Published in final edited form as:

J Stat Comput Simul. 2018 ; 88(6): 1117–1133. doi:10.1080/00949655.2017.1422127.

Feature screening in ultrahigh-dimensional additive Cox model

Guangren Yang^a, Sumin Hou^a, Luheng Wang^b, Yanqing Sun^c

^aDepartment of Statistics, School of Economics, Jinan University, Guangzhou, People's Republic of China

^bSchool of Statistics, Beijing Normal University, Beijing, People's Republic of China

^cDepartment of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA

Abstract

The additive Cox model is flexible and powerful for modelling the dynamic changes of regression coefficients in the survival analysis. This paper is concerned with feature screening for the additive Cox model with ultrahigh-dimensional covariates. The proposed screening procedure can effectively identify active predictors. That is, with probability tending to one, the selected variable set includes the actual active predictors. In order to carry out the proposed procedure, we propose an effective algorithm and establish the ascent property of the proposed algorithm. We further prove that the proposed procedure possesses the sure screening property. Furthermore, we examine the finite sample performance of the proposed procedure via Monte Carlo simulations, and illustrate the proposed procedure by a real data example.

Keywords

The additive Cox model; partial likelihood; spline approximations; ultrahigh-dimensional survival data

1. Introduction

Feature screening in the high-dimensional data model has become the most important research topics. Feature screening procedures that can effectively reduce ultrahigh dimensionality become indispensable and have attracted considerable attentions in the recent literature. Fan and Lv [1] proposed a marginal screening procedure for ultrahigh-dimensional Gaussian linear models, and demonstrated that marginal screening procedures may possess a sure screening property under certain conditions.

Survival analysis is another important subject in statistics. It has been widely used in medical science, economics, finance, social science, among others. In many studies, survival

CONTACT Guangren Yang tygrjnu@outlook.com Department of Statistics, School of Economics, Jinan University, Guangzhou 510632, People's Republic of China.

Disclosure statement

No potential conflict of interest was reported by the authors.

data have primary outcomes or responses subject to censoring. The Cox model [2] is the most commonly used regression model for survival data, and the partial likelihood method [3] has become a standard approach to parameter estimation and statistical inference. For example, Bradic et al. [4], Du et al. [5], Leng and Zhang [6] and Zhang and Lu [7] investigated variable selection and estimation in Cox regression models. Huang et al. [8] studied the penalized partial likelihood with the L_1 -penalty for the Cox model with high-dimensional covariates and established oracle inequalities. Recently, Lian et al. [9] considered shrinkage variable selection and estimation in proportional hazards models with additive structure under risk factors and high dimensionality settings. Zhang et al. [10] studied penalized variable selection in the additive Cox models based on (group) smoothly clipped absolute deviation penalty and hence widen the scope of applicability of penalized variable selection to semiparametric models for censored data. Lin et al. [11] proposed a global partial likelihood method to estimate the additive Cox model and show that the proposed estimator is consistent and asymptotically normal.

Many penalized methods have also been proposed to select the significant nonzero component for additive models. Cui et al. [12] proposed a combination of penalized regression spline approximation and group variable selection in additive models. Huang et al. [13] proposed a two-step approach to select and estimate the nonzero components simultaneously in additive models when p is fixed. It uses the group Lasso in the first stage and the adaptive group Lasso in the second stage. Meier et al. [14] considered the problem of estimating a high-dimensional additive model under $p = p_n \gg n$. Ravikumar et al. [15] proposed a new class of methods for high-dimensional non-parametric regression and classification called sparse additive models. Xue [16] considered a penalized polynomial spline method for simultaneous model estimation and variable selection in the additive model when p is fixed. The proposed method can select significant components and estimate non-parametric additive function components simultaneously. Recently, Fan et al. [17] proposed several closely related variable screening procedures in sparse ultrahigh-dimensional additive models.

In this paper, we develop a new screening procedure for the ultrahigh-dimensional additive Cox model based on potential active predictors. The proposed procedure can effectively identify active predictors. In this work, we propose a sure screening procedure for the ultrahigh-dimensional additive Cox model. An effective algorithm is proposed to carry out the screening procedure and its ascent property is demonstrated its ascent property. We establish the screening property for the proposed screening procedure with probability tending to one.

The rest of this paper is organized as follows. In Section 2, we propose a feature screening for the additive Cox model, and demonstrate the ascent property of our proposed algorithm to carry out. We also 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 data example. Some discussion and concluding remarks are in Section 4. Technical proofs are in the appendix.

2. Feature screening procedure for the additive Cox model

Let T^* and C be the event time and the censoring time, respectively. In reality, only $T = \min\{T^*, C\}$ is observed and we denote by $\delta = I\{T^* \leq C\}$ the censoring indicator, where $I\{\cdot\}$ is the indicator function. As usual, we assume T^* and C are independent condition on the p explanatory variables $\mathbf{X} = (X_1, \dots, X_p)^T$. Without loss of generality, we assume X_j has a marginal distribution supported on $[0, 1]$. In additive Cox models, the hazard function is given by

$$\lambda(t | \mathbf{X}) = \lambda_0(t) \exp \left\{ \sum_{j=1}^p \alpha_j(X_j) \right\},$$

(1)

where $\lambda_0(t)$ is an unknown baseline hazard rate for $0 \leq t \leq \tau$ and $\boldsymbol{\alpha}(\mathbf{X}) = (\alpha_1(X_1), \dots, \alpha_p(X_p))^T$ is a vector of unknown function. Here, τ is the end of study time. We assume the censoring mechanism is noninformative. That is, given $\mathbf{X}$, T and C are conditionally independent. Since any constant can be incorporated into the baseline hazard function, we impose the constraints $E\alpha_j(X_j) = 0, j = 1, \dots, p$, for identifiability.

Our main interest is in the estimation of $\boldsymbol{\alpha}(\mathbf{X}) = (\alpha_1(X_1), \dots, \alpha_p(X_p))^T$ given i.i.d. (X_i, δ_i, T_i) . Let $(\mathbf{X}_i, \delta_i, T_i), i = 1, \dots, n$, be i.i.d. copies of $(\mathbf{X}, \delta, T)$. For the non-parametric regression coefficient, we use the B-spline expansion method. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $I - 1$ and $\{\psi_{jk}(\cdot), k = 1, \dots, d_{nj}\}$ for $j = 1, \dots, p$, denote a normalized B-spline basis on $[0, 1]$ with $\|\psi_{jk}(\cdot)\|_\infty = 1$ and $d_{nj} = O(n^{1/5})$, where $\|\cdot\|_\infty$ is the sup norm. Let $\boldsymbol{\psi}_j(X_j) = (\psi_{j1}(X_j), \dots, \psi_{jd_{nj}}(X_j))^T$. For any $\alpha_{nj} \in \mathcal{S}_n$, we have

$$\alpha_{nj}(X_j) = \sum_{k=1}^{d_{nj}} \beta_{jk} \psi_{jk}(X_j) = \boldsymbol{\beta}_j^T \boldsymbol{\psi}_j(X_j), \quad j = 1, \dots, p,$$

(2)

for some coefficients $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jd_{nj}})^T$. Here, we allow d_{nj} increase with n and can be different for different j since different coefficient functions may have different smoothness. We assume that $d_{nj} \asymp d_n$ for $j = 1, \dots, p$. The non-parametric coefficient function $\{\alpha_j(\cdot)\}_{j=1}^p$ can be well approximated by functions in $\mathcal{S}_n = \cup_{j=1}^p \mathcal{S}_{nj}$.

Using the spline expansions introduced above, the problem of estimating $\alpha_j(\cdot)$ is then transformed to the problem of estimating the coefficients $\boldsymbol{\beta}_j$. Let $Y_k(t) = I\{T_k \leq t\}$. The partial likelihood can be written as

$$\ell(\beta) = \sum_{i=1}^n \delta_i \left\{ \mathbf{Z}_i^T \beta - \log \left\{ \sum_{k=1}^n Y_k(T_i) \exp(\mathbf{Z}_i^T \beta) \right\} \right\},$$

(3)

where $\mathbf{Z}_i = (\psi_1^T(X_{i1}), \dots, \psi_p^T(X_{ip}))^T$ and $\beta = (\beta_1^T, \dots, \beta_p^T)$ is the coefficients of B-spline approximations of $\mathbf{a}(\mathbf{X})$.

2.1. Feature screening procedure

Denote $\|\alpha_j(X_j)\|_2 = [E\alpha_j^2(X_j)]^{1/2}$, the L_2 -norm of $\alpha_j(X_j)$. For ease of presentation, denote by s an arbitrary subset of $[1, \dots, p]$, $\mathbf{X}_s = [X_j, j \in s]$ and $\mathbf{a}_s(\mathbf{X}_s) = [\alpha_j(X_j), j \in s]$. 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 $\mathbf{a}$ is $\mathbf{a}^*$, so β^* includes the coefficients of B-spline approximation corresponding to $\mathbf{a}^*$. Denote $s^* = [j: \|\alpha_j(X_j)\|_2 > 0]$. 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 Equation (3), we propose screening features in the additive Cox model by the constrained partial likelihood

$$\hat{\beta} = \arg \max_{\beta} \ell_p(\beta) \text{ subject to } \tau(\{j: \|\beta_j\|_2 > 0\}) \leq m$$

(4)

for a pre-specified m which is assumed to be greater than the number of nonzero elements of β^* .

For high-dimensional problems, it becomes almost impossible to solve the constrained maximization problem (4) directly. Alternatively, we consider a proxy of the partial likelihood function. It follows by the Taylor expansion for the partial likelihood function $\ell_p(\gamma)$ at β lying within a neighbour of γ that

$$\ell_p(\gamma) \approx \ell_p(\beta) + (\gamma - \beta)^T \ell'_p(\beta) + \frac{1}{2}(\gamma - \beta)^T \ell''_p(\beta)(\gamma - \beta),$$

where $\ell'_p(\beta) = \partial \ell_p(\gamma) / \partial \gamma|_{\gamma=\beta}$ and $\ell''_p(\beta) = \partial^2 \ell_p(\gamma) / \partial \gamma \partial \gamma^T|_{\gamma=\beta}$. Denote $D_n = \sum_{j=1}^p d_{nj}$. If $\ell''_p(\beta)$ is invertible, the computational complexity of calculating the inverse of $\ell''_p(\beta)$ is $O(D_n^3)$. For large D_n small n problems (i.e. $D_n \gg n$), $\ell''_p(\beta)$ becomes not invertible. Low computational costs are always desirable for feature screening. To deal with singularity of

the Hessian matrix and save computational costs, we propose to use the following approximation for $\ell_p''(\gamma)$:

$$h(\gamma \mid \beta) = \ell_p(\beta) + (\gamma - \beta)^T \ell_p'(\beta) - \frac{u}{2} (\gamma - \beta)^T W(\beta) (\gamma - \beta), \quad (5)$$

where u is a scaling constant to be specified and $W(\beta) = \text{diag}\{W_1(\beta), \dots, W_p(\beta)\}$, a block diagonal matrix with $W_j(\beta)$ being a $d_{nj} \times d_{nj}$ matrix. Here, we allow $W(\beta)$ to depend on β . Hence (5) is the second-order Taylor expansion of $\ell_p(\gamma)$ in the neighbourhood of β by replacing $\ell_p''(\beta)$ with $-uW(\beta)$. This implies that we approximate $\ell_p''(\beta)$ by $-uW(\beta)$.

Throughout this paper, we will use $W_j(\beta) = -\partial^2 \ell_p(\beta) / \partial \beta_j \partial \beta_j^T$ in order to avoid large-scale matrix inversion and carry out at low computational cost in Step 2. And there provides some insight about the choice of u in algorithm and Theorem 2.1. ■

It can be seen that $h(\beta \mid \beta) = \ell_p(\beta)$, and under some conditions, $h(\gamma \mid \beta) = \ell_p(\beta)$ for all γ . This ensures the ascent property. See Theorem 2.1 below for more details. Since $W(\beta)$ is a block diagonal matrix, $h(\gamma \mid \beta)$ is an additive function of γ_j for any given β . The additivity enables us to have a closed form solution for the following maximization problem:

$$\max_{\gamma} h(\gamma \mid \beta) \text{ subject to } \tau(\{j: \|\gamma_j\|_2 > 0\}) \leq m \quad (6)$$

for given β and m . Define $\tilde{\gamma}_j = \beta_j + u^{-1} W_j^{-1}(\beta_j) \partial \ell_p(\beta) / \partial \beta_j$ for $j = 1, \dots, p$, and

$\tilde{\gamma} = (\tilde{\gamma}_1^T, \dots, \tilde{\gamma}_p^T)^T = \beta + u^{-1} W^{-1}(\beta) \ell_p'(\beta)$. Then $\tilde{\gamma}$ is the maximizer of $h(\gamma \mid \beta)$ for fixed β .

Denote $g_j = \tilde{\gamma}_j^T W_j(\beta_j) \tilde{\gamma}_j$ for $j = 1, \dots, p$, and sort g_j so that $g_{(1)} \geq g_{(2)} \geq \dots \geq g_{(p)}$. The solution of maximization problem (6) is the hard-thresholding rule defined below

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

This method enables us to effectively screen features by using the following algorithm. Our procedure iteratively updates β at Step 2. This enables the procedure to incorporate correlation information among the predictors through updating $\ell(\beta)$ and $\ell'(\beta)$.

Feature Screening of the Additive Cox Model

Step1. Set the initial value $\beta_j^{(0)} = \mathbf{0}, j = 1, \dots, p$.

Step2. Set $t = 0, 1, 2, \dots$, iteratively conduct Step 2a and Step 2b below until the algorithm converges.

Step 2a. Calculate $\tilde{\gamma}_j^{(t)} = \beta_j^{(t)} + u_t^{-1} W_j^{-1} (\beta_j^{(t)}) \partial \ell(\beta_j^{(t)}) / \partial \beta_j$, and

$$g_j^{(t)} = \{\tilde{\gamma}_j^{(t)}\}^T W_j (\beta_j^{(t)}) \tilde{\gamma}_j^{(t)}.$$

Let $g_{(1)}^{(t)} \geq g_{(2)}^{(t)} \geq \dots \geq g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$ s. Set

$$S_t = \{j: g_j^{(t)} > g_{(m+1)}^{(t)}\},$$

the nonzero index set.

Step 2b. Update β by $\beta^{(t+1)} = (\beta_1^{(t+1)}, \dots, \beta_p^{(t+1)})^T$ as follows. If

$$j \notin S_t, \text{ set } \beta_j^{(t+1)} = \mathbf{0},$$

$$\text{otherwise, set } \beta_j^{(t+1)} = \tilde{\gamma}_j^{(t)} \text{ for } j \in S_t.$$

Theorem 2.1: Suppose that Conditions (D1) — (D4) in the appendix hold. Let $\{\beta^{(t)}\}$ be the sequence defined in Step 2b in the above algorithm. Denote

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

where $\lambda_{\max}(A)$ stands for the maximal eigenvalue of a matrix A . If $u_t \rightarrow \rho^{(t)}$, then

$$\ell_p(\beta^{(t+1)}) \geq \ell_p(\beta^{(t)}),$$

where $\beta^{(t+1)}$ is defined in Step 2b in the above algorithm.

Theorem 2.1 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. $\tau(\{j: \|\alpha_j(X_j)\|_2 > 0\}) = m$), and the resulting estimate in the current step may serve as a refinement of the last step. This theorem also provides us some insights about choosing u_t in practical implementation. In our numerical studies, the algorithm typically converged within four iterations. Still, the algorithm is not guaranteed to converge to the global optimizer.

2.2. Sure screening property

For a subset s of $\{1, \dots, p\}$ with size $\tau(s)$, recall notation $\mathbf{X}_s = \{X_j, j \in s\}$ and associated coefficients $\alpha_s = \{\alpha_j(X_j), j \in s\}$ corresponding to $\beta_s = \{\beta_j, j \in s\}$ with $\beta_j = (\beta_{j1}, \dots, \beta_{jd_{nj}})^T$.

We denote the true model by $s^* = \{j: E\alpha_j^2(X_j) > 0, \quad 1 \leq j \leq p\}$ with $\tau(s^*) = q$. The objective of feature selection 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 ultrahigh-dimensional additive Cox model feature screening procedure. The sure screening property [1] is referred to as

$$\Pr(s^* \subset \hat{s}) \rightarrow 1, \quad \text{as } n \rightarrow \infty,$$

(7)

To establish this sure screening property for the proposed additive Cox model, we introduce some additional notations as follows. For any model s , let $\ell(\beta_s) = \ell(\beta_s)/\beta_s$ and $\ell''(\beta_s) = -\partial^2 \ell(\beta_s) / \partial \beta_s \partial \beta_s^T$ be the score function and the Hessian matrix of $\ell(\cdot)$ as a function of β_s , respectively. Assume that a screening procedure retains m out of p features such that $\tau(s^*) = q < m$. So, we define

$$S_+^m = \{s: s^* \subset s; \tau(s) \leq m\} \quad \text{and} \quad S_-^m = \{s: s^* \not\subset s; \tau(s) \geq m\}$$

as the collections of the over-fitted models and the under-fitted models, respectively. We investigate the asymptotic properties of $\hat{\beta}$ under the scenario where p , q , m and β^* are allowed to depend on the sample size n . We impose the following conditions, some of which are purely technical and only serve to facilitate theoretical understanding of the proposed feature screening procedure.

(C1) The support of $X_j, j = 1, \dots, p$, is bounded and is assumed to be $[a, b]$.

(C2) The function $\{\alpha_j(X_j)\}_{j=1}^p$ belong to a class of functions $\mathcal{F}$, whose r th derivative $\alpha_j^{(r)}(\cdot)$ exists and is Lipschitz of order η ,

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

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

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

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

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

$$C_1 d_n^{-1} \leq \lambda_{\min}[-n^{-1} \ell_p''(\beta_s)] \leq \lambda_{\max}[-n^{-1} \ell_p''(\beta_s)] \leq C_2 d_n^{-1}$$

for $\beta_s \in \{\beta: \|\beta_s - \beta_s^*\|_2 \leq \delta\}$ and $s \in S_+^{2m}$, where $\lambda_{\min}[\cdot]$ and $\lambda_{\max}[\cdot]$ denotes the smallest and largest eigenvalues of a matrix.

Under the conditions (C1) and (C2), the following two properties of B-splines are valid. So, we have (a) for B-spline basis satisfying $\psi_{jk}(x) \geq 0$ and $\sum_{k=1}^d \psi_{jk}(x) = 1, x \in [a, b]$, there exist positive constants C_3 and C_4 such that $C_3 d_n^{-1} \leq E\psi_{jk}^2(x) \leq C_4 d_n^{-1}$ [18] and (b) if $\{\alpha_j(X_j), j = 1, 2, \dots, p\}$ is a set of functions in $\mathcal{F}$ described in condition (C2), there exists a positive constant C_5 that does not depend on $\alpha_j(X_j)$, then the uniform approximation error satisfies $\rho = \sup_{X_j \in [a, b]} \|\alpha_j(X_j) - \alpha_{nj}(X_j)\|_2 \leq C_5 d_n^{-\nu}, \forall j$, as $d_n \rightarrow \infty$ [19,20]. Conditions (C1) and (C2) ensure properties (a) and (b), which are required for the B-spline approximation and establishing the sure screening properties.

Note that $\|\alpha_{nj}(X_j)\|_2^2 = \beta_j^T E\{\psi_j(X_j)\psi_j(X_j)^T\} \beta_j$. Based on the properties (a)–(b) and Condition (C3), we can derive that

$$\min_{j \in s^*} \|\beta_j\|_2 \geq w_1 d_n^{-\tau} 1.$$

Condition (C3) states a few requirements for establishing the sure screening property of the proposed procedure. The first one is the sparsity of β^* which makes the sure screening possible with $\tau(\hat{s}) = m > q$. Also, it requires that the minimal component in β^* does not degenerate too fast, 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 $\tau(s) = m$. Condition (C5) assumes that p diverges with n at up to an exponential rate, which 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 by the following theorem.

Theorem 2.2: *Suppose that Conditions (C1) — (C5) and Conditions (D1) — (D7) in the appendix hold. Let $\hat{s}$ be the model obtained by the (4) of size m . We have*

$$\Pr(s^* \subset \hat{s}) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

The proof is given in the appendix. The sure screening property is an appealing property of a screening procedure since it ensures that the true active predictors are retained in the model selected by the screening procedure. One has to specify the value of m in practical

implementation. As to the choice of m , it is typical to set $m = \lfloor n/\log(n) \rfloor$ for a parametric model, where $\lfloor a \rfloor$ indicates the integer part of a [1]. Since we use a linear combination of d_n B-spline bases in our proposed screening procedure, we set $m = \lfloor (n/d_n)/\log(n/d_n) \rfloor$ through in simulation studies. Although it is an ad hoc choice, it works reasonably well in our numerical studies.

3. Numerical studies

In this section, we evaluate the finite sample performance of the proposed feature screening procedure based on the additive Cox model via Monte Carlo simulations, and applied it to an empirical analysis of a real data set. All simulations were conducted by using R codes.

3.1. Simulation studies

In this section, we investigate the performance of the proposed method with Monte Carlo simulation examples. In all the simulation examples, we generate $\mathbf{X} = (X_1, \dots, X_p)$ from a multivariate normal distribution with mean zero and covariance $\Sigma = (\sigma_{ij})$. Here, we consider two commonly used covariance matrix structures. Namely,

1. $\Sigma_1 = (\rho^{|i-j|})$, which is the AR(1) structure.
2. $\Sigma_2 = (1 - \rho)I_p + \rho \mathbf{1}_p \mathbf{1}_p^T$, where I_p is the $p \times p$ identity matrix, which is the compound symmetry structure.

We set ρ with 0.25, 0.50 and 0.75, respectively. In all the examples, we consider the sample size $n = 300, 500$, and $p = 1000, 2000$. In addition, we choose the B-spline basis number with $d_n = 4$ for each coefficient function. All the scenarios repeated with 1000 times and we adopt the following criteria to evaluate the performance of the proposed procedure.

- a. P_j : Given the model size m , the proportion that X_j is selected in 1,000 replications.
- b. P_a : Given the model size m , the proportion that all the significant variables are selected in 1,000 replications.

Example 3.1: We generate the survival time from the additive Cox model with $\lambda_0(t) = 10$ and $\alpha_j(X_j) = X_j^2, j = 1, 2, 3, \alpha_4(X_4) = -3\rho X_4^2$ and $\alpha_j(X_j) = 0$ for $j = 5, \dots, p$. There are only four variables that are significant, where the censoring time generate from an exponential distribution with mean value 0.2. Under these settings, the censoring rates of Σ_1 and Σ_2 are presented in Table 1, and show that for both Σ_1 and Σ_2 , where the censoring rate becomes larger as ρ increases. Moreover, under the compound symmetry structure Σ_2 , X_4 is jointly dependent but marginally weakly correlated with the survival time when $\rho = 0$; this setting is designed to challenge the marginal SIS procedures and the marginal screening procedure fails to identify X_4 . However, the proposed method can detect the signals that are dependent but marginally independent with the response. Tables 2 and 3 present the simulation results under Σ_1 and Σ_2 settings, respectively, which show that the proposed method performs reasonably well. When $\Sigma = \Sigma_2$ in Example 3.1, X_4 is marginally weakly correlated but jointly dependent with the survival time, the proposed method can detect the signal of X_4 quite well and the selected proportion is close to one.

Example 3.2: We generate the survival time from the additive Cox model with $\lambda_0(t) = 10$ and $\alpha_1(X_1) = (X_1 + 1)^2/2$, $\alpha_2(X_2) = \exp(X_2 + 1)/2$, $\alpha_3(X_3) = 3 \sin(X_3)/2$, $\alpha_4(X_4) = -\rho\{(X_4 + 1)^2 + \exp(X_4 + 1) + 3 \sin(X_4)\}/2$, where the first four functions are nonzero. And the censoring time generates from an exponential distribution with mean value 0.25, where the censoring rate of Example 3.2 is presented in Table 4 under Σ_1 and Σ_2 settings. The simulation results of P_j and P_a are reported in Tables 5-6. The results show that the proposed method can select the significant variables with high probability. However, X_4 can also be performed good even though it is marginal weakly correlated with the survival time under $\Sigma = \Sigma_2$. When ρ is large and sample size is 300, it is noted that the selection proportion is relatively small, since ρ and the censoring rate become large, where the effective sample size is small. So, we can find that the selected proportion will become much better when the sample size becomes large (i.e. $n = 500$).

3.2. Real data example

In this section, we illustrate the proposed methodology by an empirical analysis of a subset of data collected the microarray data with diffuse large-B-cell lymphoma (DLBCL, for short). See Lenz et al. [21] for details. This data set contains the microarray data of 181 patients with diffuse DLBCL who are treated with a combination chemotherapy with cyclophosphamide, doxorubicin, vincristine and prednisone (CHOP, for short). The DLBCL is a common type of lymphoma in adults and the survival rate after the standard chemotherapy is only about 35–40%. The original data contains 54,675 probe sets or covariates. However, most of them have smaller changes. There are 3833 genes after a filtering process, which removes the genes with lower variations that a sample variance of a gene was lower 10th percentile. One patient had survival times being close to 0, the sample size of the data set is $n = 180$ and the $p = 3833$. Given such a large number of predictors and the small sample size, feature screening is a necessary initial step for a statistical modelling procedure that cannot deal with high-dimensional survival data. All the predictors were standardized so that they had mean zero and standard deviation one.

Here, we are interested in finding the genetic factors that may influence the survival outcome. The proposed methodology might be more appropriate to analyse the feature screening problem with considering the nonlinear effect since the linear assumption is always potential under high-dimensional problem. In our analysis, we set the spline basis number $d_n = 5$ for each gene, and select $m = \lceil (n/d_n)/\log(n/d_n) \rceil = 10$ genes, where $\lceil \cdot \rceil$ is the rounding operator. The proposed method selects 6 genes probe site (i.e. 1563462_at, 210919_at, 217084_at, 229839_at, 241235_at, and 241369_at) under group SCAD penalized regression. The corresponding partial likelihood is -442.25 .

To demonstrate the significance of these selected genes for each one of these six selected genes, we remove one and refit the additive Cox model with the other five genes, then we can obtain the corresponding partial likelihood. The six partial likelihoods are -484.73 , -483.44 , -482.61 , -483.07 , -482.69 , and -484.67 , respectively. The mean value is -483.53 , and decreases 41.28 to removing one selected gene that reduces a quite significant gene. Moreover, we select m genes out of the genes that are not selected, and fit the corresponding model to select six genes plus the m randomly selected genes. Then we can get the

corresponding partial likelihood. Here, we set m from one to five and repeat each scenario with 200 times. The results are demonstrated in Table 7. Table 7 shows an average increase in the likelihood by about 2 with an additional variable, which shows no significance for the randomly selected genes. For the six selected genes, we fit the additive Cox model and obtain the additive functions of the six genes, which is displayed in Figure 1. Figure 1 that the genes 1563462_at, 217084_at and 241369_at have obviously nonlinear effect, which demonstrates that the additive Cox model is more appropriate than the Cox model.

4. Conclusion

In this work, we proposed a sure partial likelihood screening procedure for feature screening in the additive Cox model with ultrahigh-dimensional covariates and an effective algorithm to carry out the feature screening procedure. Furthermore, we show that the proposed algorithm possesses an ascent property and establish the sure screening property for the additive Cox model. We also conduct numerical study to access the empirical performance of the proposed procedure. The numerical results imply that the proposed algorithm converges quickly and computing time is reasonable.

Funding

Guangren Yang's research was supported by the National Natural Science Foundation of China grant 11471086, the National Social Science Foundation of China (grant no. 16BTJ032), the Fundamental Research Funds for the Central Universities 15JNQM019, the National Statistical Scientific Research Center Projects (2015LD02), Education Bureau of Guangdong Province (2016WTSCX007), Science and Technology Program of Guangzhou (2016201604030074) and Science and Technology Planning Project of Guangdong (2016A050503033). Yanqing Sun's research was partially supported by NIAID N1H award number R37AI054165 and NSF (grant no. DMS1513072).

Appendix

We need the following notation to present the regularity conditions for the partial likelihood and the Cox model. Most notations are adapted from Andersen and Gill [22], in which counting processes were introduced for the Cox model and the consistency and asymptotic normality of the partial likelihood estimate were established. Denote

$N_i(t) = I\{T_i^* \leq t, T_i^* \leq C_i\}$ and $R_i(t) = \{T_i^* \geq t, C_i \geq t\}$. Assume that there are no two component processes $N_i(t)$ jumping at the same time. For simplicity, we shall 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 $\Lambda_0(t) = \int_0^t \lambda_0(u)du$.

Under the additive Cox model (1),

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(u)du, \quad i = 1, \dots, n,$$

are local martingales on the time interval $[0, \tau]$, where $\lambda(t) = \lambda_0(t) \exp(\sum_{j=1}^p \alpha_j(X_j))$ is the intensity $N_\lambda(t)$ w.r.t. $\mathcal{F}_t$.

Define

$$\mathbf{A}^{(k)}(\beta, t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp\{\mathbf{X}_i^T \beta\} \mathbf{X}_i^{\otimes k}, \quad \mathbf{a}^{(k)}(\beta, t) = E[\mathbf{A}^{(k)}(\beta, t)] \quad \text{for } k = 0, 1, 2,$$

and

$$E(\beta, t) = \frac{\mathbf{A}^{(1)}(\beta, t)}{\mathbf{A}^{(0)}(\beta, t)}, \quad V(\beta, t) = \frac{\mathbf{A}^{(2)}(\beta, t)}{\mathbf{A}^{(0)}(\beta, t)} - E(\beta, t) \otimes 2.$$

where $\mathbf{X}_i^{\otimes 0} = 1$, $\mathbf{X}_i^{\otimes 1} = \mathbf{X}_i$ and $\mathbf{X}_i^{\otimes 2} = \mathbf{X}_i \mathbf{X}_i^T$. Note that $\mathbf{A}^{(0)}(\beta, t)$ is a scalar, $\mathbf{A}^{(1)}(\beta, t)$ and $E(\beta, t)$ are p -vector, and $\mathbf{A}^{(2)}(\beta, t)$ and $V(\beta, t)$ are $p \times p$ matrices.

Define

$$Q_j = \sum_{i=1}^n \int_0^{t_j} \left[\mathbf{X}_i - \frac{\mathbf{A}^{(1)}(\beta, t)}{\mathbf{A}^{(0)}(\beta, t)} \right] dM_i.$$

Here, $E[Q_j \mid \mathcal{F}_{t_{j-1}}] = Q_{j-1}$ i.e. $E[Q_j - Q_{j-1} \mid \mathcal{F}_{t_{j-1}}] = 0$. Let $b_j = Q_j - Q_{j-1}$, then $(b_j)_{j=1,2,\dots}$ is a sequence of bounded martingale differences on $(\Omega, \mathcal{F}, P)$. That is, b_j is bounded almost surely (a.s.) and $E[b_j \mid \mathcal{F}_{t_{j-1}}] = 0$ a.s. for $j = 1, 2, \dots$

(D1) (Finite interval). $\Lambda_0(\tau) = \int_0^\tau \lambda_0(t) dt < \infty$

(D2) (Asymptotic stability). There exists a neighbourhood $\mathcal{B}$ of β^* and scalar, vector and matrix functions $\mathbf{a}^{(0)}, \mathbf{s}^{(1)}$ and $\mathbf{a}^{(2)}$ defined on $\mathcal{B} \times [0, \tau]$ such that for $k = 0, 1, 2$

$$\sup_{t \in [0, \tau], \beta \in \mathcal{B}} \|\mathbf{A}^{(k)}(\beta, t) - \mathbf{a}^{(k)}(\beta, t)\| \xrightarrow{P} 0.$$

(D3) (Lindeberg's condition). There exists $\delta > 0$ such that

$$n^{-1/2} \sup_{i,t} |\mathbf{X}_i| R_i(t) I\{\beta_0^T \mathbf{X}_i > -\delta \mid \mathbf{X}_i\} \xrightarrow{P} 0,$$

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

$$\mathbf{a}^{(1)}(\beta, t) = \frac{\partial}{\partial \beta} \mathbf{a}^{(0)}(\beta, t), \quad \mathbf{a}^{(2)}(\beta, t) = \frac{\partial^2}{\partial \beta^2} \mathbf{a}^{(0)}(\beta, t),$$

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

$$\mathbf{A} = \int_0^\tau v(\beta_0, t) \mathbf{a}^{(0)}(\beta_0, t) \lambda_0(t) dt$$

is positive definite.

(D5) The function $\mathbf{A}^{(0)}(\beta^*, t)$ and $\mathbf{a}^{(0)}(\beta^*, t)$ are bounded away from 0 on $[0, \tau]$.

(D6) There exist constants $C_1, C_2 > 0$, such that $\max_{ij} |x_{ij}| < C_1$ and $\max_i |\mathbf{X}_i^T \beta^*| < C_2$.

(D7) $\{b_j\}$ is a sequence of martingale differences and there exist non-negative constants c_j such that for every real number t ,

$$E\{\exp(tb_j) \mid \mathcal{F}_{j-1}\} \leq \exp(c_j^2 t^2 / 2) \quad \text{a.s. } (j = 1, 2, \dots, N).$$

For each j , the minimum of those c_j is denoted by $\eta(b_j)$.

$$|b_j| \leq K_j \quad \text{a.s. for } j = 1, 2, \dots, N$$

with the N failures time and $E\{b_{j1}, b_{j2}, \dots, b_{jk}\} = 0$ for $b_{j1} < b_{j2} < \dots < b_{jk}$; $k = 1, 2, \dots$

Note that the partial derivative conditions on $\mathbf{a}^{(0)}$, $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ are satisfied by $\mathbf{A}^{(0)}$, $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$; and that $\mathbf{A}$ is automatically positive semidefinite. Furthermore the interval $[0, \tau]$ in the conditions may everywhere be replaced by the set $\{t: \lambda_0(t) > 0\}$.

Condition (D1) — (D5) is a standard condition for the proportional hazards model [22], which is weaker than the one required by Bradic et al. [4] and $A^{(k)}(\beta_0, t)$ converges uniformly to $\mathbf{a}^{(k)}(\beta_0, t)$. 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. [8] for discussing the Lasso estimator of the proportional hazards models. Condition (D7) is needed for the asymptotic behaviour of the score function $\ell'_p(\beta)$ 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 2.1: Applying the Taylor expansion to $\ell'_p(\gamma)$ at $\gamma = \beta$, it follows that

$$\ell'_p(\gamma) = \ell'_p(\beta) + \ell''_p(\beta)(\gamma - \beta) + \frac{1}{2}(\gamma - \beta)^T \tilde{\ell}''_p(\tilde{\beta})(\gamma - \beta),$$

where $\tilde{\beta}$ lies between γ and β .

$$\begin{aligned}
(\gamma - \beta)^T \{-\ell_p''(\tilde{\beta})\}(\gamma - \beta) &= (\gamma - \beta)^T W^{-1/2}(\beta) W^{-1/2}(\beta) \{-\ell_p''(\tilde{\beta})\} W^{-1/2}(\beta) W^{1/2}(\beta) (\gamma - \beta) \\
&\leq \lambda_{\max}[W^{-1/2}(\beta) \{-\ell_p''(\tilde{\beta})\} W^{-1/2}(\beta)] (\gamma - \beta)^T W(\beta) (\gamma - \beta),
\end{aligned}$$

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

$$u > \lambda_{\max}[W^{-1/2}(\beta) \{-\ell_p''(\tilde{\beta})\} W^{-1/2}(\beta)] \geq 0,$$

then

$$\ell_p(\gamma) \geq \ell_p(\beta) + \ell_p'(\beta)(\gamma - \beta) - \frac{u}{2}(\gamma - \beta)^T W(\beta)(\gamma - \beta) = h(\gamma | \beta).$$

Thus it follows that $\ell_p(\gamma) - h(\gamma | \beta)$ and $\ell_p(\beta) = h(\beta | \beta)$ by the definition of $h(\gamma, \beta)$. The solution of $\nabla h(\gamma | \beta) = 0$ is $\gamma = \beta + u^{-1} W(\beta) \ell_p'(\beta)$. Hence, under the conditions of Theorem 2.1, it follows that

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

The second inequality is due to the fact that $\tau(\{j: \|\beta_j^{*(t+1)}\|_2 > 0\}) = \tau(\{j: \|\beta_j^{(t)}\|_2 > 0\}) = m$ and $\beta^{*(t+1)} = \arg \max_{\gamma} h(\gamma | \beta^{(t)})$ subject to $\tau(\{j: \|\gamma_j\|_2 > 0\}) = m$. By definition of $\beta^{(t+1)}$, $\ell_p(\beta^{(t+1)}) = \ell_p(\beta^{*(t+1)})$ and $\tau(\{j: \|\beta_j^{(t+1)}\|_2 > 0\}) = m$. This proves Theorem 2.1. *

Proof of Theorem 2.2: For a given model 5, a subset of $\{1, \dots, p\}$, let $\hat{\alpha}_s$ be the partial likelihood estimate of α_s based on the spline approximation. The theorem is implied if $\Pr\{\hat{s} \in S_+^m\} \rightarrow 1$. Thus, it suffices to show that

$$\Pr \left\{ \max_{s \in S_-^m} \ell_p\{\hat{\alpha}_s\} \geq \min_{s \in S_+^m} \ell_p\{\hat{\alpha}_s\} \right\} \rightarrow 0,$$

(A1)

as $n \rightarrow \infty$.

We approximate the coefficient function $\alpha_f(X_j)$ by

$$\alpha_{nj}(X_j) = \sum_{k=1}^{d_n} \beta_{jk} \psi_{jk}(X_j) = \beta_j^T \psi_j(X_j), \quad j = 1, \dots, p,$$

(A2)

where $\psi_{jk}(X_j)$, $k = 1, \dots, d_n$ are basis functions and d_n is the number of basis functions, which is allowed to increase with the sample size n . For $\alpha_{nj}(X_j)$, define the approximation error by

$$\rho_j(X_j) = \alpha_j(X_j) - \alpha_{nj}(X_j) = \alpha_j(X_j) - \beta_j^T \psi_j(X_j), \quad j = 1, \dots, p.$$

Let $\text{dist}(\alpha_j(X_j), \mathcal{S}_j) = \inf_{\alpha_{nj}(X_j) \in \mathcal{S}_j} \sup_{X_j \in [a, b]} \|\rho_j(X_j)\|_2$, and take

$\rho = \max_{1 \leq j \leq p} \text{dist}(\alpha_j(X_j), \mathcal{S}_j)$. Let $\mathbf{a}_n(\mathbf{X}) = (\alpha_{n1}(X_1), \dots, \alpha_{np}(X_p))^T$ and $\mathbf{a} = (\alpha_1, \dots, \alpha_p)^T$.

For any s ,

$$\begin{aligned} \alpha_s(\mathbf{X}_s) &= \begin{pmatrix} \psi_1(X_1) & & \\ & \ddots & \\ & & \psi_s(X_s) \end{pmatrix}_{s \times sd_n} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_s \end{pmatrix}_{sd_n \times 1} + \begin{pmatrix} \rho_1(X_1) \\ \vdots \\ \rho_s(X_s) \end{pmatrix} \\ &\triangleq \Psi_s(\mathbf{X}_s) \beta_s + \rho_s(\mathbf{X}_s), \end{aligned}$$

where $\Psi_s(\mathbf{X}_s) = \text{diag}(\boldsymbol{\psi}_1(X_1), \dots, \boldsymbol{\psi}_s(X_s))$ with $\boldsymbol{\psi}_j(X_j) = (\psi_{j1}(X_j), \dots, \psi_{jd_n}(X_j))$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jd_n})^T$, $j = 1, \dots, s$.

For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. So, we have

$$\begin{aligned} &\ell_p\{\alpha_{s'}(\mathbf{X}_{s'})\} - \ell_p\{\alpha_{s'}^*(\mathbf{X}_{s'})\} \\ &= \ell_p\{\Psi_{s'}(\mathbf{X}_{s'})\beta_{s'} + \rho_{s'}(\mathbf{X}_{s'})\} - \ell_p\{\Psi_{s'}(\mathbf{X}_{s'})\tilde{\beta}_{s'}^* + \rho_{s'}^*(\mathbf{X}_{s'})\} \\ &= \ell_p\{\Psi_{s'}(\mathbf{X}_{s'})\beta_{s'}\} + \ell'_p\{\Psi_{s'}(\mathbf{X}_{s'})\tilde{\beta}_{s'}\} \rho_{s'}(U) - \ell_p\{\Psi_{s'}(\mathbf{X}_{s'})\beta_{s'}^*\} - \ell'_p\{\Psi_{s'}(\mathbf{X}_{s'})\tilde{\beta}_{s'}^*\} \rho_{s'}^*(\mathbf{X}_{s'}), \end{aligned}$$

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

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

Thus, we have

$$\ell_p\{\alpha_{s'}(\mathbf{X}_{s'})\} - \ell_p\{\alpha_{s'}^*(\mathbf{X}_{s'})\} = \Delta_1 + \Delta_2 + \Delta_3.$$

For Δ_2 , by the Cauchy–Schwarz inequality, we have

$$|E[\Delta_2]| = |E[\ell'_p(\tilde{\beta}_{s'})\rho_{s'}(\mathbf{X}_{s'})]| \leq \sqrt{E\|\ell'_p(\tilde{\beta}_{s'})\|^2} \sqrt{E\|\rho_{s'}(\mathbf{X}_{s'})\|^2}.$$

By condition (C5) and Corollary 1 in [23], we can obtain $\Delta_2 = o_p(1)$. Similarly Δ_3 , we can also obtain, $\Delta_3 = o_p(1)$.

Next, we consider term Δ_1 . For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. Under (C3) condition,

we consider $\beta_{s'}$ close to $\beta_{s'}^*$ such that $\|\beta_{s'} - \beta_{s'}^*\| = w_1 d_n n^{-\tau_1}$ for some $w_1, \tau_1 > 0$. Clearly, when n is sufficiently large, $\beta_{s'}$ falls into a small neighborhood of $\beta_{s'}^*$, so that Condition (C5) becomes applicable. Thus, it follows Condition (C5) and the Cauchy–Schwarz inequality that

$$\begin{aligned} \ell_p(\beta_{s'}) - \ell_p(\beta_{s'}^*) &= [\beta_{s'} - \beta_{s'}^*]^T \ell'_p(\beta_{s'}^*) + (1/2) [\beta_{s'} - \beta_{s'}^*]^T \ell''_p(\tilde{\beta}_{s'}) [\beta_{s'} - \beta_{s'}^*] \\ &\leq [\beta_{s'} - \beta_{s'}^*]^T \ell'_p(\beta_{s'}^*) - (C_1 d_n^{-1} / 2) n \|\beta_{s'} - \beta_{s'}^*\|_2^2 \\ &\leq w_1 d_n n^{-\tau_1} \|\ell'_p(\beta_{s'}^*)\|_2 - (C_1 d_n / 2) w_1^2 n^{1-2\tau_1}, \end{aligned}$$

(A3)

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

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

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

$$\begin{aligned}
\Pr\{\ell'_j(\beta_{s'}^*) \geq (2m)^{-1/2}(C_1 w_1 / 2)n^{1-\tau_1}\} &\leq \Pr\{\ell'_j(\beta_{s'}^*) \geq (2w_2 n^{\tau_2})^{-1/2}(C_1 w_1 / 2)n^{1-\tau_1}\} \\
&= \Pr\left\{\ell'_j(\beta_{s'}^*) \geq cn^{1-\tau_1-0.5\tau_2}\right\} \\
&= \Pr\left\{\ell'_j(\beta_{s'}^*) \geq nc n^{-\tau_1-0.5\tau_2}\right\},
\end{aligned}$$

(A4)

where $c = C_1 w_1 / (2\sqrt{2w_2})$ denotes some generic positive constant. Recall (3), by differentiation and rearrangement of terms, it can be shown as in [22] that the gradient of $\ell'_p(\beta)$ is

$$\ell'_p(\beta) \equiv \frac{\partial \ell_p(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^n \int_0^\infty [\mathbf{Z}_i - \bar{\mathbf{Z}}_n(\beta, t)] dN_i(t).$$

(A5)

where $\bar{\mathbf{Z}}_n(\beta, t) = \sum_{i \in R_j} \mathbf{Z}_i \exp(\mathbf{Z}_i^T \beta) / \sum_{i \in R_j} \exp(\mathbf{Z}_i^T \beta)$. As a result, the partial score function $\ell'_p(\beta)$ no longer has a martingale structure, and the large deviation results for continuous time martingale in [4,8] are not directly applicable. The martingale process associated with $N_\lambda(t)$ is given by $M_i(t) = N_i(t) - \int_0^t R_i(u) \exp(\mathbf{Z}_i^T \beta^*) d\Lambda_0(u)$.

Let t_j be the time of the j th jump of the process $\sum_{i=1}^n \int_0^\infty R_i(t) dN_i(t)$, $j = 1, \dots, N$ and $t_0 = 0$. Then, t_j are stopping times. For $j = 0, 1, \dots, N$, define

$$Q_j = \sum_{i=1}^n \int_0^{t_j} b_i(u) dN_i(u) = \sum_{i=1}^n \int_0^{t_j} b_i(u) dM_i(u),$$

(A6)

where $b_i(u) = \mathbf{Z}_i - \bar{\mathbf{Z}}_n(\beta, u)$, $i = 1, \dots, n$ are predictable, under no two component processes jumping at the same time and (D6), and satisfy $|b_\lambda(u)| \leq 1$.

Since $M_\lambda(u)$ are martingales and $b_\lambda(u)$ are predictable, $\{Q_j, j = 0, 1, \dots\}$ is a martingale with the difference $|Q_j - Q_{j-1}| \leq \max_{u,i} |b_\lambda(u)| \leq 1$. Recall definition of N in Section 2, we define

$C_0^2 n \leq N$, where C_0 is a constant. So, by the martingale version of Hoeffding's inequality [24] and under Condition (D7), we have

$$\Pr(|Q_N| > nC_0x) \leq 2 \exp\{-n^2 C_0^2 x^2 / (2N)\} \leq 2 \exp(-nx^2 / 2).$$

(A7)

By Equation (A6), $Q_N = n\ell'_p(\beta)$ if and only if $\sum_{i=1}^n R_i(t) dN_i(t) \leq N$. Thus, the left-hand side of (3.15) in Lemma 3.3 of Huang et al. [8] is no greater than $\Pr(|Q_N| > nC_0x) \leq 2 \exp(-nx^2/2)$.

Now Equation (A4) can be rewritten as follows:

$$\Pr\left\{\ell'_j(\beta_{s'}^*) \geq ncn^{-\tau_1 - 0.5\tau_2}\right\} \leq \exp\{-0.5nn^{-2\tau_1 - \tau_2}\} = \exp\{-0.5n^{1-2\tau_1 - \tau_2}\}$$

(A8)

By the same arguments, we have

$$\Pr\{\ell'_j(\beta_{s'}^*) \leq -m^{-1/2}(C_1 w_1 / 2)n^{1-\tau_1}\} \leq \exp\{-0.5n^{1-2\tau_1 - \tau_2}\}.$$

(A9)

The inequalities (A8) and (A9) imply that

$$\Pr\{\ell_p(\beta_{s'}) \geq \ell_p(\beta_{s'}^*)\} \leq 4m \exp\{-0.5n^{1-2\tau_1 - \tau_2}\}.$$

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

$$\begin{aligned}
\Pr\left\{\max_{s \in S_-^m} \ell_p(\beta_{s'}) \geq \ell_p(\beta_{s'}^*)\right\} &\leq \sum_{s \in S_-^m} \Pr\{\ell_p(\beta_{s'}) \geq \ell_p(\beta_{s'}^*)\} \\
&\leq 4mp^m \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4 \exp\{\log m + m \log p - 0.5n^{1-2\tau_1-\tau_2}\} \\
&\leq 4 \exp\{\log w_2 + \tau_2 \log n + w_2 n^{\tau_2} \tilde{c} n^\kappa - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4w_2 \exp\{\tau_2 \log n + w_2 n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= a_1 \exp\{\tau_2 \log n + a_2 n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= o(1) \quad \text{as } n \rightarrow \infty
\end{aligned}$$

(A10)

for some generic positive constants $a_1 = 4w_2$ and $a_2 = w_2 \tilde{c}$. By Condition (C5), $\ell_p(\beta_s)$ is concave in β_s ; Equation (A10) holds for any $\beta_{s'}$ such that $\|\beta_{s'} - \beta_{s'}^*\| = w_1 d_n n^{-\tau_1}$.

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

$$\begin{aligned}
\|\check{\beta}_{s'} - \beta_{s'}^*\|_2 &= \|\check{\beta}_{s^* \cup (s'/s^*)} - \beta_{s^* \cup (s'/s^*)}^*\|_2 = \|\check{\beta}_{s^* \cup (s'/s^*)} - \beta_{s^*}^*\|_2 \geq \|\beta_{s^* \cup (s'/s^*)}^* - \beta_{s^*}^*\|_2 \geq \\
\|\beta_{s'/s^*}^*\|_2 &= w_1 d_n n^{-\tau_1}
\end{aligned}$$

. consequently,

$$\Pr\left\{\max_{s \in S_-^m} \ell_p(\hat{\beta}_s) \geq \min_{s \in S_+^m} \ell_p(\hat{\beta}_s)\right\} \leq \Pr\left\{\max_{s \in S_-^m} \ell_p(\check{\beta}_{s'}) \geq \ell_p(\beta_{s'}^*)\right\} = o(1).$$

So, we have shown that

$$\Pr \left[\max_{s \in S_-^m} \ell\{\hat{\alpha}_s(\mathbf{X}_s)\} \geq \min_{s \in S_+^m} \ell\{\hat{\alpha}_s(\mathbf{X}_s)\} \right] \rightarrow 0,$$

as $n \rightarrow \infty$. The theorem is proved. ■

References

- [1]. Fan J, Lv J. Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion). *J R Stat Soc Ser B*. 2008;70:849–911.
- [2]. Cox DR. Regression models and life tables (with Discussion). *J R Stat Soc Ser B*. 1972;34:187–220.
- [3]. Cox DR. Partial likelihood. *Biometrika*. 1975;62:269–276.
- [4]. Bradic J, Fan J, Jiang J. Regularization for Cox’s proportional hazards model with NP-dimensionality. *Ann Statist*. 2011;39:3092–3120.
- [5]. Du P, Ma S, Liang H. Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann Statist*. 2010;38:2092–2117.
- [6]. Leng C, Zhang HH. The L2 rate of convergence for event history regression with time-dependent covariates. *J Nonparametr Stat*. 2006;18:417–429.
- [7]. Zhang HH, Lu W. Adaptive-Lasso for Cox’s proportional hazards model. *Biometrika*. 2007;94:1–13.
- [8]. Huang J, Sun T, Ying Z, et al. Oracle Inequalities for the Lasso in the Cox Model. *Ann Statist*. 2013;41:1142–1165.
- [9]. Lian H, Li J, Hu Y. Shrinkage variable selection and estimation in proportional hazards models with additive structure and high dimensionality. *Comput Stat Data Anal*. 2013;63:99–112.
- [10]. Zhang S, Wang L, Lian H. Estimation by polynomial splines with variable selection in additive Cox models. *Statistics*. 2014;48:67–80.
- [11]. Lin H, He Y, Huang J. A global partial likelihood estimation in the additive Cox proportional hazards model. *J Stat Plan Inference*. 2016;169:71–87.
- [12]. Cui X, Peng H, Wen S, et al. Component selection in the additive regression model. *Scand J Statist*. 2013;40:491–510.
- [13]. Huang J, Horowitz JL, Wei F. Variable Selection in Nonparametric Additive Models. *Ann Statist*. 2010;38:2282–2313.
- [14]. Meier L, Van De Geer SA, Bühlmann P. High-dimensional additive modeling. *Ann Statist*. 2009;37:3779–3821.
- [15]. Ravikumar P, Liu H, Lafferty J, et al. Sparse additive models. *J R Stat Soc Ser B*. 2009;71:1009–1030.
- [16]. Xue L. Consistent variable selection in additive models. *Stat Sin*. 2009;19:1281–1296.
- [17]. Fan J, Feng Y, Song R. Nonparametric Independence Screening in Sparse Ultra-high Dimensional Additive Models. *J Amer Stat Assoc*. 2011;116:544–557.
- [18]. de Boor C. A practical guide to splines. Springer; 1978.
- [19]. Stone CJ. Optimal global rates of convergence for nonparametric regression. *Ann Statist*. 1982;10:1040–1053.
- [20]. Stone CJ. Additive regression and other nonparametric models. *Ann Statist*. 1985;13:689–705.
- [21]. Lenz G, Wright G, Dave S, et al. Stromal gene signatures in large-b-cell lymphomas. *N Engl J Med*. 2008;359:2313–2323. [PubMed: 19038878]
- [22]. Andersen PK, Gill RD. Cox’s regression model for counting processes: a large sample study. *Ann Statist*. 1982;10:1100–1120.
- [23]. Wei F, Huang J, Li H. Variable selection and estimation in high dimensional varying-coefficient models. *Stat Sin*. 2011;21:1515–1540. [PubMed: 24478564]

- [24]. Azuma K Weighted sums of certain dependent random variables. Tohoku Math J. 1967;19:357–367.

Figure 1.
The six additive functions of the selected genes.

Table 1.

Censoring rates for model 1.

	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
Σ_1	0.155	0.235	0.302
Σ_2	0.157	0.245	0.321

Table 2.

Simulation results for Model 1 with AR(1) covariance matrix Σ_1 . P_i and P_a are the selected proportion of X_i and all the significant variables respectively. Iter is the average iteration number, and Time (s) is the computing time with seconds.

ρ	$n = 300$							$n = 500$						
	P_1	P_2	P_3	P_4	P_a	Iter	Time (s)	P_1	P_2	P_3	P_4	P_a	Iter	Time (s)
$p = 1000$														
0.25	1.00	1.00	1.00	1.00	1.00	3.0	19.4	1.00	1.00	1.00	1.00	1.00	2.1	41.9
0.50	0.99	0.98	0.99	1.00	0.97	3.5	21.8	1.00	1.00	1.00	1.00	1.00	2.5	46.4
0.75	0.94	0.93	0.92	0.99	0.85	3.5	21.1	1.00	0.99	1.00	0.99	0.98	3.4	60.8
$p = 2000$														
0.25	1.00	1.00	1.00	1.00	1.00	3.0	117.8	1.00	1.00	1.00	1.00	1.00	2.2	190.6
0.50	0.97	0.98	0.98	1.00	0.95	3.6	141.0	1.00	1.00	1.00	1.00	1.00	2.6	214.7
0.75	0.86	0.85	0.84	0.99	0.71	3.4	124.4	0.99	0.99	0.98	0.96	0.94	3.4	275.3

Table 3.

Simulation results for Model 1 with compound symmetry covariance matrix Σ_2 . P_i and P_a are the selected proportion of X_i and all the significant variables respectively. Iter is the average iteration number, and Time (s) is the computing time with seconds.

ρ	$n = 300$							$n = 500$						
	P_1	P_2	P_3	P_4	P_a	Iter	Time(s)	P_1	P_2	P_3	P_4	P_a	Iter	Time(s)
$p = 1000$														
0.25	1.00	1.00	1.00	1.00	1.00	2.6	16.8	1.00	1.00	1.00	1.00	1.00	2.0	38.9
0.50	1.00	1.00	1.00	1.00	1.00	2.3	14.7	1.00	1.00	1.00	1.00	1.00	2.0	37.3
0.75	1.00	1.00	0.99	1.00	0.99	2.7	16.5	1.00	1.00	1.00	1.00	1.00	2.5	43.6
$p = 2000$														
0.25	1.00	1.00	1.00	1.00	1.00	2.7	75.0	1.00	1.00	1.00	1.00	1.00	2.0	158.5
0.50	1.00	1.00	1.00	1.00	1.00	2.5	67.9	1.00	1.00	1.00	1.00	1.00	2.0	152.1
0.75	1.00	1.00	0.96	1.00	0.96	2.8	72.5	1.00	1.00	1.00	1.00	1.00	2.6	185.8

Table 4.

Censoring rates for Model 2.

	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
Σ_1	0.193	0.247	0.286
Σ_2	0.195	0.262	0.311

Table 5.

Simulation results for Model 2 with AR(1) covariance matrix Σ_1 . P_i and P_a are the selected proportion of X_i and all the significant variables, respectively, Iter is the average iteration number, and Time (s) is the computing time with seconds.

ρ	$n = 300$							$n = 500$						
	P_1	P_2	P_3	P_4	P_a	Iter	Time(s)	P_1	P_2	P_3	P_4	P_a	Iter	Time(s)
$p = 1000$														
0.25	1.00	1.00	1.00	1.00	1.00	2.7	10.7	1.00	1.00	1.00	1.00	1.00	3.0	32.9
0.50	0.99	0.97	0.98	1.00	0.95	3.3	12.5	1.00	1.00	1.00	1.00	1.00	3.1	33.6
0.75	0.89	0.78	0.90	0.95	0.69	2.9	10.5	1.00	1.00	1.00	0.93	0.93	3.0	31.6
$p = 2000$														
0.25	1.00	1.00	1.00	0.99	0.99	2.7	45.0	1.00	1.00	1.00	1.00	1.00	3.0	133.2
0.50	0.98	0.93	0.97	1.00	0.90	3.3	52.5	1.00	1.00	1.00	1.00	1.00	3.2	138.3
0.75	0.82	0.67	0.80	0.95	0.54	2.8	44.7	1.00	0.99	1.00	0.94	0.94	3.0	130.7

Table 6.

Simulation results for Model 2 with compound symmetry covariance matrix Σ_2 . P_i and P_a are the selected proportion of X_i and all the significant variables, respectively, Iter is the average iteration number, and Time (s) is the computing time with seconds.

ρ	$n = 300$							$n = 500$						
	P_1	P_2	P_3	P_4	P_a	Iter	Time(s)	P_1	P_2	P_3	P_4	P_a	Iter	Time(s)
$p = 1000$														
0.25	1.00	1.00	1.00	1.00	1.00	2.7	10.4	1.00	1.00	1.00	1.00	1.00	2.6	30.4
0.50	1.00	1.00	0.97	1.00	0.97	3.1	11.7	1.00	1.00	1.00	1.00	1.00	2.1	24.0
0.75	1.00	0.94	0.86	1.00	0.82	3.2	11.7	1.00	1.00	1.00	1.00	1.00	2.3	25.1
$p = 2000$														
0.25	1.00	1.00	1.00	1.00	1.00	2.6	41.2	1.00	1.00	1.00	1.00	1.00	2.7	117.2
0.50	1.00	0.98	0.94	1.00	0.93	3.2	47.7	1.00	1.00	1.00	1.00	1.00	2.2	93.9
0.75	1.00	0.91	0.76	1.00	0.70	3.3	47.2	1.00	1.00	1.00	1.00	1.00	2.4	99.6

Table 7.

The mean value of the partial likelihood fitted by the six selected genes plus the randomly added m genes. All the scenarios are repeated with 200 replications.

m	1	2	3	4	5
Partial Likelihood	-440.19	-438.23	-435.80	-433.97	-431.95
Increase	2.06	4.02	6.45	8.28	10.30