Annals of the Institute of Statistical Mathematics (2022) 74:1109–1142
https://doi.org/10.1007/s10463-022-00824-8

A sequential feature selection procedure for high-dimensional Cox proportional hazards model

Ke Yu[1] · **Shan Luo**[1]

Received: 8 December 2021 / Revised: 16 March 2022 / Accepted: 17 March 2022 /
Published online: 7 May 2022

Abstract

Feature selection for the high-dimensional Cox proportional hazards model (Cox model) is very important in many microarray genetic studies. In this paper, we propose a sequential feature selection procedure for this model. We define a novel partial profile score to assess the impact of unselected features conditional on the current model, significant features are thereby added into the model sequentially, and the Extended Bayesian Information Criteria (EBIC) is adopted as a stopping rule. Under mild conditions, we show that this procedure is selection consistent. Extensive simulation studies and two real data applications are conducted to demonstrate the advantage of our proposed procedure over several representative approaches.

Keywords Sequential feature selection · Selection consistency · Cox proportional hazards model · High-dimensional · Extended Bayesian information criteria

1 Introduction

In the contemporary era, new biotechnologies have generated numerous high-dimensional microarray data, and the role of feature selection is significant in the analysis of survival data. Various hazards regression models have been proposed in survival analysis, among which the Cox proportional hazards model (Cox model, Cox 1972) is most prevalent and has been thoroughly investigated. For feature selection, regularization approaches such as Lasso (Tibshirani 1996), adaptive Lasso (Zou 2006), the smoothly clipped absolute deviation estimator (SCAD, Fan and Li 2001), the elastic net (Zou and Hastie 2005) and so on have been studied for Cox model in Tibshirani (1997), Fan and Li (2002), Zhang and Lu (2007), Zou (2008) and Huang et al. (2013). Their oracle selection and estimation consistency properties rely on

✉ Shan Luo
sluomath@sjtu.edu.cn

[1] School of Mathematical Sciences, Shanghai Jiao Tong University, 800 Dongchuan RD, Minhang District, Shanghai 200240, China

the tuning parameter, which is available theoretically but almost infeasible computationally. As far as we know, Luo et al. (2015) is the only article discussing related issues for high-dimensional Cox model seriously. However, as demonstrated in Luo et al. (2015), the models obtained for the final selection is required to contain the true model and have comparable model sizes with the true model. If these models are generated by the above regularization approaches, it requires at least a proper range of the tuning parameter, which brings a challenge for practical application. Moreover, these methods only perform well when the number of covariate variables is moderate. Under the ultra-high-dimensional situation, these methods have problems such as algorithm instability, statistical inaccuracy and expensive computation cost (Fan et al. 2009).

In response, feature screening is proposed to reduce the dimension of variables before initiating the regularization methods. Feature screening techniques for survival data include sure independence screening based on marginal partial likelihood (Fan et al. 2010), principled sure independence screening (Zhao and Li 2012), feature aberration at survival times screening (Gorst-Rasmussen and Scheike 2013), censored rank independence screening (Song et al. 2014), sure joint screening (Yang et al. 2016) and so on. To ensure the sure screening property, these procedures require strong assumptions such as the weak dependency among relevant and irrelevant features. Besides, irrelevant features which are highly correlated with relevant features tend to be retained in the model for further feature selection, bringing new challenges for regularization methods in the ultra-high-dimensional situation.

Sequential methods have attracted much attention for feature selection in high-dimensional data for several years, such as Ing and Lai (2011), Cheng et al. (2014), Luo and Chen (2014), Luo and Chen (2021) for varying-coefficient models, linear models with main and interaction effects, etc. Sequential methods are much favorable in terms of computation time and selection accuracy. Specifically, with suitable stopping rules, they can achieve selection consistency under reasonable assumptions. The implementation is much faster comparing with regularization methods, especially for very large p. In Hong et al. (2019), the authors proposed forward regression for high-dimensional Cox model with EBIC in Chen and Chen (2008) as the stopping criterion. The procedure performs very well; however, it requires to fit almost p Cox models at each step, leading to expensive computation costs.

Motivated by the above discussions on the pros and cons of existing approaches, in this paper, we propose a novel sequential feature selection procedure for high-dimensional Cox model. It selects variables sequentially based on the partial profile score, which is a new measurement to assess the impact of unselected features conditional on the current model. The EBIC is employed as a stopping rule for our procedure. We establish the selection consistency of our procedure under regular conditions. In detail, we show that, with probability tending to one, all relevant features will be selected before irrelevant features and the procedure will stop right after all relevant features are covered in the current model. Meanwhile, our simulation studies demonstrate that our procedure has better selection accuracy than regularization approaches and faster computation speed than forward regression. The rest of this paper is organized as follows. In Sect. 2, we provide the details of this procedure. In Sect. 3, we state our main theoretical results. In Sect. 4, we conduct extensive numerical studies to assess the finite sample performance of our procedure and other methods, and the applications in analyzing two recent data sets are presented in Sect. 5. A short conclusion is provided in Sect. 6. Technical proofs of our theoretical results are relegated to the Appendix section.

2 Methodology

2.1 The Cox model

Let T represent the survival time to the event of interest, C represent the censoring time, and $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)^\top$ be a p-dimensional time independent covariate vector, respectively. Under the right-censoring situation, $X = \min(T, C)$ instead of T is observed; let $\delta = I(T \le C)$ be the indicator taking value 1 when the survival time T is observed and 0 otherwise. Define $Y(t) = I(X \ge t)$ and $N(t) = I(X \le t, \delta = 1)$. We assume that C and T are independent given covariates in $\boldsymbol{Z}$. Let F_T, f_T, and S_T be the cumulative distribution function, density function, and survival function of T, respectively. They denote the corresponding quantities of the censoring time C when the subscript T is replaced by C.

Without loss of generality, denote τ as the terminal time of observation. The Cox model proposed in Cox (1972) is

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}), \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}$ is an unknown regression coefficient vector of interest, $\lambda(t|z)$ is the conditional hazard function defined by

$$\lambda(t|z) = \lim_{\Delta t \downarrow 0} \Pr(t \le T < t + \Delta t | T \ge t, \boldsymbol{Z} = z) / \triangle\, t.$$

Given data $(X_i, \delta_i, \boldsymbol{Z}_i)$, denote $Y_i(t) = I(X_i \ge t)$, $N_i(t) = I(X_i \le t, \delta_i = 1)$ for $i = 1, \ldots, n$. We consider the situation when there are no ties in the observed event time, let $t_1 < t_2 < \cdots < t_N$ be the ordered observed survival times. Denote by $\{\boldsymbol{Z}_{(j)} : j = 1, \ldots, N\}$ the covariate vectors of the individuals with the survival times. Let $\mathcal{R}(t)$ be the risk set at time t, i.e., $\mathcal{R}(t) = \{i : X_i \ge t\}$. For the Cox model, the pseudo-likelihood function based on the observations is given by

$$L = \prod_{i:\delta_i=1} \lambda(X_i|\boldsymbol{Z}_i) \prod_{i=1}^{n} \left(1 - F_T(X_i|\boldsymbol{Z}_i)\right).$$

When the baseline cumulative hazard function $H_0(t) = \int_0^t \lambda_0(s)\mathrm{d}s$ is modeled in a nonparametric manner as $H_0(t) = \sum_{j=1}^N h_j I(t_j \le t)$, maximizing this pseudo-likelihood function with respect to the h_j's for any fixed $\boldsymbol{\beta}$ yields the so-called log partial likelihood function with expression

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{N} \left(\boldsymbol{Z}_{(j)}^{\top}\boldsymbol{\beta} - \log\left(\sum_{i\in\mathcal{R}(t_j)} \exp(\boldsymbol{Z}_i^{\top}\boldsymbol{\beta}) \right) \right). \tag{2}$$

Denote $\mathfrak{F}_t$ as the σ-algebra $\sigma(N_i(t), Y_i(t), \boldsymbol{Z}_i, i = 1, \dots, n)$, let $\boldsymbol{\beta}_0$ be the true $\boldsymbol{\beta}$, then

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_0(u)\exp(\boldsymbol{\beta}_0^{\top}\boldsymbol{Z}_i)\mathrm{d}u, \quad t \in (0, \tau]$$

is a square-integrable martingale with respect to the σ-filtration $\mathfrak{F}_t$, and the log partial likelihood function (2) can also be expressed as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left(\boldsymbol{\beta}^{\top}\boldsymbol{Z}_i - \log\left(\sum_{j=1}^{n} Y_j(u)\exp(\boldsymbol{\beta}^{\top}\boldsymbol{Z}_j) \right) \right) \mathrm{d}N_i(u). \tag{3}$$

In the high-dimensional setting where the dimension p of covariate vector $\boldsymbol{Z}$ is larger than the sample size n, it is reasonably assumed that only a few covariate variables contribute to the survival time, i.e., the set of relevant features $s_0 = \{j : \boldsymbol{\beta}_{0j} \neq 0\}$ has a size much smaller than n. The goal of feature selection is to estimate this unknown set s_0.

2.2 The partial profile score for the Cox model

For an arbitrary set $s \subset \{1, 2, \dots, p\}$, denote s^c as the complementary set of s. Then the parameter vector $\boldsymbol{\beta}$ can be divided into $(\boldsymbol{\beta}_s^{\top}, \boldsymbol{\beta}_{s^c}^{\top})^{\top}$, where $\boldsymbol{\beta}_s$ is subvector with index in s and $\boldsymbol{\beta}_{s^c}$ is subvector with index in s^c. Similarly for $\boldsymbol{Z}_i$, a vector with subscript s denotes the subvector with component indices in s. We denote

$$\hat{\ell}(\boldsymbol{\beta}_{s^c}) = \sup_{\boldsymbol{\beta}_s} \ell(\boldsymbol{\beta}_s, \boldsymbol{\beta}_{s^c}) = \ell(\hat{\boldsymbol{\beta}}_s(\boldsymbol{\beta}_{s^c}), \boldsymbol{\beta}_{s^c}),$$

where $\hat{\boldsymbol{\beta}}_s(\boldsymbol{\beta}_{s^c})$ is the maximizer given $\boldsymbol{\beta}_{s^c}$. For any $k \in s^c$, the partial profile score (PPS) is defined as

$$\psi_k(s) = n^{-1}\frac{\partial\hat{\ell}(\boldsymbol{\beta}_{s^c})}{\partial\boldsymbol{\beta}_k}|_{\boldsymbol{\beta}_{s^c}=0}.$$

By definition,

$$\frac{\partial}{\partial\boldsymbol{\beta}_s}\ell(\boldsymbol{\beta}_s, \boldsymbol{\beta}_{s^c})|_{\boldsymbol{\beta}_s=\hat{\boldsymbol{\beta}}_s(\boldsymbol{\beta}_{s^c})} = 0.$$

Therefore, for any $k \in s^c$, it holds that

$$\frac{\partial \widehat{\ell}(\boldsymbol{\beta}_{s^c})}{\partial \boldsymbol{\beta}_k} = \left\{ \frac{\partial \widehat{\boldsymbol{\beta}}_s(\boldsymbol{\beta}_{s^c})^\top}{\partial \boldsymbol{\beta}_k} \frac{\partial \ell(\boldsymbol{\beta}_s, \boldsymbol{\beta}_{s^c})}{\partial \boldsymbol{\beta}_s} + \frac{\partial \ell(\boldsymbol{\beta}_s, \boldsymbol{\beta}_{s^c})}{\partial \boldsymbol{\beta}_k} \right\}\Bigg|_{\boldsymbol{\beta}_s=\widehat{\boldsymbol{\beta}}_s(\boldsymbol{\beta}_{s^c})}$$
$$= \frac{\partial \ell(\boldsymbol{\beta}_s, \boldsymbol{\beta}_{s^c})}{\partial \boldsymbol{\beta}_k}\Big|_{\boldsymbol{\beta}_s=\widehat{\boldsymbol{\beta}}_s(\boldsymbol{\beta}_{s^c})}.$$

From straightforward calculation, we have

$$\psi_k(s) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(Z_{ik} - \frac{\sum_{j=1}^n Y_j(u) Z_{jk} \exp(\widehat{\boldsymbol{\beta}}_s^\top \boldsymbol{Z}_{js})}{\sum_{j=1}^n Y_j(u) \exp(\widehat{\boldsymbol{\beta}}_s^\top \boldsymbol{Z}_{js})} \right) \mathrm{d}N_i(u), \quad \forall k \in s^c, \tag{4}$$

where $\widehat{\boldsymbol{\beta}}_s$ is the partial likelihood estimator for model s.

As a special case, when $s = \emptyset$, $\psi_k(s)$ reduces to the Fast statistic in Gorst-Rasmussen and Scheike (2013). The partial profile score in (4) reflects the importance of Z_k conditional on covariates in s.

On the other hand, denote $\lambda_i(t) = Y_i(t)\lambda_0(t)\exp(\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i)$ as the random intensity process and $\lambda_{i(s)}(t)$ as the sub-model intensity process, when we use the Breslow type estimator to estimate the cumulative baseline hazard function

$$\widehat{\Lambda}_{0(s)}(t) = \int_0^t \frac{\sum_{i=1}^n \mathrm{d}N_i(u)}{\sum_{i=1}^n Y_i(u) \exp(\widehat{\boldsymbol{\beta}}_s^\top \boldsymbol{Z}_{is})},$$

then we have

$$\begin{aligned} \psi_k(s) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Z_{ik}\{\mathrm{d}N_i(u) - Y_i(u)\exp(\widehat{\boldsymbol{\beta}}_s^\top \boldsymbol{Z}_{is})\mathrm{d}\widehat{\Lambda}_{0(s)}(u)\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Z_{ik}\{\mathrm{d}N_i(u) - \widehat{\lambda}_{i(s)}(u)\mathrm{d}u\}. \end{aligned} \tag{5}$$

By the property of a martingale, we can see from (5) that $\psi_k(s)$ is asymptotically equivalent to $n^{-1}\sum_{i=1}^n \int_0^\tau Z_{ik}\{\lambda_i(u) - \widehat{\lambda}_{i(s)}(u)\}\mathrm{d}u$, which measures the correlation between covariate Z_k and the fitting residuals of the intensity process. It is analogous to $n^{-1}\sum_{i=1}^n Z_{ik}(y_i - \hat{\mu}_i(s))$ in linear regression models with y_i being the observed response and $\hat{\mu}_i(s)$ being the estimated $\mathrm{E}\,(y_i)$ given model s, which measures the correlation between Z_k and the fitted residuals.

2.3 A sequential feature selection procedure

In this subsection, we propose a sequential feature selection procedure based on the PPS defined in Sect. 2.2. For simplicity, we denote this procedure also by PPS. The details of this procedure are described in the following.

Initially, we let s be $\emptyset$ or a certain set of relevant features obtained from some prior knowledge. We standardize all covariates. Given s, for the next step, denote

$$k^\star = \operatorname{argmax}_{j \notin s} |\psi_j(s)|, \quad s^\star = s \cup \{k^\star\}.$$

For s and $s^\star$, calculate their EBIC values defined in Chen and Chen (2008), for a given s,

$$\mathrm{EBIC}_\gamma(s) = -2\ell(\widehat{\boldsymbol{\beta}}_s) + |s| \ln(n) + 2\gamma \ln\left(C_p^{|s|}\right) \tag{6}$$

where $\ell(\widehat{\boldsymbol{\beta}}_s)$ is the maximum value of log partial likelihood function of the submodel containing only the covariate variables in s and $C_p^{|s|}$ is the combination number. If $\mathrm{EBIC}_\gamma(s^\star) > \ \mathrm{EBIC}_\gamma(s)$, stop the procedure and output s; otherwise, let $s = s^\star$ and iterate the procedure.

It is worthy to note that this procedure can be extended to the Cox model with grouped predictors wherein the model with interaction effects is a typical example. More details of the algorithm as well as numerical and theoretical justifications are provided elsewhere.

3 Theoretical properties

In this section, we establish the selection consistency of the PPS procedure for model (1). Firstly, we give some notations. For a column vector $\boldsymbol{v}$, let $\boldsymbol{v}^{\otimes 0} = 1$, $\boldsymbol{v}^{\otimes 1} = \boldsymbol{v}$ and $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^\top$ where $\boldsymbol{v}^\top$ is the transpose vector of $\boldsymbol{v}$. Let $\|\boldsymbol{v}\|_q$ for $q \geq 1$ be the ℓ_q norm of $\boldsymbol{v}$, and denote its ℓ_2 norm by $\|\boldsymbol{v}\|$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and the largest eigenvalues of a symmetric matrix A. Let $|s|$ represent the cardinality of an index set s and let $|s_0| = p_0$.

For an index set $s \subset \{1, 2, 3, \ldots, p\}$, $k \in \{1, 2, 3, \ldots, p\}$ and $m \in \{0, 1, 2\}$, we define

$$R_s^{(m)}(\boldsymbol{\beta}_s, t) = n^{-1} \sum_{i=1}^n Y_i(t) \boldsymbol{Z}_{is}^{\otimes m} \exp\left(\boldsymbol{\beta}_s^\top \boldsymbol{Z}_{is}\right), r_s^{(m)}(\boldsymbol{\beta}_s, t) = E\left\{R_s^{(m)}(\boldsymbol{\beta}_s, t)\right\};$$

$$V_s^{(m)}(t) = n^{-1} \sum_{i=1}^n Y_i(t) \boldsymbol{Z}_{is}^{\otimes m} \lambda_0(t) \exp\left(\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i\right), v_s^{(m)}(t) = E\left\{V_s^{(m)}(t)\right\};$$

$$R_{ks}^{(m)}(\boldsymbol{\beta}_s, t) = n^{-1} \sum_{i=1}^n Y_i(t) Z_{ik}^{\otimes m} \exp\left(\boldsymbol{\beta}_s^\top \boldsymbol{Z}_{is}\right), r_{ks}^{(m)}(\boldsymbol{\beta}_s, t) = E\left\{R_{ks}^{(m)}(\boldsymbol{\beta}_s, t)\right\};$$

$$I_n(\boldsymbol{\beta}_s) = n \int_0^\tau \left\{ \frac{R_s^{(2)}(\boldsymbol{\beta}_s, t)}{R_s^{(0)}(\boldsymbol{\beta}_s, t)} - \frac{\left(R_s^{(1)}(\boldsymbol{\beta}_s, t)\right)^{\otimes 2}}{\left(R_s^{(0)}(\boldsymbol{\beta}_s, t)\right)^2} \right\} V_s^{(0)}(t) \mathrm{d}t.$$

In addition, we denote $\boldsymbol{\beta}_s^*$ as the root of equation

$$\int_0^\tau \left(v_s^{(1)}(t) - \frac{r_s^{(1)}(\boldsymbol{\beta}_s, t)}{r_s^{(0)}(\boldsymbol{\beta}_s, t)} v_s^{(0)}(t) \right) \mathrm{d}t = 0,$$

it can be shown that, $\boldsymbol{\beta}_s^*$ exists and is unique. Denote

$$\boldsymbol{\Phi}_k(s) = \int_0^\tau \left(v_k^{(1)}(t) - \frac{r_{ks}^{(1)}\left(\boldsymbol{\beta}_s^*, t\right)}{r_s^{(0)}\left(\boldsymbol{\beta}_s^*, t\right)} v_s^{(0)}(t) \right) \mathrm{d}t.$$

3.1 Assumptions

We make the following assumptions.

(A1) The study has a finite duration τ such that $\omega = \Pr(X \geq \tau) > 0$ and the baseline hazard ratio function $\lambda_0(t)$ satisfies $\int_0^\tau \lambda_0(t)\mathrm{d}t < \infty$.

(A2) The covariates Z_j's are bounded by a constant $K > 1$, and $E(Z_j) = 0$, $E(Z_j^2) = 1$ for $j = 1, \dots, p$.

(A3) There exist two positive constants $0 < \kappa_{\min} < \kappa_{\max} < \infty$, such that

$$\kappa_{\min} < \lambda_{\min}(E(\boldsymbol{Z}_s^{\otimes 2})) \leq \lambda_{\max}(E(\boldsymbol{Z}_s^{\otimes 2})) < \kappa_{\max}$$

uniformly for $s \subset \{1, \dots, p\}$, where $|s| \leq \rho$ for some $\rho > p_0$.

(A4) There exists a constant L such that $\sup_{|s| \leq \rho} \|\beta_s^*\|_1 \leq L$.

(A5) There exists a constant $\xi > 0$, such that

$$\begin{aligned}\kappa_{\min} &\leq \inf_{\|\beta_s - \beta_s^*\|_\infty \leq \xi, |s| \leq \rho} \lambda_{\min}\left\{ \int_0^\tau \left\{ \frac{r_s^{(2)}(\boldsymbol{\beta}_s, t)}{r_s^{(0)}(\boldsymbol{\beta}_s, t)} - \frac{(r_s^{(1)}(\boldsymbol{\beta}_s, t))^{\otimes 2}}{(r_s^{(0)}(\boldsymbol{\beta}_s, t))^2} \right\} v_s^{(0)}(t)\mathrm{d}t \right\} \\ &\leq \sup_{\|\beta_s - \beta_s^*\|_\infty \leq \xi, |s| \leq \rho} \lambda_{\max}\left\{ \int_0^\tau \left\{ \frac{r_s^{(2)}(\boldsymbol{\beta}_s, t)}{r_s^{(0)}(\boldsymbol{\beta}_s, t)} - \frac{(r_s^{(1)}(\boldsymbol{\beta}_s, t))^{\otimes 2}}{(r_s^{(0)}(\boldsymbol{\beta}_s, t))^2} \right\} v_s^{(0)}(t)\mathrm{d}t \right\} \leq \kappa_{\max}.\end{aligned}$$

(A6) s_{C0} denotes the set of true covariates for the censoring time C, i.e,

$$\Pr(C \leq c | Z_1, \dots, Z_p) = \Pr(C \leq c | \mathbf{Z}_{s_{C0}}).$$

Denote $S_0 = s_0 \cup s_{C0}$. $E\{Z_j S_T(t|\mathbf{Z}_{s_0}) f_C(t|\mathbf{Z}_{s_{C0}}) | \mathbf{Z}_{S_0 \setminus j}\}$ and $E\{Z_j S_T(t|\mathbf{Z}_{s_0}) S_C(t|\mathbf{Z}_{s_{C0}}) | \mathbf{Z}_{S_0 \setminus j}\}$ have the same sign across t, for $j \in s_0$.

(A7) For $\forall s \subseteq s_0$, denote $s^- = s^c \cap s_0$, we have

$$\max_{k \in s_0^c} |\boldsymbol{\Phi}_k(s)| < q \max_{k \in s^-} |\boldsymbol{\Phi}_k(s)|$$

for some constant $0 < q < 1$.

(A8) $\min_{j\in s_0} |\int_0^\tau E\{Z_j f_T(t|\mathbf{Z}_{s_0})S_C(t|\mathbf{Z}_{s_{C0}})\}\mathrm{d}t|\frac{1}{\rho}(\frac{\ln p}{n})^{-1/4} \to \infty$.

(A9) For any given $\varepsilon > 0$, there exists a constant $\delta > 0$ such that, when n is sufficiently large, $I_n(\boldsymbol{\beta}_s) \geq (1-\varepsilon)I_n(\boldsymbol{\beta}_{0s})$ for all $\boldsymbol{\beta}_s$ satisfying $s_0 \subset s$, $|s| \leq \rho$ and $\|\boldsymbol{\beta}_s - \boldsymbol{\beta}_{0s}\| \leq \delta$.

Now we provide a head-to-head comparison between our assumptions and those of existing approaches. Condition (A1)–(A6) are assumed in Hong et al. (2019). Condition (A1) is a standard assumption in survival analysis, condition (A2) and (A3) are common assumptions for variable screening and selection, see Wang (2009) and Zheng et al. (2020). The boundedness assumption is imposed for simple technical proof, which can also be found in Zhao and Li (2012), Yang et al. (2016). Condition (A4) has a similar favor to the assumption (A2) in Bühlmann (2006) for linear regression models and the Lipschitz assumption in Van de Geer (2008) for generalized linear models. It is a common assumption for the Cox proportional hazard models, see Assumption D in Kong and Nan (2014) and Assumption (D) in Hong et al. (2019). Condition (A5) is similar to condition 2 in Bradic et al. (2011), it requires that the concavity of the log partial likelihood is bounded in a neighborhood of $\boldsymbol{\beta}_s^*$. Condition (A6) is used to analyze the least false value $\boldsymbol{\beta}_s^*$. It is noteworthy that condition (A6) always holds in practice. For example, when $s_0 \cap s_{C0} = \emptyset$, condition (A6) holds according to Lemma (A) in Hong et al. (2019). Condition (A7) and (A8) are similar to the assumptions (A1) and (A3) in Luo and Chen (2014) where (A7) actually requires that the maximum "correlation" between the relevant features with the current residual should be larger than that of the irrelevant ones at the population level.

A parallel condition to (A7) is the partial orthogonality condition assumed in Zhao and Li (2012), Gorst-Rasmussen and Scheike (2013) and Hong et al. (2019) to establish the screening consistency and selection consistency. It requires the independency between $\mathbf{Z}_{s_0}$ and $\mathbf{Z}_{s_0^c}$. The partial orthogonality condition implies (A7). A simple proof is provided as follows. For $k \in (s_0 \cup s)^c$, by definition,

$$\int_0^\tau \left(v_{s\cup\{k\}}^{(1)}(t) - \frac{r_{s\cup\{k\}}^{(1)}(\boldsymbol{\beta}_{s\cup\{k\}}^*, t)}{r_{s\cup\{k\}}^{(0)}(\boldsymbol{\beta}_{s\cup\{k\}}^*, t)} v_{s\cup\{k\}}^{(0)}(t) \right) \mathrm{d}t = 0.$$

Under the partial orthogonality condition, according to Lemma B in Hong et al. (2019), we have $\boldsymbol{\beta}_{s\cup\{k\}}^* = (\boldsymbol{\beta}_s^{*T}, 0)^\top$. Hence, $r_{s\cup\{k\}}^{(0)}(\boldsymbol{\beta}_{s\cup\{k\}}^*, t) = r_s^{(0)}(\boldsymbol{\beta}_s^*, t)$ and the last element of $r_{s\cup\{k\}}^{(1)}(\boldsymbol{\beta}_{s\cup\{k\}}^*, t)$ is the $r_{ks}^{(1)}(\boldsymbol{\beta}_s^*, t)$. Similarly, $v_{s\cup\{k\}}^{(0)}(t) = v_s^{(0)}(t)$ and the last element of $v_{s\cup\{k\}}^{(1)}(t)$ is $v_k^{(1)}(t)$. Consequently,

$$\boldsymbol{\Phi}_k(s) = \int_0^\tau \left(v_k^{(1)}(t) - \frac{r_{ks}^{(1)}(\boldsymbol{\beta}_s^*, t)}{r_s^{(0)}(\boldsymbol{\beta}_s^*, t)} v_s^{(0)}(t) \right) \mathrm{d}t = 0.$$

According to the proof of Theorem 1, $\max_{k\in s^-} |\boldsymbol{\Phi}_k(s)|$ is strictly positive for $\forall s \subseteq s_0$. Condition (A7) is therefore established.

Condition (A8) is the signal strength condition, which requires that the effects of the relevant features must not tapper off too quickly. This constraint has a similar favor to the constraint on the regression coefficient for the linear model, see Zhao and Li (2012) and Hong et al. (2019). Condition (A9) can be found in Luo et al. (2015) which is used to ensure the selection consistency of EBIC in the Cox model.

3.2 Main results

Theorem 1 *Let $s_{*1}, s_{*2}, \dots, s_{*m}, \dots$ be the sequence produced by the PPS procedure in Sect. 2.3. Under assumptions $(A1)$–$(A8)$, if $p = O(n^{\kappa})$ for $\kappa > 1$, p_0 is independent of the sample size n and $\rho = Cp_0$ for some constant $C > 1$, there exists a m^* such that*:

$$\Pr\,(s_{*m^*} = s_0) \to 1, \text{ as } n \to \infty.$$

Theorem 2 *Under the assumptions in Theorem 1 and $(A9)$, for the model sequence produced by the PPS procedure, we have*

(1) *uniformly for $m < m^*$, when $\gamma > 0$,*

$$\Pr\,(\ \text{EBIC}_{\gamma}(s_{*m+1}) < \ \text{EBIC}_{\gamma}(s_{*m})) \to 1.$$

(2) *when $\gamma > 1 - \ln n/(2\ln p)$,*

$$\Pr\,(\min_{p_0<|s_{*m}|<\rho}\ \text{EBIC}_{\gamma}(s_{*m}) > \ \text{EBIC}_{\gamma}(s_0)) \to 1.$$

i.e., the PPS procedure is selection consistent.

4 Numerical studies

In this section, we present extensive numerical results to compare the finite sample performances of our procedure PPS and other methods in the literature. We let the dimension of covariates be $p = [n^{1.4}]$ and $n \in \{100, 200, 300, 400\}$, namely

$$(n,p) \in \{(100, 630),\ (200, 1665),\ (300, 2937),\ (400, 4394)\}.$$

We consider four different values of γ in EBIC, $\gamma_1 = 0$ degenerates to BIC, $\gamma_2 = 1 - \log(n)/(2\log(p))$, $\gamma_3 = 1 - \log(n)/(4\log(p))$, and $\gamma_4 = 1$. For comparison, we include forward regression (FR) proposed in Hong et al. (2019), Lasso in Tibshirani (1996) and SIS (Fan et al. 2010) followed by MCP (Zhang 2010), denoted by SIS-MCP. Since SCAD in Fan and Li (2001) and MCP in Zhang (2010) enjoy similar properties and performances, SIS-SCAD was excluded in our tables and figures.

Since estimation with MCP fails to converge without SIS, we have to implement SIS before MCP. The size of features retained at the SIS step is set to be $\lfloor n/\log(n)\rfloor$, as suggested in Fan et al. (2010). For a fair comparison, we use EBIC_γ as a unified model selection criterion for all approaches.

For the evaluation of a feature selection procedure, we consider the running time, the positive discovery rate (PDR), the false discovery rate (FDR) and Mathews correlation coefficient (MCC). Recall that s_0 is the set of true features, denote $\widehat{s}$ as the selected set of features, PDR and FDR are defined as follows:

$$\text{PDR} = \frac{|\widehat{s} \cap s_0|}{|s_0|}, \quad \text{FDR} = \frac{|\widehat{s} \cap s_0^c|}{\max\{1, |\widehat{s}|\}}.$$

For feature selection, MCC can be defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\max\{1, \sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}\}}$$

where

$$\begin{aligned} \text{TP} &= |\widehat{s} \cap s_0|, \quad \text{TN} = |s_0^c| - |\widehat{s} \cap s_0^c|, \\ \text{FN} &= |s_0| - |\widehat{s} \cap s_0|, \quad \text{FP} = |\widehat{s} \cap s_0^c|. \end{aligned}$$

PDR will converge to 1, FDR will converge to 0, and MCC will converge to 1 simultaneously if the feature selection procedure posses selection consistency.

In our simulation, the survival time is generated from a Cox model $\lambda(t|\boldsymbol{Z}) = \exp(\boldsymbol{Z}^\top \boldsymbol{\beta})$, where the baseline hazard function $\lambda_0(t) = 1$. The censoring time is independently generated from an exponential distribution with mean L. Four different data scenarios are considered.

S 1. The true coefficient vector $\boldsymbol{\beta}$ is set as $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 1, \mathbf{0}_{p-6}^\top)^\top$. The covariates vector $\boldsymbol{Z}$ follows a multivariate normal distribution with mean zero and covariance matrix I_p, where I_p is the p-dimensional identity matrix. We change the value of L to control the censoring proportion and we let $L = 1, 10$. $L = 10$ yields small censoring proportions (around 22%) and $L = 1$ yields large censoring proportions (around 50%).

S 2. The true coefficient vector $\boldsymbol{\beta}$ is set as $\boldsymbol{\beta} = (2, -2, 2, -2, 2, -2, \mathbf{0}_{p-6}^\top)^\top$. The covariate vector $\boldsymbol{Z}$ follows a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = v^{|i-j|}$ and $v \in \{0.3, 0.7\}$. Let $L = 10$.

S 3. The true coefficient vector $\boldsymbol{\beta}$ is set as $\boldsymbol{\beta} = (1, 1, 1, 1, 1, -2.5, \mathbf{0}_{p-6}^\top)^\top$. The covariate vector $\boldsymbol{Z}$ follows a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = 0.5$ for $i \neq j$ and $\sigma_{ii} = 1$. Let $L = 10$. In this setting, $Z_6 \in s_0$ but the marginal utility of Z_6 is lower than that of $j \in s_0^c$.

S 4. The true coefficient vector $\boldsymbol{\beta}$ is set as $\boldsymbol{\beta} = (1, -1, 1, -1, 1, -0.34375, \mathbf{0}_{p-6}^\top)^\top$. The covariate vector $\boldsymbol{Z}$ follows a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = 0.5^{|i-j|}$. Let $L = 10$. In this setting,

Table 1 The averaged running time (in seconds) over 100 replications

L	n	PPSγ_1	FRγ_1	PPSγ_2	FRγ_2	PPSγ_3	FRγ_3	PPSγ_4	FRγ_4	Lasso	SIS-MCP
1	100	0.25	48.65	0.04	22.36	0.03	11.42	0.02	7.21	0.17	0.06
	200	0.41	144.96	0.14	72.94	0.13	64.08	0.12	59.96	0.47	0.12
	300	0.73	288.86	0.29	136.83	0.27	123.43	0.25	118.01	0.96	0.26
	400	1.22	467.44	0.49	217.22	0.45	192.30	0.45	187.93	1.59	0.42
10	100	0.25	48.95	0.06	25.09	0.04	18.59	0.03	13.15	0.22	0.05
	200	0.47	150.92	0.16	72.72	0.15	63.80	0.15	60.80	0.60	0.12
	300	0.84	290.42	0.34	136.71	0.31	121.99	0.31	117.59	1.24	0.26
	400	1.48	473.37	0.61	212.61	0.56	195.98	0.55	189.34	2.05	0.39

$Z_6 \in s_0$ but the marginal utility of Z_6 is lower than that of $j \in s_0^c$. For other active variables in s_0, their marginal signals are also very weak.

From our simulation, we find that the running time is mainly affected by the feature dimension, sample size and the censoring proportion. We report the averaged running times of various methods over 100 replications for S 1 with $L = 1$ and $L = 10$ in Table 1. The running times of Lasso and SIS-MCP with different γ values are very close; hence, they are consolidated as Lasso and SIS-MCP in the table. From Table 1, we can see that FR takes dozens or even hundreds times longer time than PPS. PPS requires comparable running times with Lasso and SIS-MCP. For large n, PPS only requires several seconds. We observe similar phenomenon for settings S2, S3, S4, and hence, we skip their results in our table.

The averaged PDR, FDR and MCC over 100 replications for S1 are summarized in Tables 2 and 3, those for S2 are in Tables 4 and 5, and results for S3 and S4 are provided in Tables 6 and 7, respectively. The following conclusions can be drawn from these tables. (1) Regarding the effect of different γ values in the EBIC, we can see that larger γ tends to be more conservative and leads to smaller PDR and smaller FDR. Especially, PPS and FR with $\gamma_1 = 0$ (BIC) always select the largest model, indicating that BIC fails to achieve selection consistency in the high-dimensional situation. With γ values falling in the theoretical range such as $\gamma_2, \gamma_3, \gamma_4$, PPS tends to select the true model with probability tending to one as the sample size grows. (2) Higher censoring proportion and higher correlation among covariates will both result in worse performances of feature selection for all methods. The sequential methods FR and PPS are more robust than Lasso and SIS-MCP, especially when the sample size is moderate and the correlation coefficient is high. When $\nu = 0.7$ in S 2 and $\gamma \in \{\gamma_2, \gamma_3, \gamma_4\}$, most PDRs of Lasso and SIS-MCP methods stay under 40% and MCCs remain below 60% for large $n = 400$, our PPS method maintains a good performance with PDRs above 97% and MCCs around 95%. (3)For the two challenging settings S 3 and S 4 where a relevant feature has a weaker marginal effect than all irrelevant features, PPS and

Table 2 The averaged PDR, FDR and MCC over 100 replications with standard deviations in the brackets for S1 with $L = 1$

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
PPSγ_1				
PDR	0.868 (0.190)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.641 (0.094)	0.597 (0.027)	0.600 (0.003)	0.600 (0.000)
MCC	0.552 (0.131)	0.633 (0.019)	0.632 (0.002)	0.632 (0.000)
FRγ_1				
PDR	0.942 (0.117)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.623 (0.047)	0.600 (0.000)	0.600 (0.000)	0.600 (0.000)
MCC	0.590 (0.075)	0.631 (0.000)	0.631 (0.000)	0.632 (0.000)
Lassoγ_1				
PDR	0.710 (0.349)	0.997 (0.023)	1.000 (0.000)	1.000 (0.000)
FDR	0.319 (0.222)	0.352 (0.181)	0.356 (0.138)	0.342 (0.154)
MCC	0.605 (0.241)	0.795 (0.115)	0.797 (0.085)	0.805 (0.097)
SIS-MCPγ_1				
PDR	0.387 (0.188)	0.655 (0.178)	0.823 (0.138)	0.928 (0.096)
FDR	0.519 (0.289)	0.422 (0.265)	0.209 (0.230)	0.099 (0.168)
MCC	0.400 (0.192)	0.599 (0.199)	0.798 (0.169)	0.910 (0.114)
PPSγ_2				
PDR	0.577 (0.348)	0.992 (0.083)	1.000 (0.000)	1.000 (0.000)
FDR	0.155 (0.208)	0.093 (0.112)	0.082 (0.110)	0.083 (0.100)
MCC	0.631 (0.291)	0.944 (0.081)	0.956 (0.060)	0.956 (0.054)
FRγ_2				
PDR	0.768 (0.325)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.250 (0.199)	0.136 (0.134)	0.119 (0.140)	0.118 (0.125)
MCC	0.711 (0.237)	0.927 (0.074)	0.935 (0.079)	0.936 (0.069)
Lassoγ_2				
PDR	0.020 (0.086)	0.797 (0.388)	1.000 (0.000)	1.000 (0.000)
FDR	0.000 (0.000)	0.124 (0.129)	0.122 (0.128)	0.130 (0.122)
MCC	0.036 (0.137)	0.744 (0.347)	0.934 (0.070)	0.930 (0.067)
SIS-MCPγ_2				
PDR	0.202 (0.253)	0.620 (0.233)	0.823 (0.138)	0.928 (0.096)
FDR	0.057 (0.143)	0.104 (0.146)	0.053 (0.107)	0.017 (0.062)
MCC	0.276 (0.310)	0.711 (0.221)	0.879 (0.102)	0.954 (0.066)
PPSγ_3				
PDR	0.403 (0.351)	0.975 (0.145)	1.000 (0.000)	1.000 (0.000)
FDR	0.059 (0.159)	0.056 (0.090)	0.033 (0.069)	0.031 (0.070)
MCC	0.509 (0.341)	0.951 (0.130)	0.983 (0.036)	0.984 (0.037)
FRγ_3				
PDR	0.580 (0.399)	0.992 (0.083)	1.000 (0.000)	1.000 (0.000)
FDR	0.112 (0.176)	0.055 (0.088)	0.037 (0.073)	0.039 (0.076)
MCC	0.619 (0.345)	0.965 (0.073)	0.981 (0.038)	0.979 (0.041)
Lassoγ_3				
PDR	0.003 (0.033)	0.642 (0.470)	0.990 (0.100)	1.000 (0.000)

Table 2 (continued)

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
FDR	0.000 (0.000)	0.074 (0.109)	0.098 (0.110)	0.107 (0.115)
MCC	0.006 (0.058)	0.612 (0.442)	0.938 (0.112)	0.943 (0.062)
SIS-MCPγ_3				
PDR	0.138 (0.226)	0.587 (0.279)	0.823 (0.138)	0.928 (0.096)
FDR	0.029 (0.105)	0.080 (0.125)	0.042 (0.088)	0.016 (0.057)
MCC	0.199 (0.293)	0.676 (0.280)	0.884 (0.094)	0.955 (0.064)
PPSγ_4				
PDR	0.235 (0.294)	0.973 (0.153)	1.000 (0.000)	1.000 (0.000)
FDR	0.031 (0.132)	0.017 (0.047)	0.013 (0.041)	0.004 (0.024)
MCC	0.335 (0.335)	0.969 (0.130)	0.993 (0.021)	0.998 (0.013)
FRγ_4				
PDR	0.303 (0.339)	0.973 (0.153)	1.000 (0.000)	1.000 (0.000)
FDR	0.040 (0.126)	0.016 (0.045)	0.014 (0.043)	0.004 (0.024)
MCC	0.398 (0.354)	0.970 (0.130)	0.993 (0.022)	0.998 (0.013)
Lassoγ_4				
PDR	0.000 (0.000)	0.403 (0.485)	0.990 (0.100)	1.000 (0.000)
FDR	0.000 (0.000)	0.040 (0.084)	0.076 (0.097)	0.097 (0.109)
MCC	0.000 (0.000)	0.389 (0.464)	0.950 (0.109)	0.949 (0.059)
SIS-MCPγ_4				
PDR	0.107 (0.209)	0.540 (0.318)	0.823 (0.138)	0.928 (0.096)
FDR	0.020 (0.086)	0.058 (0.104)	0.037 (0.082)	0.016 (0.057)
MCC	0.156 (0.270)	0.627 (0.330)	0.887 (0.091)	0.955 (0.064)

FR are not much affected, while Lasso and SIS-MCP both deteriorate. In setting S 3, with $\gamma \in \{\gamma_2, \gamma_3, \gamma_4\}$, FDRs of Lasso almost exceed 50% for large n, PDRs of SIS-MCP are all lower than 60% and their MCCs maintain around 70%, while PPS has 100% PDR, less than 10% FDR and more than 95% MCC for large n. In setting S 4, PDRs and MCCs of Lasso and SIS-MCP almost all stay below 50% and 60% respectively. In contrast, PDRs and MCCs of PPS reach over 90% when $n = 400$. (4) Between PPS and FR, in the presence of large correlation, for example, when $v = 0.7$ in S 2, FR has slightly better performance when the sample size n is small. However, as n grows up to 400, PPS and FR have comparable performances, and the averaged MCC is 96% for PPS and 97.7%, for FR. However, FR occupies a much longer time than PPS. Similar pattern can be observed for setting S 4. For the low correlation situation, for instance, S 1 and $v = 0.3$ in S 2, PPS has slightly better performance than FR in moderate sample size according to the MCC values. From moderate to large sample sizes, they have comparable performances.

Table 3 The averaged PDR, FDR and MCC over 100 replications with standard deviations in the brackets for S1 with $L = 10$

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
PPSγ_1				
PDR	0.978 (0.102)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.602 (0.051)	0.599 (0.007)	0.600 (0.000)	0.600 (0.000)
MCC	0.619 (0.070)	0.632 (0.006)	0.631 (0.000)	0.632 (0.000)
FRγ_1				
PDR	0.992 (0.060)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.603 (0.024)	0.600 (0.000)	0.600 (0.000)	0.600 (0.000)
MCC	0.623 (0.038)	0.631 (0.000)	0.631 (0.000)	0.632 (0.000)
Lassoγ_1				
PDR	0.960 (0.128)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.472 (0.235)	0.375 (0.162)	0.356 (0.160)	0.323 (0.142)
MCC	0.678 (0.181)	0.783 (0.102)	0.796 (0.099)	0.818 (0.088)
SIS-MCPγ_1				
PDR	0.547 (0.185)	0.830 (0.111)	0.935 (0.091)	0.992 (0.037)
FDR	0.462 (0.252)	0.222 (0.230)	0.103 (0.167)	0.050 (0.124)
MCC	0.525 (0.185)	0.794 (0.153)	0.911 (0.115)	0.968 (0.082)
PPSγ_2				
PDR	0.882 (0.271)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.138 (0.161)	0.090 (0.111)	0.099 (0.109)	0.085 (0.109)
MCC	0.844 (0.224)	0.952 (0.061)	0.947 (0.059)	0.955 (0.060)
FRγ_2				
PDR	0.953 (0.191)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.203 (0.187)	0.136 (0.126)	0.121 (0.115)	0.110 (0.118)
MCC	0.852 (0.186)	0.927 (0.070)	0.936 (0.063)	0.941 (0.065)
Lassoγ_2				
PDR	0.225 (0.403)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.027 (0.071)	0.174 (0.144)	0.124 (0.112)	0.119 (0.109)
MCC	0.228 (0.383)	0.905 (0.081)	0.934 (0.061)	0.937 (0.059)
SIS-MCPγ_2				
PDR	0.448 (0.288)	0.830 (0.111)	0.935 (0.091)	0.992 (0.037)
FDR	0.130 (0.166)	0.055 (0.101)	0.009 (0.041)	0.003 (0.023)
MCC	0.524 (0.310)	0.882 (0.083)	0.961 (0.055)	0.994 (0.028)
PPSγ_3				
PDR	0.767 (0.384)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.064 (0.116)	0.049 (0.076)	0.036 (0.069)	0.025 (0.060)
MCC	0.771 (0.344)	0.975 (0.040)	0.981 (0.037)	0.987 (0.032)
FRγ_3				
PDR	0.862 (0.325)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.119 (0.149)	0.056 (0.081)	0.038 (0.072)	0.027 (0.064)
MCC	0.821 (0.290)	0.970 (0.043)	0.980 (0.038)	0.986 (0.034)
Lassoγ_3				
PDR	0.088 (0.270)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Table 3 (continued)

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
FDR	0.006 (0.028)	0.155 (0.129)	0.103 (0.103)	0.084 (0.106)
MCC	0.096 (0.272)	0.916 (0.072)	0.946 (0.055)	0.955 (0.057)
SIS-MCPγ_3				
PDR	0.355 (0.322)	0.830 (0.111)	0.935 (0.091)	0.992 (0.037)
FDR	0.066 (0.125)	0.048 (0.089)	0.008 (0.039)	0.002 (0.017)
MCC	0.429 (0.361)	0.886 (0.080)	0.962 (0.055)	0.995 (0.024)
PPSγ_4				
PDR	0.508 (0.437)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.035 (0.091)	0.025 (0.061)	0.011 (0.039)	0.009 (0.034)
MCC	0.564 (0.407)	0.987 (0.032)	0.994 (0.020)	0.996 (0.018)
FRγ_4				
PDR	0.670 (0.425)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.063 (0.116)	0.026 (0.060)	0.013 (0.044)	0.009 (0.034)
MCC	0.692 (0.381)	0.986 (0.032)	0.993 (0.023)	0.996 (0.018)
Lassoγ_4				
PDR	0.023 (0.142)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.000 (0.000)	0.137 (0.123)	0.093 (0.103)	0.068 (0.099)
MCC	0.028 (0.151)	0.926 (0.068)	0.950 (0.055)	0.964 (0.053)
SIS-MCPγ_4				
PDR	0.260 (0.322)	0.830 (0.111)	0.935 (0.091)	0.992 (0.037)
FDR	0.032 (0.082)	0.047 (0.087)	0.006 (0.036)	0.002 (0.017)
MCC	0.319 (0.372)	0.887 (0.079)	0.963 (0.054)	0.995 (0.024)

5 Real data application

In this section, we applied our proposed method PPS and forward regression (FR), Lasso and SIS-MCP to the following two recent real data examples. Before analysis, all features are standardized so that they have mean zero and variance one. For SIS-MCP, the size of features retained after the SIS procedure is set as $\lfloor n/\log n \rfloor$. SIS-MCP selects no feature when $\gamma = \gamma_4$ in EBIC, for γ_2, γ_3, Lasso and SIS-MCP select the same set of features, and hence, we denote them with $\gamma = \gamma_3$ by Lasso-EBIC and SIS-MCP-EBIC in our tables and figures. Based on the simulation study in Sect. 4, FR and PPS with BIC as the stopping rule always select too many spurious features, and hence, we use EBIC with $\gamma = \gamma_3$ for FR and PPS.

5.1 Swedish watchful waiting cohort data

The Swedish Watchful Waiting Cohort (SWWC) data were published in Sboner et al. (2010), and it was available from the Gene Expression Omnibus with accession number GSE16560. Prostate cancer is the most common neoplasm in men. This data set consists of 6144 gene expression files from 206 male samples who

Table 4 The averaged PDR, FDR and MCC over 100 replications with standard deviations in the brackets for S2 with $v = 0.3$

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
PPSγ_1				
PDR	0.937 (0.194)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.621 (0.081)	0.600 (0.000)	0.600 (0.003)	0.600 (0.000)
MCC	0.590 (0.126)	0.631 (0.000)	0.632 (0.002)	0.632 (0.000)
FRγ_1				
PDR	0.975 (0.126)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.610 (0.050)	0.600 (0.000)	0.600 (0.000)	0.600 (0.000)
MCC	0.612 (0.081)	0.631 (0.000)	0.631 (0.000)	0.632 (0.000)
Lassoγ_1				
PDR	0.833 (0.302)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.596 (0.297)	0.576 (0.122)	0.517 (0.158)	0.480 (0.157)
MCC	0.475 (0.186)	0.643 (0.094)	0.686 (0.108)	0.713 (0.106)
SIS-MCPγ_1				
PDR	0.323 (0.162)	0.517 (0.156)	0.688 (0.145)	0.775 (0.149)
FDR	0.533 (0.304)	0.498 (0.262)	0.310 (0.282)	0.171 (0.227)
MCC	0.339 (0.177)	0.488 (0.171)	0.673 (0.191)	0.792 (0.164)
PPSγ_2				
PDR	0.638 (0.433)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.171 (0.240)	0.096 (0.129)	0.089 (0.118)	0.090 (0.118)
MCC	0.636 (0.368)	0.948 (0.072)	0.952 (0.065)	0.952 (0.065)
FRγ_2				
PDR	0.815 (0.366)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.213 (0.218)	0.121 (0.140)	0.118 (0.129)	0.114 (0.126)
MCC	0.739 (0.316)	0.934 (0.079)	0.937 (0.072)	0.939 (0.069)
Lassoγ_2				
PDR	0.062 (0.215)	0.948 (0.214)	1.000 (0.000)	1.000 (0.000)
FDR	0.013 (0.069)	0.315 (0.178)	0.261 (0.169)	0.220 (0.145)
MCC	0.072 (0.208)	0.773 (0.190)	0.854 (0.100)	0.879 (0.084)
SIS-MCPγ_2				
PDR	0.147 (0.196)	0.443 (0.231)	0.678 (0.158)	0.775 (0.149)
FDR	0.038 (0.113)	0.137 (0.184)	0.059 (0.120)	0.022 (0.069)
MCC	0.229 (0.281)	0.558 (0.244)	0.792 (0.120)	0.867 (0.098)
PPSγ_3				
PDR	0.472 (0.439)	0.990 (0.100)	1.000 (0.000)	1.000 (0.000)
FDR	0.066 (0.134)	0.037 (0.078)	0.039 (0.076)	0.034 (0.068)
MCC	0.519 (0.403)	0.970 (0.107)	0.979 (0.040)	0.982 (0.036)
FRγ_3				
PDR	0.585 (0.454)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.090 (0.143)	0.043 (0.086)	0.046 (0.082)	0.040 (0.079)
MCC	0.598 (0.404)	0.977 (0.046)	0.976 (0.043)	0.979 (0.042)
Lassoγ_3				
PDR	0.038 (0.167)	0.935 (0.240)	1.000 (0.000)	1.000 (0.000)

Table 4 (continued)

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
FDR	0.006 (0.041)	0.277 (0.167)	0.230 (0.165)	0.206 (0.135)
MCC	0.048 (0.174)	0.784 (0.207)	0.872 (0.096)	0.888 (0.077)
SIS-MCPγ_3				
PDR	0.100 (0.179)	0.407 (0.252)	0.677 (0.160)	0.775 (0.149)
FDR	0.014 (0.061)	0.097 (0.159)	0.033 (0.088)	0.018 (0.057)
MCC	0.160 (0.261)	0.527 (0.281)	0.802 (0.113)	0.868 (0.094)
PPSγ_4				
PDR	0.288 (0.380)	0.950 (0.380)	1.000 (0.000)	1.000 (0.000)
FDR	0.018 (0.071)	0.014 (0.071)	0.020 (0.052)	0.006 (0.028)
MCC	0.360 (0.386)	0.943 (0.386)	0.990 (0.027)	0.997 (0.015)
FRγ_4				
PDR	0.387 (0.439)	0.960 (0.197)	1.000 (0.000)	1.000 (0.000)
FDR	0.016 (0.052)	0.013 (0.051)	0.020 (0.050)	0.007 (0.031)
MCC	0.441 (0.424)	0.953 (0.197)	0.990 (0.026)	0.996 (0.016)
Lassoγ_4				
PDR	0.005 (0.029)	0.855 (0.349)	1.000 (0.000)	1.000 (0.000)
FDR	0.000 (0.000)	0.224 (0.161)	0.207 (0.159)	0.193 (0.131)
MCC	0.012 (0.070)	0.735 (0.300)	0.886 (0.091)	0.895 (0.074)
SIS-MCPγ_4				
PDR	0.060 (0.153)	0.377 (0.267)	0.675 (0.163)	0.775 (0.149)
FDR	0.005 (0.039)	0.064 (0.132)	0.027 (0.074)	0.014 (0.052)
MCC	0.096 (0.220)	0.499 (0.309)	0.803 (0.113)	0.870 (0.093)

died from prostate cancer during follow-up and 75 survivors. The censoring rate is 26.69%, and the median observed survival time was 100 months. We are interested in finding the genes related with the survival rate of prostate cancer. The genes selected and running times by different methods are summarized in Table 8. In Fig. 1, we display the estimated baseline survival curve of the selected model in red and the Kaplan–Meier estimator in blue, and the 95% confidence regions of the Kaplan–Meier estimator are in dashed lines.

From Table 8, we can see that the PPS, FR and SIS-MCP-EBIC method select the same features, while PPS and SIS-MCP-EBIC need much less running time than FR. Lasso-EBIC fails to select any feature. Lasso-BIC and SIS-MCP-BIC method select more features, and DAP3_4041 gene is selected by all these methods but Lasso-EBIC, which is also selected in Welchowski et al. (2019). From Fig. 1, the survival curve with features selected by PPS is closer to the Kaplan–Meier estimator, while those by SIS-MCP-BIC and Lasso-BIC overflow out of its confidence region. From the above analysis, we reasonably believe

Table 5 The averaged PDR, FDR and MCC over 100 replications with standard deviations in the brackets for S2 with $v = 0.7$

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
PPSγ_1				
PDR	0.382 (0.165)	0.812 (0.241)	0.993 (0.067)	1.000 (0.000)
FDR	0.840 (0.074)	0.675 (0.096)	0.603 (0.027)	0.600 (0.000)
MCC	0.235 (0.109)	0.511 (0.154)	0.627 (0.042)	0.632 (0.000)
FRγ_1				
PDR	0.882 (0.243)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.646 (0.099)	0.600 (0.000)	0.600 (0.000)	0.600 (0.000)
MCC	0.553 (0.157)	0.631 (0.000)	0.631 (0.000)	0.632 (0.000)
Lassoγ_1				
PDR	0.248 (0.196)	0.333 (0.092)	0.395 (0.113)	0.598 (0.280)
FDR	0.305 (0.348)	0.290 (0.262)	0.287 (0.255)	0.390 (0.331)
MCC	0.298 (0.210)	0.462 (0.130)	0.513 (0.099)	0.526 (0.094)
SIS-MCPγ_1				
PDR	0.200 (0.128)	0.335 (0.112)	0.363 (0.099)	0.398 (0.111)
FDR	0.677 (0.292)	0.597 (0.267)	0.483 (0.326)	0.351 (0.336)
MCC	0.224 (0.156)	0.345 (0.139)	0.401 (0.135)	0.478 (0.151)
PPSγ_2				
PDR	0.242 (0.184)	0.577 (0.286)	0.930 (0.197)	0.987 (0.094)
FDR	0.213 (0.291)	0.182 (0.209)	0.144 (0.157)	0.088 (0.126)
MCC	0.357 (0.246)	0.662 (0.198)	0.885 (0.155)	0.946 (0.098)
FRγ_2				
PDR	0.570 (0.418)	0.990 (0.100)	1.000 (0.000)	1.000 (0.000)
FDR	0.238 (0.287)	0.122 (0.152)	0.096 (0.110)	0.084 (0.115)
MCC	0.565 (0.368)	0.929 (0.116)	0.949 (0.060)	0.955 (0.063)
Lassoγ_2				
PDR	0.033 (0.085)	0.225 (0.156)	0.345 (0.054)	0.355 (0.056)
FDR	0.000 (0.000)	0.033 (0.103)	0.029 (0.094)	0.038 (0.111)
MCC	0.069 (0.169)	0.383 (0.257)	0.576 (0.053)	0.581 (0.055)
SIS-MCPγ_2				
PDR	0.118 (0.137)	0.288 (0.132)	0.337 (0.078)	0.367 (0.098)
FDR	0.074 (0.184)	0.104 (0.187)	0.103 (0.177)	0.035 (0.097)
MCC	0.213 (0.229)	0.464 (0.191)	0.542 (0.081)	0.588 (0.070)
PPSγ_3				
PDR	0.168 (0.176)	0.523 (0.282)	0.885 (0.235)	0.980 (0.114)
FDR	0.074 (0.210)	0.095 (0.175)	0.068 (0.103)	0.043 (0.087)
MCC	0.287 (0.270)	0.658 (0.214)	0.897 (0.156)	0.966 (0.093)
FRγ_3				
PDR	0.363 (0.396)	0.950 (0.203)	1.000 (0.000)	1.000 (0.000)
FDR	0.103 (0.235)	0.073 (0.184)	0.034 (0.065)	0.033 (0.069)
MCC	0.421 (0.388)	0.935 (0.187)	0.982 (0.034)	0.983 (0.037)
Lassoγ_3				
PDR	0.020 (0.068)	0.182 (0.163)	0.318 (0.089)	0.352 (0.052)

Table 5 (continued)

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
FDR	0.000 (0.000)	0.016 (0.070)	0.023 (0.083)	0.017 (0.076)
MCC	0.042 (0.135)	0.317 (0.275)	0.541 (0.134)	0.586 (0.045)
SIS-MCPγ_3				
PDR	0.087 (0.126)	0.278 (0.136)	0.335 (0.077)	0.363 (0.096)
FDR	0.047 (0.154)	0.063 (0.144)	0.096 (0.172)	0.027 (0.084)
MCC	0.162 (0.221)	0.463 (0.202)	0.543 (0.080)	0.589 (0.067)
PPSγ_4				
PDR	0.118 (0.154)	0.483 (0.289)	0.838 (0.267)	0.970 (0.135)
FDR	0.033 (0.174)	0.038 (0.106)	0.031 (0.080)	0.028 (0.076)
MCC	0.220 (0.263)	0.634 (0.253)	0.886 (0.172)	0.968 (0.100)
FRγ_4				
PDR	0.203 (0.280)	0.877 (0.306)	1.000 (0.000)	1.000 (0.000)
FDR	0.049 (0.183)	0.013 (0.046)	0.009 (0.034)	0.014 (0.043)
MCC	0.289 (0.332)	0.886 (0.283)	0.996 (0.018)	0.993 (0.022)
Lassoγ_4				
PDR	0.017 (0.065)	0.130 (0.160)	0.287 (0.128)	0.337 (0.063)
FDR	0.000 (0.000)	0.010 (0.057)	0.009 (0.053)	0.006 (0.041)
MCC	0.034 (0.125)	0.226 (0.275)	0.491 (0.205)	0.571 (0.088)
SIS-MCPγ_4				
PDR	0.068 (0.121)	0.263 (0.144)	0.335 (0.077)	0.362 (0.095)
FDR	0.022 (0.088)	0.044 (0.123)	0.064 (0.141)	0.024 (0.079)
MCC	0.126 (0.212)	0.443 (0.225)	0.553 (0.069)	0.588 (0.065)

that DAP3_4041 could be a vital gene related to the survival of prostate cancer and the newly found gene DAP2_5670 deserves more attention for future study. DAP3_4041 was identified as an inhibitor of apoptosis and it was found in most tumor cells (Stefano et al. 2010; Xu et al. 2015, 2020).

5.2 Gastric Cancer Recurrence data

The Gastric Cancer Recurrence (GCR) data were published and analyzed in Jeeyun et al. (2014) and Oh et al. (2018); it was available from the Gene Expression Omnibus with accession number GSE26253. Gastric cancer is the second most common cause of cancer-related death worldwide. In this data, the locoregional or distant recurrence of 432 patient samples after curative surgery plus adjuvant chemoradiotherapy is collected. We are interested in the relationship between RNA expression levels and recurrence event of gastric cancer. Their RNA microarray expression measurements have 17418 probes. During the follow-up, 177 patients recurred and the other 255 patients did not relapse, which led to a censoring rate of 59%. The

Table 6 The averaged PDR, FDR and MCC over 100 replications with standard deviations in the brackets for S3

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
PPSγ_1				
PDR	0.797 (0.230)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.673 (0.097)	0.597 (0.021)	0.600 (0.000)	0.600 (0.000)
MCC	0.503 (0.149)	0.633 (0.015)	0.631 (0.000)	0.632 (0.000)
FRγ_1				
PDR	0.915 (0.173)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.634 (0.069)	0.600 (0.000)	0.600 (0.000)	0.600 (0.000)
MCC	0.573 (0.111)	0.631 (0.000)	0.631 (0.000)	0.632 (0.000)
Lassoγ_1				
PDR	0.913 (0.211)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.691 (0.162)	0.744 (0.075)	0.774 (0.057)	0.771 (0.063)
MCC	0.492 (0.135)	0.498 (0.072)	0.470 (0.059)	0.474 (0.064)
SIS-MCPγ_1				
PDR	0.168 (0.141)	0.337 (0.182)	0.467 (0.164)	0.592 (0.171)
FDR	0.641 (0.337)	0.426 (0.341)	0.336 (0.273)	0.229 (0.224)
MCC	0.235 (0.208)	0.431 (0.232)	0.548 (0.190)	0.666 (0.164)
PPSγ_2				
PDR	0.645 (0.324)	0.993 (0.067)	1.000 (0.000)	1.000 (0.000)
FDR	0.412 (0.240)	0.146 (0.149)	0.088 (0.113)	0.075 (0.116)
MCC	0.596 (0.251)	0.918 (0.100)	0.953 (0.062)	0.960 (0.064)
FRγ_2				
PDR	0.832 (0.275)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.356 (0.229)	0.128 (0.141)	0.090 (0.109)	0.086 (0.117)
MCC	0.719 (0.228)	0.930 (0.080)	0.952 (0.059)	0.954 (0.065)
Lassoγ_2				
PDR	0.092 (0.247)	0.578 (0.447)	0.962 (0.175)	1.000 (0.000)
FDR	0.037 (0.123)	0.314 (0.307)	0.607 (0.180)	0.586 (0.149)
MCC	0.108 (0.233)	0.443 (0.299)	0.581 (0.132)	0.633 (0.114)
SIS-MCPγ_2				
PDR	0.148 (0.134)	0.312 (0.167)	0.433 (0.161)	0.568 (0.164)
FDR	0.503 (0.410)	0.327 (0.345)	0.189 (0.250)	0.080 (0.167)
MCC	0.245 (0.219)	0.450 (0.224)	0.583 (0.179)	0.717 (0.149)
PPSγ_3				
PDR	0.520 (0.368)	0.993 (0.067)	1.000 (0.000)	1.000 (0.000)
FDR	0.426 (0.332)	0.119 (0.148)	0.046 (0.080)	0.031 (0.063)
MCC	0.520 (0.315)	0.932 (0.099)	0.976 (0.042)	0.984 (0.033)
FRγ_3				
PDR	0.685 (0.359)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.319 (0.274)	0.078 (0.109)	0.040 (0.074)	0.028 (0.061)
MCC	0.656 (0.284)	0.958 (0.059)	0.979 (0.039)	0.985 (0.032)
Lassoγ_3				
PDR	0.055 (0.184)	0.418 (0.444)	0.907 (0.265)	0.990 (0.100)

Table 6 (continued)

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
FDR	0.014 (0.073)	0.173 (0.248)	0.517 (0.223)	0.530 (0.169)
MCC	0.077 (0.199)	0.367 (0.343)	0.599 (0.185)	0.665 (0.134)
SIS-MCPγ_3				
PDR	0.133 (0.138)	0.307 (0.169)	0.430 (0.161)	0.565 (0.166)
FDR	0.388 (0.419)	0.298 (0.340)	0.173 (0.245)	0.072 (0.166)
MCC	0.226 (0.229)	0.450 (0.227)	0.587 (0.178)	0.718 (0.150)
PPSγ_4				
PDR	0.382 (0.368)	0.993 (0.067)	1.000 (0.000)	1.000 (0.000)
FDR	0.460 (0.404)	0.107 (0.143)	0.029 (0.065)	0.017 (0.049)
MCC	0.421 (0.346)	0.939 (0.096)	0.985 (0.034)	0.991 (0.026)
FRγ_4				
PDR	0.545 (0.390)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
FDR	0.297 (0.331)	0.050 (0.086)	0.019 (0.056)	0.014 (0.046)
MCC	0.582 (0.322)	0.973 (0.046)	0.990 (0.030)	0.993 (0.024)
Lassoγ_4				
PDR	0.030 (0.122)	0.275 (0.399)	0.782 (0.387)	0.988 (0.101)
FDR	0.004 (0.040)	0.083 (0.176)	0.392 (0.259)	0.485 (0.177)
MCC	0.049 (0.154)	0.272 (0.349)	0.567 (0.272)	0.696 (0.136)
SIS-MCPγ_4				
PDR	0.112 (0.128)	0.297 (0.167)	0.428 (0.161)	0.562 (0.165)
FDR	0.308 (0.417)	0.274 (0.333)	0.168 (0.241)	0.061 (0.148)
MCC	0.204 (0.226)	0.449 (0.220)	0.587 (0.176)	0.719 (0.143)

median observed survival time was 57.35 months. The corresponding results are summarized in Table 9 and Fig. 2.

From Table 9, we can see that the PPS and FR select the feature ILMN_1713561 which is selected by Lasso-BIC and SIS-MCP-BIC too, while Lasso-EBIC and SIS-MCP-EBIC select no feature. PPS and FR construct the most parsimonious model, but FR requires much more running time. Figure 2 demonstrates a similar pattern as Fig. 1 and provides a supportive evidence for the significance of gene ILMN_1713561. This gene promotes the metastatic potential of gastric cancer cells (Umeda et al. 2022).

6 Conclusion

In this paper, we proposed a new sequential feature selection procedure PPS for high-dimensional Cox model. We defined a novel partial profile score to measure the correlation between residuals of fitted model and unselected features. Compared with forward regression for high-dimensional Cox model, selecting important variables based on this score can release abundant running time at each step.

Table 7 The averaged PDR, FDR and MCC over 100 replications with standard deviations in the brackets for S4

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
PPSγ_1				
PDR	0.325 (0.233)	0.817 (0.211)	0.943 (0.083)	0.977 (0.058)
FDR	0.864 (0.104)	0.673 (0.084)	0.622 (0.033)	0.609 (0.023)
MCC	0.197 (0.153)	0.514 (0.134)	0.596 (0.052)	0.617 (0.037)
FRγ_1				
PDR	0.528 (0.298)	0.902 (0.174)	0.982 (0.052)	0.993 (0.033)
FDR	0.789 (0.119)	0.639 (0.070)	0.607 (0.021)	0.603 (0.013)
MCC	0.325 (0.192)	0.568 (0.111)	0.620 (0.033)	0.628 (0.021)
Lassoγ_1				
PDR	0.315 (0.279)	0.292 (0.214)	0.550 (0.240)	0.808 (0.104)
FDR	0.424 (0.448)	0.171 (0.253)	0.298 (0.292)	0.566 (0.181)
MCC	0.212 (0.187)	0.413 (0.208)	0.563 (0.122)	0.574 (0.085)
SIS-MCPγ_1				
PDR	0.203 (0.133)	0.332 (0.131)	0.407 (0.117)	0.420 (0.120)
FDR	0.726 (0.286)	0.653 (0.302)	0.571 (0.333)	0.412 (0.359)
MCC	0.196 (0.146)	0.296 (0.131)	0.379 (0.168)	0.458 (0.169)
PPSγ_2				
PDR	0.143 (0.182)	0.633 (0.334)	0.890 (0.179)	0.958 (0.073)
FDR	0.173 (0.320)	0.138 (0.189)	0.116 (0.124)	0.105 (0.121)
MCC	0.242 (0.250)	0.690 (0.285)	0.877 (0.134)	0.924 (0.081)
FRγ_2				
PDR	0.225 (0.273)	0.817 (0.269)	0.948 (0.131)	0.992 (0.037)
FDR	0.242 (0.323)	0.146 (0.164)	0.122 (0.131)	0.103 (0.116)
MCC	0.295 (0.285)	0.800 (0.241)	0.906 (0.113)	0.941 (0.067)
Lassoγ_2				
PDR	0.012 (0.049)	0.072 (0.114)	0.195 (0.197)	0.490 (0.300)
FDR	0.003 (0.033)	0.005 (0.050)	0.012 (0.064)	0.102 (0.167)
MCC	0.025 (0.100)	0.150 (0.221)	0.341 (0.265)	0.589 (0.243)
SIS-MCPγ_2				
PDR	0.043 (0.099)	0.163 (0.157)	0.290 (0.167)	0.363 (0.158)
FDR	0.024 (0.106)	0.049 (0.140)	0.072 (0.153)	0.045 (0.121)
MCC	0.081 (0.173)	0.298 (0.244)	0.475 (0.184)	0.564 (0.160)
PPSγ_3				
PDR	0.113 (0.162)	0.535 (0.356)	0.833 (0.256)	0.940 (0.135)
FDR	0.056 (0.181)	0.073 (0.153)	0.060 (0.102)	0.039 (0.078)
MCC	0.209 (0.244)	0.635 (0.313)	0.863 (0.191)	0.943 (0.119)
FRγ_3				
PDR	0.137 (0.203)	0.687 (0.358)	0.908 (0.202)	0.987 (0.202)
FDR	0.066 (0.205)	0.061 (0.132)	0.048 (0.094)	0.040 (0.094)
MCC	0.231 (0.267)	0.742 (0.317)	0.918 (0.148)	0.972 (0.148)
Lassoγ_3				
PDR	0.005 (0.029)	0.057 (0.101)	0.140 (0.158)	0.402 (0.305)

Table 7 (continued)

Method	$n = 100$	$n = 200$	$n = 300$	$n = 400$
FDR	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.057 (0.126)
MCC	0.012 (0.070)	0.122 (0.205)	0.266 (0.264)	0.523 (0.287)
SIS-MCPγ_3				
PDR	0.027 (0.078)	0.133 (0.161)	0.268 (0.180)	0.353 (0.163)
FDR	0.010 (0.070)	0.015 (0.066)	0.044 (0.117)	0.025 (0.095)
MCC	0.053 (0.146)	0.246 (0.257)	0.447 (0.215)	0.561 (0.165)
PPSγ_4				
PDR	0.058 (0.099)	0.367 (0.351)	0.728 (0.327)	0.917 (0.181)
FDR	0.005 (0.050)	0.020 (0.089)	0.022 (0.063)	0.016 (0.047)
MCC	0.128 (0.200)	0.491 (0.343)	0.800 (0.261)	0.936 (0.157)
FRγ_4				
PDR	0.073 (0.126)	0.450 (0.387)	0.812 (0.302)	0.942 (0.170)
FDR	0.005 (0.050)	0.021 (0.099)	0.017 (0.056)	0.015 (0.045)
MCC	0.147 (0.224)	0.555 (0.369)	0.857 (0.251)	0.951 (0.153)
Lassoγ_4				
PDR	0.003 (0.023)	0.040 (0.086)	0.110 (0.143)	0.305 (0.288)
FDR	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.030 (0.093)
MCC	0.008 (0.057)	0.088 (0.180)	0.215 (0.253)	0.431 (0.309)
SIS-MCPγ_4				
PDR	0.025 (0.076)	0.093 (0.130)	0.227 (0.186)	0.337 (0.176)
FDR	0.010 (0.070)	0.012 (0.058)	0.030 (0.097)	0.018 (0.079)
MCC	0.049 (0.142)	0.185 (0.234)	0.389 (0.248)	0.540 (0.192)

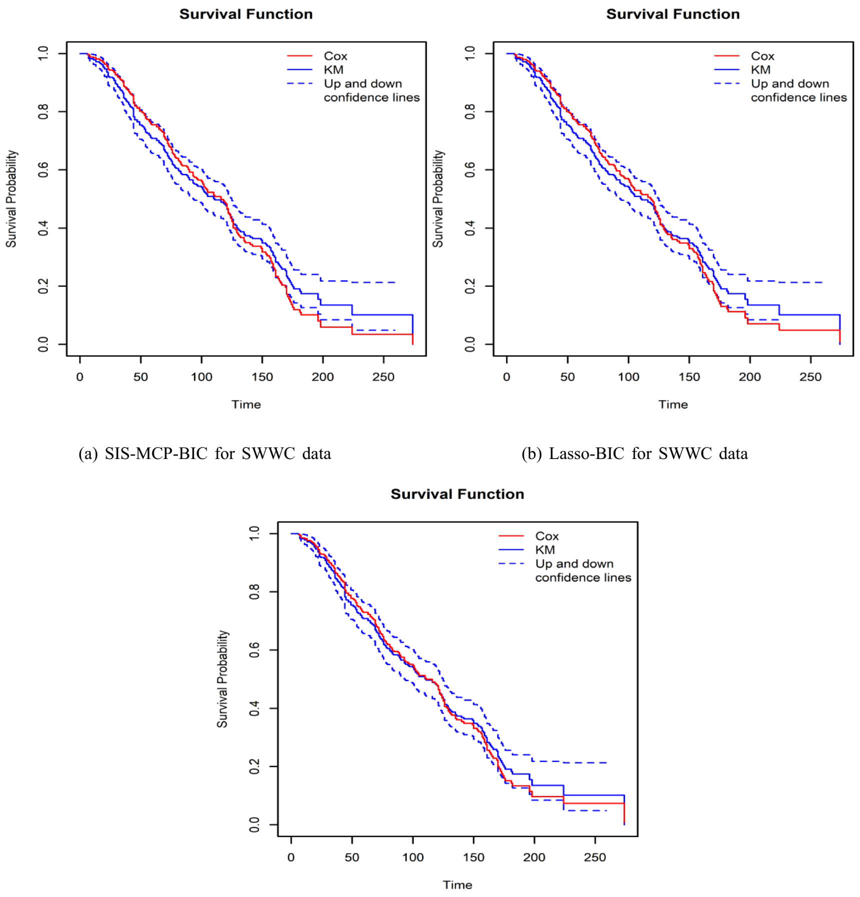
Table 8 The ID of genes selected in Swedish Watchful Waiting Cohort data

Method	The ID of genes				Running time (s)
PPS	DAP2_5670	DAP3_4041			0.18
FR	DAP2_5670	DAP3_4041			55.57
Lasso-BIC	DAP1_1759	DAP1_4857	DAP1_5047	DAP2_5670	
	DAP3_1787	DAP3_3482	DAP3_4041		1.33
Lasso-EBIC	∅				1.33
SIS-MCP-BIC	DAP1_1759	DAP1_4857	DAP3_1787	DAP3_4041	
	DAP4_1762	DAP4_1868	DAP4_2296	DAP4_6068	0.27
SIS-MCP-EBIC	DAP2_5670	DAP3_4041			0.27

Under mild conditions, we prove that all relevant features are selected before irrelevant features and EBIC will reach the minimum point when the selected features are exactly the set of true relevant features with probability converging to 1.

Table 9 The ID of genes selected in Gastric Cancer Recurrence data

Method	The ID of genes			Running time (s)
PPS	ILMN_1713561			0.40
FR	ILMN_1713561			120.47
Lasso-BIC	ILMN_1713561	ILMN_1732158	ILMN_1736078	
	ILMN_1811790			3.03
Lasso-EBIC	∅			3.03
SIS-MCP-BIC	ILMN_1673548	ILMN_1713561	ILMN_1732158	
	ILMN_1811790	ILMN_2382679		0.74
SIS-MCP-EBIC	∅			0.74

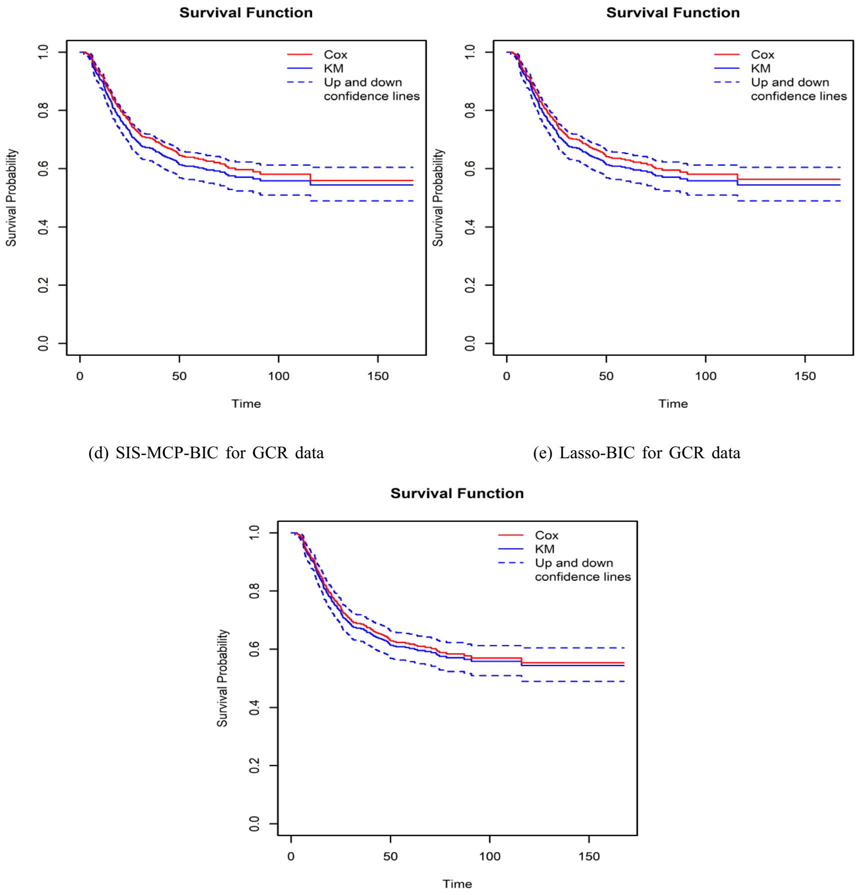


(a) SIS-MCP-BIC for SWWC data

(b) Lasso-BIC for SWWC data

(c)PPS,FR,SIS-MCP-EBIC for SWWC data

Fig. 1 Survival curves estimated by Cox model with different selected features for the SWWC data



(d) SIS-MCP-BIC for GCR data

(e) Lasso-BIC for GCR data

(f) PPS,FR for GCR data

Fig. 2 Survival curves estimated by Cox model with different selected features for the GCR data

Extensive simulation results and real data applications demonstrate that PPS has an edge over other existing approaches in terms of feature selection.

Appendix

In this section, we provide the detailed technical proofs of our main theorems in Sect. 3. Firstly, we need the following Lemma.

Lemma 1 *Under condition* (*A*1),(*A*2),(*A*4), *there exist some constants* A_1, A_2, A_3, *which do not depend on* n, *such that*

$$\Pr\left\{\sup_{|s|\le\rho,u\in[0,\tau]}|R_s^{(0)}(\boldsymbol{\beta}_s^*,u)-r_s^{(0)}(\boldsymbol{\beta}_s^*,u)|\ge A_1\sqrt{\rho\frac{\ln p}{n}}\right\}\le\exp(-3\rho\ln p),\quad(7)$$

$$\Pr\left\{\sup_{|s|\le\rho,u\in[0,\tau],k\in s^c}|R_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)-r_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)|\ge A_2\sqrt{\rho\frac{\ln p}{n}}\right\}\le\exp(-3\rho\ln p),\quad(8)$$

$$\Pr\left\{\sup_{|s|\le\rho,u\in[0,\tau],k\in s^c}\left|\frac{R_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)}{R_s^{(0)}(\boldsymbol{\beta}_s^*,u)}-\frac{r_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)}{r_s^{(0)}(\boldsymbol{\beta}_s^*,u)}\right|\ge A_3\sqrt{\rho\frac{\ln p}{n}}\right\}\le 3\exp(-3\rho\ln p).\quad(9)$$

Proof of Lemma 1 We now focus on inequality (7). By condition (A2) and (A4), we have

$$\exp(\boldsymbol{\beta}_s^{*T}\boldsymbol{Z}_s)\le\exp(\|\boldsymbol{\beta}_s^*\|_1\|\boldsymbol{Z}_s\|_\infty)\le\exp(KL).$$

Define

$$h(u;Y,\boldsymbol{Z}_s)=(\exp(KL))^{-1}Y(u)\exp(\boldsymbol{\beta}_s^{*T}\boldsymbol{Z}_s),$$

then $h(u;Y,\boldsymbol{Z}_s)\in[-1,1]$ for $u\in[0,\tau]$, the VC index (Van der Vaart and Wellner 1996) of the function class $\{h(u;Y,\boldsymbol{Z}_s),u\in[0,\tau]\}$ is two. By Theorem 1.1 in Talagrand (1994), for $\epsilon>0$, there exists a constant A independent of ϵ such that,

$$\Pr\left\{\sup_{u\in[0,\tau]}|n^{-1}\sum_{i=1}^n h(u;Y_i,\boldsymbol{Z}_{is})-E\{h(u;Y,\boldsymbol{Z}_s)\}|\ge\frac{1}{\sqrt{n}}\varepsilon\right\}\le\frac{A}{\varepsilon}(\frac{A\varepsilon^2}{2})^2e^{-2\varepsilon^2}.$$

From the combinatorical inequality $C_p^s\le(ep/s)^s$, we have

$$\Pr\left\{\sup_{|s|\le\rho,u\in[0,\tau]}|R_s^{(0)}(\boldsymbol{\beta}_s^*,u)-r_s^{(0)}(\boldsymbol{\beta}_s^*,u)|\ge\frac{\exp(KL)}{\sqrt{n}}\varepsilon\right\}\le\sum_{s=1}^{\rho}(\frac{ep}{s})^s\frac{A}{\varepsilon}(\frac{A\varepsilon^2}{2})^2e^{-2\varepsilon^2}.$$

By choosing an appropriate ε of order $\sqrt{\rho\ln p}$, we verify inequality (7). Inequality (8) can be obtained in a similar way.

We now turn to inequality (9).

$$\frac{R_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)}{R_s^{(0)}(\boldsymbol{\beta}_s^*,u)}-\frac{r_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)}{r_s^{(0)}(\boldsymbol{\beta}_s^*,u)}=\frac{R_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)-r_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)}{R_s^{(0)}(\boldsymbol{\beta}_s^*,u)}+\frac{(r_s^{(0)}(\boldsymbol{\beta}_s^*,u)-R_s^{(0)}(\boldsymbol{\beta}_s^*,u))r_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)}{r_s^{(0)}(\boldsymbol{\beta}_s^*,u)R_s^{(0)}(\boldsymbol{\beta}_s^*,u)}.$$

Under condition (A1), (A2), (A4), recall that $\omega=\Pr(X\ge\tau)$, we have

$$r_s^{(0)}(\boldsymbol{\beta}_s^*,u)=E\{Y(t)\exp(\boldsymbol{\beta}_s^{*T}\boldsymbol{Z}_s)\}\ge E\{Y(t)\exp(-\|\boldsymbol{\beta}_s^*\|_1\|\boldsymbol{Z}_s\|_\infty)\}\ge\omega\exp(-KL),$$

and $r_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)\le K\exp(KL)$ for all $|s|<\rho,u\in[0,\tau],k\in s^c$. Note that, $\sup_{|s|\le\rho,u\in[0,\tau]}|1/r_s^{(0)}(\boldsymbol{\beta}_s^*,u)|$ and $\sup_{|s|\le\rho,u\in[0,\tau],k\in s^c}|r_{ks}^{(1)}(\boldsymbol{\beta}_s^*,u)/r_s^{(0)}(\boldsymbol{\beta}_s^*,u)|$ are both bounded. Define

$$\Omega = \left\{ \sup_{|s|\le\rho, u\in[0,\tau]} |R_s^{(0)}(\boldsymbol{\beta}_s^*, u) - r_s^{(0)}(\boldsymbol{\beta}_s^*, u)| \ge A_1\sqrt{\rho\frac{\ln p}{n}} \right\},$$

under Ω^c, when n is sufficiently large, $\sup_{|s|\le\rho, u\in[0,\tau]} |1/R_s^{(0)}(\boldsymbol{\beta}_s^*, u)|$ is also bounded. Consequently, there exist constants c_1 and c_2 such that

$$\left|\frac{R_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)}{R_s^{(0)}(\boldsymbol{\beta}_s^*, u)} - \frac{r_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)}{r_s^{(0)}(\boldsymbol{\beta}_s^*, u)}\right| \le c_1\left|R_s^{(0)}(\boldsymbol{\beta}_s^*, u) - r_s^{(0)}(\boldsymbol{\beta}_s^*, u)\right| + c_2\left|R_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u) - r_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)\right|.$$

Let $A_3 = \max\{2c_1A_1, 2c_2A_2\}$; we have

$$\begin{aligned}
&\Pr\left\{ \sup_{|s|\le\rho, u\in[0,\tau], k\in s^c} \left|\frac{R_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)}{R_s^{(0)}(\boldsymbol{\beta}_s^*, u)} - \frac{r_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)}{r_s^{(0)}(\boldsymbol{\beta}_s^*, u)}\right| \ge A_3\sqrt{\rho\frac{\ln p}{n}} \right\} \\
&\le \Pr(\Omega) + \Pr\left\{ \sup_{|s|\le\rho, u\in[0,\tau], k\in s^c} \left|\frac{R_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)}{R_s^{(0)}(\boldsymbol{\beta}_s^*, u)} - \frac{r_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)}{r_s^{(0)}(\boldsymbol{\beta}_s^*, u)}\right| \ge A_3\sqrt{\rho\frac{\ln p}{n}}, \Omega^c \right\} \\
&\le \Pr(\Omega) + \Pr\left\{ \sup_{|s|\le\rho, u\in[0,\tau]} |R_s^{(0)}(\boldsymbol{\beta}_s^*, u) - r_s^{(0)}(\boldsymbol{\beta}_s^*, u)| \ge \frac{A_3}{2c_1}\sqrt{\rho\frac{\ln p}{n}} \right\} \\
&\quad + \Pr\left\{ \sup_{|s|\le\rho, u\in[0,\tau], k\in s^c} |R_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u) - r_{ks}^{(1)}(\boldsymbol{\beta}_s^*, u)| \ge \frac{A_3}{2c_2}\sqrt{\rho\frac{\ln p}{n}} \right\} \le 3\exp(-3\rho\ln p).
\end{aligned}$$

This completes the proof of Lemma 1. □

Proof of Theorem 1 For $k = 0$, $s_{*0} = \emptyset$. Define

$$\mathcal{A}_m = \{k^* : |\psi_{k^*}(s_{*m})| = \max_{k\in s_{*m}^c} |\psi_k(s_{*m})|\};$$

it suffices to show that $\Pr(\mathcal{A}_m \subset s_0) \to 1$ uniformly for m, when $s_{*m} \subset s_0$ and $|s_{*m}| < |s_0|$.

Firstly, we show that, for $s_{*m} \subset s_0$, $s_{*m}^- = s_0 \cap s_{*m}^c$,

$$\max_{k\in s_{*m}^-} |\Phi_k(s_{*m})| \ge C_n\rho\left(\frac{\ln p}{n}\right)^{1/4}, \text{ when } n \to \infty. \tag{10}$$

Using conditional independency and condition (A2), we have

$$\begin{aligned}\int_0^\tau E\{Y(u)Z_k\lambda_0(u)\exp(\beta_{0s_0}^T\mathbf{Z}_{s_0})\}du &= \int_0^\infty E\{E\{Y(u)Z_k\lambda_0(u)\exp(\beta_{0s_0}^T\mathbf{Z}_{s_0})|\mathbf{Z}\}\}\mathrm{d}u\\
&= \int_0^\infty E\{Z_k\lambda_0(u)\exp(\beta_{0s_0}^T\mathbf{Z}_{s_0})S_T(u|\mathbf{Z}_{s_0})S_C(u|\mathbf{Z}_{s_{C0}})\}\mathrm{d}u\\
&= \int_0^\infty E\{Z_kf_T(u|\mathbf{Z}_{s_0})S_C(u|\mathbf{Z}_{s_{C0}})\}\mathrm{d}u = E\{Z_k\int_0^\infty F_T(u|\mathbf{Z}_{s_0})f_C(u|\mathbf{Z}_{s_{C0}})\mathrm{d}u\}\\
&= -\int_0^\infty E\{Z_kS_T(u|\mathbf{Z}_{s_0})f_C(u|\mathbf{Z}_{s_{C0}})\}\mathrm{d}u\\
&= -\int_0^\infty E\{E\{Z_kS_T(u|\mathbf{Z}_{s_0})f_C(u|\mathbf{Z}_{s_{C0}})|\mathbf{Z}_{S_0\backslash k}\}\}\mathrm{d}u\end{aligned}$$

and for $k \in s_{*m}^-$, $s_{*m} \subset s_0$, we have

$$\begin{aligned}E\{Y(u)Z_k\exp(\beta_{s_{*m}}^{*^T}\mathbf{Z}_{s_{*m}})\} &= E\{E\{Y(u)Z_k\exp(\beta_{s_{*m}}^{*^T}\mathbf{Z}_{s_{*m}})|\mathbf{Z}\}\}\\
&= E\{Z_k\exp(\beta_{s_{*m}}^{*^T}\mathbf{Z}_{s_{*m}})S_T(u|\mathbf{Z}_{s_0})S_C(u|\mathbf{Z}_{s_{C0}})\}\\
&= E\{E\{Z_kS_T(u|\mathbf{Z}_{s_0})S_C(u|\mathbf{Z}_{s_{C0}})|\mathbf{Z}_{S_0\backslash k}\}\exp(\beta_{s_{*m}}^{*^T}\mathbf{Z}_{s_{*m}})\}.\end{aligned}$$

Hence, $\int_0^\tau v_k^{(1)}(u)du$ and $\int_0^\tau r_{ks_{*m}}^{(1)}(\boldsymbol{\beta}_{s_{*m}}^*, u)(r_{s_{*m}}^{(0)}(\boldsymbol{\beta}_{s_{*m}}^*, u))^{-1}v_s^{(0)}(u)du$ have opposite sign when condition (A6) is satisfied and $r_{s_{*m}}^{(0)}(\boldsymbol{\beta}_{s_{*m}}^*, t)$, $v_s^{(0)}(t)$, $\exp(\boldsymbol{\beta}_{s_{*m}}^{*^\top}\mathbf{Z}_{s_{*m}})$ are positive. Under condition (A8), for $k \in s_{*m}^-$ and $s_{*m} \subset s_0$, we have

$$\begin{aligned}|\boldsymbol{\Phi}_k(s_{*m})| &= \left|\int_0^\tau\left(v_k^{(1)}(u) - \frac{r_{ks_{*m}}^{(1)}(\boldsymbol{\beta}_{s_{*m}}^*, u)}{r_{s_{*m}}^{(0)}(\boldsymbol{\beta}_{s_{*m}}^*, u)}v_s^{(0)}(u)\right)\mathrm{d}u\right|\\
&\geq \left|\int_0^\tau v_k^{(1)}(u)du\right| = \left|\int_0^\tau E\{Y(u)Z_k\lambda_0(u)\exp(\boldsymbol{\beta}_{0s_0}^\top\mathbf{Z}_{s_0})\}\mathrm{d}u\right|\\
&= \left|\int_0^\tau E\{Z_kf_T(u|\mathbf{Z}_{s_0})S_C(u|\mathbf{Z}_{s_{C0}})\}\mathrm{d}u\right|\\
&\geq C_n\rho(\frac{\ln p}{n})^{1/4}\end{aligned}$$

where $C_n \to \infty$, (10) is therefore established.

Secondly, we show that there exists a constant A such that

$$\Pr\left\{\sup_{k\in s_{*m}^c, |s_{*m}|<\rho}|\psi_k(s_{*m}) - \boldsymbol{\Phi}_k(s_{*m})| \geq A\rho\left(\frac{\ln p}{n}\right)^{1/4}\right\} \to 0. \qquad (11)$$

By direct calculation, we have

$$
\begin{aligned}
&|\psi_k(s_{*m}) - \boldsymbol{\Phi}_k(s_{*m})| \\
&\quad \le \left|\psi_k(s_{*m}) - \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left(Z_{ik} - \frac{R^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}}, u)}{R^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}}, u)}\right)\mathrm{d}N_i(u)\right| \\
&\qquad + \left|\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left(Z_{ik} - \frac{R^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}}, u)}{R^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}}, u)}\right)\mathrm{d}N_i(u) - \boldsymbol{\Phi}_k(s_{*m})\right| \\
&\quad \overset{\text{def}}{=} \mathrm{I} + \mathrm{II}\,.
\end{aligned}
$$

When $|s_{*m}| < \rho$, by mean value theorem, there exists a $\tilde{\boldsymbol{\beta}}_{s_{*m}}$ between $\widehat{\boldsymbol{\beta}}_{s_{*m}}$ and $\boldsymbol{\beta}^*_{s_{*m}}$ such that

$$
\begin{aligned}
\mathrm{I} &= \left|\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left(\frac{\sum_{j=1}^{n} Y_j(u) Z_{jk}\exp(\widehat{\boldsymbol{\beta}}^T_{s_{*m}}\boldsymbol{Z}_{js_{*m}})}{\sum_{j=1}^{n} Y_j(u)\exp(\widehat{\boldsymbol{\beta}}^T_{s_{*m}}\boldsymbol{Z}_{js_{*m}})} - \frac{\sum_{j=1}^{n} Y_j(u) Z_{jk}\exp(\boldsymbol{\beta}^{*T}_{s_{*m}}\boldsymbol{Z}_{js_{*m}})}{\sum_{j=1}^{n} Y_j(u)\exp(\boldsymbol{\beta}^{*T}_{s_{*m}}\boldsymbol{Z}_{js_{*m}})}\right)\mathrm{d}N_i(u)\right| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Bigg|(\widehat{\boldsymbol{\beta}}_{s_{*m}} - \boldsymbol{\beta}^*_{s_{*m}})^T\Bigg\{\frac{\sum_{j=1}^{n} Y_j(u) Z_{jk}\boldsymbol{Z}_{js_{*m}}\exp(\tilde{\boldsymbol{\beta}}^T_{s_{*m}}\boldsymbol{Z}_{js_{*m}})}{\sum_{j=1}^{n} Y_j(u)\exp(\tilde{\boldsymbol{\beta}}^T_{s_{*m}}\boldsymbol{Z}_{js_{*m}})} \\
&\quad - \frac{\left(\sum_{j=1}^{n} Y_j(u) Z_{jk}\exp(\tilde{\boldsymbol{\beta}}^T_{s_{*m}}\boldsymbol{Z}_{js_{*m}})\right)\left(\sum_{j=1}^{n} Y_j(u)\boldsymbol{Z}_{js_{*m}}\exp(\tilde{\boldsymbol{\beta}}^T_{s_{*m}}\boldsymbol{Z}_{js_{*m}})\right)}{(\sum_{j=1}^{n} Y_j(u)\exp(\tilde{\boldsymbol{\beta}}^T_{s_{*m}}\boldsymbol{Z}_{js_{*m}}))^2}\Bigg\}\Bigg|\mathrm{d}N_i(u) \\
&\le 2K^2\sqrt{\rho}\|\widehat{\boldsymbol{\beta}}_{s_{*m}} - \boldsymbol{\beta}^*_{s_{*m}}\|.
\end{aligned}
$$

According to the Lemma G in Hong et al. (2019), there exists a A_0 such that

$$
\Pr\left\{\sup_{|s_{*m}|\le\rho}\|\widehat{\boldsymbol{\beta}}_{s_{*m}} - \boldsymbol{\beta}^*_{s_{*m}}\| \le A_0(\rho^2\frac{\ln p}{n})^{1/4}\right\} \ge 1 - 5\exp(-3\rho\ln p).
$$

Consequently,

$$
\Pr\left\{\sup_{k\in s^c_{*m}, |s_{*m}|<\rho}\mathrm{I} \ge 2K^2A_0\rho(\frac{\ln p}{n})^{1/4}\right\} \le 5p\exp(-3\rho\ln p) \le 5\exp(-2\rho\ln p).
$$

The second term

$$\begin{aligned}\mathrm{II}\ \le&\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left(Z_{ik}-\frac{R^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{R^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\right)\mathrm{d}N_i(u)-\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left(Z_{ik}-\frac{r^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{r^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\right)\mathrm{d}N_i(u)\right|\\ &+\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left(Z_{ik}-\frac{r^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{r^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\right)\mathrm{d}N_i(u)-\boldsymbol{\Phi}_k(s_{*m})\right|\\ &\overset{\mathrm{def}}{=}\ \mathrm{III}\ +\ \mathrm{IV}\ .\end{aligned}$$

For part III, define

$$\Omega=\left\{\sup_{|s|\le\rho,u\in[0,\tau],k\in s^c}\left|\frac{R^{(1)}_{ks}(\boldsymbol{\beta}^*_s,u)}{R^{(0)}_{s}(\boldsymbol{\beta}^*_s,u)}-\frac{r^{(1)}_{ks}(\boldsymbol{\beta}^*_s,u)}{r^{(0)}_{s}(\boldsymbol{\beta}^*_s,u)}\right|\ge A_3\sqrt{\rho\frac{\ln p}{n}}\right\}.$$

Under Ω^c, we have

$$\begin{aligned}\mathrm{III}\ \le&\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left|\frac{R^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{R^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}-\frac{r^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{r^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\right|\mathrm{d}N_i(u)\\ \le&\sup_{u\in[0,\tau]}\left|\frac{R^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{R^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}-\frac{r^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{r^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\right|\le A_3\sqrt{\rho\frac{\ln p}{n}}.\end{aligned}$$

By Lemma 1, we have

$$\Pr\left\{\sup_{k\in s^c_{*m},|s_{*m}|<\rho}\ \mathrm{III}\ \ge A_3\sqrt{\rho\frac{\ln p}{n}}\right\}\le 3\exp\left(-3\rho\ln p\right).$$

For part IV , when $|s_{*m}|<\rho$,

$$\begin{aligned}&\left|\int_0^{\tau}Z_{ik}-\frac{r^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{r^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\mathrm{d}N_i(u)\right|\\ &\le|Z_{ik}|+\sup_{t\in[0,\tau]}\left|\frac{r^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{r^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\right|\\ &\le K+\frac{K\exp\left(KL\right)}{\omega\exp\left(-KL\right)}\overset{\mathrm{def}}{=}Kc_3.\end{aligned}$$

Using Bernstein's inequality, when n is sufficiently large, we obtain

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left(Z_{ik}-\frac{r^{(1)}_{ks_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}{r^{(0)}_{s_{*m}}(\boldsymbol{\beta}^*_{s_{*m}},u)}\right)\mathrm{d}N_i(u)-\boldsymbol{\Phi}_k(s_{*m})\right|\geq 6Kc_3\sqrt{\rho\frac{\ln p}{n}}\right\}$$
$$\leq 2\exp\left\{-1/2\frac{36K^2c_3^2n\rho\ln p}{4nK^2c_3^2+12K^2c_3^2\sqrt{n\rho\ln p}/3}\right\}\leq 2\exp\left(-4\rho\ln p\right).$$

When n is sufficiently large,

$$\Pr\left\{\sup_{k\in s^c_{*m},|s_{*m}|<\rho}\text{IV}\ \geq 6Kc_3\sqrt{\rho\frac{\ln p}{n}}\right\}\leq\sum_{|s_{*m}|<\rho}\sum_{k=1}^{p}2\exp\left(-4\rho\ln p\right)$$
$$\leq p\sum_{s=1}^{\rho-1}(\frac{ep}{s})^s 2\exp\left(-4\rho\ln p\right)\leq 2\exp\left(-2\rho\ln p\right).$$

Consequently, we obtain

$$\Pr\left\{\sup_{k\in s^c_{*m},|s_{*m}|<\rho}\text{II}\ \geq (A_3+6Kc_3)\sqrt{\rho\frac{\ln p}{n}}\right\}\leq 3\exp\left(-3\rho\ln p\right)+2\exp\left(-2\rho\ln p\right).$$

Finally, when n is sufficiently large,

$$\Pr\left\{\sup_{k\in s^c_{*m},|s_{*m}|<\rho}|\psi_k(s_{*m})-\boldsymbol{\Phi}_k(s_{*m})|\geq(2K^2A_0+A_3+6Kc_3)\rho\left(\frac{\ln p}{n}\right)^{1/4}\right\}$$
$$\leq 7\exp\left(-2\rho\ln p\right)+3\exp\left(-3\rho\ln p\right)\to 0,$$

conclusion (11) is thus proved. According to condition (A7) and fact (10), for some constant $q\in(0,1)$,

$$\begin{aligned}&|\max_{k\in s^-_{*m}}|\boldsymbol{\Phi}_k(s_{*m})|-\max_{k\in s^c_0}|\boldsymbol{\Phi}_k(s_{*m})||\\&\geq(1-q)\max_{k\in s^-_{*m}}|\boldsymbol{\Phi}_k(s_{*m})|\\&\geq(1-q)C_n\rho(\frac{\ln p}{n})^{1/4}\end{aligned}$$

where $C_n\to\infty$; thus,

$$\Pr\left\{\max_{k\in s^-_{*m}}|\psi_k(s_{*m})|>\max_{k\in s^c_0}|\psi_k(s_{*m})|\right\}\to 1$$

uniformly when $s_{*m}\subset s_0$, which implies that $\Pr\{\mathcal{A}_m\subset s^-_{*m}\subset s_0\}\to 1$ uniformly when $s_{*m}\subset s_0$, the proof of Theorem 1 is thus completed. □

Proof of Theorem 2 Under the condition in the theorem, we have $\ln C_p^j = j \ln p(1 + o(1))$ when $j \leq Cp_0$; hence, the difference of EBIC values for two subsequent models is

$$\begin{aligned} D_m &= \mathrm{EBIC}_\gamma(s_{*m}) - \ \mathrm{EBIC}_\gamma(s_{*m+1}) \\ &= 2(\ell(\widehat{\boldsymbol{\beta}}_{s_{*m+1}}) - \ell(\widehat{\boldsymbol{\beta}}_{s_{*m}})) - |\mathcal{A}_m|(\ln n + 2\gamma \ln p)(1 + o(1)) \end{aligned}$$

where $\mathcal{A}_m \equiv s_{*m+1}/s_{*m}$ is defined in the proof of Theorem 1. Without loss of generality, we assume $|\mathcal{A}_m| = 1$. Obviously, $\rho^4 \ln p/n = (Cp_0)^4 \kappa \ln n/n \to 0$.

For conclusion (1) in the theorem, we note that, similar to Lemma B in Hong et al. (2019), if $|s| < \rho$ and $r \in s^-$, there exists $C_n \to \infty$ such that

$$\|\boldsymbol{\beta}^*_{s\cup\{r\}} - (\boldsymbol{\beta}^{*T}_s, 0)^\top\| \geq \kappa^{-1}_{\max} C_n \rho (\frac{\ln p}{n})^{1/4}.$$

Furthermore, similar to the proof of Theorem 1 in Hong et al. (2019), when $|s| < \rho$, $r \in s^-$, with probability converging to 1, there exists constant c_4 and c_5 such that

$$\ell(\widehat{\boldsymbol{\beta}}_{s\cup\{r\}}) - \ell(\widehat{\boldsymbol{\beta}}_s) \geq c_4 C_n^2 \rho^2 (n \ln p)^{1/2} - c_5 \rho^2 (n \ln p)^{1/2}$$

when $m < m^*$, $\mathcal{A}_m \in s_0$. Therefore,

$$\frac{2(\ell(\widehat{\boldsymbol{\beta}}_{s_{*m+1}}) - \ell(\widehat{\boldsymbol{\beta}}_{s_{*m}}))}{\ln p} \geq 2c_4 C_n^2 \rho^2 (\frac{n}{\ln p})^{1/2} - 2c_5 \rho^2 (\frac{n}{\ln p})^{1/2}$$

with probability converging to 1. Due to the fact that $p = O(n^\kappa)$, $\kappa > 1$, $\rho = Cp_0$ and $C_n \to \infty$, we have $2(\ell(\widehat{\boldsymbol{\beta}}_{s_{*m+1}}) - \ell(\widehat{\boldsymbol{\beta}}_{s_{*m}}))/(\ln p) \to \infty$. That is, Pr $(D_m > 0) \to 1$, for all m such that $m < m^*$, this completes the proof of (1) in Theorem 2.

Now we turn to conclusion (2) in the theorem. Note that when $s_0 \subseteq s$, $\boldsymbol{\beta}_{0s}$ is the root of

$$\int_0^\tau \left(v_s^{(1)}(t) - \frac{r_s^{(1)}(\boldsymbol{\beta}_s, t)}{r_s^{(0)}(\boldsymbol{\beta}_s, t)} v_s^{(0)}(t) \right) \mathrm{d}t = 0;$$

therefore, $\boldsymbol{\beta}_{0s} = \boldsymbol{\beta}^*_s$ for s in $\{s : s_0 \subset s,\ |s| \leq \rho\}$. By the Lemma G in Hong et al. (2019),

$$\Pr \left\{ \|\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_{0s}\| \leq A_6 \left(\rho^2 \frac{\ln p}{n} \right)^{1/4} \right\} \to 1$$

holds uniformly for s in $\{s : s_0 \subset s,\ |s| \leq \rho\}$, and then following the proof (2) of Theorem 3 in Luo et al. (2015), the desired result is obtained. □

References

Bradic, J., Fan, J., Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics, 39*(6), 3092–3120.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics, 34*(2), 559–583.

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

Cheng, M., Honda, T., Zhang, J. (2014). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association, 111*(515), 1209–1221.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *34*(2), 187–202.

Fan, J., Feng, Y., Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. *Institute of Mathematical Statistics Collections, 6,* 70–86.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

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., Samworth, R., Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research, 10*(5), 2013–2038.

Gorst-Rasmussen, A., Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75*(2), 217–245.

Hong, H. G., Zheng, Q., Li, Y. (2019). Forward regression for cox models with high-dimensional covariates. *Journal of Multivariate Analysis, 173*, 268–290.

Huang, J., Sun, T., Ying, Z., Yu, Y., Zhang, C. H. (2013). Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics, 41*(3), 1142–1165.

Ing, C., Lai, T. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica, 21*(4), 1473–1513.

Jeeyun, L., Insuk, S., In-Gu, D., Kyoung-Mee, K., Hoon, P. S., Oh, P. J., et al. (2014). Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. *PLoS ONE, 9*(3), e90133.

Kong, S., Nan, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica, 24*(1), 25–42.

Luo, S., Chen, Z. (2014). Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association, 109*(507), 1229–1240.

Luo, S., Chen, Z. (2021). Sequential interaction group selection by the principle of correlation search for high-dimensional interaction models. *Statistica Sinica, 31*(1), 197–221.

Luo, S., Xu, J., Chen, Z. (2015). Extended Bayesian information criterion in the cox model with a high-dimensional feature space. *Annals of the Institute of Statistical Mathematics, 67*(2), 287–311.

Oh, S. C., Sohn, B. H., Cheong, J. H., Kim, S. B., Lee, J. E., Park, K. C., et al. (2018). Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. *Nature Communications, 9*(1), 1–14.

Sboner, A., Demichelis, F., Calza, S., Pawitan, Y., Setlur, S. R., Hoshida, Y., et al. (2010). Molecular sampling of prostate cancer: A dilemma for predicting disease progression. *BMC Medical Genomics, 3*(1), 1–12.

Song, R., Lu, W., Ma, S., Jeng, J. X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika, 101*(4), 799–814.

Stefano, A., Iovino, F., Lombardo, Y., Eterno, V., Hger, T., Dieli, F., Stassi, G., Todaro, M. (2010). Survivin is regulated by interleukin-4 in colon cancer stem cells. *Journal of Cellular Physiology, 225*(2), 555–561.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *The Annals of Probability, 22*(1), 28–76.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *58*(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine, 16*(4), 385–395.

Umeda, S., Kanda, M., Shimizu, D., Nakamura, S., Sawaki, K., Inokawa, Y., et al. (2022). Lysosomal-associated membrane protein family member 5 promotes the metastatic potential of gastric cancer cells. *Gastric Cancer*. https://doi.org/10.1007/s10120-022-01284-y.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics, 36*(2), 614–645.

Van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association, 104*(488), 1512–1524.

Welchowski, T., Zuber, V., Schmid, M. (2019). Correlation-adjusted regression survival scores for high-dimensional variable selection. *Statistics in medicine, 38*(13), 2413–2427.

Xu, Y., Jin, Y., Liu, L., Zhang, X., Chen, Y., Wei, J. (2015). Study of circulating IgG antibodies to peptide antigens derived from BIRC5 and MYC in cervical cancer. *FEBS Open Bio, 5*(1), 198–201.

Xu, Y., Peng, P., Zhou, Q. (2020). MIR-203 mimic down-regulates baculoviral IAP repeat containing 5 expression and affects proliferation and apoptosis of gastric cancer cells. *Journal of Biomaterials and Tissue Engineering, 10*(1), 81–86.

Yang, G., Yu, Y., Li, R., Buu, A. (2016). Feature screening in ultrahigh dimensional Cox's model. *Statistica Sinica, 26*(3), 881–901.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894–942.

Zhang, H. H., Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika, 94*(3), 691–703.

Zhao, S. D., Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis, 105*(1), 397–411.

Zheng, Q., Hong, H. G., Li, Y. (2020). Building generalized linear models with ultrahigh dimensional features: A sequentially conditional approach. *Biometrics, 76*(1), 47–60.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika, 95*(1), 241–247.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society Series B* (*Statistical Methodology*), *67*(2), 301–320.