}$. For the generation of the observed data, we first generated the the follow-up time τ_i's from the uniform distribution over the interval [3, 4] and the number of observation times K_i for subject i based on the Poisson distribution with the mean function

$$\Lambda_{ih}(\tau_i \mid x_i, b_i) = \lambda_{0h}(t)\, \tau_i \, \exp(x_i^\top \boldsymbol{\gamma} + b_i)\,,$$

where $\lambda_{0h}(t) = 1/4$ and $\boldsymbol{\gamma} = \underbrace{(0.1, \ldots, 0.1)}_{p}$. For the censoring intervals, we took $U_{i1}, \ldots, U_{iK_i}$ to be the order statistics of a random sample of size K_i from the uniform distribution over $(0, \tau_i)$, $i = 1, 2, \ldots, n$. The simulation results given below are based on sample size $n = 300$ or 500 with 100 replications.

Table 1 presents the results obtained on the covariate selection with $p = 10$,$q = 3$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_a^\top, \beta_b)^\top$, $\boldsymbol{\beta}_a = (0.6, 0.6, 0, 0, 0, 0, 0, 0, 0.6)^\top$, $\beta_b = 0.2$. The results

include the averaged number of non-zero estimates of the parameters whose true values are not zero (TP), and the averaged number of non-zero estimates of parameters whose true values are zero (FP), the median (MMSE) of the mean weighted squared errors (MSE), and the standard deviation (SD) of the MSE. And define MSE to be $(\hat{\boldsymbol{\beta}}_a^* - \boldsymbol{\beta}_a)^\top E(x^\top x)(\hat{\boldsymbol{\beta}}_a^* - \boldsymbol{\beta}_a)$, where $\hat{\boldsymbol{\beta}}_a^*$ denote the BAR estimator of $\boldsymbol{\beta}_a$. It is easy to see that TP and FP provide the estimates of the true and false positive probabilities, respectively. In addition, we considered other penalty functions (LASSO, ALASSO, SCAD, SICA, SELO or MCP) and for the results, we took m, the degree of Bernstein polynomials, to be 3 and used BIC criterion to select the tuning parameter λ_n. From Table 1, one can see that the proposed procedure seems to perform well no matter which penalty function was used, especially in terms of TP which measuring the true positive selection. As expected, the proposed BAR approach presents the smallest MMSE and FP in all methods considered. Also the proposed approach generally yields the largest TP among all except the procedure based on the LASSO penalty. In addition, the performance does not seem to depend on the cumulative baseline hazard function for all methods. The results given in Table 2 were obtained in the similar set-ups as Table 1 except that $p = 30$ and $\boldsymbol{\beta}_a = (0.6, 0.6, \underbrace{0, \ldots, 0}_{p-3}, 0.6)^\top$, $\beta_b = 0.2$. The covariate selection performs similar to those shown in Table 1 and indicates that the estimation and variable selection of the proposed method seems to be robust for number of variables.

To see the performance of the proposed approach for different observation process, we repeated the set-ups in Table 1 but generated the number of observation times K_i from the mixed-Poisson process. Specifically, we first generated a random sample $\{\omega_1, \omega_2, \ldots, \omega_n\}$ from $\{-0.25, 0, 0.25\}$ with $P(\omega_i = -0.25) = P(\omega_i = 0.25) = 0.25$ and $P(\omega_i = 0) = 0.5$. For each i, given ω_i, b_i and x_i, the K_i were then generated from the Poisson process with the mean function

$$\Lambda_{ih}(\tau_i \mid x_i, b_i) = (1 - \omega_i)\lambda_{0h}(t)\, \tau_i \, \exp(x_i^\top \boldsymbol{\gamma} + b_i)\,.$$

The results on variable selection of the proposed method are presented in Table 3, and indicate that the proposed variable selection procedure seems to work well for the situations considered.

Note that in the above variable selection procedure, another interest is the performance of the proposed approach for the number of different non-zero variables. We considered the number of two different non-zero variables, namely, $q = 5$, $\boldsymbol{\beta}_a = (0.6, 0.6, 0.6, \underbrace{0, \ldots, 0}_{p-5}, 0.6, 0.6)^\top$ and $q = 8$, $\boldsymbol{\beta}_a = (0.6, 0.6, 0.6, 0.6, \underbrace{0, \ldots, 0}_{p-8}, 0.6, 0.6, 0.6, 0.6)^\top$, under the setting of $\rho = 0.5$, $\Lambda_0(t) = t^2$ and $n = 500$. These results will be summarized in Table 4. It is apparent that they all gave similar conclusions to those given by Tables 1-3. Morever, the estimation and variable selection of the proposed method also seems to be robust for the number of different non-zero variables.

In addition, to investigate the sensitivity of the proposed method for latent variables b_i and covariates x_i, we considered that the latent variables $b_i^* = \exp(b_i)$

Table 1 Results on covariate selection with $p = 10$ based on nonhomogeneous Poisson process

	Penalty	TP	FP	MMSE	SD	TP	FP	MMSE	SD
		$n = 300$				$n = 500$			
		$\Lambda_0(t) = \log(t+1) + t^{1.5}$							
$\rho = 0.2$	BAR	2.97	0.17	0.048	0.085	3.00	0.20	0.038	0.043
	LASSO	2.95	1.68	0.170	0.193	3.00	1.95	0.111	0.093
	ALASSO	2.95	0.60	0.110	0.144	3.00	0.39	0.065	0.074
	SCAD	2.93	0.14	0.039	0.108	3.00	0.09	0.029	0.038
	SICA	2.96	0.14	0.044	0.094	3.00	0.09	0.029	0.038
	SELO	2.97	0.14	0.045	0.090	3.00	0.10	0.029	0.041
	MCP	2.96	0.20	0.043	0.104	3.00	0.10	0.029	0.042
$\rho = 0.5$	BAR	2.92	0.17	0.060	0.092	3.00	0.18	0.041	0.048
	LASSO	2.98	1.68	0.181	0.135	3.00	2.17	0.121	0.081
	ALASSO	2.93	0.50	0.144	0.135	3.00	0.52	0.085	0.076
	SCAD	2.82	0.17	0.061	0.121	2.98	0.10	0.036	0.049
	SICA	2.82	0.10	0.059	0.113	2.97	0.10	0.036	0.054
	SELO	2.84	0.11	0.059	0.109	2.97	0.10	0.036	0.054
	MCP	2.85	0.16	0.056	0.109	2.97	0.11	0.037	0.056
		$\Lambda_0(t) = t^2$							
$\rho = 0.2$	BAR	2.97	0.22	0.058	0.083	3.00	0.17	0.033	0.036
	LASSO	2.88	1.45	0.187	0.213	3.00	1.38	0.126	0.091
	ALASSO	2.97	0.55	0.122	0.141	3.00	0.34	0.058	0.058
	SCAD	2.90	0.12	0.050	0.119	3.00	0.07	0.023	0.036
	SICA	2.88	0.13	0.052	0.127	3.00	0.06	0.024	0.035
	SELO	2.88	0.14	0.054	0.127	3.00	0.06	0.024	0.035
	MCP	2.91	0.20	0.063	0.115	3.00	0.06	0.023	0.035
$\rho = 0.5$	BAR	2.95	0.23	0.049	0.082	3.00	0.17	0.030	0.036
	LASSO	2.97	1.51	0.180	0.183	3.00	1.43	0.103	0.084
	ALASSO	2.91	0.45	0.147	0.189	3.00	0.30	0.068	0.060
	SCAD	2.76	0.17	0.048	0.253	3.00	0.08	0.025	0.036
	SICA	2.85	0.12	0.044	0.121	3.00	0.06	0.026	0.035
	SELO	2.87	0.14	0.044	0.118	3.00	0.07	0.027	0.037
	MCP	2.82	0.16	0.048	0.182	3.00	0.09	0.027	0.039

were generated from $\Gamma(2, 1/2)$, where $\Gamma(\nu_1, \nu_2)$ stands for the gamma distribution with shape parameter ν_1 and rate parameter ν_2, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$, where $x_{i1}, x_{i2}, \ldots, x_{ip}$ are independently generated from the uniform distribution over $(-3, 3)$ with $p = 10, n = 300$ and $\Lambda_0(t) = t^2$. Table 5 summarizes the estimation results based on 100 replications. The parameter estimates perform similarly to those shown in Tables 1 and 2, indicating that the proposed method is robust to the different distribution of the latent variables and covariates.

Table 2 Results on covariate selection with $p = 30$ based on nonhomogeneous Poisson process

	Penalty	TP	FP	MMSE	SD	TP	FP	MMSE	SD
		$n = 300$				$n = 500$			
		$\Lambda_0(t) = \log(t+1) + t^{1.5}$							
$\rho = 0.2$	BAR	2.97	0.39	0.081	0.098	3.00	0.28	0.061	0.052
	LASSO	2.87	1.51	0.432	0.261	3.00	1.49	0.285	0.139
	ALASSO	2.90	0.89	0.193	0.224	3.00	0.66	0.098	0.089
	SCAD	2.97	0.29	0.060	0.101	3.00	0.17	0.028	0.045
	SICA	2.95	0.32	0.067	0.109	3.00	0.20	0.029	0.046
	SELO	2.95	0.31	0.069	0.105	3.00	0.21	0.029	0.047
	MCP	2.95	0.37	0.063	0.108	3.00	0.20	0.028	0.048
$\rho = 0.5$	BAR	2.94	0.49	0.083	0.158	3.00	0.50	0.057	0.055
	LASSO	2.84	1.77	0.409	0.306	3.00	2.31	0.253	0.144
	ALASSO	2.85	0.94	0.207	0.244	3.00	0.95	0.123	0.095
	SCAD	2.86	0.24	0.054	0.175	3.00	0.12	0.027	0.046
	SICA	2.88	0.21	0.053	0.171	3.00	0.22	0.032	0.055
	SELO	2.87	0.32	0.065	0.175	3.00	0.23	0.034	0.052
	MCP	2.82	0.27	0.061	0.184	3.00	0.18	0.030	0.052
		$\Lambda_0(t) = t^2$							
$\rho = 0.2$	BAR	2.95	0.57	0.086	0.129	3.00	0.47	0.063	0.061
	LASSO	2.77	1.66	0.422	0.276	3.00	1.72	0.244	0.123
	ALASSO	2.89	0.98	0.197	0.198	3.00	0.69	0.096	0.076
	SCAD	2.66	0.22	0.084	0.263	3.00	0.20	0.027	0.056
	SICA	2.83	0.32	0.068	0.181	3.00	0.27	0.032	0.061
	SELO	2.83	0.36	0.071	0.187	3.00	0.25	0.034	0.061
	MCP	2.85	0.53	0.097	0.187	3.00	0.28	0.028	0.067
$\rho = 0.5$	BAR	2.95	0.58	0.086	0.129	3.00	0.49	0.057	0.059
	LASSO	2.94	1.80	0.347	0.228	3.00	1.99	0.217	0.137
	ALASSO	2.88	1.03	0.184	0.190	2.98	0.73	0.110	0.127
	SCAD	2.81	0.29	0.051	0.162	2.99	0.15	0.030	0.060
	SICA	2.85	0.35	0.057	0.154	3.00	0.25	0.035	0.057
	SELO	2.85	0.45	0.064	0.158	3.00	0.26	0.037	0.054
	MCP	2.85	0.44	0.063	0.174	2.99	0.28	0.037	0.067

5 An application

Now we apply the proposed methodology to a set of informatively case K interval-censored data arising from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI study is an ongoing, prospective, longitudinal multicenter study designed to investigate clinical, imaging, genetic and biochemical biomarkers for early detecting of the Alzheimer's Disease (AD) and tracking its progression. In the study, each participant is visited intermittently sev-

Table 3 Results on covariate selection with $p = 10$ and $n = 300$ based on nonhomogeneous Mixed Poisson process

	Penalty	TP	FP	MMSE	SD	TP	FP	MMSE	SD
		$\Lambda_0(t) = \log(t+1) + t^{1.5}$				$\Lambda_0(t) = t^2$			
$\rho = 0.2$	BAR	2.95	0.18	0.051	0.092	2.98	0.14	0.056	0.072
	LASSO	3.00	1.59	0.183	0.157	2.94	1.49	0.169	0.192
	ALASSO	2.96	0.56	0.110	0.111	2.97	0.45	0.124	0.159
	SCAD	2.91	0.14	0.042	0.107	2.83	0.08	0.048	0.203
	SICA	2.92	0.15	0.043	0.108	2.91	0.09	0.053	0.123
	SELO	2.93	0.15	0.044	0.102	2.93	0.08	0.050	0.107
	MCP	2.92	0.16	0.045	0.106	2.90	0.12	0.053	0.133
$\rho = 0.5$	BAR	2.91	0.18	0.057	0.093	2.92	0.18	0.050	0.100
	LASSO	3.00	1.71	0.162	0.135	3.00	1.33	0.169	0.108
	ALASSO	2.95	0.47	0.131	0.124	2.90	0.41	0.148	0.165
	SCAD	2.86	0.08	0.051	0.100	2.74	0.07	0.047	0.203
	SICA	2.87	0.07	0.053	0.093	2.81	0.08	0.045	0.120
	SELO	2.87	0.08	0.053	0.094	2.81	0.09	0.046	0.122
	MCP	2.88	0.17	0.056	0.097	2.82	0.16	0.047	0.127

Table 4 Results on covariate selection for the number of different non-zero variables based on the setting of $\rho = 0.5$, $\Lambda_0(t) = t^2$ and $n = 500$

	Penalty	TP	FP	MMSE	SD	TP	FP	MMSE	SD
		$b \sim N(0, 1)$				$\exp(b) \sim \Gamma(2, 1/2)$			
		Poisson process							
$p = 30, q = 5$	BAR	4.94	0.21	0.099	0.091	5.00	0.26	0.083	0.060
	LASSO	4.99	2.58	0.423	0.263	5.00	2.69	0.389	0.205
	ALASSO	4.94	1.03	0.232	0.192	4.98	1.08	0.192	0.125
	SCAD	4.90	0.20	0.081	0.106	4.97	0.21	0.061	0.068
	SICA	4.85	0.24	0.087	0.115	4.97	0.23	0.062	0.073
	SELO	4.86	0.19	0.087	0.111	4.97	0.29	0.066	0.076
	MCP	4.89	0.19	0.082	0.107	4.98	0.33	0.064	0.081
$p = 50, q = 8$	BAR	7.90	0.31	0.195	0.123	7.97	0.44	0.185	0.091
	LASSO	8.00	4.03	1.018	0.457	8.00	4.72	0.896	0.379
	ALASSO	7.87	2.18	0.503	0.297	7.93	2.32	0.442	0.279
	SCAD	7.81	0.19	0.168	0.123	7.92	0.16	0.116	0.095
	SICA	7.61	0.37	0.240	0.155	7.86	0.31	0.134	0.105
	SELO	7.61	0.33	0.230	0.157	7.86	0.44	0.142	0.109
	MCP	7.72	0.28	0.202	0.136	7.90	0.39	0.137	0.103
		Mixed Poisson process							
$p = 30, q = 5$	BAR	4.95	0.22	0.102	0.076	4.99	0.34	0.093	0.077
	LASSO	5.00	2.59	0.447	0.257	5.00	3.11	0.313	0.206

Table 4 continued

	Penalty	TP	FP	MMSE	SD	TP	FP	MMSE	SD
		$b \sim N(0, 1)$				$\exp(b) \sim \Gamma(2, 1/2)$			
	ALASSO	4.95	1.08	0.234	0.155	4.97	1.17	0.165	0.133
	SCAD	4.94	0.09	0.068	0.075	4.98	0.28	0.065	0.087
	SICA	4.92	0.11	0.075	0.079	4.95	0.28	0.067	0.082
	SELO	4.93	0.17	0.073	0.084	4.95	0.33	0.076	0.083
	MCP	4.94	0.16	0.073	0.079	4.97	0.35	0.071	0.086
$p = 50, q = 8$	BAR	7.97	0.49	0.184	0.111	7.97	0.42	0.158	0.102
	LASSO	8.00	5.02	1.017	0.435	8.00	4.46	0.828	0.402
	ALASSO	7.94	2.34	0.502	0.284	7.94	2.12	0.423	0.264
	SCAD	7.85	0.19	0.138	0.139	7.92	0.21	0.115	0.095
	SICA	7.70	0.45	0.200	0.151	7.82	0.30	0.147	0.112
	SELO	7.67	0.34	0.188	0.152	7.83	0.32	0.142	0.115
	MCP	7.78	0.45	0.178	0.164	7.88	0.43	0.135	0.125

Table 5 Results on covariate selection with $p = 10, \Lambda_0(t) = t^2$ and $n = 300$

Penalty	TP	FP	MMSE	SD	TP	FP	MMSE	SD
	$b \sim N(0, 1)$				$\exp(b) \sim \Gamma(2, 1/2)$			
Poisson process								
BAR	3.00	0.22	0.076	0.106	3.00	0.22	0.077	0.094
LASSO	3.00	1.82	0.304	0.230	3.00	1.95	0.267	0.161
ALASSO	3.00	0.33	0.151	0.149	3.00	0.31	0.125	0.115
SCAD	2.95	0.42	0.078	0.238	2.99	0.09	0.062	0.137
SICA	3.00	0.15	0.064	0.112	3.00	0.10	0.064	0.097
SELO	3.00	3.04	0.064	0.113	3.00	3.01	0.062	0.098
MCP	3.00	0.16	0.064	0.113	3.00	0.12	0.064	0.103
Mixed Poisson process								
BAR	3.00	0.28	0.084	0.109	3.00	0.20	0.080	0.101
LASSO	3.00	1.92	0.317	0.211	3.00	2.39	0.218	0.180
ALASSO	3.00	0.33	0.163	0.157	3.00	0.30	0.131	0.124
SCAD	2.94	0.33	0.083	0.257	3.00	0.20	0.072	0.108
SICA	2.98	0.13	0.075	0.168	3.00	0.14	0.070	0.102
SELO	2.98	3.01	0.081	0.169	3.00	3.00	0.069	0.107
MCP	3.00	0.10	0.073	0.098	3.00	0.14	0.068	0.110

eral times, and all patients are grouped into three groups based on their cognitive conditions, cognitively normal (CN), mild cognitive impairment (MCI) and AD. One purpose of the study is to estimate the AD conversion time, which is usually used to monitor the progress of participants and also to identify the risk factors for AD

Table 6 The selected and estimated covariate effects for the ADNI data

Covariate	BAR	LASSO	ALASSO	SCAD	SICA	SELO	MCP
PTGENDER	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
AGE	− 0.096	− 0.328	− 0.300	− 0.451	− 0.390	− 0.390	− 0.456
	(0.215)	(0.171)	(0.196)	(0.298)	(0.276)	(0.275)	(0.289)
PTEDUCAT	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
APOE4	0.255	0.209	0.157	0.241	0.228	0.232	0.266
	(0.175)	(0.122)	(0.153)	(0.207)	(0.201)	(0.202)	(0.211)
ADAS11	–(–)	0.002	–(–)	–(–)	–(–)	–(–)	–(–)
		(0.106)					
ADAS13	0.425	0.370	0.462	0.432	0.476	0.475	0.422
	(0.260)	(0.173)	(0.269)	(0.254)	(0.305)	(0.298)	(0.311)
ADASQ4	–(–)	0.032	–(–)	–(–)	–(–)	–(–)	–(–)
		(0.156)					
CDRSB	0.209	0.170	0.044	0.236	0.185	0.195	0.222
	(0.180)	(0.127)	(0.152)	(0.191)	(0.202)	(0.203)	(0.200)
MMSE	–(–)	− 0.068	–(–)	–(–)	–(–)	–(–)	–(–)
		(0.114)					
RAVLT_i	− 0.588	− 0.468	− 0.395	− 0.641	− 0.608	− 0.611	− 0.596
	(0.247)	(0.188)	(0.277)	(0.257)	(0.268)	(0.271)	(0.291)
RAVLT_l	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
RAVLT_f	–(–)	–(–)	− 0.064	–(–)	–(–)	–(–)	–(–)
			(0.258)				
RAVLT_pf	–(–)	0.061	0.209	–(–)	–(–)	–(–)	–(–)
		(0.207)	(0.327)				

Table 6 continued

Covariate	BAR	LASSO	ALASSO	SCAD	SICA	SELO	MCP
DIGITSCOR	–(–)	− 0.075	–(–)	–(–)	–(–)	–(–)	–(–)
		(0.122)					
TRABSCOR	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
FAQ	0.415	0.385	0.417	0.442	0.461	0.458	0.415
	(0.197)	(0.159)	(0.215)	(0.187)	(0.216)	(0.217)	(0.200)
Ventricles	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
Hippocampus	− 0.305	− 0.216	− 0.248	− 0.266	–(–)	–(–)	− 0.316
	(0.256)	(0.184)	(0.218)	(0.232)			(0.268)
WholeBrain	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
Entorhinal	− 0.260	− 0.209	− 0.030	− 0.234	− 0.337	− 0.338	− 0.273
	(0.228)	(0.142)	(0.185)	(0.222)	(0.241)	(0.241)	(0.248)
Fusiform	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)	–(–)
MidTemp	− 0.638	− 0.500	− 0.576	− 0.425	− 0.494	− 0.495	− 0.648
	(0.268)	(0.230)	(0.278)	(0.243)	(0.282)	(0.281)	(0.270)
ICV	0.412	0.227	0.222	–(–)	–(–)	–(–)	0.434
	(0.286)	(0.197)	(0.218)				(0.294)

conversion time. One variable of interest is the time (in year) from the baseline visit date to the AD conversion.

For the analysis, by following Li et al. (2017), we also focus on the data from 298 participants in the MCI group who had at least one follow-up and underwent 2421 examinations in total at random observation points for whom the complete information on 23 demographic, clinical and genetic factors are available. These 23 demographic and clinical covariates are identified as possible important factors associated with the AD conversion by Li et al. (2017). In the application, let T_i denote the AD conversion time for subject i. The observed information for each participant includes the number of examinations or observation number K_i, the observation points U_{ij} and the AD conversion indicators δ_{ij}. To identify the important covariates or risk factors that have effects on the risk of developing AD and estimate their effects, we considered proportional odds frailty model (1). And the cumulative baseline hazard function corresponding to model (1) can then be denoted by

$$\Lambda(t \mid x_i, b_i) = \log[1 + \Lambda_0(t) \exp(x_i^\top \boldsymbol{\beta}_a + b_i \beta_b)].$$

Table 6 presents the variable selection results for the ADNI data and it also gives the estimated covariate effects along with the estimated standard errors in brackets based on the bootstrap procedure, which are based on 100 bootstrap samples randomly drawn with replacement from the data. In the simulation study, besides the BAR penalty function, we also apply LASSO, ALASSO, SCAD, SICA, SELO and MCP penalty functions to the real data set and present the obtained results in the table for comparison. The results suggest that the AD conversion is clearly related to clinical factors, including AGE, APOES4, ADAS13, CDRSB, RAVLT_i, FAQ, Hippocampus, Entorhinal, MidTemp and ICV. This performance is consistent with using other penalty functions except for variable ICV. And seven factors, PTGENDER, PTENDUCAT, RAVLT_l, TRABSCOR, Ventricles, WholeBrain and Fusitorm, are not selected by any penalty functions, thereby indicating that these seven factors had no relationship on the hazards function of AD conversion time. In addition, the results suggest that six factors, ADAS11, ADASQ4, MMSE, RAVLT_f, RAVLT_pf, DIGITSCOR, have no relationship with or significant influence on the hazards function of AD conversion time except using the LASSO- and ALASSO-based method.

Compared with the conclusions obtained by Li et al. (2017) based on the PH model, although we obtain the similar important variables, the significance of the variables is different. For instance, although the selected the AGE, APOES4, ADAS13, CDRSB, Hippocampus, Entorhinal, and ICV as non-zero covariates, the estimation suggests that it did not have any significant effect on the development of the AD conversion. In addition, in the selected non-zero variables, FAQ has a significantly positive effect on the hazards of AD conversion, suggesting that FAQ increases the risk of AD conversion. Hence, maintaining a normal FAQ is beneficial to the prevention of AD conversion. And RAVLT_i and MidTemp have significantly negative effects on the hazards of the AD conversion, thereby indicating that maintaining a high level of fRAVLT_i and MidTemp inhibits the development of the AD conversion. Moreover, we also obtain effect $\hat{\beta}_b = -0.610$ of latent or frailty variable with the estimated standard error

0.127. It seems to exist a strong correlation between the AD conversion time and the observation process. In other words, the participant's examination process seems to be a negative association with the AD conversion time.

6 Discussion and concluding remarks

This paper discussed the variable or covariate selection problem for the informatively case K interval-censored failure time data and a unified variable selection procedure was proposed under the proportional odds model. Specifically, the proposed variable selection was implemented by following a borrow strength idea. We first estimated the frailty variables for the observation process model, and then proposed a unified penalized variable selection procedure. The proposed method involves minimizing a negative sieve log-likelihood function plus a broken adaptive ridge penalty to simultaneous estimation and covariate selection under the proportional odds model. The major advantage of the proposed method is that it is for the general type of failure time data, case K interval-censored data, which includes all of other types (interval-censored data, current status data or right-censored data). Another advantage is that it models the proportional odds model for the dependent or informative censoring, which often occurs in medical studies as well as other studies. In addition, like some existing methods(SCAD, ALASSO, SICA), the oracle property of the proposed approach is established and the numerical studies indicate that the proposed approach works well for practical situations.

Some work remains to be done for further research. In the previous section, it is assumed that the failure time of interest coming from the proportional odds model. However, some other models such as the additive hazards model or the linear transformation model may be more appropriate sometimes. An optimal penalty parameter needs to be selected through the cross validation or the BIC criteria, but these implementation processes are very time-consuming. A variable selection procedure, which overcomes the challenge of tuning parameter selection, is also an important direction for future research. Although Wang et al. (2020a, b) proposed a novel approach to tuning parameter selection of Lasso for the high-dimensional regression, however, their method focused only on complete data and linear regression model, and is inapplicable to the semi- or non-parametric model for the censoring data. Constructing a new variable selection procedure, which overcomes the challenge of tuning parameter selection, is highly challenging and worthy of further investigation, in the presence of informatively case K interval-censored failure time data.

Acknowledgements This work was partly supported by the National Natural Science Foundation of China Grant (No. 12271060), China Postdoctoral Science Foundation (No. 2021M700536), Outstanding Youth Fund Project of Jilin Natural Science Foundation (No. 20230101371JC) and Major science and technology projects of Jilin Provincial Department of science and technology (No. 20210301038GX).

Declarations

Conflict of interest The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Appendix A Asymptotic properties of $\hat{\boldsymbol{\beta}}^*$

In this Appendix, we will sketch the proof of the asymptotic properties of the proposed BAR estimator $\hat{\boldsymbol{\beta}}^*$ described in Theorem 1. For this, the required regularity conditions are given as follows.

(C1) For the latent variable b, the variance of $\exp(b)$ is bound and there exists a positive small constant $\varepsilon > 0$, such that $\exp(b) > \varepsilon$ almost surely.

(C2) For the follow-up time τ and latent variable b, (I) $P(\tau \geq \tau_0, \exp(b) > 0) > 0$, (II) the function $Q(s) = E[\exp(b) I(\tau \geq s)]$ is continuous for $s \in [0, \tau_0]$.

(C3) (I) The matrix $E(xx^\top)$ is non singular with x being bounded. That is, there exists $x_0 > 0$ such that $P(||x|| \leq x_0) = 1$. (II) The set $\mathcal{B}$ is a compact of $\mathcal{R}^{p+1}$ and $\boldsymbol{\beta}_0$ is an interior point of $\mathcal{B}$.

(C4) (I) For subject i, the union of the support of $U_{i,j}$ is contained in the interval $[a^*, b^*]$, where $0 < a^* < b^* < \infty$ and $j = 1, \ldots, K_i$. (II) There exists a positive η^* such that $P((U_{i,j} - U_{i,j-1}) \geq \eta^*) = 1$ for subject i,where $j = 2, \ldots, K_i$.

(C5) The function $\lambda_0(.)$ is continuously differentiable up to order r in $[u, v]$ and $a^{-1} < \Lambda_0(u) < \Lambda_0(v) < a$ for some positive constant a.

(C6) (I) There exists a compact neighborhood of $\mathcal{B}_0$ of the true value of $\boldsymbol{\beta}_0$ such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}_0} \left\| n^{-1} \boldsymbol{\Omega}_n(\boldsymbol{\beta}) - I(\boldsymbol{\beta}_0) \right\| \xrightarrow{as} 0.$$

Where $I(\boldsymbol{\beta}_0)$ is a positive definite $(p+1) \times (p+1)$ matrix, which is defined in the Appendix. (II) There exists a constant $c > 1$ such that $c^{-1} < \lambda_{min}(n^{-1}\boldsymbol{\Omega}_n) \leq \lambda_{max}(n^{-1}\boldsymbol{\Omega}_n) < c$ for sufficiently large n, where λ_{min} and λ_{max} stand for the smallest and largest eigenvalues of the matrix, respectively.

(C7) The exist positive constant a_0 and a_1 such that $a_0 \leq | \beta_{0,k} | \leq a_1, k = 1, 2, \ldots, q$.

(C8) As $n \to \infty$, $(p+1)^2 q/\sqrt{n} \to 0$, $\lambda_n/\sqrt{n} \to 0$, $\varrho_n/\sqrt{n} \to 0$, $\lambda_n \sqrt{(q+1)/n} \to 0$ and $\lambda_n^2/((p+1)\sqrt{n}) \to \infty$.

It has been point out that if the conditions (C1)–(C3) hold, $\lambda_{0h}(t)$ and α are consistent and asymptotic normal (Huang and Wang 2004). And latent variable b's as the functions of the above terms, Based on the delta method, one can easily shows that as $n \to \infty$, b's are consistent and converge in distribution to a normal random variable. Conditions (C3)–(C5) are necessary for the existence and consistence of the sieve maximum likelihood estimator of $\Lambda_0(t)$ and usually satisfied in practice (Huang and Rossini 1997). Conditions (C6) assume that $n^{-1}\boldsymbol{\Omega}_n(\boldsymbol{\beta})$ is positive definite almost surely and its eigenvalues are bounded away from zero and infinity. Condition (C7) assumes that the nonzero coefficients are uniformly bounded away from zero and

infinity, and condition (C8) gives some sufficient, but not necessary, conditions need to prove the numerical convergence and asymptotic properties of the BAR estimator.

In the next, given $\hat{b}$, we will sketch the proof of the asymptotic properties of the proposed BAR estimator $\hat{\boldsymbol{\beta}}^*$. The following Lemmas are needed.

Lemma 1 *Let the ridge estimator*

$$\boldsymbol{\beta}^o = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_p(\boldsymbol{\beta}) + \varrho_n \sum_{s=1}^{p} \beta_s^2 \right\},$$

and suppose that the conditions (C1)–(C8) hold. Then we have that

$$||\boldsymbol{\beta}^o - \boldsymbol{\beta}_0|| = O_p(\sqrt{(p+1)/n}).$$

Proof of Lemma 1 Let

$$\mathcal{L}(\boldsymbol{\beta}) = 2l_p(\boldsymbol{\beta}) - \varrho_n \sum_{s=1}^{p} \beta_s^2,$$

and $p_\varrho(\beta_{0s}) = \beta_{0s}^2 \varrho_n/n, s = 1, 2, \ldots, p, p+1$. The first and second derivation of $p_\varrho(\beta_{0s})$ are $\dot{p}_\varrho(\beta_{0\,s}) = 2\beta_{0\,s}\varrho_n/n$ and $\ddot{p}_\varrho(\beta_{0\,s}) = 2\varrho_n/n$, respectively.

Define

$$a_n = \max\{|\ \dot{p}_\varrho(\beta_{0s})\ |\colon \beta_{0s} \neq 0, s = 1, 2, \ldots, p+1\},$$
$$b_n = \max\{|\ \ddot{p}_\varrho(\beta_{0s})\ |\colon \beta_{0s} \neq 0, s = 1, 2, \ldots, p+1\}.$$

According to Conditions (C7)–(C8), we obtain that $a_n \leq 2a_1\varrho_n/n = o(n^{-1/2})$, and $b_n \leq 2\varrho_n/n = (n^{-1/2})$. Therefore, $a_n \to 0, b_n \to 0$. Let $\varphi_n = \sqrt{p+1}(n^{-1/2} + a_n)$, then using the similar manipulation as those in Cai et al. (2005), we have prove that for any $\epsilon > 0$, there exists a large constant C_0 such that

$$\sup_{||\nu||=C_0} \left\{ \mathcal{L}(\boldsymbol{\beta} + \varphi_n \nu) < \mathcal{L}(\boldsymbol{\beta}_0) \right\} \geq 1 - \epsilon,$$

which implies that there exists a local maximiser $\boldsymbol{\beta}^o$ with that $||\boldsymbol{\beta}^o - \boldsymbol{\beta}_0|| = O_p(\sqrt{(p+1)/n})$. We complete this proof. □

To prove Theorem 1, we need to describe the following notations and Lemma 2 and 3. Define

$$\begin{pmatrix} \vartheta_1(\boldsymbol{\beta}) \\ \vartheta_2(\boldsymbol{\beta}) \end{pmatrix} \equiv g(\boldsymbol{\beta}) = (\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta}))^{-1} \boldsymbol{\xi}_n \boldsymbol{\Omega}_n, \tag{G1}$$

where $\boldsymbol{D}(\boldsymbol{\beta}) = diag(\beta_1, \ldots, \beta_{p-1}, \beta_p, 0)$, $\boldsymbol{\Omega}_n = \boldsymbol{\Omega}_n(\boldsymbol{\beta}) = \boldsymbol{Z}^\top \boldsymbol{Z}$, and $\boldsymbol{\xi}_n = \boldsymbol{\xi}_n(\boldsymbol{\beta}) = \boldsymbol{Z}^\top \boldsymbol{y}$, and partition the matrix $n^{-1}\boldsymbol{\Omega}_n$ into

$$(n^{-1}\boldsymbol{\Omega}_n)^{-1} = \begin{pmatrix} A & B \\ B^\top & G \end{pmatrix},$$

where A is a $(q+1) \times (q+1)$ matrix.

Note that since $\boldsymbol{\Omega}_n$ is nonsingular, it follows by multiplying $\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta}))$ and subtracting $\boldsymbol{\beta}_0$ on both sides of (G1) that we have

$$\begin{pmatrix} \vartheta_1(\boldsymbol{\beta}) - \boldsymbol{\beta}_{(01)} \\ \vartheta_2(\boldsymbol{\beta}) \end{pmatrix} + \frac{\lambda_n}{n} \begin{pmatrix} A\boldsymbol{D}_1(\boldsymbol{\beta}_1)\vartheta_1 & B\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2 \\ B^\top \boldsymbol{D}_1(\boldsymbol{\beta}_1)\vartheta_1 & G\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2 \end{pmatrix} = \boldsymbol{\beta}^o - \boldsymbol{\beta}_0, \quad \text{(G2)}$$

where $\boldsymbol{D}_1(\boldsymbol{\beta}_1) = diag(\beta_1^{-2}, \ldots, \beta_q^{-2}, 0)$, and $\boldsymbol{D}_2(\boldsymbol{\beta}_2) = diag(\beta_{q+1}^{-2}, \ldots, \beta_p^{-2})$. In the next, we will describe the following Lemma 2 and 3, and complete this proofs.

Lemma 2 *Let $\{\Delta_n\}$ be a sequence of positive real numbers,with that $\Delta_n \to \infty$ and $\Delta_n^2(p+1)/\lambda_n \to 0$. Let $H_n \equiv \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top : \boldsymbol{\beta}_1 \in [1/K_0, K_0]^{q+1}, ||\boldsymbol{\beta}_2|| \le \Delta_n\sqrt{(p+1)/n}\}$, where $K_0 > 1$ is a constant such that $\boldsymbol{\beta}_{01} \in [1/K_0, K_0]^{q+1}$. Suppose that the regular conditions (C1)–(C8) hold. Then, with probability tending to 1, we have*

(I) $\sup_{\boldsymbol{\beta} \in H_n} \left(\frac{||\vartheta_2(\boldsymbol{\beta})||}{||\boldsymbol{\beta}_2||} \right) < \frac{1}{C_0}$ *for some constant $C_0 > 1$.*

(II) $g(.)$ *is a mapping from H_n to H_n.* □

Proof of Lemma 2 Let $\vartheta_1 = \vartheta_1(\boldsymbol{\beta})$, $\vartheta_2 = \vartheta_2(\boldsymbol{\beta})$. Based on the Lemma 1, it follows from (G2) that

$$\sup_{\boldsymbol{\beta} \in H_n} \left\| \vartheta_2 + \frac{\lambda_n}{n} B^\top \boldsymbol{D}_1(\boldsymbol{\beta}_1)\vartheta_1 + \frac{\lambda_n}{n} G\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2 \right\| = O_p(\sqrt{(p+1)/n}). \quad \text{(G3)}$$

For given a constant C, and the matrix $n^{-1}\Omega_n$ has fact that

$$||B^\top B|| - ||A^2|| \le ||BB^\top + A^2|| \le ||(n^{-1}\boldsymbol{\Omega}_n(\boldsymbol{\beta}))^{-2}|| \le C^2$$

by conditions (C6)(II) and (G2). Thus, we derive matrix B with that

$$||B|| \le \sqrt{2}C. \quad \text{(G4)}$$

Furthermore, note that $\boldsymbol{\beta}_1 \in [1/K_o, K_0]^{q+1}$ and $||\vartheta_1|| \le ||g(\boldsymbol{\beta})|| \le ||\boldsymbol{\beta}^o|| = O_p(\sqrt{p+1})$. Combining (G4), Condition (C6)(II) and (C7), we have

$$\sup_{\boldsymbol{\beta} \in H_n} \left\| \frac{\lambda_n}{n} B^\top \boldsymbol{D}_1(\boldsymbol{\beta}_1)\vartheta_1 \right\| = O_p(\sqrt{(p+1)/n}). \quad \text{(G5)}$$

Since $\lambda_{min}(G) > C^{-1}$, it follows from (G2) that with probability tending to 1,

$$C^{-1}\left\|\frac{\lambda_n}{n}\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2\right\| - ||\vartheta_2|| \le \sup_{\boldsymbol{\beta}\in H_n}\left\|\vartheta_2 + \frac{\lambda_n}{n}B^\top \boldsymbol{D}_1(\boldsymbol{\beta}_1)\vartheta_1\right\| = O_p(\sqrt{(p+1)/n}) \le \Delta_n\sqrt{(p+1)/n}. \tag{G6}$$

Let $m_{\vartheta_2} = \vartheta_2/\boldsymbol{\beta}_2$. It then follows from the Cauchy–Schwarz inequality and the assumption $||\boldsymbol{\beta}_2|| \le \Delta_n\sqrt{(p+1)/n}$ that

$$||m_{\vartheta_2}|| \le ||\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2||\Delta_n\sqrt{(p+1)/n} \tag{G7}$$

and

$$||\vartheta_2|| = ||\boldsymbol{D}_2(\boldsymbol{\beta}_2)^{-1/2}m_{\vartheta_2}|| \le ||m_{\vartheta_2}||||\boldsymbol{\beta}_2|| \le ||m_{\vartheta_2}||\Delta_n\sqrt{(p+1)/n} \tag{G8}$$

for all large n.

Thus, from (G6) and (G8), we have the following inequality

$$\frac{\lambda_n}{nC}\frac{\sqrt{n}}{\Delta_n\sqrt{p+1}}||m_{\vartheta_2}|| - ||m_{\vartheta_2}||\frac{\Delta_n\sqrt{p+1}}{\sqrt{n}} \le \frac{\Delta_n\sqrt{p+1}}{\sqrt{n}}.$$

Immediately based on $p\Delta_n^2/\lambda_n \to 0$, we can derive

$$||m_{\vartheta_2}|| \le \frac{1}{\frac{\lambda_n}{(p+1)\Delta_n^2 C} - 1} < \frac{1}{c_0}, (c_0 > 1), \tag{G9}$$

with probability tending to 1. Hence, as $n \to \infty$, it from (G8) and (G9) that

$$||\vartheta_2|| \le ||\boldsymbol{\beta}_2|| \le \Delta_n\sqrt{(p+1)/n} \to 0. \tag{G10}$$

It implies that conclusion (I) of Lemma 2 holds.

Here, we prove conclusion (II) and need to verity that $\vartheta_1 \in [1/K_0, K_0]^{q+1}$ with probability tending to 1, since (G10) has showed that $||\vartheta_1|| \le \Delta_n\sqrt{(p+1)/n}$ with probability tending to 1. Following arguments similar to those in the proof of conclusion (I), based on the conditions C6(II), $\boldsymbol{\beta}_1 \in [1/K_0, K_0]^{q+1}$ and $||\vartheta_2|| < O_p(\sqrt{p+1})$, we obtain

$$\sup_{\boldsymbol{\beta}\in H_n}\left\|\frac{\lambda_n}{n}A\boldsymbol{D}_1(\boldsymbol{\beta}_1)\vartheta_1\right\| = o_p(\sqrt{(p+1)/n}).$$

According to the formula (G2), we obtain

$$\sup_{\boldsymbol{\beta}\in H_n}\left\|\vartheta_1 - \boldsymbol{\beta}_{(01)} + \frac{\lambda_n}{n}B\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2\right\| = O_p(\sqrt{(p+1)/n}) \le \Delta_n\sqrt{(p+1)/n}. \tag{G11}$$

And from (G6) and (G10), we have

$$\left\| \frac{\lambda_n}{n} \boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2 \right\| \leq 2c\Delta_n\sqrt{(p+1)/n}. \tag{G12}$$

Hence, according to condition (C6)(II), we have that as $n \to \infty$ and with probability tending to 1,

$$\sup_{\boldsymbol{\beta}\in H_n} \left\| \frac{\lambda_n}{n} B\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2 \right\| \leq \frac{\lambda_n}{n}||B|| \sup_{\boldsymbol{\beta}\in H_n} \|\boldsymbol{D}_2(\boldsymbol{\beta}_2)\vartheta_2\| \leq \frac{2\sqrt{2}c^2\Delta_n\sqrt{p+1}}{\sqrt{n}}. \tag{G13}$$

Therefore, based on the above formula (G11) and (G13), we can obtain

$$\sup_{\boldsymbol{\beta}\in H_n} \|\vartheta_1 - \boldsymbol{\beta}_{(01)}\| \leq \frac{(2\sqrt{2}c^2+1)\Delta_n\sqrt{p+1}}{\sqrt{n}} \to 0$$

with probability tending to 1, which implies that $P\left(||\vartheta_1 - \boldsymbol{\beta}_{(01)}|| \leq \epsilon\right) \to 1$, for any $\epsilon > 0$. Thus it follows from $\boldsymbol{\beta}_{(01)} \in [1/K_0, K_0]^{q+1}$ that $\vartheta_1 \in [1/K_0, K_0]^{q+1}$ holds for large n, which implies that the conclusion (II) of Lemma 2 holds. This completes the proof. □

Lemma 3 *Suppose that the regularity conditions (C1)–(C8) holds. Then, the equation* $\vartheta_1^* = (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\vartheta_1^*))^{-1}\boldsymbol{\xi}_n^{(1)}$ *has a unique fixed-point* $\hat{\vartheta}_1$ *in the domain* $[1/K_0, K_0]^{q+1}$ *with probability tending to 1.*

Proof of Lemma 3 Define $\boldsymbol{Z}_1$ is the first $q+1$ columns of matrix $\boldsymbol{Z}$, measurement error $\varepsilon = \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}$, and

$$f(\vartheta_1^*) = (f_1(\vartheta_1^*), \ldots, f_q(\vartheta_1^*), f_{q+1}(\vartheta_1^*))^\top \equiv (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\vartheta_1^*))^{-1}\boldsymbol{\xi}_n^{(1)}. \tag{G14}$$

By multiply $(\boldsymbol{\Omega}_n^{(1)})^{-1}(\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\vartheta_1^*))$ and then minus $\boldsymbol{\beta}_{(01)}$ on both sides of (G14), we have

$$f(\vartheta_1^*) - \boldsymbol{\beta}_{(01)} + \lambda_n(\boldsymbol{\Omega}_n^{(1)})^{-1}\boldsymbol{D}_1(\vartheta_1^*)f(\vartheta_1^*) = (\boldsymbol{\Omega}_n^{(1)})^{-1}\boldsymbol{\xi}_n^{(1)} - \boldsymbol{\beta}_{(01)} = (\boldsymbol{Z}_1^\top \boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^\top \varepsilon, \tag{G15}$$

Therefore,

$$\sup_{\vartheta_1^*\in[1/K_0,K_0]^{q+1}} \|f(\vartheta_1^*) - \boldsymbol{\beta}_{(01)} + \lambda_n(\boldsymbol{\Omega}_n^{(1)})^{-1}\boldsymbol{D}_1(\vartheta_1^*)f(\vartheta_1^*)\| = O_p(\sqrt{(q+1)/n}).$$

Following arguments similar to those in the proof of (G5), we can obtain

$$\sup_{\vartheta_1^*\in[1/K_0,K_0]^{q+1}} \left\| \frac{\lambda_n}{n}(n^{-1}\boldsymbol{\Omega}_n^{(1)})^{-1}\boldsymbol{D}_1(\vartheta_1^*)f(\vartheta_1^*) \right\| = o_p(\sqrt{(q+1)/n}).$$

Thus,

$$\sup_{\vartheta_1^* \in [1/K_0, K_0]^{q+1}} \| f(\vartheta_1^*) - \boldsymbol{\beta}_{(01)} \| \leq \Delta_n o_p(\sqrt{(q+1)/n}) \to 0, \tag{G16}$$

which implies that $f(\vartheta_1^*) \in [1/K_0, K_0]^{q+1}$ with probability tending to 1. That is $f(\vartheta_1^*)$ is a mapping from $[1/K_0, K_0]^{q+1}$ to itself.

Let $\vartheta_1^* = (\vartheta_{1,1}^*, \ldots, \vartheta_{1,q}^*, \vartheta_{1,q+1}^*)^\top$ and $\dot{f}(\vartheta_1^*) = \frac{\partial f(\vartheta_1^*)}{\partial \vartheta_1^*}$. Similar to those in the proof of (G15), by multiplying $\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\vartheta_1^*)$ and taking derivative with respect to ϑ_1^* on the both sides of (G14), we can obtain

$$n^{-1}(\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\vartheta_1^*))\dot{f}(\vartheta_1^*) + \frac{\lambda_n}{n} diag\left(\frac{-2f(\vartheta_1^*)}{\vartheta_{1,1}^{*3}}, \ldots, \frac{-2f(\vartheta_1^*)}{\vartheta_{1,q+1}^{*3}}\right) = 0.$$

Then

$$\begin{aligned} &\sup_{\vartheta_1^* \in [1/K_0, K)]^{q+1}} \| n^{-1}(\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\vartheta_1^*))\dot{f}(\vartheta_1^*) \| \\ &= \sup_{\vartheta_1^* \in [1/K_0, K)]^{q+1}} \left\| \frac{2\lambda_n}{n} diag\left(\frac{f(\vartheta_1^*)}{\vartheta_{1,1}^{*3}}, \ldots, \frac{f(\vartheta_1^*)}{\vartheta_{1,q+1}^{*3}}\right) \right\| = o_p(1). \end{aligned}$$

According to condition (C8) and the fact $\vartheta_1^* \in [1/K_0, K_0]^{q+1}$, we can obtain

$$\begin{aligned} \| n^{-1}(\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\vartheta_1^*))\dot{f}(\vartheta_1^*) \| &\geq \| n^{-1}\boldsymbol{\Omega}_n^{(1)}\dot{f}(\vartheta_1^*) \| - \| n^{-1}\boldsymbol{D}_1(\vartheta_1^*)\dot{f}(\vartheta_1^*) \| \\ &\geq \left(\frac{1}{c} - \frac{\lambda_n}{n}K_0^2\right) \| \dot{f}(\vartheta_1^*) \|. \end{aligned} \tag{G17}$$

Thus, we have that $\sup_{\vartheta_1^* \in [1/K_0, K_0]^{q+1}} \| \dot{f}(\vartheta_1^*) \| \to 0$, which implies that $f(.)$ is a contraction mapping from $[1/K_0, K_0]^{q+1}$ to itself with probability tending to 1. Hence, according to the contraction mapping theorem, there exists one unique fixed-point $\hat{\vartheta}_1 \in [1/K_0, K_0]^{q+1}$ such that

$$\hat{\vartheta}_1 = (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\vartheta}_1))^{-1}\boldsymbol{\xi}_n^{(1)}.$$

We complete the proof of Lemma 3. □

Proof of Theorem 1 Firstly, based on the definitions of $\hat{\boldsymbol{\beta}}^*$ and $\boldsymbol{\beta}_2^{(k)}$, it follows from Lemma 1 and 2 that

$$\hat{\boldsymbol{\beta}}_2^* \equiv \lim_{k \to \infty} \boldsymbol{\beta}_2^{(k)} = 0$$

holds, with the probability tending to 1. And the conclusion (I) holds.

Secondly, to prove the conclusion (II), we need show that $P(\hat{\boldsymbol{\beta}}_1^* = \hat{\vartheta}_1) \to 1$. For this consider (G2) and define $\vartheta_2 = 0$ if $\boldsymbol{\beta}_2 = 0$. Note that we obtain $\lim_{\boldsymbol{\beta}_2 \to 0} \vartheta_2 = 0$ from the formula (G2) for fixed large n.

Furthermore, based on the formula (G15), we can obtain

$$\lim_{\boldsymbol{\beta}_2 \to 0} \vartheta_1 = (\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}_1(\boldsymbol{\beta}_1))^{-1} \boldsymbol{\xi}_n^{(1)} = f(\boldsymbol{\beta}_1), \tag{G18}$$

by multiplying $(\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}_1(\boldsymbol{\beta}))$ on both sides of (G1). Combining conclusion (I) and formula (G18), as $k \to \infty$, it follows that

$$\psi_k \equiv \sup_{\boldsymbol{\beta}_1^* \in [1/K_0, K_0]^{q+1}} \| f(\boldsymbol{\beta}_1) - \vartheta_1 \| \to 0. \tag{G19}$$

Since $f(.)$ is a contract mapping, Lemma 3 yields

$$\| f(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\vartheta}_1 \| = \| f(\hat{\boldsymbol{\beta}}_1^{(k)}) - f(\hat{\vartheta}_1) \| \le \frac{1}{C} \| \hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\vartheta}_1 \|, C > 1 \tag{G20}$$

Let $h_k = \| \hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\vartheta}_1 \|$. It then follows from (G19) and (G20) that

$$h_{k+1} = \| \vartheta_1 - \hat{\vartheta}_1 \| \le \| \vartheta_1 - f(\hat{\boldsymbol{\beta}}_1^{(k)}) \| + \| f(\hat{\boldsymbol{\beta}}_1^{(k)}) - \hat{\vartheta}_1 \| \le \psi_k + \frac{h_k}{C}. \tag{G21}$$

From equation (G19), there exists constant $N > 0$ such that $| \psi_k | < \varepsilon, k > N$ for any $\varepsilon \ge 0$. According to the above formula (G21), we have $h_k \to 0$ by using recursive calculation, as $k \to \infty$. Hence, we obtain $\| \hat{\boldsymbol{\beta}}_1^{(k)} - \hat{\vartheta}_1 \|$ as $k \to \infty$.

Since $\hat{\boldsymbol{\beta}}_1^* \equiv \lim_{k \to \infty} \hat{\boldsymbol{\beta}}_1^{(k)}$, based on the Lemma 3, we have $P(\hat{\boldsymbol{\beta}}_1^* = \hat{\vartheta}_1) \to 1, k \to \infty$. And we complete this proof of the conclusion (II).

Finally, we will prove the conclusion (III). Define

$$\Psi_1 = \sqrt{n}[(\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\vartheta}_1))^{-1} \boldsymbol{\Omega}_n^{(1)} - I_{q+1}] \boldsymbol{\beta}_{(01)}$$

and

$$\Psi_2 = \sqrt{n}(\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\vartheta}_1))^{-1} (\boldsymbol{\xi}_n^{(1)} - \boldsymbol{\Omega}_n^{(1)} \boldsymbol{\beta}_{(01)}),$$

where I_{q+1} denotes $q+1$ dimensional identity matrix. According to the Lemma 3, we have $\sqrt{n}(\hat{\vartheta}_1 - \boldsymbol{\beta}_{(01)}) = \Psi_1 + \Psi_2$. Furthermore, It follows from Woodbury matrix identity and condition (C7)–(C8) that

$$\begin{aligned} \Psi_1 &= \frac{\lambda_n}{\sqrt{n}} (n^{-1} \boldsymbol{\Omega}_n^{(1)})^{-1} \boldsymbol{D}_1(\hat{\vartheta}_1) (n^{-1} \boldsymbol{\Omega}_n^{(1)} + n^{-1} \lambda_n \boldsymbol{D}_1(\hat{\vartheta}_1))^{-1} n^{-1} \boldsymbol{\Omega}_n^{(1)} \boldsymbol{\beta}_{(01)} \\ &= O_p(\lambda_n \sqrt{(q+1)/n}) \to 0 \end{aligned} \tag{G22}$$

Under the assumption $\lambda_n/\sqrt{n}$ of condition (C8), following arguments similar to those in the proof of (G22),

$$\begin{aligned}\Psi_2 &= \sqrt{n}((n^{-1}\boldsymbol{\Omega}_n^{(1)})^{-1} - o_p(1/\sqrt{n}))(n^{-1}\boldsymbol{\xi}_n^{(1)} - n^{-1}\boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{(01)}) \\ &= n^{-1}\boldsymbol{\Omega}_n^{(1)})^{-1}\frac{1}{\sqrt{n}}(\boldsymbol{\xi}_n^{(1)} - \boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{(01)}) + o_p(1)\end{aligned} \tag{G23}$$

where $n^{-1/2}(\boldsymbol{\xi}_n^{(1)} - \boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{(01)}) = n^{-1/2}\dot{l}_n^{(1)}(\hat{\boldsymbol{\beta}}^* \mid \hat{\boldsymbol{\alpha}}) + o_p(1)$ with $\dot{l}_n^{(1)}(\hat{\boldsymbol{\beta}}^* \mid \hat{\boldsymbol{\alpha}})$ denoting the first $q+1$ components of $\dot{l}_n(\hat{\boldsymbol{\beta}}^* \mid \hat{\boldsymbol{\alpha}})$. Let the Fisher information matrix $I(\boldsymbol{\beta}) = -E(\ddot{l}_n(\boldsymbol{\beta} \mid \hat{\boldsymbol{\alpha}}))$ and $I^{(1)}(\boldsymbol{\beta}_0)$ denotes the leading $(q+1)\times(q+1)$ sub-matrix of $I(\boldsymbol{\beta}_0)$. The law of large numbers and the multivariate central limit theorem show that $n^{-1/2}\dot{l}_n(\hat{\boldsymbol{\beta}}^* \mid \hat{\boldsymbol{\alpha}}) \to N(0, n^{-1}I(\boldsymbol{\beta}_0))$, we have $\sqrt{n}(\hat{\vartheta}_1 - \boldsymbol{\beta}_{(01)}) \to N(0, \Sigma)$ with $\Sigma = n(\boldsymbol{\Omega}_n^{(1)}(\boldsymbol{\beta}_0))^{-1}I^{(1)}(\boldsymbol{\beta}_0)(\boldsymbol{\Omega}_n^{(1)}(\boldsymbol{\beta}_0))^{-1}$. This completes the proof of the Theorem 1. □

References

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

Cook R, Lawless J (2007) The statistical analysis of recurrent events. Springer, New York

Cox DR (1972) Regression models and life-tables. J R Stat Soc B 34:187–202

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, Sun J (2022) Variable selection for interval-censored failure time data. Int Stat Rev 90:193–215

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

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 Coxs proportional hazards model and frailty model. Ann Stat 30:74–99

Hu T, Zhou Q, Sun J (2017) Regression analysis of bivariate current status data under the proportional hazards model. Can J Stat 45:410–424

Huang J (1996) Effcient estimation for the proportional hazards model with interval censoring. Ann Stat 24:540–568

Huang J, Rossini A (1997) Sieve estimation for the proportional-odds failure-time regression model with interval censoring. J Am Stat Assoc 92:960–967

Huang C, Wang M (2004) Joint modeling and estimation for recurrent event processes and failure time data. J Am Stat Assoc 99:1153–1165

Kalbfleisch J, Prentice R (2002) The statistical analysis of failure time data, 2nd edn. Wiley, New York

Klein J, Moeschberger M (2003) Survival analysis: techniques for censored and truncated data, 2nd edn. Springer, New York

Li K, Chan W, Doody RS, Quinn J, Luo S, Initiative ADN (2017) Prediction of conversion to Alzheimers disease with longitudinal measures and time-to-event data. J Alzheimers Dis 58:361–371

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

Lu W, Zhang HH (2007) Variable selection for proportional odds model. Stat Med 26:3771–3781

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
Murphy SA, Rossini AJ, van der Vaart AW (1997) Maximum likelihood estimation in the proportional odds model. J Am Stat Assoc 92:968–976
Rossini AJ, Tsiatis AA (1996) A semiparametric proportional odds regression model for the analysis of current status data. J Am Stat Assoc 91:713–721
Scolas S, El Ghouch A, Legrand C, Oulhaj A (2016) Variable selection in a flexible parametric mixture cure model with interval-censored data. Stat Med 35:1210–1225
Shen X (1998) Propotional odds regression and sieve maximum likelihood estimation. Biometrika 85:165–177
Sun J (2006) The statistical analysis of interval-censored failure time data. Springer, New York
Sun L, Li S, Wang L, Song X, Sui X (2022a) Simultaneous variable selection in regression analysis of multivariate interval-censored data. Biometrics 78:1402–1413
Sun Z, Liu Y, Chen K, Li G (2022b) Broken adaptive ridge regression for right-censored survival data. Ann Inst Stat Math 74:69–91
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288
Tibshirani R (1997) The Lasso method for variable selection in the Cox model. Stat Med 16:385–395
Wang J, Ghosh SK (2012) Shape restricted nonparametric regression with Bernstein polynomials. Comput Stat Data Anal 56:2729–274
Wang H, Leng C (2007) Unified LASSO estimation by least squares approximation. J Am Stat Assoc 102:1039–1048
Wang L, Wang L (2021) Regression analysis of arbitrarily censored survival data under the proportional odds model. Stat Med 40:3724–3739
Wang H, Li R, Tsai CL (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94:553–568
Wang H, Li B, Leng C (2009) Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc B 71:671–683
Wang P, Zhao H, Sun J (2016) Regression analysis of case K interval-censored failure time data in the presence of 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 L, Peng B, Bradic J, Li R, Wu Y (2020a) A tuning-free robust and efficient approach to high-dimensional regression. J Am Stat Assoc 115:1700–1714
Wang S, Wang C, Wang P, Sun J (2020b) Estimation of the additive hazards model with case K interval-censored failure time data in the presence of informative censoring. Comput Stat Data Anal 144:1–9
Wang S, Xu D, Wang C, Sun J (2023) Estimation of linear transformation cure models with informatively interval-censored failure time data. J Nonparametric Stat 35:283–301
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
Yang S, Prentice RL (1999) Semiparametric inference in the proportional odds regression model. J Am Stat Assoc 94:125–136
Zhang H, Lu WB (2007) Adaptive Lasso for Coxs 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 115:204–216
Zhao B, Wang S, Wang C, Sun J (2021) New methods for the additive hazards model with the informatively interval-censored failure time data. Biom J 63:1507–1525
Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429