Ann Inst Stat Math (2015) 67:287–311
DOI 10.1007/s10463-014-0448-y

# Extended Bayesian information criterion in the Cox model with a high-dimensional feature space

**Shan Luo · Jinfeng Xu · Zehua Chen**

Received: 25 June 2013 / Revised: 2 November 2013 / Published online: 6 March 2014


**Abstract** Variable selection in the Cox proportional hazards model (the Cox model) has manifested its importance in many microarray genetic studies. However, theoretical results on the procedures of variable selection in the Cox model with a high-dimensional feature space are rare because of its complicated data structure. In this paper, we consider the extended Bayesian information criterion (EBIC) for variable selection in the Cox model and establish its selection consistency in the situation of high-dimensional feature space. The EBIC is adopted to select the best model from a model sequence generated from the SIS-ALasso procedure. Simulation studies and real data analysis are carried out to demonstrate the merits of the EBIC.

**Keywords** Variable selection · Cox model · Extended Bayesian information criterion · Selection consistency

## 1 Introduction

In many microarray genetic studies, a primary goal is to identify the genes which are associated with a phenotype. When the outcomes of interest are censored time-to-

S. Luo (✉)
Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, China
e-mail: sluo2012@gmail.com

J. Xu
Division of Biostatistics, Department of Population Health, School of Medicine,
New York University, New York 10016, USA

Z. Chen
Department of Statistics and Applied Probability, National University of Singapore,
Singapore 117546, Singapore

event data, the Cox model is usually used to model the covariate–response association (Cookson et al. 2009). Usually, the responsible genes occupy only a small proportion of the whole genome, which is the so-called "sparsity" phenomenon (Barabási et al. 2011). The objective of variable selection is to extract these responsible genes from an enormous number of genes. Various techniques have been proposed for variable selection in the Cox model with high-dimensional feature space. Regularization methods such as the Lasso (Tibshirani 1997), the adaptive lasso (Zhang and Lu 2007; Zou 2008) and the SCAD (Fan and Li 2002) have been proposed and shown to have a so-called oracle property, i.e., the ability to identify exactly the responsible features and estimate the model parameters as if the responsible features were known in advance. For an overview of such techniques, see Fan et al. (2005). For a more general case, in Du et al. (2010), the authors investigated the estimation and variable selection problem in cox model with semiparametric relative risk, where either the SCAD or the adaptive LASSO penalty is applied for variable selection for the parametric part. The resulting estimator of the parametric part was also shown to possess the oracle property. It is important to note that the oracle property is obtained under certain theoretical assumptions on the penalty parameter involved in the methods. In practice, the penalty parameter needs to be chosen by a certain model selection criterion.

Traditionally, model selection criteria such as cross validation (CV), generalized cross validation (GCV), Bayesian information criterion (BIC, Schwarz 1978), etc. have been used for model selection. However, in the case of high-dimensional feature space, it has been observed that these model selection criteria tend to select models with many spurious covariates and hence do not have the property of selection consistency, see Broman and Speed (2002), Siegmund (2004) and Bogdan et al. (2004). To tackle the problem of model selection with high-dimensional feature space, various variants of the BIC have been considered in the literature. Among them are the modified BIC (mBIC) proposed in Bogdan et al. (2004) and the extended BIC (EBIC) developed in Chen and Chen (2008). The mBIC is asymptotically a sepcial case of EBIC. The selection consistency of the EBIC has been established under various settings of high-dimensional feature space for model selection in linear and generalized linear models, see Chen and Chen (2008), Chen and Chen (2012), Luo and Chen (2013a) and Luo and Chen (2013b).

In this paper, we consider the EBIC for variable selection in the Cox model when the dimension of the feature space is high, particularly when it is much larger than the sample size. The selection consistency of the EBIC in the Cox model is rigorously established under fairly mild conditions. This is technically challenging since it involves the uniform convergence rate of the partial likelihood, which has never been treated in the literature. Our results allow the dimension of the feature space to be of order $O(n^{\kappa})$ for any $\kappa > 1$, and the number of relevant features is finite. The performance of the EBIC is demonstrated in simulation studies as well as in real data analysis with a microarray dataset.

The rest of the paper is organized as follows. The main theorems are provided in Sect. 2. The simulation studies are presented in Sect. 3. The real data analysis is given in Sect. 4. Some explanations on the conditions assumed for the theorems and the technical proofs are relegated to the Appendix.

## 2 Selection consistency of the EBIC in the Cox model

In this section, a brief review of the Cox model and the definition of the EBIC in the case of Cox model are given at first. Then the conditions and the theoretical results are stated. The theoretical results include a large deviation result on the score function of the partial likelihood of the Cox model (Theorem 1), a uniform convergence rate for the partial likelihood estimator of the Cox model (Theorem 2) and the selection consistency of the EBIC for Cox model (Theorem 3). Besides their application in the proof of Theorem 3, Theorem 1 and Theorem 2 are of their own interests in studies of high-dimensional data analysis.

Denote by $T$ and $C$ the survival and censoring times, respectively. Let $F$ and $G$, $f$ and $g$ be the cumulative distribution functions and density functions of $T$ and $C$, respectively. Let $z$ be a $p$-dimensional covariate vector which might depend on time $t$. The survival time $T$ and censoring time $C$ are assumed independent given the $z$.

We consider the right-censored model. The observations from $n$ individuals are $n$ triplets $\{(X_i, \delta_i, z_i) : i = 1, \ldots, n\}$ where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$ and $z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^\tau$. Suppose there are no ties in the observed times. The likelihood function based on the observations is given by

$$L = \prod_{i:\delta_i=1} f(X_i|z_i) \prod_{i:\delta_i=0} (1 - F(X_i|z_i)) = \prod_{i:\delta_i=1} h(X_i|z_i) \prod_{i=1}^{n} (1 - F(X_i|z_i)),$$

where

$$h(t|z) = \lim_{\Delta t \downarrow 0} P(t \leq T < t + \Delta t | T \geq t, z = z) = f(t|z)/(1 - F(t|z))$$

is the hazard function at $t$ given $z$. The Cox model assumes that

$$h(t|z) = h_0(t) \exp(z^\tau \boldsymbol{\beta}),$$

where $h_0(t)$ is a baseline hazard rate with cumulative hazard function

$$H_0(t) = \int_0^t h_0(u) \mathrm{d}u.$$

Without loss of generality, assume that the support of $h_0(t)$ is $[0, 1]$ and $\int_0^1 h_0(t)\mathrm{d}t < +\infty$.

Let $t_1 < t_2 < \cdots < t_N$ be the ordered observed survival times. Denote by $\{z_{(j)} : j = 1, \ldots, N\}$ the covariate vectors of the individuals with the survival times. Let $\mathscr{R}(t)$ be the risk set at time $t$, i.e., $\mathscr{R}(t) = \{i : X_i \geq t\}$. In the Cox model, the baseline cumulative hazard function $H_0(t)$ is modeled nonparametrically as $H_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. Hence, $h_0(t_j) = h_j$ and the second product of the likelihood function is expressed as

$$\prod_{i=1}^{n}(1-F(X_i|z_i)) = \exp\left(-\sum_{i=1}^{n} H(X_i|z_i)\right)$$
$$= \exp\left(-\sum_{j=1}^{N} h_j \sum_{i=1}^{n} I(i \in \mathscr{R}(t_j)) \exp(z_i^\tau \boldsymbol{\beta})\right)$$
$$= \exp\left(-\sum_{j=1}^{N} h_j \sum_{i\in\mathscr{R}(t_j)} \exp(z_i^\tau \boldsymbol{\beta})\right).$$

The likelihood function $L$ reduces to

$$L = \left(\prod_{j=1}^{N} h_j\right)\left(\prod_{j=1}^{N} \exp(z_{(j)}^\tau \boldsymbol{\beta})\right) \exp\left(-\sum_{j=1}^{N} h_j \sum_{i\in\mathscr{R}(t_j)} \exp(z_i^\tau \boldsymbol{\beta})\right).$$

Maximizing the log of the likelihood function with respect to the $h_j$'s for any fixed $\boldsymbol{\beta}$ yields the log partial likelihood:

$$\ell_n(\boldsymbol{\beta}) = \sum_{j=1}^{N}\left(z_{(j)}^\tau \boldsymbol{\beta} - \ln\left(\sum_{i\in\mathscr{R}(t_j)} \exp(z_i^\tau \boldsymbol{\beta})\right)\right).$$

Let $s$ be a subset of $\{1, 2, \cdots, p\}$. Let $|s|$ be the number of indices in $s$. By convention, we denote by $z(s)$ and $\boldsymbol{\beta}(s)$ the vectors obtained from the components of $z$ and $\boldsymbol{\beta}$ whose indices are contained in $s$. For convenience, we refer to $s$ as both the index set and the model consisting of the covariate vector $z(s)$. Let $\mathscr{S}_j$ be the set of all models consisting of $j$ covariates and $\tau(\mathscr{S}_j) = \binom{p}{j}$. In the context of Cox model, for model $s$, the EBIC proposed in Chen and Chen (2008) is defined as follows:

$$\text{EBIC}_\gamma(s) = -2\ell_n(\widehat{\boldsymbol{\beta}}(s)) + |s|\ln n + 2\gamma \ln \tau(\mathcal{S}_{|s|}), \quad 0 \le \gamma \le 1,$$

where $\widehat{\boldsymbol{\beta}}(s)$ is the maximum partial likelihood estimator of $\boldsymbol{\beta}(s)$. When $\gamma = 0$, the EBIC reduces to the original BIC. The EBIC with $\gamma = 1$ is equivalent to the mBIC mentioned in Sect. 1.

In order to establish the selection consistency of the EBIC above, we need the tool of counting processes. In the following, we define some needed counting processes and express the partial log likelihood function in terms of the counting processes. Without loss of generality, we confine time $t$ to the interval [0, 1]. At time $t$, define

$$N_i(t) = I(X_i \le t, \delta_i = 1), \quad Y_i(t) = I(X_i \ge t).$$

Let

$$S_n(\boldsymbol{\beta}(s), t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) \exp\left(z_i^\tau(s)\boldsymbol{\beta}(s)\right).$$

Define

$$l_n(\boldsymbol{\beta}(s), t) = \sum_{i=1}^{n} \int_0^t z_i^\tau(s)\boldsymbol{\beta}(s)\mathrm{d}N_i(u) - \int_0^t \ln\left(nS_n(\boldsymbol{\beta}(s), u)\right) \mathrm{d}\bar{N}(u),$$

where $\bar{N}(u) = \sum_{i=1}^{n} N_i(u)$.

In terms of the counting processes defined above, the log partial likelihood function is expressed as

$$\begin{aligned}\ell_n(\boldsymbol{\beta}(s)) &= \sum_{i=1}^{n} \delta_i \left( z_i^\tau(s)\boldsymbol{\beta}(s) - \ln\left( \sum_{k=1}^{n} Y_k(X_i) \exp(z_k^\tau(s)\boldsymbol{\beta}(s)) \right) \right) \\ &= \ln\left( \prod_{i=1}^{n} \prod_{0\le u\le 1} \left\{ \frac{\exp(z_i^\tau(s)\boldsymbol{\beta}(s))}{\sum_{k=1}^{n} Y_k(u)\exp(z_k^\tau(s)\boldsymbol{\beta}(s)))} \right\}^{\Delta N_i(u)} \right) \\ &= l_n(\boldsymbol{\beta}(s), 1),\end{aligned}$$

where $\Delta N_i(u) = 1$, if $N_i(u) - N_i(u-) = 1$; 0, otherwise.

The conditions required for the selection consistency of EBIC are in terms of the first and second derivatives of the processes $S_n(\boldsymbol{\beta}(s), t)$ and $l_n(\boldsymbol{\beta}(s), t)$. These derivatives are given below:

$$\begin{aligned}
S_n^{(1)}(\boldsymbol{\beta}(s), t) &= \frac{\partial S_n(\boldsymbol{\beta}(s), t)}{\partial \boldsymbol{\beta}(s)} = \frac{1}{n}\sum_{i=1}^{n} z_i(s) Y_i(t) \exp\left(z_i^\tau(s)\boldsymbol{\beta}(s)\right), \\
S_n^{(2)}(\boldsymbol{\beta}(s), t) &= \frac{\partial^2 S_n(\boldsymbol{\beta}(s), t)}{\partial \boldsymbol{\beta}(s)\partial \boldsymbol{\beta}^\tau(s)} = \frac{1}{n}\sum_{i=1}^{n} z_i(s) z_i^\tau(s) Y_i(t) \exp\left(z_i^\tau(s)\boldsymbol{\beta}(s)\right), \\
E_n(\boldsymbol{\beta}(s), t) &= \frac{\partial \ln(S_n(\boldsymbol{\beta}(s), t))}{\partial \boldsymbol{\beta}(s)} = \frac{S_n^{(1)}(\boldsymbol{\beta}(s), t)}{S_n(\boldsymbol{\beta}(s), t)}, \\
V_n(\boldsymbol{\beta}(s), t) &= \frac{\partial^2 \ln(S_n(\boldsymbol{\beta}(s), t))}{\partial \boldsymbol{\beta}(s)\partial \boldsymbol{\beta}^\tau(s)} = \frac{S_n^{(2)}(\boldsymbol{\beta}(s), t)}{S_n(\boldsymbol{\beta}(s), t)} - E_n(\boldsymbol{\beta}(s), t) E_n^\tau(\boldsymbol{\beta}(s), t) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{[z_i(s) - E_n(\boldsymbol{\beta}(s), t)][z_i(s) - E_n(\boldsymbol{\beta}(s), t)]^\tau Y_i(t) \exp\left(z_i^\tau(s)\boldsymbol{\beta}(s)\right)}{S_n(\boldsymbol{\beta}(s), t)}, \\
U_n(\boldsymbol{\beta}(s), t) &= \frac{\partial l_n(\boldsymbol{\beta}(s), t)}{\partial \boldsymbol{\beta}(s)} = \sum_{i=1}^{n} \int_0^t (z_i(s) - E_n(\boldsymbol{\beta}(s), u))\, \mathrm{d}M_i(u), \\
I_n(\boldsymbol{\beta}(s), t) &= -\frac{\partial l_n(\boldsymbol{\beta}(s), t)}{\partial \boldsymbol{\beta}(s)\partial \boldsymbol{\beta}^\tau(s)} = n\int_0^t V(\boldsymbol{\beta}(s), u)\, S_n(\boldsymbol{\beta}(s), u)\, h_0(u)\mathrm{d}u,
\end{aligned}$$

where $M_i(u) = N_i(u) - \int_0^u Y_i(w)\exp\left(z_i^\tau(s)\boldsymbol{\beta}(s)\right)h_0(w)\mathrm{d}w$ is a martingale with respect to the $\sigma$-fields $\{\mathscr{F}_u^{(n)} : u \geq 0\}$ defined by

$$\mathscr{F}_u^{(n)} = \sigma\left\{N_i(w), I(X_i \leq w, \delta_i = 0) : 0 \leq w \leq u, 1 \leq i \leq n\right\}.$$

Let

$$\begin{aligned} s(\boldsymbol{\beta}(s), t) &= \lim_{a.s.} S_n(\boldsymbol{\beta}(s), t), \quad s^{(l)}(\boldsymbol{\beta}(s), t) = \lim_{a.s.} S_n^{(l)}(\boldsymbol{\beta}(s), t), l = 1, 2, \\ e(\boldsymbol{\beta}(s), t) &= \lim_{a.s.} E_n(\boldsymbol{\beta}(s), t), \quad v(\boldsymbol{\beta}(s), t) = \lim_{a.s.} V_n(\boldsymbol{\beta}(s), t), \\ \Sigma(\boldsymbol{\beta}(s), t) &= \lim_{a.s.} \frac{1}{n} I_n((\boldsymbol{\beta}(s), t). \end{aligned}$$

We now state the conditions below. Let $s_0 = \{j : \beta_j \neq 0\}$ and $p_0 = |s_0|$. Let $\boldsymbol{\beta}_0$ denote the true values of $\boldsymbol{\beta}$. Let $C > 1$ be any fixed constant. Define

$$\mathscr{A}_0 = \{s : s_0 \subset s, |s| \leq Cp_0\}, \quad \mathscr{A}_1 = \{s : s_0 \not\subset s, |s| \leq Cp_0\}.$$

**Assumption A0** $p = p_n = O(n^\kappa)$ for some $\kappa > 1$ and $\boldsymbol{\beta}_0(s_0)$ is independent of the sample size $n$.

**Assumption A1** A1.1. For $u$ in a neighborhood of zero,

$$\begin{aligned} &\sup_{t\in[0,1]} E\exp\left[uY_i(t)\exp\left(z_i^\tau\boldsymbol{\beta}_0\right)\right] < +\infty; \\ &\sup_{t\in[0,1]} E\exp\left[uz_{ij}Y_i(t)\exp\left(z_i^\tau\boldsymbol{\beta}_0\right)\right] < +\infty, \quad \forall j \in \{1, 2, \ldots, p_n\}; \\ &\sup_{t\in[0,1]} E\exp\left[uz_{ij}z_{il}Y_i(t)\exp\left(z_i^\tau\boldsymbol{\beta}_0\right)\right] < +\infty, \quad \forall j, l \in \{1, 2, \ldots, p_n\}. \end{aligned}$$

A1.2. As functions of $t$, $s(\boldsymbol{\beta}_0, t), s^{(1)}(\boldsymbol{\beta}_0, t), s^{(2)}(\boldsymbol{\beta}_0, t)$ are element-wise bounded and $s(\boldsymbol{\beta}_0, t)$ is bounded away from 0; the family of functions (as functions of $\boldsymbol{\beta}$) $\{s(\boldsymbol{\beta}, t), s^{(1)}(\boldsymbol{\beta}, t), s^{(2)}(\boldsymbol{\beta}, t) : 0 \leq t \leq 1\}$ is an equi-continuous family.

A1.3. $\Sigma(\boldsymbol{\beta}_0, 1)$ is positive definite.

A1.4. The process $Y(t) = (Y_1(t), \ldots, Y_n(t))^\tau$ is left continuous with right-hand limits and satisfies $P\left(Y(t) = 1, 0 \leq t \leq 1\right) > 0$; the covariate vector $Z(t)$ is left continuous if $Z$ depends on $t$.

**Assumption A2** Let $\xi_{ij} = \int_0^1 \left(z_{ij}(t) - e_j(\boldsymbol{\beta}_0, t)\right)\mathrm{d}M_i(t)$, where $e_j(\boldsymbol{\beta}_0, t)$ is the $j$th component of $e(\boldsymbol{\beta}_0, t)$. For any set $s \in \mathscr{A}_0$, any fixed $|s|-$dimensional vector $\boldsymbol{a}$ satisfying $\boldsymbol{var}(\sum_{i=1}^n \sum_{j\in s} \boldsymbol{a}_j\xi_{ij}/\sqrt{n}) = 1$, and any $u$ in a neighborhood of zero,

$$E\exp\left[u\sum_{j\in s}\boldsymbol{a}_j\xi_{ij}\right] < +\infty. \tag{1}$$

**Assumption A3** A3.1 . Let $\lambda_{1n} = \inf_{s\in\mathscr{A}_1} \lambda_{\min}\left(I\left(\boldsymbol{\beta}_0(s), 1\right)\right)$. There exists a positive constant $c$ such that $\lambda_{1n} \geq cn^{2/3-\delta}$ for some $\delta \in (0, 1/6)$.

A3.2. For any given $\varepsilon > 0$, there exists a constant $\delta > 0$ such that, when $n$ is sufficiently large,

$$I\left(\boldsymbol{\beta}(s), 1\right) \geq (1-\varepsilon) I\left(\boldsymbol{\beta}_0(s), 1\right)$$

for all $\boldsymbol{\beta}(s)$ such that $s \in \mathscr{A}_0$ and $\|\boldsymbol{\beta}(s) - \boldsymbol{\beta}_0(s)\|_2 \leq \delta$.

Some explanations and remarks on the above assumptions are given in the Appendix.

We now present our main theoretical results. First we give two theorems which are needed in the proof of the selection consistency of the EBIC.

**Theorem 1** (Large deviation of the score function) *Under Assumptions* A0, A1 *and* A2, *for any positive sequence* $u_n$ *satisfying* $(\ln n)^{-1/2} u_n \to +\infty$ *and* $n^{-1/6} u_n \to 0$, *as* $n \to +\infty$, *and for any arbitrary* $\varepsilon > 0$, *there exists positive constant* $c_0$ *such that* $U_{nj}(\boldsymbol{\beta}_0, 1)$, *the* $j$*th component of* $U_n(\boldsymbol{\beta}_0, 1)$, *satisfies*

$$P\left(|U_{nj}(\boldsymbol{\beta}_0, 1)| > \sqrt{n} u_n\right) \leq c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right), \quad j = 1, 2, \ldots, p. \tag{2}$$

*Furthermore, for any unit vector* $\boldsymbol{u}$ *and* $s \in \mathscr{A}_0$,

$$P(|\boldsymbol{u}^\tau [\Sigma(\boldsymbol{\beta}_0(s), 1)]^{-1/2} U_n(\boldsymbol{\beta}_0(s), 1)| > \sqrt{n} u_n) \leq c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right). \tag{3}$$

**Theorem 2** (Uniform convergence of the partial likelihood estimator) *Under Assumptions* A0, A1, A2 *and* A3,

$$P(\|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2 = O(\psi_n)) \to 1,$$

*uniformly for* $s \in \mathscr{A}_0$, *where* $\hat{\boldsymbol{\beta}}(s)$ *is the maximum partial likelihood estimator of* $\boldsymbol{\beta}(s)$, *and*

$$\psi_n \to 0, \lambda_{1n}\psi_n(n \ln n)^{-1/2} \to +\infty, \ \lambda_{1n}\psi_n n^{-2/3} \to 0. \tag{4}$$

The selection consistency of the EBIC is given in the following theorem:

**Theorem 3** (Selection consistency of the EBIC) *Under Assumptions* A0, A1, A2 *and* A3, *when* $\gamma > 1 - \frac{\ln n}{2\ln p} = 1 - \frac{1}{2\kappa}$,

$$P\left(\min_{s: s\neq s_0, |s|\leq Cp_0} \text{EBIC}_\gamma(s) \leq \text{EBIC}_\gamma(s_0)\right) \to 0, \quad \text{as } n \to \infty,$$

*where* $C$ *is any fixed constant >1.*

Theorem 3 implies that, if the model selection is made among the models with size smaller than or comparable with that of the true model $s_0$, the EBIC is able to identify the true model with probability 1 asymptotically. The confining of the model size to the range $< Cp_0$ is reasonable in practice, since in practical problems people will never consider models whose size is much larger than that of the true model.

The proofs of the theorems are given in the Appendix.

## 3 Simulation study

It is a common practice that the variable selection with high-dimensional feature spaces is carried out in multiple stages. In the first stage, a simple screening procedure is used to screen out obviously non-important variables to reduce the computational burden. In the second stage, a sophisticated approach is used for the final selection. In our simulation studies, we use the marginal utility ranking procedure (SIS) elaborated in Fan et al. (2010) for the screening in the first stage; that is, consider the maximum likelihood for each of the $p$ models, each consisting of only one variable, and retain the variables with the higher maximum likelihoods. In the second stage, we use the adaptive Lasso considered in Zhang and Lu (2007) to generate candidate models with a sequence of penalty parameter values. The EBIC is then used to select among those candidate models. The whole procedure is denoted by SIS-ALasso-EBIC.

In the simulation study, we aim at evaluating the performance of SIS-ALasso-EBIC in the Cox model when the dimension of the feature space is high. Specifically, we let $\kappa = 1.25$, $p_n = [n^\kappa]$. We consider $n$ = 100, 150, 200, 250 with corresponding $p_n$ = 316, 524, 752, 994. We compare the performance of SIS-ALasso-EBIC with different $\gamma$ values in terms of the positive discovery rate (PDR) and the false discovery rate (FDR) averaged over the simulation replicates. The PDR and FDR are defined by

$$\text{PDR} = \frac{|s^* \cap s_0|}{|s_0|}, \quad \text{FDR} = \frac{|s^* \cap s_0^c|}{|s^*|},$$

where $s_0$ and $s^*$ are the set of true and selected features respectively. The model selection consistency implies that PDR converges to 1 and FDR converges to 0 simultaneously as $n$ goes to $+\infty$. The $\gamma$ values considered are $(\gamma_1, \gamma_2, \gamma_3) = (0, 1 - 1/(4\kappa), 1)$. The EBIC with $\gamma_1$ corresponds to the original BIC. The EBIC with $\gamma_3$ corresponds to mBIC. The value $\gamma_2$ is slightly larger than the lower bound of the consistency range given in Theorem 3.

For a comparison with other benchmark analysis, we have also included the results from other procedures designed specifically for $n < p$ scenario. For instance, the SIS-SCAD procedure proposed in Fan et al. (2010), which can be realized directly using R package `SIS`. The regularization parameters and model selection method are all chosen to be the values by default in `SIS`. We note that, in Zhang and Lu (2007), the Generalized cross validation (GCV) was proposed as a model selection criterion after ALasso. However, it was defined under the $n > p$ situation. Therefore, we excluded it from our simulation study for a fair comparison.

**Table 1** The censoring proportion averaged over 200 replicates

| $L$ | $n = 100$ | | $n = 150$ | | $n = 200$ | | $n = 250$ | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| $L = 2$ | 0.470 | (0.061) | 0.471 | (0.043) | 0.467 | (0.042) | 0.468 | (0.037) |
| $L = 3$ | 0.448 | (0.065) | 0.451 | (0.048) | 0.446 | (0.047) | 0.448 | (0.042) |
| $L = 4$ | 0.433 | (0.068) | 0.435 | (0.051) | 0.429 | (0.051) | 0.432 | (0.046) |
| $L = 50$ | 0.277 | (0.086) | 0.275 | (0.073) | 0.273 | (0.070) | 0.275 | (0.066) |

### 3.1 Simulation 1

In this subsection, we examine the influence of the censoring proportion on the performance of the procedure. The data setting in this simulation is adapted from that considered in Fan and Li (2002) and Zou (2008).

1. The predictors in $X$ are normally distributed with mean $\mathbf{0}$ and covariances $\sigma_{ij} = 0.5^{|i-j|}$.
2. The true parameter vector $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_9 = 0.8, \boldsymbol{\beta}_4 = \boldsymbol{\beta}_{12} = 1, \boldsymbol{\beta}_7 = \boldsymbol{\beta}_{15} = 0.6$ and 0 for other components.
3. The survival time $T$ is generated as $\ln T = -X^\tau \boldsymbol{\beta} + \ln \varepsilon$, where $\varepsilon \sim \exp(1)$, therefore, $h(t|X) = \exp(X^\tau \boldsymbol{\beta})$. The censoring time was simulated from an exponential distribution with mean $U \exp(X^\tau \boldsymbol{\beta}_0)$, where $U$ follows a uniform distribution $U(1, L)$ and it is independent of $\varepsilon$.

The value of $L$ controls the censoring proportion in the data. We let $L = 2, 3, 4, 50$. The sample mean (mean) and standard deviations (sd) of the corresponding censoring proportions simulated from 200 replicates are summarized in Table 1. The larger the $L$, the less the censoring proportion. The averaged PDR and FDR over 200 replicates and their standard deviations (in brackets) are given in Table 2. The censoring proportion does have an effect on the accuracy of the variable selection. It is obvious from Table 2 that the higher the censoring proportion, the lower the PDR and that the FDRs are comparable.

### 3.2 Simulation 2

In this subsection, we examine the performance of the procedure when the predictors have different linear correlations. The data settings in this simulation are adapted from Fan et al. (2010).

1. The predictors in $X$ are normally distributed with mean $\mathbf{0}$ the variances and covariances: $\sigma_{ii} = 1,\ \sigma_{ij} = \rho,\ i \neq j$.
2. The first 6 components of $\boldsymbol{\beta}$ are given below:

$$(\boldsymbol{\beta}_{0,1}, \boldsymbol{\beta}_{0,2}, \boldsymbol{\beta}_{0,3}, \boldsymbol{\beta}_{0,4}, \boldsymbol{\beta}_{0,5}, \boldsymbol{\beta}_{0,6}) = (-1.6328, 1.3988, -1.6497, 1.6353, -1.4209, 1.7022).$$

**Table 2** The PDR and FDR averaged over 200 replicates for Simulation 1

| L | | $n = 100$ | | $n = 150$ | | $n = 200$ | | $n = 250$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | PDR | FDR | PDR | FDR | PDR | FDR | PDR | FDR |
| 2 | $\text{EBIC}_{\gamma_1}$ | 0.713 (0.183) | 0.551 (0.176) | 0.873 (0.135) | 0.488 (0.193) | 0.953 (0.082) | 0.425 (0.184) | 0.969 (0.067) | 0.422 (0.196) |
| | $\text{EBIC}_{\gamma_2}$ | 0.345 (0.269) | 0.192 (0.275) | 0.659 (0.267) | 0.170 (0.164) | 0.844 (0.186) | 0.174 (0.142) | 0.933 (0.104) | 0.151 (0.129) |
| | $\text{EBIC}_{\gamma_3}$ | 0.241 (0.256) | 0.115 (0.242) | 0.600 (0.285) | 0.132 (0.174) | 0.811 (0.211) | 0.148 (0.143) | 0.902 (0.138) | 0.122 (0.118) |
| | SCAD | 0.435 (0.134) | 0.478 (0.161) | 0.614 (0.140) | 0.474 (0.120) | 0.723 (0.124) | 0.518 (0.083) | 0.835 (0.066) | 0.545 (0.122) |
| 3 | $\text{EBIC}_{\gamma_1}$ | 0.728 (0.179) | 0.540 (0.197) | 0.883 (0.128) | 0.470 (0.197) | 0.951 (0.105) | 0.407 (0.197) | 0.964 (0.075) | 0.399 (0.198) |
| | $\text{EBIC}_{\gamma_2}$ | 0.384 (0.289) | 0.204 (0.282) | 0.683 (0.259) | 0.172 (0.186) | 0.871 (0.174) | 0.173 (0.142) | 0.936 (0.100) | 0.133 (0.122) |
| | $\text{EBIC}_{\gamma_3}$ | 0.275 (0.280) | 0.099 (0.228) | 0.628 (0.270) | 0.127 (0.164) | 0.838 (0.198) | 0.143 (0.133) | 0.916 (0.121) | 0.111 (0.112) |
| | SCAD | 0.449 (0.132) | 0.461 (0.158) | 0.619 (0.139) | 0.469 (0.119) | 0.734 (0.130) | 0.511 (0.086) | 0.845 (0.110) | 0.539 (0.060) |
| 4 | $\text{EBIC}_{\gamma_1}$ | 0.737 (0.182) | 0.523 (0.195) | 0.893 (0.134) | 0.460 (0.196) | 0.957 (0.093) | 0.414 (0.193) | 0.973 (0.068) | 0.418 (0.202) |
| | $\text{EBIC}_{\gamma_2}$ | 0.398 (0.296) | 0.202 (0.281) | 0.713 (0.251) | 0.170 (0.178) | 0.882 (0.165) | 0.167 (0.140) | 0.938 (0.102) | 0.133 (0.123) |
| | $\text{EBIC}_{\gamma_3}$ | 0.309 (0.289) | 0.121 (0.236) | 0.664 (0.267) | 0.136 (0.163) | 0.849 (0.187) | 0.132 (0.128) | 0.925 (0.117) | 0.112 (0.115) |
| | SCAD | 0.458 (0.134) | 0.450 (0.161) | 0.627 (0.144) | 0.463 (0.123) | 0.747 (0.137) | 0.502 (0.091) | 0.853 (0.110) | 0.535 (0.060) |
| 50 | $\text{EBIC}_{\gamma_1}$ | 0.866 (0.141) | 0.427 (0.173) | 0.956 (0.079) | 0.372 (0.197) | 0.982 (0.057) | 0.356 (0.197) | 0.994 (0.031) | 0.353 (0.204) |
| | $\text{EBIC}_{\gamma_2}$ | 0.668 (0.255) | 0.162 (0.162) | 0.898 (0.147) | 0.142 (0.142) | 0.957 (0.090) | 0.114 (0.111) | 0.985 (0.051) | 0.074 (0.101) |
| | $\text{EBIC}_{\gamma_3}$ | 0.575 (0.292) | 0.119 (0.292) | 0.869 (0.170) | 0.124 (0.140) | 0.950 (0.102) | 0.102 (0.106) | 0.980 (0.062) | 0.062 (0.092) |
| | SCAD | 0.530 (0.118) | 0.364 (0.141) | 0.698 (0.124) | 0.402 (0.106) | 0.811 (0.118) | 0.459 (0.079) | 0.898 (0.101) | 0.510 (0.055) |

All the other components of $\boldsymbol{\beta}$ are 0.

3. The survival time $T$ is generated as $\ln T = -X^\tau \boldsymbol{\beta}_0 + \ln \varepsilon$, where $\varepsilon \sim \exp(1)$. The censoring time was simulated from an exponential distribution with mean $U \exp(X^\tau \boldsymbol{\beta}_0)$, where $U$ follows a uniform distribution $U(1, 3)$ and it is independent of $\varepsilon$.

The $\rho$ controls the linear associations among the predictors and we consider $\rho = 0, 0.3, 0.5$. The averaged PDR and FDR over 200 replicates and their standard deviations (in brackets) are presented in Table 3.

The following points manifest themselves in Table 3: (i) for the EBIC with $\gamma_2$ and $\gamma_3$ (both are in the consistency range), the PDR rapidly approaches 1 and the FDR steadily approaches 0, as $n$ increases, which reflects the theoretical property of selection consistency; (ii) the original BIC (corresponding to $\gamma_1$) does not seem to be selection consistent, since the FDR stays away from zero as $n$ increases; (iii) the mBIC (corresponding to $\gamma_3$) is more conservative than the EBIC (corresponding to $\gamma_2$); (iv) the correlation structure affects the accuracy of the variable selection: the more correlated the covariates, the less accurate the variable selection. The SIS-SCAD procedure performs the worst in the sense that, their FDRs are close to or even greater than the original BIC, which has the highest FDR among $\text{EBIC}_{\gamma_1}$, $\text{EBIC}_{\gamma_2}$, $\text{EBIC}_{\gamma_3}$, while its PDRs are much smaller than the original BIC.

## 4 Real data analysis

In this section, we consider a data set published and analyzed in Rosenwald et al. (2002). The data set has also been studied in Gui and Li (2005) and Sha et al. (2006). In this dataset, 240 patients were monitored using a Lyphochip cDNA microarray with 7399 probes. Their follow-up time after chemotherapy and survival status at the follow-up time were also recorded. The censoring proportion is 0.425. Many of the gene expression measurements of the 7399 genes (genes sharing the same name but having different predictor values are considered different) are missing. In our study, we applied the technique of Troyanskaya et al. (2001) to impute the missing values; that is, the missing values are imputed by the average expression levels of their physically nearest 8 neighboring genes with complete observations.

We applied the SIS-ALasso-EBIC method to the data set. First, the 7399 genes are screened using the maximum partial likelihoods of uni-covariate models. The $0.6n$ genes with larger maximum partial likelihoods are retained. Then the adaptive Lasso is applied to these genes. The EBIC with $\gamma = 0.7$ (which is slightly larger than $1 - \frac{\ln n}{2 \ln p} = 0.6924$) is used for the model selection. The following two genes are identified to be related to diffuse large B-cell lymphoma: HLA-DQ$\alpha$ and HLA-DP$\alpha$. Note that HLA-DQ$\alpha$ is the second important gene selected by the LARS-Cox procedure in Gui and Li (2005) and one of the representative genes selected in Rosenwald et al. (2002), and that HLA-DP$\alpha$ is detected in Sha et al. (2006) and Rosenwald et al. (2002) but not in Gui and Li (2005). By using smaller $\gamma$ values, more genes are selected. The $\gamma$ values together with the selected genes are given in Table 4.

**Table 3** The PDR and FDR averaged over 200 replicates for Simulation 2

| $\rho$ | | $n = 100$ | | $n = 150$ | | $n = 200$ | | $n = 250$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | PDR | FDR | PDR | FDR | PDR | FDR | PDR | FDR |
| 0 | $\mathrm{EBIC}_{\gamma_1}$ | 0.780 (0.184) | 0.548 (0.188) | 0.926 (0.104) | 0.460 (0.198) | 0.959 (0.076) | 0.314 (0.216) | 0.984 (0.064) | 0.263 (0.208) |
| | $\mathrm{EBIC}_{\gamma_2}$ | 0.479 (0.366) | 0.177 (0.226) | 0.871 (0.220) | 0.279 (0.185) | 0.955 (0.088) | 0.145 (0.136) | 0.982 (0.074) | 0.084 (0.119) |
| | $\mathrm{EBIC}_{\gamma_3}$ | 0.348 (0.367) | 0.106 (0.198) | 0.817 (0.299) | 0.240 (0.197) | 0.952 (0.101) | 0.135 (0.133) | 0.978 (0.101) | 0.081 (0.114) |
| | SCAD | 0.454 (0.149) | 0.455 (0.179) | 0.608 (0.151) | 0.479 (0.129) | 0.779 (0.141) | 0.481 (0.094) | 0.865 (0.122) | 0.528 (0.067) |
| 0.3 | $\mathrm{EBIC}_{\gamma_1}$ | 0.591 (0.262) | 0.525 (0.243) | 0.776 (0.242) | 0.437 (0.218) | 0.858 (0.199) | 0.324 (0.203) | 0.914 (0.156) | 0.271 (0.187) |
| | $\mathrm{EBIC}_{\gamma_2}$ | 0.274 (0.323) | 0.108 (0.216) | 0.620 (0.374) | 0.196 (0.210) | 0.814 (0.260) | 0.161 (0.158) | 0.898 (0.197) | 0.110 (0.148) |
| | $\mathrm{EBIC}_{\gamma_3}$ | 0.202 (0.306) | 0.047 (0.141) | 0.527 (0.412) | 0.144 (0.192) | 0.804 (0.277) | 0.147 (0.145) | 0.893 (0.209) | 0.101 (0.143) |
| | SCAD | 0.366 (0.140) | 0.561 (0.169) | 0.502 (0.153) | 0.569 (0.131) | 0.646 (0.169) | 0.569 (0.112) | 0.748 (0.177) | 0.592 (0.096) |
| 0.5 | $\mathrm{EBIC}_{\gamma_1}$ | 0.454 (0.286) | 0.455 (0.321) | 0.626 (0.311) | 0.406 (0.278) | 0.732 (0.290) | 0.292 (0.239) | 0.818 (0.257) | 0.262 (0.208) |
| | $\mathrm{EBIC}_{\gamma_2}$ | 0.180 (0.263) | 0.071 (0.185) | 0.437 (0.385) | 0.145 (0.208) | 0.665 (0.347) | 0.137 (0.157) | 0.784 (0.304) | 0.121 (0.159) |
| | $\mathrm{EBIC}_{\gamma_3}$ | 0.125 (0.226) | 0.028 (0.113) | 0.376 (0.386) | 0.093 (0.163) | 0.635 (0.373) | 0.121 (0.151) | 0.763 (0.326) | 0.096 (0.130) |
| | SCAD | 0.334 (0.133) | 0.599 (0.160) | 0.453 (0.155) | 0.611 (0.133) | 0.569 (0.175) | 0.621 (0.117) | 0.673 (0.195) | 0.633 (0.107) |

**Table 4** Genes selected via the EBIC

| GenBank ID | Signature | $0 \le \gamma \le 0.1$ | $\gamma = 0.2$ | $0.3 \le \gamma \le 0.7$ | $\gamma \ge 0.8$ |
|---|---|---|---|---|---|
| AA278718 | | + | + | − | − |
| AA004687 | | + | + | − | − |
| LC_24432 | Proli | + | + | − | − |
| AA824616 | Proli | + | + | + | − |
| X00452 | MHC | + | + | + | − |
| AA490586 | Germ | + | − | − | − |
| AA731721 | | + | − | − | − |
| X02530 | Lymph | + | − | − | − |
| AA193262 | | + | + | − | − |
| AA469973 | | + | − | − | − |

1, Germ, Germinal-cancer B-cell signature; MHC, MHC class II signature; Lymph = lymph-node signature; Proli, proliferation_signature. 2, + / −, selected/not selected

## Appenidx A: remarks on the assumptions

Remark on assumption A1

Note that $S_n$, $S_n^{(1)}$ and $S_n^{(2)}$ are summations of i.i.d random variables; it is verified in Fill (1983) that, when the associated random variable satisfies A1.1, for instance, when the components in $\boldsymbol{Z}$ are bounded or Gaussian random variables, there exists positive constants $C_0$, $C_1$ such that

$$P\left(\sup_{t\in[0,1]} \left|S_n(\boldsymbol{\beta}_0,t) - s(\boldsymbol{\beta}_0,t)\right| \ge \frac{C_1 u_n}{\sqrt{n}}\right) \le \frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right),$$

$$P\left(\sup_{t\in[0,1]} \left|S_{nj}^{(1)}(\boldsymbol{\beta}_0,t) - s_j^{(1)}(\boldsymbol{\beta}_0,t)\right| \ge \frac{C_1 u_n}{\sqrt{n}}\right) \le \frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right),$$

$$P\left(\sup_{t\in[0,1]} \left|S_{nij}^{(2)}(\boldsymbol{\beta}_0,t) - s_{ij}^{(2)}(\boldsymbol{\beta}_0,t)\right| \ge \frac{C_1 u_n}{\sqrt{n}}\right) \le \frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right)$$

hold for any positive $u_n$ such that $u_n \to +\infty$, $n^{-1/6}u_n \to 0$ as $n \to +\infty$. These inequalities and A1.2 are similar to Condition (2.2) and (2.5) in Section 8.2 of Fleming and Harrington (1991). However, it is worth noting that they assume the convergence of $S_n$, $S_n^{(l)}$ to $s$, $s^{(l)}$ holds for a neighborhood $\mathscr{B}$ of $\boldsymbol{\beta}_0$. That is, $\sup_{t\in[0,1],\boldsymbol{\beta}\in\mathscr{B}} \|S_n(\boldsymbol{\beta},t) - s(\boldsymbol{\beta},t)\| \to 0$, $\sup_{t\in[0,1]\boldsymbol{\beta}\in\mathscr{B}} \|S_n^{(l)}(\boldsymbol{\beta},t) - s^{(l)}(\boldsymbol{\beta},t)\| \to 0$ for $l = 1, 2$ in probability. Similarly for the boundedness of $s$, $s^{(l)}$. But our assumptions are made at the true value $\boldsymbol{\beta}_0$. Moreover, with condition A1.2, it can be deduced that

$$P\left(\sup_{t\in[0,1]} \left|E_{nj}(\boldsymbol{\beta}_0,t) - e_j(\boldsymbol{\beta}_0,t)\right| \ge \frac{C_1 u_n}{\sqrt{n}}\right) \le \frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right) \tag{5}$$

and

$$P\left(\sup_{t\in[0,1]}\left|\frac{I_{ij}\left(\boldsymbol{\beta}_0,t\right)}{n}-\Sigma_{ij}\left(\boldsymbol{\beta}_0,t\right)\right|\geq\frac{C_1u_n}{\sqrt{n}}\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right). \tag{6}$$

The detailed proofs of inequalities (5) and (6) are provided in the Appendix 4. A1.3 is Condition (2.6) in Section 8.2 of Fleming and Harrington (1991). A1.4 is assumed in Theorem 4.1 in Andersen and Gill (1982); they are regular conditions in counting process theory.

Remark on assumption A2

Under Assumption A2, we have, for any positive $u_n$ such that $u_n \to +\infty$, $n^{-1/6}u_n \to 0$ as $n \to +\infty$, there exists positive constant $C_0$ such that

$$P\left(\left|\sum_{i=1}^{n}\sum_{j\in s}\boldsymbol{a}_j\xi_{ij}\right|\geq\sqrt{n}u_n\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right). \tag{7}$$

Without loss of generality, we assume all the diagonal elements of $\Sigma\left(\boldsymbol{\beta}_0,1\right)$ are 1. Then when $\boldsymbol{a}_j = 1$ for any fixed $j$ and 0 otherwise, (7) reduces to

$$P\left(\left|\sum_{i=1}^{n}\xi_{ij}\right|\geq\sqrt{n}u_n\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right),\ \forall j\in\{1,2,\ldots,p_n\}.$$

Now let us see how A2 is related to A1.1 in the following: Denote $\xi_{ij}(t) = \int_0^t \left(Z_{ij}(u) - e_j(\boldsymbol{\beta}_0,u)\right)\mathrm{d}M_i(u)$; it can be shown that $Cov(\xi_{ij}(t),\xi_{ik}(t)) = \left[\Sigma\left(\boldsymbol{\beta}_0,t\right)\right]_{jk}$ in the following:

$$\begin{aligned}
<\xi_{ij},\xi_{ik}>(t) &= \int_0^t (Z_{ij}-e_j(\boldsymbol{\beta}_0,u))(Z_{ik}-e_k(\boldsymbol{\beta}_0,u))d<M_i,M_i>(u)\\
&= \int_0^t (Z_{ij}-e_j(\boldsymbol{\beta}_0,u))(Z_{ik}-e_k(\boldsymbol{\beta}_0,u))Y_i(u)\exp\left(\boldsymbol{z}_i^\tau\boldsymbol{\beta}_0\right)h_0(u)\mathrm{d}u;\\
E<\xi_{1j},\xi_{1k}>(t) &= \int_0^t EZ_{ij}Z_{ik}Y_i(u)\exp\left(\boldsymbol{z}_i^\tau\boldsymbol{\beta}_0\right)h_0(u)\mathrm{d}u\\
&\quad -\int_0^t e_j(\boldsymbol{\beta}_0,u)EZ_{ik}Y_i(u)\exp\left(\boldsymbol{z}_i^\tau\boldsymbol{\beta}_0\right)h_0(u)\mathrm{d}u\\
&\quad -\int_0^t e_k(\boldsymbol{\beta}_0,u)EZ_{ij}Y_i(u)\exp\left(\boldsymbol{z}_i^\tau\boldsymbol{\beta}_0\right)h_0(u)\mathrm{d}u\\
&\quad +\int_0^t e_j(\boldsymbol{\beta}_0,u)e_k(\boldsymbol{\beta}_0,u)EY_i(u)\exp\left(\boldsymbol{z}_i^\tau\boldsymbol{\beta}_0\right)h_0(u)\mathrm{d}u
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^t \left[ \frac{E\left[Z_{ij} Z_{ik} Y_i(u) \exp\left(z_i^\tau \boldsymbol{\beta}_0\right)\right]}{s(\boldsymbol{\beta}_0, u)} - e_j(\boldsymbol{\beta}_0, u) e_k(\boldsymbol{\beta}_0, u) \right] \\
&\quad \times s(\boldsymbol{\beta}_0, u) h_0(u) \mathrm{d}u \\
&= \left[\Sigma\left(\boldsymbol{\beta}_0, t\right)\right]_{jk}.
\end{aligned}
$$

For any fixed set $s$, denote $\xi_i(s) = (\xi_{ij})_{j \in s}$, note that $\boldsymbol{var}(\sum_{i=1}^n \sum_{j \in s} \boldsymbol{a}_j \xi_{ij}/\sqrt{n}) = 1$ implies $\boldsymbol{a}^\tau \Sigma(\boldsymbol{\beta}_0(s), 1)\boldsymbol{a} = 1$. Let $\lambda_{\min}$ denote the smallest eigenvalue. Since for $u > 0$, we have

$$
\begin{aligned}
E \exp\left(u \sum_{j \in s} \boldsymbol{a}_j \xi_{ij}\right) &\leq E \exp(u \|\boldsymbol{a}\|_2 \|\xi_i(s)\|_2) \\
&\leq \lambda_{\min}^{-1/2}(\Sigma(\boldsymbol{\beta}_0(s), 1)) |s| \max_j E \exp(u|\xi_{ij}|).
\end{aligned}
$$

Therefore, when $\lambda_{\min}(\Sigma(\boldsymbol{\beta}_0(s), 1))$ is bounded from below and $|s|$ is bounded from above, $E \exp(|u\xi_{ij}|) < +\infty$ for all $j$, inequality (1) holds.

Remark on assumption A3

The more strict counterpart of A3.1 in linear regression models is the Sparse Riesz Condition. Similar conditions were also assumed in Chen and Chen (2012) for generalized linear regression models. As was relaxed technically in linear regression models, a weaker version of A3.1 can be expected in the Cox models.

## Appenidx B: proofs of the main results

*Proof of inequality (5)* By definition, for a fixed $j$,

$$
\begin{aligned}
E_{nj}(\boldsymbol{\beta}_0, t) - e_j(\boldsymbol{\beta}_0, t) &= \frac{S_{nj}^{(1)}(\boldsymbol{\beta}_0, t)}{S_n(\boldsymbol{\beta}_0, t)} - \frac{s_j^{(1)}(\boldsymbol{\beta}_0, t)}{s(\boldsymbol{\beta}_0, t)} \\
&= \frac{1}{S_n(\boldsymbol{\beta}_0, t)} \left(S_{nj}^{(1)}(\boldsymbol{\beta}_0, t) - s_j^{(1)}(\boldsymbol{\beta}_0, t)\right) \\
&\quad - \frac{s_j^{(1)}(\boldsymbol{\beta}_0, t)}{S_n(\boldsymbol{\beta}_0, t) s(\boldsymbol{\beta}_0, t)} \left(S_n(\boldsymbol{\beta}_0, t) - s(\boldsymbol{\beta}_0, t)\right) \\
&= \mathscr{I}_1(t) - \mathscr{I}_2(t).
\end{aligned}
$$

Assumption A1.2 implies $\sup_{t \in [0,1]} \left| \frac{s_j^{(1)}(\boldsymbol{\beta}_0, t)}{s(\boldsymbol{\beta}_0, t)} \right|$ and $\sup_{t \in [0,1]} \left| \frac{1}{s(\boldsymbol{\beta}_0, t)} \right|$ are bounded from above.

Note that $\sup_{t\in[0,1]}\left|\frac{1}{S_n(\boldsymbol{\beta}_0,t)}\right|$ is bounded from above when

$$\sup_{t\in[0,1]}\left|S_n(\boldsymbol{\beta}_0,t)-s(\boldsymbol{\beta}_0,t)\right|\le\frac{C_1u_n}{\sqrt{n}}$$

and $n$ is sufficiently large. That is, under this condition, there exists constants $c_1>0$, $c_2>0$ such that

$$|\mathscr{I}_1(t)|\le c_1\left|S_{nj}^{(1)}(\boldsymbol{\beta}_0,t)-s_j^{(1)}(\boldsymbol{\beta}_0,t)\right|;\ |\mathscr{I}_2(t)|\le c_2\left|S_n(\boldsymbol{\beta}_0,t)-s(\boldsymbol{\beta}_0,t)\right|.$$

Hence,

$$\begin{aligned}
&P\left(\sup_{t\in[0,1]}\left|E_{nj}(\boldsymbol{\beta}_0,t)-e_j(\boldsymbol{\beta}_0,t)\right|\ge\frac{C_1u_n}{\sqrt{n}}\right)\\
&\quad\le P\left(\sup_{t\in[0,1]}\left|E_{nj}(\boldsymbol{\beta}_0,t)-e_j(\boldsymbol{\beta}_0,t)\right|\ge\frac{C_1u_n}{\sqrt{n}},\ \sup_{t\in[0,1]}|S_n(\boldsymbol{\beta},t)-s(\boldsymbol{\beta},t)|\le\frac{C_1u_n}{\sqrt{n}}\right)\\
&\qquad+P\left(\sup_{t\in[0,1]}|S_n(\boldsymbol{\beta},t)-s(\boldsymbol{\beta},t)|\ge\frac{C_1u_n}{\sqrt{n}}\right)\\
&\quad\le P\left(\sup_{t\in[0,1]}\left|S_{nj}^{(1)}(\boldsymbol{\beta}_0,t)-s_j^{(1)}(\boldsymbol{\beta}_0,t)\right|\ge\frac{C_1u_n}{2c_1\sqrt{n}}\right)\\
&\qquad+P\left(\sup_{t\in[0,1]}\left|S_n(\boldsymbol{\beta}_0,t)-s(\boldsymbol{\beta}_0,t)\right|\ge\frac{C_1u_n}{2c_2\sqrt{n}}\right)\\
&\qquad+P\left(\sup_{t\in[0,1]}|S_n(\boldsymbol{\beta},t)-s(\boldsymbol{\beta},t)|\ge\frac{C_1u_n}{\sqrt{n}}\right)\le\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right).
\end{aligned}$$

□

*Proof of inequality (6)* By definition, for fixed $i$, $j$,

$$\begin{aligned}
&V_{ij}\left(\boldsymbol{\beta}_0,t\right)S_n\left(\boldsymbol{\beta}_0,t\right)-v_{ij}\left(\boldsymbol{\beta}_0,t\right)s\left(\boldsymbol{\beta}_0,t\right)\\
&\quad=[S_{nij}^{(2)}\left(\boldsymbol{\beta}_0,t\right)-s_{ij}^{(2)}\left(\boldsymbol{\beta}_0,t\right)]-[E_{ni}\left(\boldsymbol{\beta}_0,t\right)S_{nj}^{(1)}\left(\boldsymbol{\beta}_0,t\right)-e_i\left(\boldsymbol{\beta}_0,t\right)s_j^{(1)}\left(\boldsymbol{\beta}_0,t\right)]\\
&\quad=[S_{nij}^{(2)}\left(\boldsymbol{\beta}_0,t\right)-s_{ij}^{(2)}\left(\boldsymbol{\beta}_0,t\right)]-[E_{ni}\left(\boldsymbol{\beta}_0,t\right)-e_i\left(\boldsymbol{\beta}_0,t\right)]S_{nj}^{(1)}\left(\boldsymbol{\beta}_0,t\right)\\
&\quad-e_i\left(\boldsymbol{\beta}_0,t\right)[S_{nj}^{(1)}\left(\boldsymbol{\beta}_0,t\right)-s_j^{(1)}\left(\boldsymbol{\beta}_0,t\right)].
\end{aligned}$$

By following the steps in the proof of inequality (5), we can obtain inequality (6). □

*Proof of Theorem 1* Here we decompose the $j$th component of the score function $U(\boldsymbol{\beta}_0,t)$ defined in Sect. 2 as

$$U_j(\boldsymbol{\beta}_0, t) = \sum_{i=1}^{n} \int_0^t \left(z_{ij} - e_j\left(\boldsymbol{\beta}_0, u\right)\right) \mathrm{d}M_i(u) - \sum_{i=1}^{n} \int_0^t \left(E_{nj}\left(\boldsymbol{\beta}_0, u\right)\right.$$
$$\left. - e_j\left(\boldsymbol{\beta}_0, u\right)\right) \mathrm{d}M_i(u)$$
$$= \xi_{1j}(t) - \xi_{2j}(t).$$

To avoid confusion, let $\xi_j = \xi_j(1),\ \xi_{1j} = \xi_{1j}(1),\ \xi_{2j} = \xi_{2j}(1)$. For any fixed $s \in \mathscr{A}_0$, note that for any $j \in s$, $E_{nj}\left(\boldsymbol{\beta}_0, u\right) = E_{nj}\left(\boldsymbol{\beta}_0(s), u\right)$, $e_j\left(\boldsymbol{\beta}_0, u\right) = e_j\left(\boldsymbol{\beta}_0(s), u\right)$, for any unit vector $\boldsymbol{u}$, let $\boldsymbol{a} = \boldsymbol{u}^\tau \Sigma^{-1/2}(\boldsymbol{\beta}_0(s), 1)$. Then

$$\boldsymbol{u}^\tau \Sigma^{-1/2}\left(\boldsymbol{\beta}_0(s), 1\right) U\left(\boldsymbol{\beta}_0(s), 1\right) = \sum_{j\in s} \boldsymbol{a}_j \xi_{1j} - \sum_{j\in s} \boldsymbol{a}_j \xi_{2j}.$$

Also, from the remark on Assumption A2, we have $\boldsymbol{var}(\sum_{j\in s} \boldsymbol{a}_j \xi_{1j}/\sqrt{n}) = 1$ and $\|\boldsymbol{a}\|_2^2 \le \lambda_{\min}^{-1}\left(\Sigma(\boldsymbol{\beta}_0(s), 1)\right)$. Let $u_n$ satisfy $n^{-1/6}u_n \to 0$, $u_n(\ln n)^{-1/2} \to +\infty$ as $n \to +\infty$; note that for any positive constant $c \in (0, 1)$ independent of $n$,

$$P(|\sum_{j\in s} \boldsymbol{a}_j \xi_{1j} - \sum_{j\in s} \boldsymbol{a}_j \xi_{2j}| > \sqrt{n}u_n) \le P\left(|\sum_{j\in s} \boldsymbol{a}_j \xi_{1j}| > c\sqrt{n}u_n\right)$$
$$+P\left(|\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}| > (1-c)\sqrt{n}u_n\right),$$

the large deviation result of $\sum_{j\in s} \boldsymbol{a}_j \xi_{1j}$ is already given in the remark on Assumption A2, that is, there exists a constant $C_0$ such that

$$P\left(|\sum_{j\in s} \boldsymbol{a}_j \xi_{1j}| > c\sqrt{n}u_n\right) \le C_0 \exp\left(-\frac{c^2 u_n^2}{2} - \ln u_n\right). \tag{8}$$

Now it suffices to show the large deviation of $\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}$. Let $C_1$ be a positive constant, denote

$$\mathscr{C} = \left\{\| \sup_{u\in[0,1]} [E_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty} \le \frac{C_1 u_n}{\sqrt{n}},\ \sup_{u\in[0,1]} |S_n(\boldsymbol{\beta}_0, u)\right.$$
$$\left. - s(\boldsymbol{\beta}_0, u)| \le \frac{C_1 u_n}{\sqrt{n}}\right\},$$

then

$$
\begin{aligned}
&P\left(\left|\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}\right|>(1-c)\sqrt{n}u_n\right)\\
&\quad\le P\left(\|\sup_{u\in[0,1]}\left[E_n\left(\boldsymbol{\beta}_0,u\right)-e\left(\boldsymbol{\beta}_0,u\right)\right]\|_{+\infty}\ge\frac{C_1u_n}{\sqrt{n}}\right)\\
&\qquad+P\left(\sup_{u\in[0,1]}|S_n(\boldsymbol{\beta}_0,u)-s(\boldsymbol{\beta}_0,u)|\ge\frac{C_1u_n}{\sqrt{n}}\right)\\
&\qquad+P\left(\left|\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}\right|>(1-c)\sqrt{n}u_n\mid\mathscr{C}\right)\\
&\quad\equiv P_{2,1}+P_{2,2,1}+P_{2,2,2}.
\end{aligned}
$$

Inequality (5) and the remark on Assumption A1 demonstrate that there exists positive constant $C_0$ such that

$$
P_{2,1}\le C_0\exp\left(-\frac{u_n^2}{2}+\kappa\ln n-\ln u_n\right)\ ;\ P_{2,2,1}\le C_0\exp\left(-\frac{u_n^2}{2}-\ln u_n\right). \quad (9)
$$

In the following, we verify that condition on $\mathscr{C}$, the new martingale $\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)$ has bounded jumps by following the steps in the proof of Theorem 3.1 in ?. Let $\bar{M}(t)=\sum_{i=1}^{n}M_i(t)$, $\bar{N}(t)=\sum_{i=1}^{n}N_i(t)$, then $|\triangle(\bar{M}(t))|=|\triangle(\bar{N}(t))|\le 1$.

First,

$$
\left|\triangle\left(n^{-1/2}\xi_{2j}(t)\right)\right|\le n^{-1/2}\|\sup_{u\in[0,1]}\left[E_n\left(\boldsymbol{\beta}_0,u\right)-e\left(\boldsymbol{\beta}_0,u\right)\right]\|_{+\infty}\equiv n^{-1/2}c_n\le\frac{C_1u_n}{n};
$$

therefore,

$$
\left|\triangle\left(n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right)\right|\le\sum_{j\in s}|\boldsymbol{a}_j|\left|\triangle\left(n^{-1/2}\xi_{2j}(t)\right)\right|\le\frac{|s|C_1u_n}{n}. \quad (10)
$$

Second, the predictable quadratic variation of $n^{-1/2}\xi_{2j}(t)$, denoted by $\left\langle n^{-1/2}\xi_{2j}(t)\right\rangle$ is bilinear and for all $j\in\{1,2,\ldots,p_n\}$,

$$
\begin{aligned}
\left\langle n^{-1/2}\xi_{2j}(t)\right\rangle&=n^{-1}\int_0^t\left(E_{nj}\left(\boldsymbol{\beta}_0,u\right)-e_j\left(\boldsymbol{\beta}_0),u\right)\right)^2\mathrm{d}\left\langle\bar{M}(u)\right\rangle\\
&=\int_0^t\left(E_{nj}\left(\boldsymbol{\beta}_0,u\right)-e_j\left(\boldsymbol{\beta}_0,u\right)\right)^2S_n(\boldsymbol{\beta}_0,u)h_0(u)\mathrm{d}u\\
&\le\|\sup_{u\in[0,1]}\left[E_n\left(\boldsymbol{\beta}_0,u\right)-e\left(\boldsymbol{\beta}_0,u\right)\right]\|_{+\infty}^2\int_0^tS_n(\boldsymbol{\beta}_0,u)h_0(u)\mathrm{d}u\\
&\equiv b_n^2(t).
\end{aligned}
$$

$$\left\langle n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right\rangle \le |s|\sum_{j\in s}\boldsymbol{a}_j^2\left\langle n^{-1/2}\xi_{2j}(t)\right\rangle \le |s|^2 b_n^2(t).$$

Obviously, $b_n^2(t) \le b_n^2(1) \le c_n^2\int_0^1 S_n(\boldsymbol{\beta}_0,u)h_0(u)\mathrm{d}u$. Note that

$$\int_0^1 S_n(\boldsymbol{\beta}_0,u)h_0(u)\mathrm{d}u \le \int_0^1 s(\boldsymbol{\beta}_0,u)h_0(u)\mathrm{d}u + \sup_{u\in[0,1]}|S_n(\boldsymbol{\beta}_0,u)-s(\boldsymbol{\beta}_0,u)|\int_0^1 h_0(u)\mathrm{d}u.$$

Assumption A1.2 and Eq. (10) imply that

$$\sup_{t\in[0,1]} b_n^2(t) \le c_n^2\left(C_1 + C_2\frac{C_1 u_n}{\sqrt{n}}\right) \le C\frac{u_n^2}{n}.$$

That is, when $|s| = O(1)$, condition on $\mathscr{C}$, there exists constants $b^2 = O(\frac{u_n^2}{n})$, $K = O(\frac{u_n}{n})$ such that

$$\left|\triangle\left(n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right)\right| \le K;\ \left\langle n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right\rangle \le b^2.$$

According to Lemma 2.1 in Van de Geer (1995), we have

$$\begin{aligned} P_{2,2,2} &\le 2\exp\left(-\frac{(1-c)^2u_n^2}{2(K(1-c)u_n+b^2)}\right)\\ &= 2\exp\left(-\frac{u_n^2}{2(K(1-c)^{-1}u_n+(1-c)^{-2}b^2)}\right), \end{aligned}$$

since $u_n^2/n \to 0$, when $n$ is sufficiently large, there exists an arbitrarily large positive constant $M$ such that

$$P_{2,2,2} \le 2\exp(-Mu_n^2).$$

Hence, together with (8) and (9), because of the arbitrariness of $c$, we know that there exists positive constants $c_0$ independent of $j$ and an arbitrarily small positive $\varepsilon$ such that

$$P\left(\left|\boldsymbol{u}^\tau\Sigma^{-1/2}\left(\boldsymbol{\beta}_0(s),1\right)U\left(\boldsymbol{\beta}_0(s),1\right)\right| > \sqrt{n}u_n\right) \le c_0\exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right).$$

When $\boldsymbol{a}_j = 1$ and 0 otherwise, we have

$$P\left(|U_j\left(\boldsymbol{\beta}_0, 1\right)| > \sqrt{n}u_n\right) \leq c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right)$$

over $j \in \{1, 2, \ldots, p_n\}$. □

*Proof of Theorem 2* For any unit vector $\boldsymbol{w}(s)$, let $\boldsymbol{\beta}(s) = \boldsymbol{\beta}_0(s) + \psi_n \boldsymbol{w}(s)$ where $\psi_n$ satisfies (4). Under Assumption A3, for all $s \in \mathscr{A}_0$, the mean value theorem implies that there exists $\tilde{\boldsymbol{\beta}}(s)$ satisfying $\|\tilde{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2 \leq \|\psi_n \boldsymbol{w}(s)\|_2$ such that

$$\begin{aligned}
l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) &= \psi_n \boldsymbol{w}^\tau(s) U(\boldsymbol{\beta}_0(s), 1) - \frac{1}{2}\psi_n^2 \boldsymbol{w}(s)^\tau \{I(\tilde{\boldsymbol{\beta}}(s), 1)\} \boldsymbol{w}(s) \\
&\leq \psi_n \boldsymbol{w}^\tau(s) U(\boldsymbol{\beta}_0(s), 1) - \frac{1-\varepsilon}{2}\lambda_{1,n}\psi_n^2 \\
&\leq \psi_n \sqrt{\boldsymbol{w}^\tau(s)\boldsymbol{w}(s)} \sqrt{U^\tau(\boldsymbol{\beta}_0(s), 1) U(\boldsymbol{\beta}_0(s), 1)} - \frac{1-\varepsilon}{2}\lambda_{1,n}\psi_n^2 \\
&\leq \psi_n \sqrt{k_n} \max_{j \in s, s \in \mathscr{A}_0} \left|U_j(\boldsymbol{\beta}_0(s), 1)\right| - \frac{1-\varepsilon}{2}\lambda_{1,n}\psi_n^2.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&P(l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s) > 0 : \quad \text{for some } \ \boldsymbol{w}(s)) \\
&\leq P\left(\max_{j \in s, s \in \mathscr{A}_0} \left|U_j(\boldsymbol{\beta}_0(s), 1)\right| \geq \frac{1-\varepsilon}{2\sqrt{k_n}}\lambda_{1,n}\psi_n\right).
\end{aligned}$$

By noting that $k_n = O(1)$, $p_n = O(n^\kappa)$ and letting $u_n = \frac{1-\varepsilon}{2\sqrt{nk_n}}\lambda_{1,n}\psi_n$, $n^{-1/6}u_n \to 0$, $u_n(\ln n)^{-1/2} \to +\infty$. According to (2), it follows that

$$\begin{aligned}
&P\left(\max_{j \in s, s \in \mathscr{A}_0} \left|U_j(\boldsymbol{\beta}_0(s), 1)\right| \geq \frac{1-\varepsilon}{2\sqrt{k_n}}\lambda_{1,n}\psi_n\right) \\
&\leq \sum_{j \in s, s \in \mathscr{A}_0} P\left(\left|U_j(\boldsymbol{\beta}_0(s), 1)\right| \geq \frac{1-\varepsilon}{2\sqrt{k_n}}\lambda_{1,n}\psi_n\right) \\
&\leq k_n p_n^{k_n} C_0 \exp\left(-C_1 \frac{\lambda_{1,n}^2 \psi_n^2}{n}\right) \\
&\leq \tilde{C}_0 \exp\left(-C_1 \frac{\lambda_{1,n}^2 \psi_n^2}{n} + C_2 \kappa \ln n\right)
\end{aligned}$$

for some positive constants $C_0, C_1, C_2, \tilde{C}_0$. It converges to 0 as $n$ goes to infinity. Because $l_n(\boldsymbol{\beta}(s))$ is a concave function for any $\boldsymbol{\beta}(s)$, we get the desired result. □

*Proof of Theorem 3* Note that $\{s : s \neq s_0, |s| \leq Cp_0\} = \mathscr{A}_1 \cup \mathscr{A}_0$, if we can prove that when $\gamma > 1 - \frac{1}{2\kappa}$, as $n \to +\infty$,

$$P\left(\min_{s:s\in\mathscr{A}_1} \mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_0)\right) \to 0, \tag{11}$$

and

$$P\left(\min_{s:s\in\mathscr{A}_0} \mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_0)\right) \to 0, \tag{12}$$

then we will have completed the proof. Since asymptotically, $\ln \tau(\mathscr{S}_j) = j\kappa \ln n(1 + o(1))$,

$$\mathrm{EBIC}_\gamma(s_{0n}) - \mathrm{EBIC}_\gamma(s) = 2\left(l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n}))\right) + (1 + 2\gamma\kappa)\,(|s_{0n}| - |s|)\ln n,$$

$\mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})$ implies

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) \geq -\frac{1 + 2\gamma\kappa}{2}\,(|s_{0n}| - |s|)\ln n.$$

(1) When $s \in \mathscr{A}_1$, note that

$$-\frac{1 + 2\gamma\kappa}{2}\,(|s_{0n}| - |s|)\ln n \geq -\frac{1 + 2\gamma\kappa}{2}|s_{0n}|\ln n \geq -C\ln n$$

for some positive constant $C$ when $-\dfrac{1}{2\kappa} < \gamma \leq 1$ and $\kappa$ is a positive constant. Therefore, if we can show that

$$P(\sup\{l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) : s \in \mathscr{A}_1\} \geq -C\ln n) \to 0, \tag{13}$$

then we will have (11). Now, consider $\tilde{s} = s \cup s_{0n}$ and $\boldsymbol{\beta}(\tilde{s})$ near $\boldsymbol{\beta}_0(\tilde{s})$. Taylor expansion shows that

$$\begin{aligned} l_n\left(\boldsymbol{\beta}(\tilde{s})\right) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \leq {} & \left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau U(\boldsymbol{\beta}_0(s)) \\ & -\frac{(1-\varepsilon)\lambda_{1,n}}{2}\left\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right\|_2^2. \end{aligned}$$

Let $\breve{\boldsymbol{\beta}}(\tilde{s})$ be augmented $\hat{\boldsymbol{\beta}}(s)$ with components in $\tilde{s} \cap s^c$ being 0, then $l_n\left(\hat{\boldsymbol{\beta}}(s)\right) = l_n\left(\breve{\boldsymbol{\beta}}(\tilde{s})\right)$ and $\|\breve{\boldsymbol{\beta}}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \geq |\boldsymbol{\beta}_{0,\min}|$, where $|\boldsymbol{\beta}_{0,\min}| = \min\left\{|\boldsymbol{\beta}_{0,j}| : j \in s_{0n}\right\}$. The concavity of $l_n\left(\boldsymbol{\beta}(s)\right)$ implies

$$\begin{aligned} \mathscr{M}_n &= \sup\left\{l_n\left(\boldsymbol{\beta}(\tilde{s})\right) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) : s \in \mathscr{A}_1, \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \geq |\boldsymbol{\beta}_{0,\min}|\right\} \\ &\leq \sup\left\{l_n\left(\boldsymbol{\beta}(\tilde{s})\right) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) : s \in \mathscr{A}_1, \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = |\boldsymbol{\beta}_{0,\min}|\right\}. \end{aligned}$$

Since for any fixed $\tilde{s}$, when $\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = |\boldsymbol{\beta}_{0,\min}|$,

$$l_n(\boldsymbol{\beta}(\tilde{s})) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \leq |\boldsymbol{\beta}_{0,\min}| \|U_j(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} - \boldsymbol{\beta}_{0,\min}^2 \frac{(1-\varepsilon)\lambda_{1,n}}{2}.$$

Therefore,

$$\begin{aligned} P\left(\mathscr{M}_n \geq -\boldsymbol{\beta}_{0,\min}^2 \frac{(1-\varepsilon)\lambda_{1,n}}{4}\right) &\leq k_n p_n^{k_n} P(\|U_j(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} \\ &\geq \frac{|\boldsymbol{\beta}_{0,\min}|(1-\varepsilon)\lambda_{1,n}}{4}). \end{aligned}$$

When $n^{1/6-\delta} = O(\lambda_{1,n}/\sqrt{n})$ for some $0 < \delta < 1/6$.

$$\begin{aligned} &P(\sup\{l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) : s \in \mathscr{A}_1\} \geq -C \ln n) \\ &\leq P\left(\mathscr{M}_n \geq -C \ln n\right) \leq P\left(\mathscr{M}_n \geq -\boldsymbol{\beta}_{0,\min}^2 \frac{(1-\varepsilon)\lambda_{1,n}}{4}\right) \\ &\leq k_n p_n^{k_n} P(\|U_j(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} \geq \sqrt{n} n^{1/6-\delta}) \leq c_0 \exp\left(-c_1 n^{1/3-2\delta} + \kappa \ln n\right). \end{aligned}$$

It converges to 0 when $n$ goes to $\infty$; inequality (13) is thus obtained.

(2) When $s \in \mathscr{A}_0$ and $s \neq s_{0n}$, let $m = |s| - |s_{0n}|$, $\text{EBIC}_\gamma(s) \leq \text{EBIC}_\gamma(s_{0n})$ if and only if

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) \geq m[0.5 \ln n + \gamma \ln p_n] \approx \frac{m(1+2\gamma\kappa)\ln n}{2}.$$

From the assumptions, we can see that

$$\begin{aligned} l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) &\leq l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\boldsymbol{\beta}(s_{0n})) = l_n\left(\hat{\boldsymbol{\beta}}(s)\right) - l_n\left(\boldsymbol{\beta}_0(s)\right) \\ &\leq \left(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s_{0n})\right)^\tau U\left(\boldsymbol{\beta}_0(s), 1\right) \\ &\quad -\frac{1}{2}\left(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s_{0n})\right)^\tau I\left(\widetilde{\boldsymbol{\beta}}(s), 1\right)\left(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s_{0n})\right) \\ &\leq \left(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s_{0n})\right)^\tau U\left(\boldsymbol{\beta}_0(s), 1\right) \\ &\quad -\frac{1-\varepsilon}{2}\left(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s_{0n})\right)^\tau I\left(\boldsymbol{\beta}_0(s), 1\right)\left(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s_{0n})\right) \\ &\leq \max_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}^\tau U\left(\boldsymbol{\beta}_0(s), 1\right) - \frac{1-\varepsilon}{2}\boldsymbol{\beta}^\tau I\left(\boldsymbol{\beta}_0(s), 1\right)\boldsymbol{\beta}\right] \\ &\leq \left[\boldsymbol{\beta}^\tau U\left(\boldsymbol{\beta}_0(s), 1\right)\right]|_{\boldsymbol{\beta}=[(1-\varepsilon)I(\boldsymbol{\beta}_0(s),1)]^{-1}U(\boldsymbol{\beta}_0(s),1)} \\ &= \frac{1}{2n(1-\varepsilon)} U^\tau\left(\boldsymbol{\beta}_0(s), 1\right)\left[\frac{I\left(\boldsymbol{\beta}_0(s), 1\right)}{n}\right]^{-1} U\left(\boldsymbol{\beta}_0(s), 1\right), \end{aligned}$$

where $\varepsilon$ is an arbitrary positive value. Note that $m$ is finite; therefore, if we can show that for any fixed positive integer $m$, when $\gamma > 1 - \frac{1}{2\kappa}$,

$$P\left(\max_{s\in\mathscr{A}_0,|s|=m+|s_{0n}|} \frac{1}{2n(1-\varepsilon)} U^{\tau}\left(\boldsymbol{\beta}_0(s),1\right)\left[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}\right]^{-1} U\left(\boldsymbol{\beta}_0(s),1\right) \geq \frac{m(1+2\gamma\kappa)\ln n}{2}\right) \to 0, \tag{14}$$

then we will have (12). Denote

$$\mathscr{T}_1 = \left\{\max_{s\in\mathscr{A}_0} \|[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}]^{-1} - \Sigma^{-1}\left(\boldsymbol{\beta}_0(s),1\right)\|_{+\infty} \leq \frac{C_1 u_n}{\sqrt{n}}\right\}$$
$$\mathscr{T}_2 = \left\{\max_{s\in\mathscr{A}_0} \frac{U^{\tau}\left(\boldsymbol{\beta}_0(s),1\right) U\left(\boldsymbol{\beta}_0(s),1\right)}{|s|} \leq n u_n^2\right\}.$$

Inequalities (6) and (2) show that

$$P\left(\mathscr{T}_1^c\right) \leq \frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2} + 2\kappa\ln n\right);\ P\left(\mathscr{T}_2^c\right) \leq c_0\exp\left(-\frac{(1-\varepsilon)u_n^2}{2} + \kappa\ln n\right). \tag{15}$$

Therefore, we have

$$\begin{aligned}
&P\left(\max_{s\in\mathscr{A}_0,|s|=m+|s_{0n}|} \frac{1}{2n(1-\varepsilon)} U^{\tau}\left(\boldsymbol{\beta}_0(s),1\right)\left[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}\right]^{-1} U\left(\boldsymbol{\beta}_0(s),1\right) \geq \frac{m(1+2\gamma\kappa)\ln n}{2}\right)\\
&\leq P\left(\max_{s\in\mathscr{A}_0,|s|=m+|s_{0n}|} U^{\tau}\left(\boldsymbol{\beta}_0(s),1\right)\left[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}\right]^{-1} U\left(\boldsymbol{\beta}_0(s),1\right)\right.\\
&\left.\geq mn(1-\varepsilon)(1+2\gamma\kappa)\ln n \mid \mathscr{T}_1,\mathscr{T}_2\right)\\
&\quad + P\left(\mathscr{T}_1^c\right) + P\left(\mathscr{T}_2^c\right).
\end{aligned}$$

Since under $\mathscr{T}_1, \mathscr{T}_2$,

$$\begin{aligned}
&\max_{s\in\mathscr{A}_0,|s|=m+|s_{0n}|}\left[U^{\tau}\left(\boldsymbol{\beta}_0(s),1\right)\left|[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}]^{-1} - \Sigma^{-1}\left(\boldsymbol{\beta}_0(s),1\right)\right| U\left(\boldsymbol{\beta}_0(s),1\right)\right]\\
&\leq C\sqrt{n}u_n^3 = C\frac{(n^{-1/6}u_n)^3}{\ln n}(n\ln n) = o(n\ln n),
\end{aligned}$$

the two terms in (15) both converge to 0 as $n$ goes to $+\infty$ and

$$\begin{aligned}
&P\left(\max_{s\in\mathscr{A}_0,|s|=m+|s_{0n}|} U^{\tau}\left(\boldsymbol{\beta}_0(s),1\right)\Sigma^{-1}\left(\boldsymbol{\beta}_0(s),1\right)U\left(\boldsymbol{\beta}_0(s),1\right)\right.\\
&\quad\geq mn(1-\varepsilon)(1+2\gamma\kappa)\ln n \mid \mathscr{T}_1,\mathscr{T}_2)\\
&\quad\leq C\,P\left(\max_{s\in\mathscr{A}_0,|s|=m+|s_{0n}|} \boldsymbol{u}^{\tau}\Sigma^{-1/2}\left(\boldsymbol{\beta}_0(s),1\right)U\left(\boldsymbol{\beta}_0(s),1\right)\right.\\
&\quad\left.\geq(1-\delta)\sqrt{mn(1-\varepsilon)(1+2\gamma\kappa)\ln n}\mid\mathscr{T}_1,\mathscr{T}_2\right),
\end{aligned}$$

where $\|\boldsymbol{u}\|_2 = 1$, $\delta$ is an arbitrary positive value. According to (3), it can be further bounded by $c_0^{\star}\exp\left[-\frac{1-\varepsilon^{\star}}{2}(1+2\gamma\kappa)m\ln n + m\kappa\ln n\right]$ where $c_0^{\star}$ is a positive constant. It converges to 0 when $\gamma > \frac{1}{1-\varepsilon^{\star}} - \frac{1}{2\kappa}$, where $\varepsilon^{\star}$ is an arbitrary positive value; inequality (14) is thus obtained.

□

## References


Andersen, P., Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, *10*(4), 1100–1120.

Barabási, A., Gulbahce, N., Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, *12*(1), 56–68.

Bogdan, M., Ghosh, J. K., Doerge, R. (2004). Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, *167*(2), 989–999.

Broman, K. W., Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 641–656.

Chen, J., Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Chen, J., Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, *22*(2), 555.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, *10*(3), 184–194.

Du, P., Ma, S., Liang, H. (2010). Penalized variable selection procedure for cox models with semiparametric relative risk. *Annals of statistics*, *38*(4), 2092.

Fan, J., Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*(1), 74–99.

Fan, J., Li, G., Li, R. (2005). An overview on variable selection for survival analysis. *Contemporary multivariate analysis and design of experiments* (p. 315). New Jersey: World Scientific.

Fan, J., Feng, Y., Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. *Borrowing strength: theory powering applications—a Festschrift for Lawrence D Brown*, vol. 6 (pp. 70–86). Beachwood: IMS Collections.

Fill, J. (1983). Convergence rates related to the strong law of large numbers. *The Annals of Probability*, *11*(1), 123–142.

Fleming, T., Harrington, D. (1991). Counting processes and survival analysis, vol 8. Wiley Online Library.

Gui, J., Li, H. (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, *21*(13), 3001–3008.

Luo, S., Chen, Z. (2013a). Extended bic for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, *143*, 494–504.

Luo, S., Chen, Z. (2013b). Selection consistency of ebic for glim with non-canonical links and diverging number of parameters. *Statistics and Its Interface*, *6*, 275–284.
Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, E., Fisher, R., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, *346*(25), 1937–1947.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Sha, N., Tadesse, M., Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, *22*(18), 2262–2268.
Siegmund, D. (2004). Model selection in irregular problems: Application to mapping quantitative trait loci. *Biometrika*, *91*, 785–800.
Tibshirani, R., et al. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, *16*(4), 385–395.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, *17*(6), 520–525.
Van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, *23*(5), 1779–1801.
Zhang, H., Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, *94*(3), 691–703.
Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, *95*(1), 241–247.