第50卷第10期 中国科学技术大学学报 Vol. 50, No. 10
2020年10月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Oct. 2020
Received: 2020-09-12; Revised: 2020-10-20 doi:10.3969/j.issn.0253-2778.2020.10.003

Bayesian variable selection for proportional hazards model with current status data

Cui Di, Zhang Weiping
Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, *China*

Abstract: A Bayesian proportional hazards (PH) model is proposed for analyzing current status data based on Expectation-Maximization Variable Selection (EMVS) method. This model can estimate parameters and select variables simultaneously, which efficiently improves model interpretability and predictive ability. To identify risk factors, appropriate priors are assigned on the indicator variables that denote the existence of covariates. The baseline cumulative hazard function is approximated via monotone splines. A novel Expectation-Maximization (EM) algorithm is developed for model fitting by using a two-stage data augmentation procedure involving latent Poisson variables. Finally, the performance of proposed method is investigated by simulations and a real data analysis.

Key words: proportional hazards model; Bayesian variable selection; current status data; EM algorithm; spline
CLC number: O212.8 **Document code**: A **2010 Mathematics Subject Classification**: 62F15, 62F07

0 Introduction

Current status data naturally arise in cross-sectional studies across a range of disciplines from epidemiological to clinical and social science, where the failure time of interest is not directly observed, but instead is known whether or not it exceeds the observation time. For example, as part of a community-wide study examining time trends associated with risk factors of patients hospitalized with acute myocardial infarction (MI) in the Worcester, Massachusetts metropolitan area, the survival state of patients were followed up once after hospital discharge. In Worcester Heart Attack Study (WHAS), the survival days of a patient after hospital discharge was not determined accurately but was known to be earlier or later than the date of the last follow-up, leading to either left-censored or right-censored observations.

The goal of analyzing current status data mainly focuses on the estimation of the covariates effect and survival functions. For example, millions of people suffer from MI that has been a major cause of morbidity and mortality[1]. And the survival rate of patients with MI has been found to be highly involved with the patients′ age and the congestive heart complications[2-3]. Thus, a key objective of WHAS is to identify the risk factors that threaten the lives of patients. The proportional hazards (PH) model[4] is one of the most popular frameworks for the regression analysis of time-to-event data. The majority of available methods for PH model have focused on the estimation in the presence of right-censored data. However, due to the more complex structure of current status data, the partial likelihood[5] allowing one to estimate the parameters without specifying the baseline hazard function no longer exists, and counting process and martingale techniques designed for right-censored data fail to work[6]. Therefore, a number of studies have been developed to analyze current status data under PH assumption and its numerous variants. This topic was first studied by Finkelstein[7] who proposed a Newton-Raphson based algorithm to estimate the baseline function and the parameters simultaneously. Huang et al[8] established a profile likelihood and proved the large sample properties of the nonparametric maximum likelihood estimation. Pan[9] adopted Tanner and Wong's data[10] augmentation scheme to transfer the interval-censored data to right-censored data. As the baseline cumulative hazard function is unspecified, serval authors utilized monotone splines[11] in order to

Citation: Cui Di, Zhang Weiping. Bayesian variable selection for proportional hazards model with current status data. J. Univ. Sci. Tech. China, 2020,50(10):1303-1314.

reduce the dimension. Cai et al[12] discussed the regression analysis for current status data using monotone splines in Bayesian framework. McMahan et al[13] developed an expectation-maximization (EM) type algorithm under the same model specification. Zeng et al[14] extended McMahan et al's work[13] to accommodate a broad class of semi-parametric transformation models. Lu et al[15] made use of monotone B-spline to model the baseline hazard function.

WHAS recorded a dozen covariates of hospitalized patients. Keeping all covariates may result in overfitting, which poses a problem in estimation accuracy and model interpretability. A number of statistical methods were proposed for variable selection, such as penalization procedures LASSO[16], SCAD[17] and adaptive LASSO[18]. In addition, Bayesian methods have also gained popularity[19-22]. Despite the extensive literature on the regression analysis for current status data, little attention has yet been paid to variable selection in Bayesian framework. The latest work concentrated on variable selection for such outcomes are mainly developed via frequentist-based penalization methods[23-28], which may be highly challenging when model or data structures are complicated. Expectation maximization variable selection (EMVS)[27] inspired by stochastic search variable selection (SSVS) approach[20] has been shown to be a deterministic Bayesian variable selection method as its efficiency at identifying associated covariates and the capability of accommodating multifarious data structures[28-30]. Since this method selects all covariates simultaneously, it avoids the issue of multiple model comparisons that the main challenge faced by the traditional pairwise comparison method using Bayesian model comparison statistics. Compared with SSVS, where the posterior inference is drawn using Markov Chain Monte Carlo (MCMC) algorithm, EMVS utilizes EM algorithm to derive posterior estimates with enormous computational savings.

In this paper, we make use of EMVS's validity in selecting covariates, by developing it for PH model to identify the relationship between risk factors and survival time with current status data. To reduce the model dimension, monotone I-splines are utilized to approximate the baseline cumulative hazard function. We develop an efficient EM algorithm that achieves parameters estimation through a two-stage data augmentation procedure involving latent Poisson variables and variable selection through latent index variables simultaneously. Furthermore, additional constrained optimization procedures are avoided since the constraints of monotonicity of baseline cumulative hazard function are satisfied directly during their closed-form updates.

The rest of the paper is organized as follows. In Section 1, we specify the PH model with current status data and develop the EMVS method for the covariates selection. Section 2 reports the results of simulated studies. A real data analysis is presented in Section 3 to illustrate the performance of the proposed method. Section 4 concludes with a summary discussion.

1 The model framework

1.1 Model setup

For the event time T, the cumulative hazard function of an additive Cox model takes the form

$$\Lambda(t \mid x) = \Lambda_0(t)\exp(x^{\mathrm{T}}\beta) \tag{1}$$

where $\Lambda_0(\cdot)$ is unknown baseline cumulative hazard function, x is $p \times 1$ covariates vector, β is the vector of regression coefficient. Given x, the conditional cumulative distribution function (CDF) of event time T is written as

$$F(t \mid x) = 1 - \exp\{-\Lambda_0(t)\exp(x^{\mathrm{T}}\beta)\} \tag{2}$$

To hold the conditions of a proper CDF, it is required that $\Lambda_0(\cdot)$ is a non-negative and monotone function with $\Lambda_0(0) = 0$.

In the setting of current status data, the event time T cannot be observed directly. Let $t_i, i = 1, \cdots,$

n, denote the event time for n subjects, and $c_i, i = 1,\cdots,n$, the observation time. The censoring indicator is defined as $\delta_i = I(t_i \leqslant c_i)$, $i = 1,\cdots,n$. Then the likelihood function of all observed data $\mathcal{D} = \{(c_i, \delta_i, x_i^{\mathrm{T}})^{\mathrm{T}},\ i = 1,\cdots,n\}$ is given by

$$L = \prod_{i=1}^{n} F(c_i \mid x_i)^{\delta_i}[1 - F(c_i \mid x_i)]^{1-\delta_i} \quad (3)$$

which is under the assumption that the failure time and censoring time are independent given covariates.

In the above likelihood, the baseline cumulative hazard function $\Lambda_0(\cdot)$ is totally unspecified with infinite dimension of parameters. Spline is a feasible technique to reduce the dimension while maintaining model flexibility as it makes no assumption with the shape of fitted curve. In this article, $\Lambda_0(\cdot)$ is approximated via I-splines[11], that is,

$$\Lambda_0(\cdot) = \sum_{k=1}^{K} \alpha_k I_k(\cdot) \quad (4)$$

where $\{I_k(\cdot)\}_{k=1}^{K}$ are integrated spline basis function, each of which is nondecreasing from 0 to 1, and spline coefficients $\{\alpha_k\}_{k=1}^{K}$ are taken to be non-negative to guarantee the monotonicity. To construct the basis functions, the number and location of interior knots need to be specified to determine the shape, and the degree controls the smoothness of the model. The number of all basis functions K is the summation of the number of interior knots and the degree.

It is known that the order as well as the number and location of interior knots have an impact on model fitting. In general, cubic spline (the order is 3) is smooth enough to fit the curve, whereas too many (few) knots lead to over (under) fitting. Yu and Ruppert[31] made the strategy that 5 to 10 knots are adequate for unimodal or monotone functions while more than 10 knots are necessary to capture the characteristics for multimodal functions. In simulation studies, we tried different numbers of interior knots to evaluate the performance. Once the number is determined, the knots are equidistantly placed at the quantiles of support of splines.

Substituting the splines approximation into $\Lambda_0(\cdot)$, the baseline cumulative hazard function Eq. (3) can be written in vector notation as

$$\Lambda(t \mid x_i) = \alpha^{\mathrm{T}} I(c_i) \exp(x_i^{\mathrm{T}}\beta) \quad (5)$$

where $\alpha = (\alpha_1,\cdots,\alpha_K)^{\mathrm{T}}$ and $I(c_i) = (I_1(c_i),\cdots, I_K(c_i))^{\mathrm{T}}$.

1.2 Data augmentation for the EM algorithm

Direct maximization of Eq. (3) is intractable because of the complex form. In the spirit of Wang et al[32], an EM algorithm is proposed to identify the maximizer of unknown parameters. In order to derive the algorithm, a two-stage data augmentation involving latent Poisson random variables is utilized based on the relationship between the Cox model and a nonhomogeneous Poisson process.

The first stage is to associate the censoring indicator δ_i with nonhomogeneous Poisson process w_i with mean $\alpha^{\mathrm{T}} I(c_i)\exp(x_i^{\mathrm{T}}\beta)$. The second stage w_i is decomposed into a sum of independent Poisson process w_{ik}. Therefore, the data augmentation procedure is as follows:

$$\left.\begin{aligned} &\delta_i = I(w_i > 0), \\ &w_i \sim \text{Poisson}(\alpha^{\mathrm{T}} I(c_i)\exp\{x_i^{\mathrm{T}}\beta\}) \\ &w_i = \sum_{k=1}^{K} w_{ik},\ w_{ik} \sim \text{Poisson}(\alpha_k I_k(c_i)\exp\{x_i^{\mathrm{T}}\beta\}), \\ &\qquad\qquad\qquad\qquad k = 1,\cdots,K \end{aligned}\right\} \quad (6)$$

Denote $P_{w_ik}(\cdot)$ the Poisson mass function associating with the random variables w_i and w_{ik} respectively, the augmented data likelihood can be expressed as

$$L_{\text{aug}} = \prod_{i=1}^{n}\prod_{k=1}^{K} P_{w_{ik}}(w_{ik})(\delta_i I(w_i > 0) + (1-\delta_i) I(w_i = 0)) \quad (7)$$

1.3 Prior specification

To facilitate Bayesian variable selection, the well-known spike-and-slab prior is assigned to the regression coefficients β. An indicator variable $\gamma = (\gamma_1,\cdots,\gamma_p)^{\mathrm{T}}$ is introduced to identify β, i. e., $\beta_m = 0$ if $\gamma_m = 0$ and $\beta_m \neq 0$ otherwise for $m = 1,\cdots,p$. Thus, the prior we assigned to β is

$$P(\beta \mid \gamma, \sigma, v_1) = N(0, \Sigma_\beta) \text{ with}$$
$$\Sigma_\beta = \sigma^2 \operatorname{diag}(d_1,\cdots,d_p),$$

where N is normal density function and $d_m = (1 - \gamma_m) \cdot v_0 + \gamma_m \cdot v_1$ for $0 < v_0 < v_1$. Though v_0 is frequently set to be 0 in practice, George and McCulloch[20] recommended setting small but positive v_0 to exclude unimportant nonzero effects. v_0 of the spike distribution serves to pull coefficients estimates toward zero. The increase of v_0 enlarges the variance of the spike component, which has the effect to shrink the small effect without much affecting the significant effects. To leave large coefficients possibly unaffected by the shrinkage of spike prior, a heavy-tailed slab prior suggested by Ref. [27] is induced for v_1,

$$P(v_1) = \frac{v_1^{b_2}(1+v_1)^{-a_2-b_2-2}}{B(a_2+1, b_2+1)} I(v_1 > 0),$$

where $B(\cdot,\cdot)$ is Beta function. Referring to Refs. [21,33,34], hyper-parameters a_2 and b_2 are set to be 0 and $-3/4$ for better performance. The setting of a_2 and b_2 makes more flat proper prior, resulting in stable estimation.

Without extra structural information about the predictors, an independent and identically distributed (i. i. d) Bernoulli prior is chosen for γ_m,

$$P(\gamma \mid \omega) = \omega^{|\gamma|} (1-\omega)^{p-|\gamma|},$$

where $|\gamma| = \sum_{m=1}^{q} \gamma_m$ and ω is a hyperparameter following uniform distribution $U(0,1)$.

As α is constraint with non-negativity, the i. i. d exponential priors are assigned to α,

$$P(\alpha_k \mid \lambda) = \lambda e^{-\lambda \alpha_k}, \ k = 1, \cdots, K,$$

where λ is a hyperparameter. This specification allows the hyper prior for λ providing information for the spline coefficients, more important, can penalize the coefficients of unnecessary spline basis functions toward zero[12].

The priors for σ^2 and Λ are chosen as $\mathrm{IG}(a_1, b_1)$ and $\mathrm{Ga}(a_3, b_3)$ respectively, resulting in uninformative priors, where IG and Ga denote inverse γ distribution and γ distribution, respectively. a_1, b_1, a_3, b_3 are set to be 0.5 in all numerical experiments.

Lemma 1.1 The joint posterior distribution $P(\beta, \sigma^2 \mid \mathscr{D})$ given d_m, $m = 1, \cdots, p$, fixed is unimodal.

Remark 1.1 The unimodality is established for any value of γ and any choice of v_0 and v_1, in the sense that for every $c > 0$, the upper level set $\{(\beta, \sigma^2 \mid P(\beta, \sigma^2))\}$ is connected.

Finally, the joint posterior distribution of all the parameters is given by

$$L_c(\alpha, \beta, \gamma, \sigma^2, v_1, \omega, \lambda \mid \mathscr{D}, w) \propto P(\mathscr{D}, w \mid \alpha, \beta) \cdot P(\beta \mid \gamma, \sigma^2, v_1) P(\alpha \mid \lambda) P(\gamma \mid \omega) \cdot P(\sigma^2) P(v_1) P(\lambda) P(\omega) \tag{8}$$

where $h = (h_1^{\mathrm{T}}, \cdots, h_1^{\mathrm{T}})^{\mathrm{T}}$, $w = (w_{11}, \cdots, w_{1K}, \cdots, w_{n1}, \cdots, w_{nK})$ and the first term in the right side is L_{aug}.

1.4 EM algorithm

An EM algorithm is derived to find the posterior maximizer of parameters iteratively as an alternative to the conventional MCMC approach, which possesses much computational efficiency over stochastic search alternatives.

The EM algorithm begins with the expectation (E-step) of the logarithm of L_c with respect to the latent variables (w and γ) conditional on the observed data $\mathscr{D}$ and current parameter estimate, whereafter, the maximum (M-step) likelihood estimators of the expected log-posterior likelihood resulting from E-step are calculated. Each parameter is estimated under the condition that the remaining parameters are fixed in M-step. The two steps are repeated until the convergence is achieved.

Denote $\theta = \{\alpha, \beta, \sigma^2, v_1, \omega, \lambda\}$. In $(u+1)$th iteration, the expected log likelihood in E-step is given by

$$E[\log L_c \mid \mathscr{D}, \theta^{(u)}] = Q_1(\alpha, \beta, \sigma^2, v_1, \lambda \mid \mathscr{D}, \theta^{(u)}) + Q_2(\omega \mid \mathscr{D}, \theta^{(u)}) + C,$$

where C is a constant and

$$\begin{aligned} &Q_1(\alpha, \beta, \sigma^2, v_1, \lambda \mid \mathscr{D}, \theta^{(u)}) = \\ &\sum_{i=1}^{n} \sum_{k=1}^{K} \{E(w_{ik} \mid \mathscr{D}, \theta^{(u)}) [\log \alpha_k + x_i^{\mathrm{T}} \beta] - \alpha_k I_k(c_i) \exp(x_i^{\mathrm{T}} \beta)\} - \\ &\frac{1}{2} \sum_{m=1}^{p} E(\log d_m \mid \mathscr{D}, \theta^{(u)}) - \left(\frac{p}{2} + a_1 + 1\right) \log \sigma^2 - \\ &\frac{1}{\sigma^2} \left[b_1 + \frac{1}{2} \sum_{m=1}^{p} E\left(\frac{1}{d_m} \mid \mathscr{D}, \theta^{(u)}\right) \beta_m{}^2\right] + \\ &b_2 \log v_1 - (a_2 + b_2 + 2) \log(1 + v_1) + \\ &(K + a_3 - 1) \log \lambda - \left(\sum_{k=1}^{K} \alpha_k + b_3\right) \lambda, \end{aligned}$$

$$Q_2(\omega \mid \mathscr{D}, \theta^{(u)}) = \sum_{m=1}^{p} E(\gamma_m \mid \mathscr{D}, \theta^{(u)}) \log \omega + \left(p - \sum_{m=1}^{p} E(\gamma_m \mid \mathscr{D}, \theta^{(u)})\right) \log(1 - \omega).$$

The E-steps proceeds by computing the conditional expectations $E(w_{ik} \mid \mathscr{D}, \theta^{(u)})$, $E(\gamma_m \mid \mathscr{D}, \theta^{(u)})$, $E(\log d_m \mid \mathscr{D}, \theta^{(u)})$ and $E(1/d_m \mid \mathscr{D}, \theta^{(u)})$ from Q_1 and Q_2. Noting that w_{ik} follows a multinomial distribution given w_i and w_i is truncated Poisson distributed, the expected values can be expressed

$$E(w_{ik} \mid \mathcal{D},\theta^{(u)}) = \frac{\alpha_k^{(u)} I_k(c_i)}{\sum_{k=1}^{K} \alpha_k^{(u)} I_k(c_i)} E(w_i \mid \mathcal{D},\theta^{(u)}),$$

$$E(w_i \mid \mathcal{D},\theta^{(u)}) = \frac{\sum_{k=1}^{K} \alpha_k^{(u)} I_k(c_i) \exp\{x_i^{\mathrm{T}}\beta^{(u)}\}\delta_i}{1-\exp\{-\exp\{x_i^{\mathrm{T}}\beta^{(u)}\}}.$$

And γ_m is multinomial . Following Rockova and George[27], the conditional expectations are

$$E(\gamma_m \mid \mathcal{D},\theta^{(u)}) = P(\gamma_m = 1 \mid \cdot) = p_m^* = \frac{r_m}{r_m + s_m},$$

$$E(\log d_m \mid \mathcal{D},\theta^{(u)}) = (1-p_m^*)\log v_0 + p_m^* \log v_1^{(u)},$$

$$E\left[\frac{1}{d_m} \mid \mathcal{D},\theta^{(u)}\right] = \frac{1-p_m^*}{v_0} + \frac{p_m^*}{v_1^{(u)}},$$

where $r_m = P(\beta_m \mid (\sigma^2)^{(u)}, v_1^{(u)}, r_m = 1) P(r_m = 1 \mid \omega^{(u)})$ and

$$s_m = P(\beta_m \mid (\sigma^2)^{(u)}, v_1^{(u)}, r_m = 0) P(r_m = 0 \mid \omega^{(u)}).$$

The next step is to find $\theta^{(u+1)}$ that maximizes Q_1 and Q_2. Consider the partial derivation of Q_1 and Q_2 with respect to each parameter, for β ,

$$\frac{\partial Q_1}{\beta} = \sum_{i=1}^{n}\sum_{k=1}^{K} x_i [E(w_{ik} \mid \mathcal{D},\theta^{(u)}) - \alpha_k I_k(c_i) \exp\{x_i^{\mathrm{T}}\beta + h^{\mathrm{T}} B(z_i)\}] - \Sigma\beta,$$

where

$$\Sigma = \frac{1}{\sigma^2}\mathrm{diag}\Big(E\Big(\frac{1}{d_1} \mid \mathcal{D},\theta^{(u)}\Big), \cdots, E\Big(\frac{1}{d_p} \mid, \mathcal{D},\theta^{(u)}\Big)\Big).$$

Setting the partial derivation 0, the new maximizers $\beta^{(u+1)}$ can be obtained from equation. The root of above equation can be found using standard root finding routine. As the new maximizers in regard to the remaining parameters have closed-form expressions, the parameters are directly updated by

$$\alpha_k^{(u+1)} = \frac{\sum_{i=1}^{n} E(w_{ik} \mid \mathcal{D},\theta^{(u)})}{\lambda^{(u)} + \sum_{i=1}^{n} I_k(c_i)\exp\{x_i^{\mathrm{T}}\beta^{(u+1)}\}},$$

$$(\sigma^2)^{(u+1)} = \frac{b_1 + \frac{1}{2}\sum_{m=1}^{p} E\Big(\frac{1}{d_m} \mid \mathcal{D},\theta^{(u)}\Big)(\beta_m^{(u+1)})^2}{p/2 + a_1 + 1},$$

$$v_1^{(u+1)} = \frac{A - B \pm \sqrt{(B-A)^2 + 4A(5/4+B)}}{2(5/4+B)},$$

$$\lambda^{(u+1)} = \frac{K + a_3 - 1}{\sum_{k=1}^{K} \alpha_k^{(u+1)} + b_3},$$

$$\omega^{(u+1)} = \sum_{m=1}^{p} p_m^* / p,$$

where $A = \sum_{m=1}^{p} p_m^* (\beta_m^{(u+1)})^2 / (2(\sigma^2)^{(u+1)})$ and $B = \sum_{m=1}^{p} p_m^* / 2$.

Thus, the EMVS algorithm proceeds as follows:

Step 1 Initialize the parameters $\alpha^{(0)}$, $\beta^{(0)}$, $\sigma^{(0)}$, $v_1^{(0)}$, $\omega^{(0)}$ and $\lambda^{(0)}$;

Step 2 Evaluate the conditional expectations $E(\gamma_m \mid \mathcal{D},\theta^{(u)})$, $E(\log d_m \mid \mathcal{D},\theta^{(u)})$, $E(1/d_m \mid \mathcal{D}, \theta^{(u)})$ and $E(w_{ik} \mid \mathcal{D},\theta^{(u)})$;

Step 3 Obtain $\alpha^{(u+1)}$, $\beta^{(u+1)}$, $\sigma^{(u+1)}$, $v_1^{(u+1))}$, $\omega^{(u+1)}$ and $\lambda^{(u+1)}$ by maximizing $Q(\theta \mid \mathcal{D},\theta^{(u)})$;

Step 4 Iterate between Steps 2 and 3 until the maximum absolute difference of β between two successive iterations is smaller than 10^{-6}.

1.5 Variance estimation

The covariance matrix of $\hat{\beta}$ can be estimated based on the profile likelihood. Denote $\psi = \{\alpha, \sigma^2, v_1, \omega, \lambda\}$, then the profile log-likelihood is defined as

$$l_{pn}(\beta) = \max_{\psi} \log(L \times P(\beta \mid \gamma, \sigma^2, v_1) \cdot P(\sigma^2)P(v_1)\ P(\gamma \mid \omega)P(\omega)P(\alpha \mid \lambda)P(\lambda)).$$

Referring to Ref. [14], the covariance matrix of $\hat{\beta}$ is calculated by the inverse of the information matrix $I(\hat{\beta})$, where the (s,t) th of $I(\hat{\beta})$ element is approximated by

$$\frac{l_{pn}(\hat{\beta}) - l_{pn}(\hat{\beta} + r_n e_s) - l_{pn}(\hat{\beta} + r_n e_t) + l_{pn}(\hat{\beta} + r_n e_s + r_n e_t)}{r_n^2},$$

where e_s is a p-dimensional vector with the sth element 1 and the remaining is 0, and r_n a tuning constant of order $n^{-1/2}$. The value of $l_{pn}(\beta)$ can be calculated using the EM algorithm again with β held fixed.

2 Simulation studies

Simulation studies are conducted in this section to illustrate the performance of the proposed method in different scenarios. We independently generate 200 data sets from the following model

$$F(t \mid x)=1-\exp[-\Lambda_0(t)\exp\{x^{\mathrm{T}}\beta\}].$$

We consider three cases: (i) The cumulative baseline hazards function $\Lambda_0(t)=\log(1+t)$ and $\beta=(1,0,2,0,-1,0,0,0)$ is a 8×1 vector. For the discrete covariates vector $x_d=(x_1,x_2)^{\mathrm{T}}$, first we sample $x_1\sim$ Bernoulli(p) with $p=0.5$ then sample $x_2=k\mid\{x_1=0\}\sim p_{0k}(k=1,2,3)$ with $p_0=(0.5, 0.4, 0.1)$ and $x_2=k\mid\{x_1=1\}\sim p_{1k}(k=1,2,3)$ with $p_{1k}=(0.4,0.4,0.2)$. The remaining continuous covariates vector $x_c=(x_3,\cdots,x_8)^{\mathrm{T}}$ follows a multi-normal distribution with mean 0 and covariance matrix $(0.5^{|k-l|})_{1\leqslant k,l\leqslant 6}$. The censoring time C is generated from exponential distribution with mean 1, then the censoring indicator is determined as $\Delta=I(t\leqslant c)$; (ii) $\Lambda_0(t)=\sqrt{t}/2$. The remaining setup is the same as case (i); (iii) A high dimensional case with $n=200$ and $p=100$, where $\lambda_0(t)=\log(1+t)$ and $\beta=(1,1,1,0,\cdots,0)$ with only the first three nonzero elements. The covariate vector $x=(x_1,\cdots,x_{100})^{\mathrm{T}}$ is generated from a multi-normal distribution with mean 0 and covariance matrix $(0.5^{|k-l|})_{1\leqslant k,l\leqslant 100}$. The censoring time C follows the exponential distribution with mean 1. In each case, the proposed EMVS method PH model is taken into account.

In specifying the monotone splines to estimate $\Lambda_0(\cdot)$, cubic basis functions are utilized for adequate smoothness. We try different numbers of equally spaced interior knots for the cumulative baseline hazard function within the minimum and maximum of c. v_0 of the spike distribution is set to be 0.01 in the first two cases, and we empirically find that the estimation results are quite robust for the variation of v_0 within [0.001,0.1]. In the third case, as the dimension is high, v_0 is set to be 1 within [0.5,2] Iteration of the proposed EM algorithm is terminated if the maximum absolute difference of the parameters between two successive iterations is smaller than 10^{-6}. The decision of variable selection is based on the probability $p(\gamma_m=1)$, that is, the conditional expectation p_m^* calculated in E-step. The default threshold value is 0.5, i.e., x_m is selected if $p_m^*>0.5$ and is not selected otherwise.

We use LASSO as a benchmark method to compare the variable selection accuracy with the proposed method under the same model specifications. We calculate the LASSO estimator based on Eq. (7) using EM algorithm, where w_i and w_{ik} are treated as missing data and $\{\alpha,\beta\}$ are the parameters to be estimated. In $(u+1)$th iteration, the expected values of w_i and w_{ik} in E-step take the same value of the proposed method. In M-step, first we update α as

$$\alpha_k^{(u+1)}=\frac{\sum_{i=1}^{n}E(w_{ik}\mid\mathcal{D},\beta^{(u)})}{\sum_{i=1}^{n}I_k(c_i)\exp\{x_i^{\mathrm{T}}\beta^{(u+1)}\}}.$$

Then we update β by maximizing

$$Q(\beta\mid\beta^{(u)},\alpha^{(u+1)})-n\tau\Big(\sum_{m=1}^{p}\mid\beta_m\mid\Big),$$

where

$$Q(\beta\mid\beta^{(u)},\alpha^{(u+1)})=\sum_{i=1}^{n}\sum_{k=1}^{K}E(w_{ik}\mid\mathcal{D},\beta^{(u)})[\log\alpha_k^{(u+1)}+x_i^{\mathrm{T}}\beta]-\alpha_k^{(u+1)}I_k(c_i)\exp\{x_i^{\mathrm{T}}\beta\}$$

and τ is a tuning parameter. Define $\nabla Q(\beta^{(u)})=-\partial Q/\partial\beta\mid_{\beta=\beta^{(u)}}$ and $\nabla^2 Q(\beta^{(u)})=-\partial^2 Q/\partial\beta\partial\beta^{\mathrm{T}}\mid_{\beta=\beta^{(u)}}$. Through a second-order Taylor expansion around $\beta^{(u+1)}$, $-Q(\beta\mid\beta^{(u)},\alpha^{(u+1)})$ can be written as $1/2(Y-R\beta)^{\mathrm{T}}(Y-R\beta)$, where R is from Cholesky decomposition of $\nabla^2 Q(\beta^{(u)})$ satisfying $R^{\mathrm{T}}R=\nabla^2 Q(\beta^{(u)})$ and pseudo response $Y=(R^{\mathrm{T}})^{-1}\{\nabla^2 Q(\beta^{(u)})\beta-\nabla Q(\beta^{(u)})\}$[38]. Thus, we need to minimize

$$\frac{1}{2}(Y-R\beta)^{\mathrm{T}}(Y-R\beta)+n\tau\sum_{m=1}^{p}\beta_m.$$

To obtain the LASSO regressor $\beta^{(u+1)}$, we use GLM net package in R. The EM algorithm stops if the maximum absolute difference of the parameters between two successive iterations is smaller than 10^{-6}.

The estimation results based on the 200 data sets of EMVS of the first two cases are summarized in Tabs. 1 and 2, including bias and mean square error (MSE) between the estimated parameters and the true values, the Monte Carlo standard error (MCE), the average of the numerical standard error (SEE). The results indicate that our method performs satisfactorily under different situations.

And the performance of the estimates becomes improved when the sample size increases. It seems that the estimation results rarely depend on the number of interior knots. For the variance estimation, we set $h_n = 10n^{-1/2}$ for all cases. The variance method is quite accurate even in small samples.

The false positive rate (FPR) and false negative rate (FNR) are important indexes to evaluate the variable selection accuracy. They are given by FPR = FP/(FP + TN) and FNR = FN/(FN + TP), where FP is the number of false positives, FN is the number of false negatives, TP is the number of true positives and TN is the number of true negatives. We independently generated 200 data sets for each case. Tabs. 3 and 4 reported the average of FPR and FNR of the first two cases with different numbers of interior knots. It is shown that EMVS outperforms LASSO in different settings. EMVS exhibits a considerable accuracy of variable selection even in small sample size. The results of LASSO are more conservative as it inclines to reserve more variables, possibly affected by the correlation of the covariates.

Tab. 1 Simulation results on the estimation of the non-zero coefficients for Case 1

		True effect	Bias	MSE	MCE	SEE
$n=200$	$K=5$	β_1	−0.0183	0.1231	0.2913	0.2277
		β_3	−0.1674	0.1063	0.2806	0.2306
		β_5	0.0983	0.0525	0.2074	0.1667
	$K=10$	β_1	0.0394	0.1665	0.3071	0.2410
		β_2	−0.1356	0.0412	0.2886	0.2406
		β_5	0.1000	0.0279	0.1918	0.1705
$n=500$	$K=5$	β_1	−0.0465	0.0469	0.2121	0.1647
		β_3	−0.1175	0.0416	0.1673	0.1411
		β_5	0.0650	0.0192	0.1229	0.1408
	$K=10$	β_1	0.0089	0.0318	0.1785	0.1652
		β_3	−0.1070	0.0346	0.1524	0.1663
		β_5	0.0679	0.0170	0.1116	0.1378

Tab. 2 Simulation results on the estimation of the non-zero coefficients for Case 2

		True effect	Bias	MSE	MCE	SEE
$n=200$	$K=5$	β_1	−0.0393	0.1121	0.2934	0.2138
		β_3	−0.1643	0.1148	0.2970	0.2519
		β_5	0.1006	0.0772	0.2598	0.1648
	$K=10$	β_1	0.0529	0.1192	0.2921	0.2197
		β_3	−0.1205	0.1127	0.2641	0.2543
		β_5	0.0868	0.0516	0.2105	0.1668
$n=500$	$K=5$	β_1	0.0284	0.0459	0.2129	0.1496
		β_3	0.0461	0.0360	0.1846	0.1425
		β_5	−0.0225	0.0166	0.1272	0.1044
	$K=10$	β_1	0.0619	0.0473	0.2091	0.1504
		β_3	0.0349	0.0372	0.1900	0.1430
		β_5	−0.0139	0.0138	0.1170	0.1086

Tab. 3 FPR and FNR of EMVS and LASSO for Case 1

		FPR		FNR	
		EMVS	LASSO	EMVS	LASSO
$n=200$	$K=5$	0.006	0.105	0.017	0.080
	$K=10$	0.033	0.092	0.027	0.077
$n=500$	$K=5$	0.012	0.092	0.007	0.005
	$K=10$	0.026	0.072	0.000	0.000

Tab. 4 FPR and FNR of EMVS and LASSO for Case 2

		FPR		FNR	
		EMVS	LASSO	EMVS	LASSO
$n=200$	$K=5$	0.008	0.097	0.025	0.048
	$K=10$	0.033	0.087	0.013	0.097
$n=500$	$K=5$	0.008	0.077	0.000	0.007
	$K=10$	0.017	0.073	0.002	0.002

Tab. 5 Variable selection of EMVS and LASSO for Case 3

	Method	MSE			FPR	FNR
		β_1	β_2	β_3		
$K=5$	EMVS	0.2443	0.2655	0.2351	0.000	0.036
	LASSO	0.2035	0.1751	0.2192	0.104	0.000
$K=10$	EMVS	0.1970	0.2462	0.2275	0.000	0.030
	LASSO	0.2073	0.2070	0.2207	0.099	0.002

This phenomenon gets more obvious in high-dimensional case. From Tab. 5, EMVS and LASSO show comparable MSE with different interior knots. While the two methods produce higher FNR and FPR than each other respectively. Consequently, EMVS performs well whether in common or in high-dimensional situations.

3 Real data analysis

In this section, we applied the proposed Bayesian variable selection procedures for PH model to the Worcester Heart Attack Study (WHAS) data set used in Ref. [36]. The goal of WHAS is to describe time trend associated with risk factors in long-term survival among residents following acute myocardial infraction (MI). A total of 500 patients were followed up from the hospital admission years 1997, 1999 and 2001. This dataset contains 22 attributes, identification code (id), age at hospital admission (age), gender (0 = male, 1 = female), initial heart rate (hr), initial systolic blood pressure (sysbp), initial diastolic blood pressure (diasbp), body mass index (bmi), history of cardiovascular disease (cvd, 0 = no, 1 = yes), atrial fibrillation (afb, 0=no, 1=yes), cardiogenic shock (sho, 0=no, 1=yes), congestive heart complications (chf, 0=no, 1=yes), complete heart block (av3, 0=no, 1 = yes), MI order (miord, 0 = first, 1 = recurrent), MI type (mitype, 0= non Q-wave, 1= Q-wave), cohort year (year, 1 = 1997, 2 = 1999, 3= 2001), hospital admission date (admitdate), hospital discharge date (disdate), date of last follow up (fdate), length of hospital stay (los), discharge status from hospital (dstat, 0 = alive, 1 = dead), total length of follwo-up (lenfol) and vital status at last follow-up (fstat, 0=alive, 1=dead).

We aim to identify the risk factors that affect the survival days after hospital discharge. As the data set only gives the date of last follow up and the vital status, for those people died before the last follow up, the accurate survival days is unknown but is known to be earlier than the date of last follow-up, which is left-censored. Without regard to those died during the hospitalization, the number of target subjects is 461. Among these 461 patients, there are 176 deaths, about 38% left-censoring.

Tab. 6　Results of the analysis of WHAS data set

Factor	EM			EMVS		
	Estimate	SEE	P-value	stimate	SEE	Index
age	4.9835	0.4004	0.0000	4.5401	0.7817	1
gender	−0.5086	0.1753	0.0037	−0.3112	0.1603	0
hr	2.1324	0.5228	0.0000	1.6006	0.4911	1
sysbp	1.4633	0.6940	0.0350	0.2169	0.1182	0
diasbp	−3.6527	1.0582	0.0006	−1.7694	0.7877	1
bmi	−0.4455	0.5718	0.4359	−0.1356	0.3748	0
cvd	−0.2340	0.1898	0.2178	−0.1244	0.2142	0
afb	0.2153	0.1951	0.2699	0.1950	0.1829	0
sho	0.4535	0.5115	0.3754	0.0848	0.7201	0
chf	0.8968	0.1756	0.0000	0.8658	0.3110	1
av3	0.7177	0.4060	0.0771	0.2350	0.3740	0
miord	0.0059	0.1654	0.9717	0.0420	0.1620	0
mitype	−0.3609	0.2028	0.0751	−0.3221	0.1720	0
year	−0.5249	0.2070	0.0112	−0.3439	0.1828	0

[Note] index=1 means that the factor is selected, index=0 otherwise.

To model survival time, Hosmer Jr et al[36] suggested fitting the PH model with six explanatory variables: age, bmi, hr, diasbp, gender and chf. For better modeling performance, all the explanatory variables are linearly transformed to [0, 1]. Based on this result, the PH model for WHAS data set is given by

$$\Lambda(t \mid \text{covariates}) = \Lambda_0(t)\exp\{\text{age}\beta_1 + \text{gender}\beta_2 + \text{hr}\beta_3 + \text{sysbp}\beta_4 + \text{diasbp}\beta_5 + \text{bmi}\beta_6 + \text{cvd}\beta_7 + \text{afb}\beta_8 + \text{sho}\beta_9 + \text{chf}\beta_{10} + \text{av3}\beta_{11} + \text{miord}\beta_{12} + \text{mitype}\beta_{13} + \text{year}\beta_{14}\}.$$

We applied the proposed method and the EM methods[13] without Bayesian variable selection to WHAS data set. The data analysis results are reported in Tab. 6. The left column shows the estimates, the estimated standard errors and the p-values of EM algorithm. The factors age, gender, hr, sysbp, diasbp, chf and year are shown to be significant. Our method selected four factors, age, hr, diasbp and chf, a subset of the significant covariates by EM method into the PH model. Old age, high heart rate and the existence of congestive heart complications will decrease the survival rate of people, and higher diastolic blood pressure has positive effect of the survival rate, which agrees with common sense. Fig. 1 provides the fitted curves of $\Lambda_0(\cdot)$. It is can be seen that the cumulative hazard no longer increases after reaching a threshold value.

4 Conclusion

In this paper, we developed an EMVS algorithm for variable selection of proportional hazards model in the context of survival analysis. Based on spike-and-slab prior and the two-stage data augmentation procedure, the proposed method is

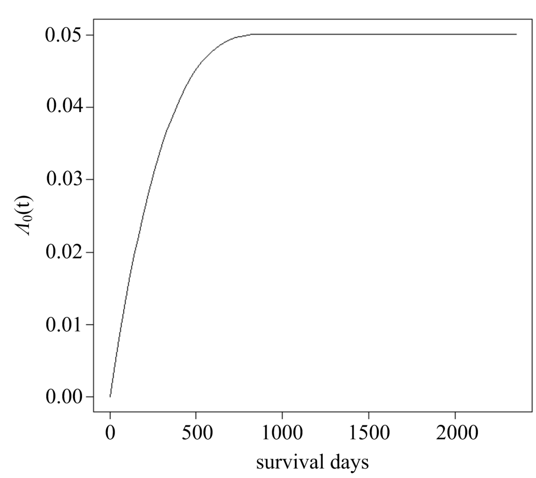

Fig. 1　The estimated baseline cumulative hazard function of WHAS data set

efficient and easy to implement. Both the simulation studies and the WHAS data analysis demonstrate the good performance of the proposed method. The method can be extended to other types of censored data, for example, the case Ⅱ interval-censored data, where the failure time is known to be lied in an interval. The two-stage data augmentation procedure can be applied to case Ⅱ interval-censored data directly. Furthermore, the linear assumption of the covariates effect can be softened. The nonparametric model may describe the more complex relationship between the survival time and the explanatory variables.

References

[1] White H D, Chew D P. Acute myocardial infarction. The Lancet, 2008, 372(9638): 570-584.

[2] Lirette S T, Aban I. Quantifying predictive accuracy in survival models. Journal of Nuclear Cardiology, 2017, 24(6): 1998-2003.

[3] Li R, Ren J J, Yang G, et al. Asymptotic behavior of Cox's partial likelihood and its application to variable selection. Statistica Sinica, 2018, 28(4):2713.

[4] Cox D R. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(2):187-202.

[5] Cox D R. Partial likelihood. Biometrika, 1975, 62(2): 269-276.

[6] Sun Jianguo. The Statistical Analysis of Interval-Censored Failure Time data. Springer, 2007.

[7] Finkelstein D M. A proportional hazards model for interval-censored failure time data. Biometrics, 1986, 42 (4): 845-854.

[8] Huang J. Efficient estimation for the proportional hazards model with interval censoring. The Annals of Statistics, 1996, 24(2):540-568.

[9] Pan W. A multiple imputation approach to cox regression with interval-censored data. Biometrics, 2000, 56(1):199-203.

[10] Tanner M A, Wong W H. The calculation of posterior distributions by data augmentation. Journal of the American statistical Association, 1987, 82(398):528-540.

[11] Ramsay O J. Monotone regression splines in action. Statistical science, 1988: 3(4): 425-441.

[12] Cai B, Lin X, Wang L. Bayesian proportional hazards model for current status data with monotone splines. Computational Statistics & Data Analysis, 2011, 55 (9):2644-2651.

[13] McMahan C S, Wang L, Tebbs J M. Regression analysis for current status data using the EM algorithm. Statistics in medicine, 2013, 32(25):4452-4466.

[14] Zeng D, Mao L, Lin D Y. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. Biometrika, 2016, 103(2): 253-271.

[15] Lu M, Li C S. Penalized estimation for proportional hazards models with current status data. Statistics in medicine, 2017, 36(30):4893-4907.

[16] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1):267-288.

[17] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 2001, 96(456):1348-1360.

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

[19] George E I, McCulloch R E. Variable selection via Gibbs sampling. Journal of the American Statistical Association, 1993, 88(423):881-889.

[20] George E I, McCulloch R E. Approaches for Bayesian variable selection. Statistica Sinica, 1997, 7: 339-373.

[21] Liang F, Paulo R, Molina G, et al. Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association, 2008, 103 (481):410-423.

[22] Narisetty N N, He X. Bayesian variable selection with shrinking and diffusing priors. The Annals of Statistics, 2014, 42(2):789-817.

[23] Wu Y, Cook R J. Penalized regression for interval-censored times of disease progression: Selection of hla markers in psoriatic arthritis. Biometrics, 2015, 71(3): 782-791.

[24] Scolas S, Ghouch A E, Legrand G, et al. Variable selection in a flexible parametric mixture cure model with interval-censored data. Statistics in medicine, 2016, 35(7):1210-1225.

[25] Zhao H, Wu Q, Li G, et al. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. Journal of the American Statistical Association, 2019, 115(529): 1-13.

[26] Li C, Pak D, Todem D. Adaptive lasso for the cox regression with interval censored and possibly left truncated data. Statistical Methods in Medical Research, 2020, 29(4): 1243-1255.

[27] Ročková V, George E I. EMVS: The em approach to Bayesian variable selection. Journal of the American Statistical Association, 2014, 109(506):828-846.

[28] Zhao K, Lian H. The expectation-maximization approach for Bayesian quantile regression. Computational Statistics & Data Analysis, 2016, 96 (C):1-11.

[29] Kendzor D E, Businelle M S. Bayesian variable selection for multistate Markov models with interval-censored data in an ecological momentary assessment study of smoking cessation. Biometrics, 2018a, 74(2):636-644.

[30] Koslovsky M D , Swartz M D, WChan L L N, et al. Using the em algorithm for Bayesian variable selection in

logistic regression models with related covariates. Journal of Statistical Computation and Simulation, 2018b, 88(3):575-596.
[31] Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association, 2002, 97(460):1042-1054.
[32] Wang L, McMahan C S, Hudgens M G, et al. A flexible, computationally efficient method for fitting the proportional hazards model to interval censored data. Biometrics, 2016, 72(1):222-231.
[33] Cui W, George E I. Empirical Bayes vs. fully Bayes variable selection. Journal of Statistical Planning and Inference, 2008, 138(4):888-900.
[34] Maruyama Y, George E I. Fully Bayes factors with a generalized g-prior. The Annals of Statistics, 2011, 39(5):2740-2765.
[35] Zhang H H, Lu W. Adaptive lasso for Cox's proportional hazards model. Biometrika, 2007, 94(3):691-703.
[36] Hosmer Jr D W, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time-to-Event Data, vol. 618, 2ed. John Wiley & Sons, 2011.

当前状态数据中比例风险模型的一种贝叶斯变量选择方法

崔 笛,张伟平
中国科学技术大学统计与金融系,安徽合肥 230026

摘要: 针对当前状态数据中的比例风险模型提出了一种基于期望-最大化的贝叶斯变量选择方法.该模型能够同时进行参数估计和变量选择,有效地增强了模型的可解释性和预测能力.为了识别风险因素,首先对表示协变量是否存在的指示变量赋予合适的先验分布,使用单调样条来近似基准累积风险函数;然后 通过使用基于泊松隐变量的两阶段数据扩充技术提出了一种有效的期望-最大化对模型拟合算法;最后通过模拟研究和一个实例分析证明了所提方法的有效性.

关键词: 比例风险模型;贝叶斯变量选择;当前状态数据;EM 算法;样条

Cui Di: PhD cadidate. Research field: Statistical model. E-mail: cuidi@mail.ustc.edu.cn
Zhang Weiping: Corresponding author, PhD/professor. Research field: Statistical learning theory.
E-mail: zwp@ustc.edu.cn

Appendix

Proof of Lemma 1.1

The marginal posterior distribution $P(\beta,\sigma^2 \mid \mathscr{D})$ is obtained by integrating α from $P(\beta,\alpha,\sigma^2 \mid \mathscr{D})$ as follows.

$$P(\beta,\sigma^2 \mid \mathscr{D}) = \int_{\mathbb{R}^K} P(\beta,\alpha,\sigma^2 \mid \mathscr{D})\mathrm{d}\alpha \propto P(\beta \mid \sigma^2)P(\sigma^2)\int_{\mathbb{R}^K} L(\mathscr{D} \mid \beta,\alpha)P(\alpha)\mathrm{d}\alpha.$$

By using the first-order Taylor expansion we have

$$L(\mathscr{D} \mid \beta,\alpha) = \prod_{i=1}^{n} [1-\exp\{-\alpha^{\mathrm{T}} I(c_i)\exp(x_i^{\mathrm{T}}\beta)\}]^{\delta_i} [\exp\{-\alpha^{\mathrm{T}} I(c_i)\exp(x_i^{\mathrm{T}}\beta)\}]^{1-\delta_i} \approx \prod_{i=1}^{n} \delta_i\alpha^{\mathrm{T}} I(c_i)\exp(x_i^{\mathrm{T}}\beta)\exp\{-(1-\delta_i)\alpha^{\mathrm{T}} I(c_i)\exp(x_i^{\mathrm{T}}\beta)\}.$$

Therefore,

$$\int_{\mathbb{R}^K} L(\mathscr{D} \mid \beta,\alpha)P(\alpha)\mathrm{d}\alpha = \int_{\mathbb{R}^K} \prod_{i=1}^{n} \delta_i\alpha^{\mathrm{T}} I(c_i)\exp(x_i^{\mathrm{T}}\beta)\exp\{-(1-\delta_i)\alpha^{\mathrm{T}} I(c_i)\exp(x_i^{\mathrm{T}}\beta)\}\lambda^K e^{-\lambda\alpha^{\mathrm{T}}}\mathrm{d}\alpha = \lambda^K \prod_{i=1}^{n} \delta_i I(c_i)\exp(x_i^{\mathrm{T}}\beta) \prod_{k=1}^{K} \frac{I_k(c_i)}{[(1-\delta_i)I_k(c_i)\exp(x_i^{\mathrm{T}}\beta)+\lambda/n]^3}.$$

The marginal joint posterior distribution of (β,σ^2) can be written as

$$P(\beta,\sigma^2 \mid \mathscr{D}) \propto \Big(\prod_{m=1}^{p} d_m\Big)^{-\frac{1}{2}} (\sigma^2)^{-\frac{p}{2}+a_1+1} \exp\left\{-\frac{1}{\sigma^2}\Big(\sum_{m=1}^{p}\frac{\beta_m{}^2}{d_m}+b_1\Big)\right\} \times$$

$$\lambda^K \prod_{i=1}^{n} \delta_i I(c_i)\exp(x_i^{\mathrm{T}}\beta) \prod_{k=1}^{K} \frac{I_k(c_i)}{[(1-\delta_i)I_k(c_i)\exp(x_i^{\mathrm{T}}\beta)+\lambda/n]^3}.$$

Then, the log posterior is given by

$$\log P(\beta,\sigma^2 \mid \mathscr{D}) = -(\frac{p}{2}+a_1+1)\log\sigma^2 - \frac{1}{\sigma^2}\Big(\sum_{m=1}^{p}\frac{{\beta_m}^2}{d_m}+b_1\Big)+ \sum_{i=1}^{n}\left[x_i^{\mathrm{T}}\beta+\frac{I_k(c_i)}{[(1-\delta_i)I_k(c_i)\exp(x_i^{\mathrm{T}}\beta)+\lambda/n]^3}\right]+C,$$

where C is the term not involving either β and σ^2. Denote

$$g(\beta)=\sum_{i=1}^{n}[x_i^{\mathrm{T}}\beta+I_k(c_i)/[(1-\delta_i)I_k(c_i)\exp(x_i^{\mathrm{T}}\beta)+\lambda/n]^3].$$

To show $g(\beta)$ is concave in each component of β, it suffices to show that $\frac{\partial^2 g(\beta)}{\partial{\beta_m}^2}<0$. It can be checked that

$$\frac{\partial^2 g(\beta)}{{\beta_m}^2}=-\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{x_i^2\exp(x_i^{\mathrm{T}}\beta[1-(1-\delta_i I_k(c_i))]}{[(1-\delta_i)I_k(c_i)\exp(x_i^{\mathrm{T}}\beta)+\lambda/n]^2}<0$$

as $I_k(\cdot)$, $k=1,\cdots,K$, are bounded in $[0,1]$. Therefore, $g(\beta)$ is concave in each component of β.

Denote

$$h(\beta,\sigma^2)=-(\frac{p}{2}+a_1+1)\log\sigma^2-\frac{1}{\sigma^2}\Big(\sum_{m=1}^{p}\frac{{\beta_m}^2}{d_m}+b_1\Big),$$

we introduce the coordinate transformation, which is defined by

$$r=1/\sigma^2,\ \phi_m=\beta_m\sqrt{\sigma^2},$$

where ϕ_m, $m=1,\cdots,p$, and r are continuous. Thus, unimodality is the original coordinates is equivalent in the transformed coordinates. Let $\phi=(\phi_1,\cdots,\phi_p)^{\mathrm{T}}$. With the above transformed coordinates, the above formula becomes

$$h(\phi,r)=\Big(\frac{p}{2}+a+1\Big)\log r-b_1 r-\frac{1}{2}\sum_{m=1}^{p}\frac{{\phi_m}^2}{d_m}.$$

It can be checked that $\frac{\partial^2 h(\phi,r)}{\partial r^2}<0$ and $\frac{\partial^2 h(\beta,\sigma^2)}{\partial{\phi_m}^2}<0$ for each m. Thus, $h(\beta,\sigma^2)$ is concave. Consequently, the joint posterior distribution of β and σ^2 is unimodal.