Journal of Multivariate Analysis 171 (2019) 284–297

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Feature screening in ultrahigh-dimensional varying-coefficient Cox model

Guangren Yang [a,*], Ling Zhang [b], Runze Li [c], Yuan Huang [d]

[a] Department of Statistics, School of Economics, Jinan University, Guangzhou, 510632, China
[b] Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA
[c] Department of Statistics and the Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA
[d] Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA

ARTICLE INFO

Article history:
Received 17 March 2018
Available online 28 December 2018

AMS 2010 subject classifications:
62N01
62N02

Keywords:
Cox model
Partial likelihood
Penalized likelihood
Ultrahigh-dimensional survival data

ABSTRACT

The varying-coefficient Cox model is flexible and useful for modeling the dynamic changes of regression coefficients in survival analysis. In this paper, we study feature screening for varying-coefficient Cox models in ultrahigh-dimensional covariates. The proposed screening procedure is based on the joint partial likelihood of all predictors, thus different from marginal screening procedures available in the literature. In order to carry out the new procedure, we propose an effective algorithm and establish its ascent property. We further prove that the proposed procedure possesses the sure screening property. That is, with probability tending to 1, the selected variable set includes the actual active predictors. We conducted simulations to evaluate the finite-sample performance of the proposed procedure and compared it with marginal screening procedures. A genomic data set is used for illustration purposes.

1. Introduction

Feature screening can effectively reduce ultrahigh dimensionality and therefore has attracted considerable attention in the recent literature. Fan and Lv [12] proposed a marginal screening procedure for ultrahigh-dimensional Gaussian linear model, and further showed that marginal screening procedures may possess a sure screening property under certain conditions. Feature screening procedures for varying-coefficient models (VCM) with ultrahigh-dimensional covariates have been proposed in the literature. Liu et al. [21] developed a sure independence screening (SIS) procedure for ultrahigh-dimensional VCM by taking conditional Pearson correlation coefficients as a marginal utility for ranking the importance of predictors. Fan et al. [13] proposed an SIS procedure for ultrahigh-dimensional VCM by extending B-spline techniques in Fan et al. [10] for additive models. Xia et al. [26] further extended the SIS procedure proposed in [13] to generalized varying-coefficient models (GVCM). Cheng et al. [5] proposed a forward variable selection procedure for ultrahigh-dimensional VCM based on techniques related to B-splines regression and grouped variable selection. Song et al. [22] extended the procedure in [13] to longitudinal data without taking into account within-subject correlation, while Chu et al. [6] proposed an SIS procedure for longitudinal data based on a weighted residual sum of squares to use within-subjection correlation to improve accuracy of feature screening. Kong et al. [17] proposed a new screening method that leaves a variable in the active set if it has, jointly with some other variables, a high canonical correlation with the response.

* Corresponding author.
E-mail address: tygr@jnu.edu.cn (G. Yang).

https://doi.org/10.1016/j.jmva.2018.12.009
0047-259X/

Survival analysis has been widely used in medical science, economics, finance, and social science, among others. In many studies, survival data have primary outcomes or responses that are subject to censoring. The Cox model [7,8] is the most commonly used regression model for survival data, and the partial likelihood method has become a standard approach to parameter estimation and statistical inference. Recently, variable selection and parameter estimation in Cox regression models have been considered by various authors (see, e.g., [4,9,14,18,19,30]). Huang et al. [15] studied the penalized partial likelihood with the ℓ_1-penalty for the Cox model with high-dimensional covariates. Yan and Huang [28] proposed the adaptive group Lasso in a Cox regression model with time-varying coefficients. However, they have not considered varying-coefficient models.

In this paper, we propose a new feature screening procedure for ultrahigh-dimensional varying-coefficient Cox models. It is distinguished from SIS procedures [11,32] in that the proposed procedure is based on the joint partial likelihood of potentially important features, rather than the marginal partial likelihood of individual features. Xu and Chen [27] proposed a joint screening procedure and showed its advantage over SIS procedures in the context of generalized linear models. Yang et al. [29] extended the procedures in [27] to the Cox models. This work further extends the joint screening strategy and develops a feature screening procedure for varying-coefficient Cox models, which are natural extensions of Cox models and can be useful to explore nonlinear interaction effects between a primary covariate and other covariates.

The asymptotic properties of the proposed procedure are studied systematically. It is technically challenging to establish its sure screening property. The techniques used in [29] and other works related to SIS procedures cannot be applied for the present setting. We first develop Hoeffding's inequality for a sequence of martingale differences and then establish a concentration inequality for the score function of a partial likelihood. Based on the concentration inequality, we prove the screening property for our proposed sure joint screening procedure. We also conduct simulation studies to assess the finite-sample performance of the proposed procedure and compare its performance with existing sure screening procedures for ultrahigh-dimensional survival data. The proposed methodology is demonstrated through an empirical analysis of a genomic data set.

The rest of this paper is organized as follows. In Section 2, we propose a new feature screening procedure for the varying-coefficient Cox model, develop an algorithm to carry it out, and demonstrate the ascent property of the proposed algorithm. We study the sampling property of the proposed procedure and establish its sure screening property. In Section 3, we present numerical comparisons and an empirical analysis of a real data set. Discussion is in Section 4. Technical proofs are in the Appendix.

2. New feature screening procedure for varying-coefficient Cox model

Let T be the survival time and $\mathbf{x}$ and U be p-dimensional covariate vector and univariate covariate, respectively. Throughout this paper, we consider the varying-coefficient Cox proportional hazard model given by

$$h(t|\mathbf{x}, U) = h_0(t)\exp\{\mathbf{x}^\top \boldsymbol{\alpha}(U)\}, \tag{1}$$

where $h_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{\alpha}(U) = (\alpha_1(U), \ldots, \alpha_p(U))^\top$ consists of the unknown nonparametric coefficient functions. It is assumed that the support of U is finite and denoted by $[a, b]$. In survival data analysis, survival times are subject to a censoring time C. Denote the observed time by $Z = \min(T, C)$ and the event indicator by $\delta = \mathbf{1}(T \le C)$. It is assumed throughout this paper that the censoring mechanism is noninformative. That is, given $\mathbf{x}$ and U, T and C are conditionally independent.

Suppose that $(\mathbf{x}_1, U_1, Z_1, \delta_1), \ldots, (\mathbf{x}_n, U_n, Z_n, \delta_n)$ is a random sample from model (1). Let $t_1^0 < \cdots < t_N^0$ be the ordered observed failure times. Let (j) be the label for the subject failing at time t_j^0, so that the covariates associated with the N failures are $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}$ and $U_{(1)}, \ldots, U_{(N)}$. Denote the risk set right before time t_j^0 by $R_j = \{i : Z_i \ge t_j^0\}$. The partial likelihood function [8] of the random sample is

$$\ell_p\{\boldsymbol{\alpha}(U)\} = \sum_{j=1}^{N}\left[\mathbf{x}_{(j)}^\top \boldsymbol{\alpha}(U_{(j)}) - \ln\left[\sum_{i\in R_j}\exp\{\mathbf{x}_i^\top \boldsymbol{\alpha}(U_i)\}\right]\right]. \tag{2}$$

To estimate the nonparametric regression, we use a B-spline basis. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $\ell \ge 1$ and $\{\psi_{j1}, \ldots, \psi_{jd_{nj}}\}$ denote a normalized B-spline basis with $\|\psi_{jk}\|_\infty \le 1$ and $d_{nj} = O(n^{1/5})$, where $\|\cdot\|_\infty$ is the supremum norm. For any $j \in \{1, \ldots, p\}$ and $\alpha_{nj}(U) \in \mathcal{S}_n$, we have

$$\alpha_{nj}(U) = \sum_{k=1}^{d_{nj}} \beta_{jk}\psi_{jk}(U) = \boldsymbol{\beta}_j^\top \boldsymbol{\psi}_j(U) \tag{3}$$

for some coefficients $\beta_{j1}, \ldots, \beta_{jd_{nj}}$. Here we allow d_{nj} to increase with n and differ for different j because different coefficient functions may have different smoothness. Under some conditions, the nonparametric coefficient functions $\alpha_1(U), \ldots, \alpha_p(U)$ can be well approximated by functions in $\mathcal{S}_n$.

Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$ and $\mathbf{z}_i = (x_{i1}\boldsymbol{\psi}_1(U_i)^\top, \ldots, x_{ip}\boldsymbol{\psi}_p(U_i)^\top)^\top$, and define $\mathbf{z}_{(j)}$ similarly to $\mathbf{x}_{(j)}$. Substituting (3) into (2), the maximum partial likelihood estimate of (2) is to maximize

$$\ell_p(\boldsymbol{\beta}) \hat{=} \sum_{j=1}^{N} \left[\mathbf{z}_{(j)}^\top \boldsymbol{\beta} - \ln \left\{ \sum_{i \in R_j} \exp(\mathbf{z}_i^\top \boldsymbol{\beta}) \right\} \right] \tag{4}$$

with respect to $\boldsymbol{\beta}$. We next propose a feature screening procedure based on (4).

2.1. A new feature screening procedure

Denote $\|\alpha_j(\cdot)\|_2 = \{\mathrm{E}\alpha_j^2(U)\}^{1/2}$, the L_2-norm of $\alpha_j(\cdot)$. For ease of presentation, denote s as an arbitrary subset of $\{1, \ldots, p\}$, $\mathbf{x}_s = \{x_j : j \in s\}$ and $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U) : j \in s\}$. For a set s, $\tau(s)$ stands for the cardinality of s. Suppose the effect of $\mathbf{x}$ is sparse and the true value of $\boldsymbol{\alpha}(U)$ is $\boldsymbol{\alpha}^*(U)$, where $\boldsymbol{\beta}^*$ is the corresponding coefficients of $\boldsymbol{\alpha}^*(U)$. Denote $s^* = \{j : \|\alpha_j(\cdot)\|_2 > 0\}$. By sparsity, we mean that $\tau(s^*)$ is much less than p. The goal of feature screening is to identify a subset s such that $s^* \subset s$ with overwhelming probability and $\tau(s)$ is also much less than p. According to (4), we propose screening features for the varying-coefficient Cox model by the constrained partial likelihood

$$\hat{\boldsymbol{\beta}}_m = \arg\max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}) \quad \text{subject to} \quad \tau(\{j : \|\boldsymbol{\beta}_j\|_2 > 0\}) \le m \tag{5}$$

for a pre-specified m, which is assumed to be greater than the number of nonzero elements of $\boldsymbol{\beta}^*$.

For high-dimensional problems, it becomes almost impossible to solve the constrained maximization problem (5) directly. Alternatively, we consider a proxy of the partial likelihood function. It follows by the Taylor expansion for the partial likelihood function $\ell_p(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$ lying within a neighborhood of $\boldsymbol{\gamma}$ that

$$\ell_p(\boldsymbol{\gamma}) \approx \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p'(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p''(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta})/2,$$

where $\ell_p'(\boldsymbol{\beta}) = \partial \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$ and $\ell_p''(\boldsymbol{\beta}) = \partial^2 \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$. Denote $P_t = d_{n1} + \cdots + d_{np}$. If $\ell_p''(\boldsymbol{\beta})$ is invertible, the computational complexity of calculating the inverse of $\ell_p''(\boldsymbol{\beta})$ is $O(P_t^3)$. For large P_t, small n problems (i.e., $P_t \gg n$), $\ell_p''(\boldsymbol{\beta})$ becomes not invertible. Low computational cost is always desirable for feature screening. To deal with singularity of the Hessian matrix and save computational cost, we propose using the approximation

$$h(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p'(\boldsymbol{\beta}) - u(\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta})/2 \tag{6}$$

for $\ell_p''(\boldsymbol{\gamma})$, where u is a scaling constant to be specified and $W(\boldsymbol{\beta}) = \mathrm{diag}\{W_1(\boldsymbol{\beta}), \ldots, W_p(\boldsymbol{\beta})\}$, a block diagonal matrix with $W_j(\boldsymbol{\beta})$ being a $d_{nj} \times d_{nj}$ matrix. Here (6) is the minimization of the original objective function, $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \le \ell_p(\boldsymbol{\beta})$, for all $\boldsymbol{\gamma}$ under some conditions. Due to the properties of the majorization and minorization algorithm, using (6) we can obtain the same estimates as the original objective function. The two functions themselves, however, are not numerically equal. Here we allow $W(\boldsymbol{\beta})$ to depend on $\boldsymbol{\beta}$. This implies that we approximate $\ell_p''(\boldsymbol{\beta})$ by $-uW(\boldsymbol{\beta})$. Throughout this paper, we will use $W_j(\boldsymbol{\beta}) = -\partial^2 \ell_p(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^\top$.

It can be seen that $h(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta})$, and, under some conditions, $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \le \ell_p(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 for more details. Since $W(\boldsymbol{\beta})$ is a block diagonal matrix, $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of $\boldsymbol{\gamma}_j$ for any given $\boldsymbol{\beta}$. The additivity enables us to have a closed form solution for the maximization problem

$$\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \quad \text{subject to} \quad \tau(\{j : \|\boldsymbol{\gamma}_j\|_2 > 0\}) \le m \tag{7}$$

for given $\boldsymbol{\beta}$ and m. Define $\tilde{\boldsymbol{\gamma}}_j = \boldsymbol{\beta}_j + u^{-1} W_j^{-1}(\boldsymbol{\beta}_j) \partial \ell_p(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_j$ for $j \in \{1, \ldots, p\}$, and $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1^\top, \ldots, \tilde{\boldsymbol{\gamma}}_p^\top)^\top = \boldsymbol{\beta} + u^{-1} W^{-1}(\boldsymbol{\beta}) \ell_p'(\boldsymbol{\beta})$ is the maximizer of $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$. Denote $g_j = \tilde{\boldsymbol{\gamma}}_j^\top W_j(\boldsymbol{\beta}_j) \tilde{\boldsymbol{\gamma}}_j$ for $j \in \{1, \ldots, p\}$, and sort g_j so that $g_{(1)} \ge \cdots \ge g_{(p)}$. The solution of the maximization problem (7) is the hard-thresholding rule defined below:

$$\hat{\boldsymbol{\gamma}}_j = \tilde{\boldsymbol{\gamma}}_j \mathbf{1}\{g_j > g_{(m+1)}\}.$$

This enables us to effectively screen features by using the following algorithm.

Feature Screening Algorithm of Varying Coefficient Cox's Models

Step 1. Set the initial value $\boldsymbol{\beta}_1^{(0)} = \cdots = \boldsymbol{\beta}_p^{(0)} = \mathbf{0}$.

Step 2. For $t \in \{0, 1, \ldots\}$, iteratively conduct Step 2a and Step 2b below until the algorithm converges:

Step 2a. Calculate $\tilde{\boldsymbol{\gamma}}_j^{(t)} = \boldsymbol{\beta}_j^{(t)} + u_t^{-1} W_j^{-1}(\boldsymbol{\beta}_j) \partial \ell(\boldsymbol{\beta}^{(t)})/\partial \boldsymbol{\beta}_j$, and $g_j^{(t)} = \{\tilde{\boldsymbol{\gamma}}_j^{(t)}\}^\top W_j(\boldsymbol{\beta}^{(t)}) \tilde{\boldsymbol{\gamma}}_j^{(t)}$. Let $g_{(1)}^{(t)} \ge \cdots \ge g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : g_j^{(t)} \ge g_{(m+1)}^{(t)}\}$, the nonzero index set.

Step 2b. Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{\beta}_1^{(t+1)}, \ldots, \boldsymbol{\beta}_p^{(t+1)})^\top$ as follows. If $j \notin S_t$, set $\boldsymbol{\beta}_j^{(t+1)} = \mathbf{0}$, otherwise, set $\{\boldsymbol{\beta}_j^{(t+1)} : j \in S_t\}$ be the partial likelihood estimate of the submodel S_t.

Theorem 1. *Suppose that Conditions (D1)–(D4) in the Appendix hold. Let $\boldsymbol{\beta}^{(t)}$ be the sequence defined in Step 2b in the above algorithm. Denote*

$$\rho^{(t)} = \sup_{\boldsymbol{\beta}} \left[\lambda_{\max}\{W^{-1/2}(\boldsymbol{\beta}^{(t)})\{-\ell_p''(\boldsymbol{\beta})\}W^{-1/2}(\boldsymbol{\beta}^{(t)})\}\right],$$

where $\lambda_{\max}(A)$ stands for the maximal eigenvalue of a matrix A. If $u_t \geq \rho^{(t)}$, then $\ell_p(\boldsymbol{\beta}^{(t+1)}) \geq \ell_p(\boldsymbol{\beta}^{(t)})$, where $\boldsymbol{\beta}^{(t+1)}$ is defined in Step 2b in the above algorithm.

Theorem 1 claims the ascent property of the proposed algorithm if u_t is appropriately chosen. That is, the proposed algorithm may improve the current estimate within the feasible region (i.e., $\tau(\{j : \|\alpha_j(U)\|_2 > 0\}) \leq m$), and the resulting estimate in the current step may serve as a refinement of the last step. This theorem provides us with some insights into how to choose u_t in practical implementation.

2.2. Sure screening property

For a subset s of $\{1, \ldots, p\}$ with size $\tau(s)$, recall the notation $\mathbf{x}_s = \{x_j : j \in s\}$ and associated coefficients $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U) : j \in s\}$ corresponding to $\boldsymbol{\beta}_s = \{\boldsymbol{\beta}_j : j \in s\}$ with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_{n_j}})^\top$. We denote the true model by $s^* = \{j : \mathrm{E}\alpha_j^2(U) > 0, 1 \leq j \leq p\}$ with $\tau(s^*) = q$. The objective of feature screening is to obtain a subset $\hat{s}$ such that $s^* \subset \hat{s}$ with very high probability.

We now provide some theoretical justifications for the proposed screening procedure for the ultrahigh-dimensional varying-coefficient Cox model. The sure screening property [12] is referred to as

$$\lim_{n\to\infty} \Pr(s^* \subset \hat{s}) = 1. \tag{8}$$

To establish this sure screening property for the proposed screening procedure, we introduce some additional notation as follows. For any model s, let $\ell'(\boldsymbol{\beta}_s) = \partial\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s$ and $\ell''(\boldsymbol{\beta}_s) = \partial^2\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^\top$ be the score function and the Hessian matrix of ℓ as a function of $\boldsymbol{\beta}_s$, respectively. Assume that a screening procedure retains m out of p features such that $\tau(s^*) = q < m$. So, we define

$$S_+^m = \{s : s^* \subset s, \|s\|_0 \leq m\} \quad \text{and} \quad S_-^m = \{s : s^* \not\subset s, \|s\|_0 \leq m\}$$

as the collections of the over-fitted models and the under-fitted models, respectively. We investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_m$ under the scenario where p, q, m and $\boldsymbol{\beta}^*$ are allowed to depend on the sample size n. We impose the following conditions, some of which are purely technical and merely serve to facilitate theoretical understanding of the proposed feature screening procedure. For ease of presentation and without loss of generality, it is assumed that $d_{n1} = \cdots = d_{np} \hat{=} d_n$.

(C1) The support of U is bounded on $[a, b]$.

(C2) The functions $\alpha_1(U), \ldots, \alpha_p(U)$ belong to a class of functions $\mathcal{F}$, whose rth derivative $\alpha_j^{(r)}$ exists and is Lipschitz of order η,

$$\mathcal{F} = \{\alpha_j : |\alpha_j^{(r)}(s) - \alpha_j^{(r)}(t)| \leq K|s - t|^\eta \text{ for } s, t \in [a, b]\},$$

for some positive constant K, where r is a nonnegative integer and $\eta \in (0, 1]$ such that $\nu = r + \eta > 0.5$.

(C3) There exist $w_1, w_2 > 0$ and some non-negative constants τ_1, τ_2 such that $\tau_1 + \tau_2 < 1/2$ and

$$\min_{j\in s^*} \|\alpha_j(U)\|_2 \geq w_1 n^{-\tau_1} \quad \text{and} \quad q < m \leq w_2 n^{\tau_2}.$$

(C4) $\ln p = O(n^\kappa)$ for some $0 \leq \kappa < 1 - 2(\tau_1 + \tau_2)$.

(C5) There exist constants $C_1, C_2 > 0$, $\delta > 0$, such that for sufficiently large n,

$$C_1 d_n^{-1} \leq \lambda_{\min}\{-n^{-1}\ell_p''(\boldsymbol{\beta}_s)\} \leq \lambda_{\max}\{-n^{-1}\ell_p''(\boldsymbol{\beta}_s)\} \leq C_2 d_n^{-1},$$

for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \leq \delta\}$ and $s \in S_+^{2m}$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of a matrix, respectively.

Under Conditions (C1)–(C2), the following two properties of B-splines are valid.

(a) de Boor [3]: For $k \in \{1, \ldots, d_n\}$, $\psi_{jk}(U) \geq 0$ and $\psi_{j1}(U) + \cdots + \psi_{jd_n}(U) = 1$, $U \in [a, b]$. In addition, there exist positive constants C_3 and C_4 such that $C_3 d_n^{-1} \leq \mathrm{E}\psi_{jk}^2(U) \leq C_4 d_n^{-1}$.

(b) Stone [23,24]: If $\{\alpha_1, \ldots, \alpha_p\}$ is a set of functions in $\mathcal{F}$ described in Condition (C2), there exists a positive constant C_5 that does not depend on $\alpha_j(U)$; then the uniform approximation error satisfies $\rho = \sup_{U\in[a,b]} \|\alpha_j(U) - \alpha_{nj}(U)\|_2 \leq C_5 d_n^{-\nu}$ for all $j \in \{1, \ldots, p\}$, as $d_n \to \infty$.

Conditions (C1)–(C2) ensure properties (a) and (b), which are required for the B-spline approximation and establishing the sure screening properties. Note that $\|\alpha_{nj}(U)\|_2^2 = \boldsymbol{\beta}_j^\top \mathrm{E}\{\boldsymbol{\psi}_j(U)\boldsymbol{\psi}_j(U)^\top\}\boldsymbol{\beta}_j$. Based on properties (a) and (b) and Condition (C3), we can derive that

$$\min_{j\in s^*} \|\boldsymbol{\beta}_j\|_2 \geq w_1 d_n n^{-\tau_1}.$$

Table 1
Censoring rates.

	$\rho = 0.25$			$\rho = 0.5$			$\rho = 0.75$		
Σ	(a1)	(a2)	(a3)	(a1)	(a2)	(a3)	(a1)	(a2)	(a3)
S1	.276	.367	.223	.277	.356	.260	.277	.340	.248
S2	.275	.365	.265	.279	.358	.283	.278	.347	.245

Condition (C3) states a few requirements for establishing the sure screening property of the proposed procedure. The first one is the sparsity of $\boldsymbol{\alpha}^*(U)$, which makes the sure screening possible with $\tau(\hat{s}) = m > q$. Also, it requires that the minimal component in $\boldsymbol{\alpha}^*(U)$ does not degenerate too quickly, so that the signal is detectable in the asymptotic sequence. Meanwhile, together with (C4), it confines an appropriate order of m that guarantees the identifiability of s^* over s for $\tau(s) \leq m$. Condition (C5) assumes that p diverges from n at up to an exponential rate; it implies that the number of covariates can be substantially larger than the sample size.

We establish the sure screening property of the quasi-likelihood estimation in the following theorem.

Theorem 2. *Suppose that Conditions (C1)–(C5) and Conditions (D1)–(D7) in the Appendix hold. Let* $\hat{s}$ *be the model obtained by Eq. (5) of size m. We have* $\Pr(s^* \subset \hat{s}) \to 1$ *as* $n \to \infty$.

The proof is given in the Appendix. The sure screening property is an appealing property of a screening procedure because it ensures that the true active predictors are retained in the model selected by the screening procedure. To be distinguished from the SIS procedure, the proposed procedure is referred to as a sure joint screening (SJS) procedure.

3. Numerical studies

In this section, we assess the finite-sample performance of the proposed procedure, compare it with existing procedures via simulation, and illustrate the proposed procedure by an empirical analysis of a genomic data set.

3.1. Simulation studies

The main purpose of our simulation studies is to assess the performance of the proposed procedure by comparing it with the SIS [11] and the SJS [29] procedures for the Cox model. The model sizes selected by the three methods are set to be the same for comparison. We vary the dimension of predictors p, sample size n and sample correlation ρ to examine their impact on the performance of the proposed procedure. We use the success rate of active predictors being selected and computing time as our criteria to compare the performance of screening procedures.

In our simulation, the predictors $\mathbf{x}$ are generated from a p-dimensional normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$. Two commonly used covariance structures are used in our simulation:

(S1) Σ is compound symmetric, (i.e., $\sigma_{ij} = \rho$ for $i \neq j$ and equal 1 for $i = j$). We choose $\rho \in \{0.25, 0.5, 0.75\}$.
(S2) Σ has autoregressive structure with AR(1), (i.e., $\sigma_{ij} = \rho^{|i-j|}$). We choose $\rho \in \{0.25, 0.5, 0.75\}$.

We generate the survival time from the Cox model with $h_0(t) = 1$ and the censoring time from a uniform distribution $\mathcal{U}[0, 10]$. Three different coefficient function settings $\boldsymbol{\alpha}(u)$s are considered:

(a1): $\alpha_1^{(1)}(u) = 1 + 2\sin(2\pi u), \alpha_2^{(1)}(u) = 1 - 2\cos(2\pi u), \alpha_3^{(1)}(u) = 0.5 + 2u^2$;
(a2): $\alpha_1^{(2)}(u) = 5\sin(2\pi u), \alpha_2^{(2)}(u) = 5\cos(2\pi u), \alpha_3^{(2)}(u) = 2.5 + 5u^2$;
(a3): $\alpha_1^{(3)}(u) = e^{0.5u}, \alpha_2^{(3)}(u) = 2(u^3 + 1.5(u - 0.5)^2), \alpha_3^{(3)}(u) = 2u$.

We consider $n \in \{200, 400\}$, and $p \in \{2000, 5000\}$. For the feature screening model size, we follow Liu et al. [21] and set $m = \lfloor n^{0.8}/\ln(n^{0.8}) \rfloor$, where $\lfloor a \rfloor$ denotes the integer part of a. For each combination of different inputs, we conduct 1000 repetitions.

To illustrate the performance of a statistical procedure in survival data analysis, we want the censoring rates to lie within a reasonable range. Table 1 depicts the censoring rates for the 18 combinations of covariance structure, sample correlation ρ and the values of $\boldsymbol{\alpha}(u)$. The censoring rates range from 22% to 37%, which is reasonable for simulation studies.

We compare the performance of feature screening procedures using the following two criteria: P_s, the probability that an individual active predictor is selected, and P_a, the probability that all active predictors are selected. It is expected that the performance of the proposed varying-coefficient SJS (VSJS) procedure depends on the following factors: the structure of the covariance matrix, the values of $\boldsymbol{\alpha}(u)$, the dimension of all candidate features p, the sample correlation ρ and the sample size n.

Tables 2–3 report P_s and P_a of VSJS, SIS and SJS for the active predictors under (S1). Overall, VSJS outperforms both SIS and SJS for all the three sets of $\boldsymbol{\alpha}(u)$ in terms of P_s and P_a. For (a1), VSJS achieves a high success rate in detecting signals of $\alpha_1^{(1)}$ and $\alpha_2^{(1)}$, while SIS and SJS fail from time to time.

Table 2
Comparison between VSJS, SIS and SJS with $\Sigma = (1-\rho)I + \rho 11^\top$ ($n = 200$).

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n = 200, p = 2000$ and $\rho = .25$														
$\alpha^{(1)}$	.989	1	1	.989	74.5	.796	.747	.990	.580	9.5	.499	.419	.936	.190	3.6
$\alpha^{(2)}$	.999	.998	.999	.996	67.7	.016	.002	1	0	8.3	.018	.037	.999	.002	2.4
$\alpha^{(3)}$	1	.810	.993	.803	82.2	1	.771	.992	.763	6.0	1	.785	.996	.781	2.8
	$n = 200, p = 2000$ and $\rho = .5$														
$\alpha^{(1)}$	.970	.976	.915	.868	68.9	.621	.557	.968	.325	9.2	.392	.311	.863	.092	2.9
$\alpha^{(2)}$	.922	.922	.990	.848	66.8	.006	.003	1	0	7.8	.020	.052	.997	0	2.5
$\alpha^{(3)}$	.998	.617	.938	.581	74.8	.999	.611	.932	.573	5.3	1	.574	.932	.542	3.2
	$n = 200, p = 2000$ and $\rho = .75$														
$\alpha^{(1)}$	.628	.670	.682	.259	62.4	.357	.316	.879	.093	9.4	.247	.211	.701	.031	3.0
$\alpha^{(2)}$	.485	.535	.738	.204	67.3	.005	.001	1	0	6.8	.018	.059	.935	0	3.4
$\alpha^{(3)}$	.910	.361	.686	.247	62.5	.987	.341	.736	.250	5.3	.958	.286	.644	.181	3.4
	$n = 200, p = 5000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	.993	.993	464.0	.721	.649	.983	.456	15.4	.391	.326	.865	.097	32.9
$\alpha^{(2)}$	.996	.994	1	.990	416.3	.004	.004	1	0	18.1	.007	.016	.994	0	17.6
$\alpha^{(3)}$	1	.708	.984	.694	451.5	1	.684	.974	.667	15.2	1	.627	.980	.615	16.8
	$n = 200, p = 5000$ and $\rho = .5$														
$\alpha^{(1)}$	.925	.930	.845	.725	412.7	.496	.430	.954	.199	22.9	.281	.224	.779	.040	16.8
$\alpha^{(2)}$	.856	.876	.976	.740	423.7	.005	.002	1	0	16.1	.007	.030	.968	0	18.9
$\alpha^{(3)}$	.992	.508	.884	.446	390.4	.999	.455	.866	.38	15.2	.998	.435	.878	.383	24.0
	$n = 200, p = 5000$ and $\rho = .75$														
$\alpha^{(1)}$	.510	.501	.504	.121	398.1	.261	.218	.803	.042	15.3	.135	.140	.541	.010	20.3
$\alpha^{(2)}$	.372	.399	.625	.093	396.6	.002	0	.999	0	14.9	.006	.022	.867	0	22.2
$\alpha^{(3)}$	.892	.276	.597	.158	369.5	.977	.258	.624	.159	13.3	.909	.164	.493	.075	24.7

We next consider the performance of VSJS under (a2). For the zero-centered $\alpha_1^{(2)}$ and $\alpha_2^{(2)}$, VSJS successfully detects their variation signal and achieves high success rates. As a comparison, SIS and SJS fail to identify $\alpha_1^{(2)}$ and $\alpha_2^{(2)}$ as active predictors completely in (a2). In general, VSJS still performs better to some extent in (a3), though SIS slightly outperforms VSJS in a few cases.

Tables 2–3 clearly show how performance is affected by sample correlation ρ, predictor dimension p, and sample size n. When ρ increases, n decreases, or p increases, the three methods perform worse under (S1). Compared to SIS and SJS, the performance of VSJS is more resistant to these changes. Also, Tables 2–3 suggest that VSJS is more computationally inefficient than SIS and SJS.

Tables 4–5 report P_s and P_a of VSJS, SIS, and SJS for the active predictors under (S2). Overall, VSJS still outperforms SIS and SJS. It is worth noting that the three methods have much better performance under (S2) than previous cases under (S1), especially when the correlation ρ is larger. In (a1), VSJS and SIS both perform perfectly and slightly better than SJS. When we consider (a2), SIS and SJS perform better under (S2) and successfully identify $\alpha_1^{(2)}$ and $\alpha_2^{(2)}$ from time to time. However, VSJS again outperforms them in (a2). For (a3), the three methods achieve almost 100% success rate for selecting active predictors. SJS misses some active predictors in a few cases.

We can conclude from Tables 4 and 5 that SIS and SJS tend to perform better when ρ increases, n increases, or p decreases. For VSJS, it performs perfectly in all three settings under (S1). Similarly, Tables 4 and 5 show that VSJS is more computationally intensive than SIS and SJS.

3.2. Real data analysis

We analyze The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) data on liver hepatocellular carcinoma to illustrate the proposed procedure. Liver hepatocellular carcinoma is the most common form of liver cancer and the third cancer death cause worldwide. Zhang and Sun [31] studied 17,255 patients in the Surveillance, Epidemiology, and End Results Program (SEER ;https://seer.cancer.gov/) cancer registry and suggested that age is a prognostic factor for liver cancer. Therefore, we consider age as the univariate covariate for coefficient functions, allowing the effects of gene expression on survival time to vary with age. After removing five subjects whose survival time is zero, we obtain 354 subjects with gene expressions (IlluminaHiSeq RNA-seq v2 platform), age at diagnosis, and survival months. We apply a log 2 transformation to gene expressions and analyze 14,683 genes that have more than 90% nonzero observations.

For VSJS, we use a linear combination of five B-spline basis functions to approximate the varying-coefficient functions. As a result, VSJS retains $23 = \lfloor 354^{0.8}/\ln(354^{0.8}) \rfloor$ genes and the partial likelihood function value for the corresponding model is

Table 3
Comparison between VSJS, SIS and SJS with $\Sigma = (1-\rho)I + \rho 11^\top$ ($n = 400$).

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n = 400$, $p = 2000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	217.7	1	.960	1	.960	8.8	.859	.805	.999	.686	5.8
$\alpha^{(2)}$	1	1	1	1	205.9	.020	.001	1	0	7.9	.010	.076	1	0	5.6
$\alpha^{(3)}$	1	1	1	1	215.3	1	.974	1	.974	8.3	1	.997	1	.997	4.9
	$n = 400$, $p = 2000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	190.2	.900	.871	.999	.779	8.5	.736	.607	.998	.437	4.6
$\alpha^{(2)}$	1	1	1	1	184.3	.010	.001	1	0	8.5	.023	.133	1	.002	6.3
$\alpha^{(3)}$	1	.988	1	.988	199.5	1	.918	.997	.916	8.2	1	.944	1	.944	5.1
	$n = 400$, $p = 2000$ and $\rho = .75$														
$\alpha^{(1)}$	.984	.991	.976	.955	169.0	.655	.566	.997	.349	8.6	.474	.356	.955	.155	6.3
$\alpha^{(2)}$	.998	.995	1	.994	162.2	.001	0	1	0	9.5	.035	.193	.999	.004	6.6
$\alpha^{(3)}$	1	.733	.982	.719	162.8	1	.676	.968	.657	8.2	1	.576	.938	.540	6.1
	$n = 400$, $p = 5000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	1202	.963	.957	1	.920	21.6	.963	.957	1	.920	21.6
$\alpha^{(2)}$	1	1	1	1	1164	.006	.001	1	0	20.6	.004	.038	1	.001	31.1
$\alpha^{(3)}$	1	1	1	1	1180	1	.960	1	.960	18.2	1	.993	1	.993	36.5
	$n = 400$, $p = 5000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	1086	.849	.798	.999	.669	21.0	.849	.798	.999	.669	21.1
$\alpha^{(2)}$	1	1	1	1	1101	.001	0	1	0	22.2	.011	.071	1	.002	32.1
$\alpha^{(3)}$	1	.975	1	.975	1071	1	.840	.998	.838	19.6	1	.872	1	.872	40.3
	$n = 400$, $p = 5000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	.980	.980	929.0	.562	.426	.994	.224	21.0	.336	.267	.933	.073	35.9
$\alpha^{(2)}$	.994	.992	.997	.988	936.7	.001	0	1	0	20.8	.016	.109	1	.001	35.3
$\alpha^{(3)}$	.995	.621	.926	.586	909.6	1	.580	.935	.535	18.3	.999	.446	.900	.401	46.1

−544.9. With the same number of genes retained, the resulting partial likelihood function values for SIS and SJS are −589.2 and −588.4, respectively. Simultaneous modeling of the 23 retained genes shows a clear advantage of VSJS in terms of higher partial likelihood value.

To better understand the screening result of VSJS, we apply the backward selection procedure to those 23 genes and obtain a more parsimonious model. Specifically, each backward elimination step removes a gene with the smallest likelihood ratio test statistic until all the genes are significant at level 0.05. Table 6 provides the final list of 11 genes after applying the backward elimination, and Fig. 1 depicts their varying coefficients.

Our literature search reveals that those 11 genes are all associated with cancer risk and some genes. For example, GTPBP4 [20] and SLC2A2 [16] are promising prognostic factors for hepatocellular carcinoma. To test whether those 11 genes have varying coefficients versus constant coefficients, a test of $\mathcal{H}_0$: $\alpha_j(u) = \alpha_j$ for some constant α_j versus $\mathcal{H}_1$: $\alpha_j(u) \not\equiv \alpha_j$ can be conducted for each j in the selected gene set. The test result is shown in Table 7, and all the genes except DYNC1LI1 have significant varying-coefficient functions of age at the 5% level of significance. There is no evidence of their time-varying effects in the current medical literature, but our study may suggest some evidence for potential granular investigation on those genes.

4. Discussion

We have proposed an SJS procedure for the varying-coefficient Cox model with ultrahigh-dimensional covariates based on partial likelihood. The proposed SJS is distinguished from the existing SIS procedure in that the proposed procedure is based on the joint likelihood of potential candidate features. We also proposed an effective algorithm to carry out the feature screening procedures and show that the proposed algorithm possesses an ascent property. We studied the sampling property of SJS and established the sure screening property for SJS.

Theorem 1 ensures the ascent property of the proposed algorithm under certain conditions, but it does not imply that the proposed algorithm converges to the global optimizer. If the proposed algorithm converges to a global maximizer of (5), then Theorem 2 shows that such a solution enjoys the sure screen property.

Acknowledgments

Yang's research was supported by the National Nature Science Foundation of China (NNSFC) grants 11471086 and 11871173, the National Social Science Foundation of China (NSSFC) grant 16BTJ032, the National Statistical Scientific Center

Table 4
Comparison between VSJS, SIS and SJS with $\Sigma = (\rho^{|i-j|})$ ($n = 200$).

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n = 200, p = 2000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	76.9	1	1	.988	.988	5.2	.856	.809	.997	.684	2.6
$\alpha^{(2)}$	1	1	1	1	70.8	.042	.116	1	.008	5.9	.047	.027	1	0	2.4
$\alpha^{(3)}$	1	1	1	1	86.6	1	1	1	1	7.1	1	.981	1	.981	3.0
	$n = 200, p = 2000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	73.1	1	1	.999	.999	8.2	.889	.792	.990	.690	2.5
$\alpha^{(2)}$	1	1	1	1	67.6	.166	.611	1	.145	5.8	.052	.065	1	.011	2.4
$\alpha^{(3)}$	1	1	1	1	82.8	1	1	1	1	7.7	1	.977	1	.977	3.1
	$n = 200, p = 2000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	75.5	1	1	1	1	5.2	.877	.768	.990	.642	3.0
$\alpha^{(2)}$	1	1	1	1	68.6	.722	.968	1	.720	5.8	.125	.417	.997	.076	2.6
$\alpha^{(3)}$	1	.997	1	.997	79.4	1	1	1	1	8.4	1	.926	.991	.917	3.1
	$n = 200, p = 5000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	456.4	.968	.997	1	.965	15.4	.785	.734	.989	.559	16.1
$\alpha^{(2)}$	1	1	1	1	463.8	.016	.067	1	.004	14.6	.016	.022	.999	0	14.9
$\alpha^{(3)}$	1	1	.998	.998	477.1	1	.999	1	.999	16.2	1	.967	1	.967	20.1
	$n = 200, p = 5000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	451.1	1	1	1	1	13.1	.799	.730	.979	.543	13.2
$\alpha^{(2)}$	1	1	1	1	439.9	.121	.501	1	.103	14.3	.030	.025	1	.003	16.0
$\alpha^{(3)}$	1	1	1	1	475.4	1	1	1	1	15.8	1	.966	.997	.963	20.3
	$n = 200, p = 5000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	448.2	1	1	1	1	15.4	.844	.685	.987	.538	19.0
$\alpha^{(2)}$	1	1	1	1	427.3	.627	.938	1	.626	14.8	.062	.327	1	.040	15.9
$\alpha^{(3)}$	1	.996	1	.996	453.9	1	1	1	1	14.4	1	.916	.980	.896	23.3

grant 2015LD02, China; and the Fundamental Research Funds for the Central Universities of Jinan University Qimingxing Plan 15JNQM019, China. Zhang and Li's research was supported by National Institute on Drug Abuse (NIDA) grant P50 DA039838, USA; National Science Foundation (NSF) grant DMS 1820702, USA; and NNSFC grants 11690014 and 11690015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA, the NNSFC, the NSF, or the NSSFC.

Appendix

We use the following notation to present the regularity conditions for the partial likelihood and the Cox model. Most notations are adapted from Andersen and Gill [1], in which counting processes were introduced for the Cox model and the consistency and asymptotic normality of the partial likelihood estimate were established. Denote $\overline{N}_i(t) = \mathbf{1}(T_i \le t, T_i \le C_i)$ and $R_i(t) = \{T_i \ge t, C_i \ge t\}$. Assume that there are no two component processes $N_i(t)$ jumping at the same time. For simplicity, we work on the finite interval $[0, \tau]$.

In Cox's model, properties of stochastic processes, such as being a local martingale or a predictable process, are relative to a right-continuous nondecreasing family $\{\mathcal{F}_t : t \in [0, \tau]\}$ of sub σ-algebras on a sample space $(\Omega, \mathcal{F}, \mathcal{P})$; $\mathcal{F}_t$ represents everything that happens up to time t. Throughout this section, we define $\Lambda_0(t) = \int_0^t h_0(u)\,du$.

By stating that $\overline{N}_i(t)$ has intensity process $h_i(t) \hat{=} h(t|\mathbf{x}_i, U_i)$, we mean that the processes $M_i(t)$ defined, for each $i \in \{1, \ldots, n\}$, by

$$M_i(t) = \overline{N}_i(t) - \int_0^{\top} h_i(u)du,$$

are local martingales on the time interval $[0, \tau]$. For $k \in \{0, 1, 2\}$, define

$$\mathbf{S}^{(k)}(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^{n} R_i(t)\exp(\mathbf{z}_i^{\top}\boldsymbol{\beta})\mathbf{z}_i^{\otimes k}, \quad \mathbf{s}^{(k)}(\boldsymbol{\beta}, t) = \mathrm{E}\{\mathbf{S}^{(k)}(\boldsymbol{\beta}, t)\}$$

and

$$\mathrm{E}(\boldsymbol{\beta}, t) = \mathbf{S}^{(1)}(\boldsymbol{\beta}, t)/\mathbf{S}^{(0)}(\boldsymbol{\beta}, t), \quad \mathrm{V}(\boldsymbol{\beta}, t) = \mathbf{S}^{(2)}(\boldsymbol{\beta}, t)/\mathbf{S}^{(0)}(\boldsymbol{\beta}, t) - \mathrm{E}(\boldsymbol{\beta}, t)^{\otimes 2},$$

Table 5
Comparison between VSJS, SIS and SJS with $\Sigma = (\rho^{|i-j|})$ $(n = 400)$.

	VSJS					SIS					SJS				
	P_s			P_a	Time	P_s			P_a	Time	P_s			P_a	Time
$\alpha(U)$	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)	X_1	X_2	X_3	all	(s)
	$n = 400, p = 2000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	229.6	1	1	1	1	8.6	.991	.979	1	.970	6.3
$\alpha^{(2)}$	1	1	1	1	223.3	.083	.251	1	.036	8.5	.047	.040	1	.001	5.2
$\alpha^{(3)}$	1	1	1	1	240.1	1	1	1	1	11.9	1	1	1	1	7.0
	$n = 400, p = 2000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	225.9	1	1	1	1	7.5	.992	.959	1	.951	5.3
$\alpha^{(2)}$	1	1	1	1	226.1	.387	.922	1	.382	8.8	.070	.263	1	.031	5.2
$\alpha^{(3)}$	1	1	1	1	236.8	1	1	1	1	8.5	1	1	1	1	7.3
	$n = 400, p = 2000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	217.9	1	1	1	1	8.9	.979	.907	1	.886	6.4
$\alpha^{(2)}$	1	1	1	1	218.4	.969	1	1	.969	9.1	.139	.598	1	.080	5.8
$\alpha^{(3)}$	1	.999	1	.999	227.8	1	1	1	1	11.9	1	.997	1	.997	7.6
	$n = 400, p = 5000$ and $\rho = .25$														
$\alpha^{(1)}$	1	1	1	1	1264	1	1	1	1	20.6	.988	.962	1	.952	29.5
$\alpha^{(2)}$	1	1	1	1	1265	.054	.183	1	.018	18.7	.029	.032	1	0	28.8
$\alpha^{(3)}$	1	1	1	1	1215	1	1	1	1	20.8	1	1	1	1	33.8
	$n = 400, p = 5000$ and $\rho = .5$														
$\alpha^{(1)}$	1	1	1	1	1274	1	1	1	1	20.5	.976	.924	1	.900	32.5
$\alpha^{(2)}$	1	1	1	1	1256	.318	.884	1	.312	19.9	.038	.162	1	.017	29.1
$\alpha^{(3)}$	1	1	1	1	1194	1	1	1	1	20.6	1	.999	1	.999	35.6
	$n = 400, p = 5000$ and $\rho = .75$														
$\alpha^{(1)}$	1	1	1	1	1202	1	1	1	1	20.7	.969	.902	1	.871	36.9
$\alpha^{(2)}$	1	1	1	1	1225	.954	1	1	.954	21.9	.085	.548	1	.051	29.9
$\alpha^{(3)}$	1	1	1	1	1139	1	1	1	1	29.5	1	.995	1	.995	34.6

Table 6
Genes selected by backward elimination.

Gene Name	ANLN	CEP55	DYNC1LI1	GTPBP4
LRT Stat	15.869	14.137	18.171	22.658
p-value	0.00723	0.0148	0.00274	< 0.001
Gene Name	SLC2A1	KIF2C	KIF20A	KPNA2
LRT Stat	18.465	26.261	15.839	14.511
p-value	0.00241	< 0.001	0.00731	0.0127
Gene Name	LIMS2	TRIP13	UCK2	
LRT Stat	23.093	17.517	14.671	
p-value	< 0.001	0.00361	0.0119	

Table 7
LRT statistics and p-values for the varying coefficients of the final selected genes.

Gene Name	ANLN	CEP55	DYNC1LI1	GTPBP4
LRT Stat	15.058	10.495	8.268	19.036
p-value	0.00458	0.0328	0.0822	0.000773
Gene Name	SLC2A1	KIF2C	KIF20A	KPNA2
LRT Stat	17.473	24.253	15.183	14.238
p-value	0.00156	0.000071	0.00433	0.00657
Gene Name	LIMS2	TRIP13	UCK2	
LRT Stat	23.097	16.191	13.803	
p-value	0.000121	0.00277	0.00795	

where $\mathbf{z}_i^{\otimes 0} = 1$, $\mathbf{z}_i^{\otimes 1} = \mathbf{z}_i$ and $\mathbf{z}_i^{\otimes 2} = \mathbf{z}_i\mathbf{z}_i^\top$. Note that $\mathbf{S}^{(0)}(\boldsymbol{\beta}, t)$ is a scalar, $\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)$ and $\mathrm{E}(\boldsymbol{\beta}, t)$ are p-vector, and $\mathbf{S}^{(2)}(\boldsymbol{\beta}, t)$ and $\mathrm{V}(\boldsymbol{\beta}, t)$ are $p \times p$ matrices. Define

$$Q_j = \sum_{i=1}^{n} \int_0^{t_j} \left\{ \mathbf{z}_i - \sum_{i \in R_j} \mathbf{z}_i \exp(\mathbf{z}_i^\top \boldsymbol{\beta}) \Big/ \sum_{i \in R_j} \exp(\mathbf{z}_i^\top \boldsymbol{\beta}) \right\} dM_i.$$

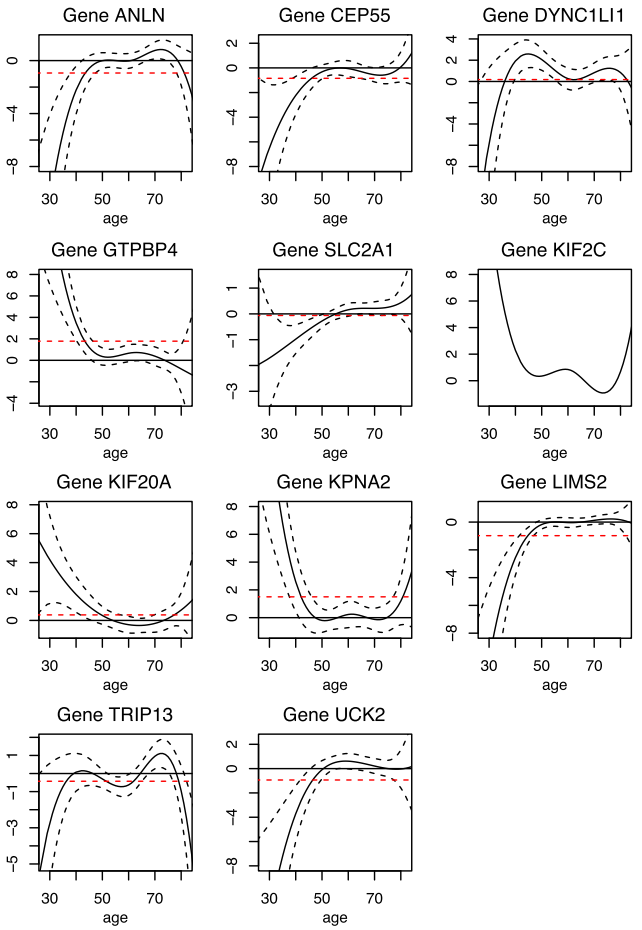

Fig. 1. Estimated coefficient functions and the pointwise conference intervals of selected genes. The red line represents the average level of the varying-coefficient functions.

Here, $\mathrm{E}(Q_j|\mathcal{F}_{j-1}) = Q_{j-1}$, i.e., $\mathrm{E}(Q_j - Q_{j-1}|\mathcal{F}_{j-1}) = 0$. Let $b_j = Q_j - Q_{j-1}$, then $b_1, b_2, \ldots$ is a sequence of bounded martingale differences on $(\Omega, \mathcal{F}, P)$. That is, b_j is bounded almost surely and $\mathrm{E}(b_j|\mathcal{F}_{j-1}) = 0$ as for $j \in \{1, 2, \ldots\}$.

(D1) Finite interval: $\Lambda_0(\tau) = \int_0^\tau h_0(t)dt < \infty$.

(D2) Asymptotic stability: There exist a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$ and scalar, vector and matrix functions $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ defined on $\mathcal{B} \times [0, \tau]$ such that for $k \in \{0, 1, 2\}$,

$$\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B}} \|\mathbf{S}^{(k)}(\boldsymbol{\beta}, t) - \mathbf{s}^{(k)}(\boldsymbol{\beta}, t)\| \xrightarrow{p} 0.$$

(D3) Lindeberg condition: There exists $\delta > 0$ such that

$$n^{-1/2} \sup_{i,t} |\mathbf{z}_i| R_i(t) \mathbf{1}\{\boldsymbol{\beta}_0^\top \mathbf{z}_i > -\delta|\mathbf{z}_i|\} \xrightarrow{p} 0,$$

(D4) Asymptotic regularity conditions: Let $\mathcal{B}$, $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ be as in Condition (D2) and define $e = \mathbf{s}^{(1)}/\mathbf{s}^{(0)}$ and $v = \mathbf{s}^{(2)}/\mathbf{s}^{(0)} - e^{\otimes 2}$. For all $\boldsymbol{\beta} \in \mathcal{B}, t \in [0, \tau]$,

$$\mathbf{s}^{(1)}(\boldsymbol{\beta}, t) = \partial \mathbf{s}^{(0)}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}, \quad \mathbf{s}^{(2)}(\boldsymbol{\beta}, t) = \partial^2 \mathbf{s}^{(0)}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}^2,$$

$\mathbf{s}^{(0)}(\cdot, t)$, $\mathbf{s}^{(1)}(\cdot, t)$ and $\mathbf{s}^{(2)}(\cdot, t)$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0, \tau]$, $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ are bounded on $\mathcal{B} \times [0, \tau]$; $\mathbf{s}^{(0)}$ is bounded away from zero on $\mathcal{B} \times [0, \tau]$, and the matrix

$$\mathbf{S} = \int_0^\tau v(\boldsymbol{\beta}_0, t)\mathbf{s}^{(0)}(\boldsymbol{\beta}_0, t)h_0(t)dt$$

is positive definite.

(D5) The function $\mathbf{S}^{(0)}(\boldsymbol{\beta}^*, t)$ and $\mathbf{s}^{(0)}(\boldsymbol{\beta}^*, t)$ are bounded away from 0 on $[0, \tau]$.
(D6) There exist constants $C_1, C_2 > 0$, such that $\max_{ij} |z_{ij}| < C_1$ and $\max_i |\mathbf{z}_i^\top \boldsymbol{\beta}^*| < C_2$.
(D7) $b_1, b_2, \ldots$ is a sequence of martingale differences and there exist nonnegative constants $c_1, \ldots, c_N$ such that for every real number t and all $j \in \{1, \ldots, N\}$, $\mathrm{E}\{\exp(tb_j)|\mathcal{F}_{j-1}\} \le \exp(c_j^2 t^2/2)$ almost surely. For each $j \in \{1, \ldots, N\}$, the minimum of those c_j is denoted by $\eta(b_j)$ and $|b_j| \le K_j$ as and $\mathrm{E}(b_{j_1}, b_{j_2}, \ldots, b_{j_k}) = 0$ for $b_{j_1} < \cdots < b_{j_k}$.

Note that the partial derivative conditions on $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ are satisfied by $\mathbf{S}^{(0)}$, $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$; furthermore, $\mathbf{S}$ is automatically positive semidefinite. Moreover, the interval $[0, \tau]$ in the conditions may everywhere be replaced by the set $\{t : h_0(t) > 0\}$.

Conditions (D1)–(D5) are standard requirements for the proportional hazards model [1], which are weaker than those required by Bradic et al. [4], and $\mathbf{S}^{(k)}(\boldsymbol{\beta}_0, t)$ converges uniformly to $\mathbf{s}^{(k)}(\boldsymbol{\beta}_0, t)$. Condition (D6) is a routine one, which is needed to apply the concentration inequality for general empirical processes. For example, the bounded covariate assumption is used by Huang et al. [15] for discussing the Lasso estimator of proportional hazards models. Condition (D7) is needed for the asymptotic behavior of the score function $\ell_p'(\boldsymbol{\beta})$ of partial likelihood because the score function cannot be represented as a sum of independent random vectors, but it can be represented as sum of a sequence of martingale differences.

Proof of Theorem 1. Applying the Taylor expansion to $\ell_p(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma} = \boldsymbol{\beta}$, one finds

$$\ell_p(\boldsymbol{\gamma}) = \ell_p(\boldsymbol{\beta}) + \ell_p'(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top \ell_p''(\tilde{\boldsymbol{\beta}})(\boldsymbol{\gamma} - \boldsymbol{\beta})/2,$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

$$\begin{aligned}(\boldsymbol{\gamma} - \boldsymbol{\beta})^\top\{-\ell_p''(\tilde{\boldsymbol{\beta}})\}(\boldsymbol{\gamma} - \boldsymbol{\beta}) &= (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W^{1/2}(\boldsymbol{\beta}) W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-/2}(\boldsymbol{\beta}) W^{1/2}(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) \\ &\le \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-1/2}(\boldsymbol{\beta})](\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}),\end{aligned}$$

where $W(\boldsymbol{\beta})$ is a block diagonal matrix with $W_j(\boldsymbol{\beta})$ being a $d_{n_j} \times d_{n_j}$ matrix. Given that $-\ell''(\boldsymbol{\beta})$ is non-negative definite, $\lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-1/2}(\boldsymbol{\beta})] \ge 0$. Thus, if $u > \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\} W^{-1/2}(\boldsymbol{\beta})] \ge 0$, then

$$\ell_p(\boldsymbol{\gamma}) \ge \ell_p(\boldsymbol{\beta}) + \ell_p'(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) - u(\boldsymbol{\gamma} - \boldsymbol{\beta})^\top W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta})/2 = h(\boldsymbol{\gamma}|\boldsymbol{\beta}).$$

Thus it follows that $\ell_p(\boldsymbol{\gamma}) \ge h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ and $\ell_p(\boldsymbol{\beta}) = h(\boldsymbol{\beta}|\boldsymbol{\beta})$ by the definition of $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$. The solution of $\partial h(\boldsymbol{\gamma}|\boldsymbol{\beta})/\partial \boldsymbol{\gamma} = 0$ is $\boldsymbol{\gamma} = \boldsymbol{\beta} + u^{-1} W(\boldsymbol{\beta})\ell'(\boldsymbol{\beta})$. Hence, under the conditions of Theorem 1, it follows that

$$\ell_p(\boldsymbol{\beta}^{*(t+1)}) \ge h(\boldsymbol{\beta}^{*(t+1)}|\boldsymbol{\beta}^{(t)}) \ge h(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t)}) = \ell(\boldsymbol{\beta}^{(t)}).$$

The second inequality is due to the fact that

$$\tau[\{j : \|\boldsymbol{\beta}_j^{*(t+1)}\|_2 > 0\}] = \tau[\{j : \|\boldsymbol{\beta}_j^{(t)}\|_2 > 0\}] = m$$

and $\boldsymbol{\beta}^{*(t+1)} = \arg\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})$ subject to $\tau[\{j : \|\boldsymbol{\gamma}_j\|_2 > 0\}] \le m$. By definition of $\boldsymbol{\beta}^{(t+1)}$, $\ell_p(\boldsymbol{\beta}^{(t+1)}) \ge \ell_p(\boldsymbol{\beta}^{*(t+1)})$ and $\tau[\{j : \|\boldsymbol{\beta}_j^{(t+1)}\|_2 > 0\}] = m$. This proves Theorem 1. □

Proof of Theorem 2. For a given model s, a subset of $\{1, \ldots, p\}$, let $\hat{\boldsymbol{\alpha}}_s(U)$ be the partial likelihood estimate of $\boldsymbol{\alpha}_s(U)$ based on the spline approximation. The theorem is implied if $\Pr(\hat{s} \in S_+^m) \to 1$. Thus, it suffices to show that

$$\lim_{n\to\infty} \Pr\left[\max_{s \in S_-^m} \ell_p\{\hat{\boldsymbol{\alpha}}_s(U)\} \ge \min_{s \in S_+^m} \ell_p\{\hat{\boldsymbol{\alpha}}_s(U)\}\right] = 0, \tag{A.1}$$

For each $j \in \{1, \ldots, p\}$, we approximate the coefficient function $\alpha_j(U)$ by

$$\alpha_{nj}(U) = \sum_{k=1}^{d_n} \beta_{jk}\psi_{jk}(U) = \boldsymbol{\beta}_j^\top \boldsymbol{\psi}_j(U), \tag{A.2}$$

where $\psi_{j1}(U), \ldots, \psi_{jd_n}(U)$ are basis functions and d_n is the number of basis functions, which is allowed to increase with the sample size n. For $\alpha_{nj}(U)$, define the approximation error for each $j \in \{1, \ldots, p\}$, by

$$\rho_j(U) = \alpha_j(U) - \alpha_{nj}(U) = \alpha_j(U) - \boldsymbol{\beta}_j^\top \boldsymbol{\psi}_j(U).$$

Let $\mathrm{dist}\{\alpha_j(U), \mathcal{S}_j\} = \inf_{\alpha_{nj}(U) \in \mathcal{S}_j} \sup_{U \in [a,b]} \|\rho_j(U)\|_2$, and take $\rho = \max_{1 \le j \le p} \mathrm{dist}\{\alpha_j(U), \mathcal{S}_j\}$. Let $\boldsymbol{\alpha}_n(U) = (\alpha_{n1}(U), \ldots, \alpha_{np}(U))^\top$ and $\boldsymbol{\alpha}(U) = (\alpha_1(U), \ldots, \alpha_p(U))^\top$. For any s,

$$\boldsymbol{\alpha}_s(U) = \begin{pmatrix} \boldsymbol{\psi}_1(U) & & \\ & \ddots & \\ & & \boldsymbol{\psi}_s(U) \end{pmatrix}_{s \times sd_n} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_s \end{pmatrix}_{sd_n \times 1} + \begin{pmatrix} \rho_1(U) \\ \vdots \\ \rho_s(U) \end{pmatrix} \hat{=} \Psi_s(U)\boldsymbol{\beta}_s + \rho_s(U),$$

where $\Psi_s(U) = \text{diag}\{\boldsymbol{\psi}_1(U), \ldots, \boldsymbol{\psi}_s(U)\}$ with $\boldsymbol{\psi}_j(U) = (\psi_{j1}(U), \ldots, \psi_{jd_n}(U))$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_n})^\top$ for all $j \in \{1, \ldots, s\}$. For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. So, we have

$$\begin{aligned}\ell_p\{\boldsymbol{\alpha}_{s'}(U)\} - \ell_p\{\boldsymbol{\alpha}^*_{s'}(U)\} &= \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}_{s'} + \rho_{s'}(U)\} - \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}^*_{s'} + \rho^*_{s'}(U)\} \\ &= \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}_{s'}\} + \ell'_p\{\Psi_{s'}(U)\tilde{\boldsymbol{\beta}}_{s'}\}\rho_{s'}(U) - \ell_p\{\Psi_{s'}(U)\boldsymbol{\beta}^*_{s'}\} - \ell'_p\{\Psi_{s'}(U)\tilde{\boldsymbol{\beta}}^*_{s'}\}\rho^*_{s'}(U),\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ and $\tilde{\boldsymbol{\beta}}^*_{s'}$ are two immediate values. Denote

$$\Delta_1 = \{\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'})\}, \quad \Delta_2 = \ell'_p(\tilde{\boldsymbol{\beta}}_{s'})\rho_{s'}(U), \quad \Delta_3 = \ell'_p(\tilde{\boldsymbol{\beta}}^*_{s'})\rho^*_{s'}(U).$$

Thus, we have $\ell_p\{\boldsymbol{\alpha}_{s'}(U)\} - \ell_p\{\boldsymbol{\alpha}^*_{s'}(U)\} = \Delta_1 + \Delta_2 + \Delta_3$. For Δ_2, by the Cauchy–Schwarz inequality, we have

$$\text{E}|\Delta_2| = \text{E}|\ell'_p(\tilde{\boldsymbol{\beta}}_{s'})\rho_{s'}(U)| \le \sqrt{\text{E}\|\ell'_p(\tilde{\boldsymbol{\beta}}_{s'})\|^2}\sqrt{\text{E}\|\rho_{s'}(U)\|^2}.$$

By Condition (C5) and Corollary 1 in [25], we obtain $\Delta_2 = o_p(1)$. Similarly to Δ_2, we can also conclude that $\Delta_3 = o_p(1)$.

Next, we consider the term Δ_1. For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. Under Condition (C3), we consider $\boldsymbol{\beta}_{s'}$ close to $\boldsymbol{\beta}^*_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| = w_1 d_n n^{-\tau_1}$ for some $w_1, \tau_1 > 0$. Clearly, when n is sufficiently large, $\boldsymbol{\beta}_{s'}$ falls into a small neighborhood of $\boldsymbol{\beta}^*_{s'}$, so that Condition (C5) becomes applicable. Thus, it follows from Condition (C5) and the Cauchy–Schwarz inequality that

$$\begin{aligned}\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) &= (\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'})^\top \ell'_p(\boldsymbol{\beta}^*_{s'}) + (1/2)(\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'})^\top \ell''_p(\tilde{\boldsymbol{\beta}}_{s'})(\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}) \\ &\le (\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'})^\top \ell'_p(\boldsymbol{\beta}^*_{s'}) - (C_1 d_n^{-1}/2)n\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2^2 \\ &\le w_1 d_n n^{-\tau_1}\|\ell'_p(\boldsymbol{\beta}^*_{s'})\|_2 - (C_1 d_n/2)w_1^2 n^{1-2\tau_1}, \end{aligned} \tag{A.3}$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ is an intermediate value between $\boldsymbol{\beta}_{s'}$ and $\boldsymbol{\beta}^*_{s'}$. Thus, we have

$$\begin{aligned}\Pr\{\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) \ge 0\} \le \Pr\{\|\ell'_p(\boldsymbol{\beta}^*_{s'})\|_2 \ge (C_1 w_1/2)n^{1-\tau_1}\} &= \Pr\left[\sum_{j\in s'}\{\ell'_j(\boldsymbol{\beta}^*_{s'})\}^2 \ge (C_1 w_1/2)^2 n^{2-2\tau_1}\right] \\ &\le \sum_{j\in s'}\Pr[\{\ell'_j(\boldsymbol{\beta}^*_{s'})\}^2 \ge (2m)^{-1}(C_1 w_1/2)^2 n^{2-2\tau_1}].\end{aligned}$$

Also, by (C3), we have $m \le w_2 n^{\tau_2}$, and also the following probability inequality

$$\begin{aligned}\Pr\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \ge (2m)^{-1/2}(C_1 w_1/2)n^{1-\tau_1}\} &\le \Pr\{\ell'_j(\boldsymbol{\beta}^*_s) \ge (2w_2 n^{\tau_2})^{-1/2}(C_1 w_1/2)n^{1-\tau_1}\} \\ &= \Pr\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \ge cn^{1-\tau_1-0.5\tau_2}\} = \Pr\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \ge ncn^{-\tau_1-0.5\tau_2}\},\end{aligned} \tag{A.4}$$

where $c = C_1 w_1/(2\sqrt{2w_2})$ denotes some generic positive constant. Recall (2), by differentiation and rearrangement of terms, it can be shown as in [1] that the gradient of $\ell_p(\boldsymbol{\beta})$ is

$$\ell'_p(\boldsymbol{\beta}) \equiv \partial\ell_p(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^n \int_0^\infty \{\mathbf{z}_i - \bar{\mathbf{z}}_n(\boldsymbol{\beta}, t)\}d\overline{N}_i(t), \tag{A.5}$$

where $\bar{\mathbf{z}}_n(\boldsymbol{\beta}, t) = \sum_{i\in R_j}\mathbf{z}_i \exp(\mathbf{z}_i^\top\boldsymbol{\beta})/\sum_{i\in R_j}\exp(\mathbf{z}_i^\top\boldsymbol{\beta})$. As a result, the partial score function $\ell'_p(\boldsymbol{\beta})$ no longer has a martingale structure, and the large deviation results for continuous time martingale in [4] and [15] are not directly applicable. The martingale process associated with $\overline{N}_i(t)$ is given by

$$M_i(t) = \overline{N}_i(t) - \int_0^\top R_i(u)\exp(\mathbf{z}^\top\boldsymbol{\beta}^*)d\Lambda_0(u).$$

For each $j \in \{1, \ldots, N\}$, let t_j be the time of the jth jump of the process $\sum_{i=1}^n \int_0^\infty R_i(t)d\overline{N}_i(t)$, and set $t_0 = 0$. Then, t_j are stopping times. For $j \in \{0, \ldots, N\}$, further define

$$Q_j = \sum_{i=1}^n \int_0^{t_j} b_i(u)d\overline{N}_i(u) = \sum_{i=1}^n \int_0^{t_j} b_i(u)dM_i(u), \tag{A.6}$$

where $b_i(u) = \mathbf{z}_i - \bar{\mathbf{z}}_n(\boldsymbol{\beta}, u)$ for all $i \in \{1, \ldots, n\}$ are predictable, provided that no two component processes jump at the same time, (D6 holds), and $|b_i(u)| \le 1$.

Since $M_i(u)$ are martingales and $b_i(u)$ are predictable, $\{Q_0, Q_1, \ldots\}$ is a martingale with the difference $|Q_j - Q_{j-1}| \le \max_{u,i}|b_i(u)| \le 1$. Recall definition of N in Section 2, we define $C_0^2 n \le N$, where C_0 is a constant. So, by the martingale version of Hoeffding's inequality [2] and under Condition (D7), we have

$$\Pr(|Q_N| > nC_0 x) \le 2\exp\{-n^2 C_0^2 x^2/(2N)\} \le 2\exp(-nx^2/2).$$

By (A.6), $Q_N = n\ell'_p(\boldsymbol{\beta})$ if and only if $\sum_{i=1}^n \int_0^\infty R_i(t)d\overline{N}_i(t) \le N$. Thus, the left-hand side of (3.15) in Lemma 3.3 of [15] is no greater than $\Pr(|Q_N| > nC_0x) \le 2\exp(-nx^2/2)$. Now (A.4) can be rewritten as follows:

$$\Pr\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \ge ncn^{-\tau_1-0.5\tau_2}\} \le \exp\{-0.5nn^{-2\tau_1-\tau_2}\} = \exp\{-0.5n^{1-2\tau_1-\tau_2}\}. \tag{A.7}$$

By the same arguments, we have

$$\Pr\{\ell'_j(\boldsymbol{\beta}^*_{s'}) \le -m^{-1/2}(C_1w_1/2)n^{1-\tau_1}\} \le \exp\{-0.5n^{1-2\tau_1-\tau_2}\}. \tag{A.8}$$

Inequalities (A.7) and (A.8) imply that,

$$\Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \ge \ell_p(\boldsymbol{\beta}^*_{s'})\} \le 4m\exp\{-0.5n^{1-2\tau_1-\tau_2}\}.$$

Consequently, by Bonferroni's inequality and under conditions (C3)–(C4), we have

$$\begin{aligned}
\Pr\left\{\max_{s\in S_-^m} \ell_p(\boldsymbol{\beta}_{s'}) \ge \ell_p(\boldsymbol{\beta}^*_{s'})\right\} &\le \sum_{s\in S_-^m} \Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \ge \ell_p(\boldsymbol{\beta}^*_{s'})\} \\
&\le 4mp^m\exp\{-0.5n^{1-2\tau_1-\tau_2}\} = 4\exp\{\log m + m\log p - 0.5n^{1-2\tau_1-\tau_2}\} \\
&\le 4\exp\{\log w_2 + \tau_2\log n + w_2n^{\tau_2}\tilde{c}n^{\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4w_2\exp\{\tau_2\log n + w_2\tilde{c}n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= a_1\exp\{\tau_2\log n + a_2n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} = o(1),
\end{aligned} \tag{A.9}$$

as $n \to \infty$ for some generic positive constants $a_1 = 4w_2$ and $a_2 = w_2\tilde{c}$. By Condition (C5), $\ell_p(\boldsymbol{\beta}_{s'})$ is concave in $\boldsymbol{\beta}_{s'}$ and (A.9) holds for any $\boldsymbol{\beta}_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| = w_1d_nn^{-\tau_1}$.

For any $s \in S_-^m$, let $\breve{\boldsymbol{\beta}}_{s'}$ be $\hat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in s'/s^*, i.e., $s' = \{s \cup (s^*/s)\} \cup (s'/s^*)$. By Condition (C3),

$$\|\breve{\boldsymbol{\beta}}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*\cup(s'/s^*)}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \ge \|\boldsymbol{\beta}^*_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \ge \|\boldsymbol{\beta}^*_{s'/s^*}\|_2 = w_1d_nn^{-\tau_1}.$$

Consequently,

$$\Pr\left\{\max_{s\in S_-^m} \ell_p(\hat{\boldsymbol{\beta}}_s) \ge \min_{s\in S_+^m} \ell_p(\hat{\boldsymbol{\beta}}_s)\right\} \le \Pr\left\{\max_{s\in S_-^m} \ell_p(\breve{\boldsymbol{\beta}}_{s'}) \ge \ell_p(\boldsymbol{\beta}^*_{s'})\right\} = o(1).$$

So, we have shown that

$$\lim_{n\to\infty} \Pr\left[\max_{s\in S_-^m} \ell\{\hat{\boldsymbol{\alpha}}_s(U)\} \ge \min_{s\in S_+^m} \ell\{\hat{\boldsymbol{\alpha}}_s(U)\}\right] = 0.$$

Therefore, the theorem is proved. □

References

[1] P.K. Andersen, R.D. Gill, Cox's regression model for counting processes: A large sample study, Ann. Statist. 10 (1982) 1100–1120.
[2] K. Azuma, Weighted sums of certain dependent random variables, Tohoku Math. J. 19 (1967) 357–367.
[3] C. de Boor, A Practical Guide to Splines, Springer, New York, 1978.
[4] J.F.J. Bradic, J. Fan, J. Jiang, Regularization for Cox's proportional hazards model with NP-dimensionality, Ann. Statist. 39 (2011) 3092–3120.
[5] M.-Y. Cheng, T. Honda, J.-T. Zhang, Forward variable selection for sparse ultra-high dimensional varying-coefficient models, J. Amer. Statist. Assoc. 111 (2016) 1209–1221.
[6] W. Chu, R. Li, M. Reimherr, Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data, Ann. Appl. Statist. 10 (2016) 596–617.
[7] D.R. Cox, Regression models and life tables (with discussion), J. R. Stat. Soc. Ser. B 34 (1972) 187–220.
[8] D.R. Cox, Partial likelihood, Biometrika 62 (1975) 269–276.
[9] P. Du, S. Ma, H. Liang, Penalized variable selection procedure for Cox models with semiparametric relative risk, Ann. Statist. 38 (2010) 2092–2117.
[10] J. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high-dimensional additive models, J. Amer. Statist. Assoc. 106 (2011) 544–557.
[11] J. Fan, Y. Feng, Y. Wu, High-dimensional variable selection for Cox's proportional hazards model, in: Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D Brown, in: Inst. Math. Stat. (IMS) Collect. 6, Inst. Math. Statist, Beachwood, OH, 2010, pp. 70–86.
[12] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion), J. R. Stat. Soc. Ser. B 70 (2008) 849–911.
[13] J. Fan, Y. Ma, W. Dai, Nonparametric independence screening in sparse ultra-high dimensional varying-coefficient models, J. Amer. Statist. Assoc. 109 (2014) 1270–1284.
[14] Y. Hu, H. Liang, Variable selection in a partially linear proportional hazards model with a diverging dimensionality, Statist. Probab. Lett. 83 (2013) 61–69.
[15] J. Huang, T. Sun, Z. Ying, Y. Yu, C.-H. Zhang, Oracle inequalities for the LASSO in the Cox model, Ann. Statist. 41 (2013) 1142–1165.
[16] Y.H. Kim, D.C. Jeong, K. Pak, M.-E. Han, J.-Y. Kim, L. Liangwen, H.J. Kim, T.W. Kim, T.H. Kim, D.W. Hyun, S.-O. Oh, Slc2a2 (glut2) as a novel prognostic factor for hepatocellular carcinoma, Oncotarget 8 (2017) 68381–68392.
[17] X.-B. Kong, Z. Liu, Y. Yao, W. Zhou, Sure screening by ranking the canonical correlations, Test 26 (2017) 46–70.
[18] C. Leng, H. Zhang, Model selection in nonparametric hazard regression, J. Nonparametr. Stat. 18 (2006) 417–429.

[19] H. Lian, J. Li, Y. Hu, Shrinkage variable selection and estimation in proportional hazards models with additive structure and high dimensionality, Comput. Statist. Data Anal. 63 (2013) 99–112.
[20] W.-B. Liu, W.-D. Jia, J.-L. Ma, G.-L. Xu, H.-C. Zhou, Y. Peng, W. Wang, Krockdown of gtpbp4 inhibits cell growth and survival in human hepatocellular carcinoma and its prognostic significance, Oncotarget 8 (2017) 93984–93997.
[21] J. Liu, R. Li, R. Wu, Feature selection for varying-coefficient models with ultrahigh-dimensional covariates, J. Amer. Statist. Assoc. 109 (2014) 266–274.
[22] R. Song, F. Yi, H. Zou, On varying-coefficient independence screening for high dimensional varying-coefficient models, Statist. Sinica 24 (2014) 1735–1752.
[23] C.J. Stone, Optimal global rates of convergence for nonparametric regression, Ann. Statist. 10 (1982) 1040–1053.
[24] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[25] F. Wei, J. Huang, H. Li, Variable selection and estimation in high-dimensional varying-coefficient models, Statist. Sinica 21 (2011) 1515–1540.
[26] X. Xia, H. Yang, J. Li, Feature screening for generalized varying-coefficient models with application to dichotomous response, Comput. Statist. Data Anal. 102 (2016) 85–97.
[27] C. Xu, J. Chen, The sparse MLE for ultrahigh-dimensional feature screening, J. Amer. Statist. Assoc. 109 (2014) 1257–1269.
[28] J. Yan, J. Huang, Model selection for Cox models with time-varying coefficients, Biometrics 68 (2012) 419–428.
[29] G. Yang, Y. Yu, R. Li, A. Buu, Feature screening in ultrahigh dimensional Cox's model, Statist. Sinica 26 (2016) 881–901.
[30] H. Zhang, W. Lu, Adaptive lasso for Cox's proportional hazards model, Biometrika 94 (2007) 691–703.
[31] W. Zhang, B. Sun, Impact of age on the survival of patients with liver cancer: An analysis of 27, 255 patients in the seer database, Oncotarget 6 (2015) 633–641.
[32] S. Zhao, Y. Li, Principled sure independence screening for cox models with ultra-high-dimensional covariates, J. Multivariate Anal. 105 (2012) 397–411.