Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802211023957
journals.sagepub.com/home/smm
SAGE

Variable selection for a mark-specific additive hazards model using the adaptive LASSO

Dongxiao Han[1],*, Lianqiang Qu[2],*, Liuquan Sun[3,4] and Yanqing Sun[5]

Abstract

In HIV vaccine efficacy trials, mark-specific hazards models have important applications and can be used to evaluate the strain-specific vaccine efficacy. Additive hazards models have been widely used in practice, especially when continuous covariates are present. In this article, we conduct variable selection for a mark-specific additive hazards model. The proposed method is based on an estimating equation with the first derivative of the adaptive LASSO penalty function. The asymptotic properties of the resulting estimators are established. The finite sample behavior of the proposed estimators is evaluated through simulation studies, and an application to a dataset from the first HIV vaccine efficacy trial is provided.

Keywords

Adaptive LASSO, additive hazards model, competing risks, continuous mark, mark-specific vaccine effects, survival data

1 Introduction

In preventive HIV vaccine efficacy trials, the trial population is exposed to many HIV genotypes but the vaccine contains only a few, and the vaccine may only provide protection for HIV strains genetically similar to the HIV virus or viruses represented in the vaccine. The similarity between the infecting virus and the virus contained in the vaccine construct can be measured by the genetic distance (or mark), which is defined as the weighted percent mismatch of amino acids between two aligned HIV sequences. Due to the extensive genetic diversity of HIV, this distance may be unique for all infected subjects. Thus, it is natural to consider such distance as a continuous mark variable.

The cause-specific hazard function is a commonly used tool for the analysis of failure time data with finitely many competing risks. The mark-specific hazard function is an extension of the cause-specific hazard function defined in a competing risks setting, where the cause of failure is replaced by a continuous mark only observed at the failure time.[1–3] Recently, many statistical methods have been developed for the analysis of survival data with a

[1]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China
[2]School of Mathematics and Statistics, Central China Normal University, Hubei, China
[3]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[4]School of Economics and Statistics, Guangzhou University, Guangdong, China
[5]Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA

*Dongxiao Han and Lianqiang Qu are the cofirst authors of the article.

Corresponding author:
Liuquan Sun, Chinese Academy of Sciences, No. 55, Zhongguancun East Road, Haidian District, Beijing 100190, P. R. China.
Email: slq@amt.ac.cn

continuous mark variable with important applications in HIV vaccine efficacy trials.[1–6] For example, Gilbert et al.[1] developed a statistical method to evaluate the dependence of the mark-specific hazard rate on the mark. Gilbert et al.[2] developed a statistical approach for assessing mark-specific HIV vaccine efficacy. This work was extended in several studies[3–5] to assess the mark-specific HIV vaccine efficacy adjusting for covariates under the mark-specific proportional hazards model. Han et al.[6] presented estimation and hypothesis testing methods for the mark-specific additive hazards model with a univariate continuous mark.

In many biomedical applications, identifying potential risk factors out of many available covariates is of scientific interest. An illustrative example is from a preventive vaccine efficacy trial. There were 5403 HIV-negative subjects enrolled in a 36-month randomized trial. Subjects were randomly assigned to receive either a recombinant glycoprotein 120 vaccine (AIDSVAX) or placebo in a 2:1 ratio and were monitored for HIV infection. During the trial, 368 individuals were infected with HIV, but 32 individuals had missing marks. Each of the remaining 336 samples (217 vaccine and 119 placebo samples) had a unique mark. The dataset includes covariates such as treatment indicator, age at enrollment, sex, region, race, country, education, and behavioral risk score. Our goal is to identify variables that are related to the mark-specific risk of infection. Sparse estimation via regularization or penalization is a popular variable selection method with many advantages. Commonly used penalized methods include least absolute shrinkage and selection operator (LASSO),[7] Smoothly clipped absolute deviation penalty (SCAD),[8] adaptive LASSO (ALASSO),[9] and Minimax concave penalty (MCP).[10] Some of these methods have been extended to deal with varying coefficient models.[11–17] Motivated by the HIV vaccine efficacy trial, we develop a variable selection method to identify the risk factors under the mark-specific additive hazards model. To the best of our knowledge, no study has been conducted for variable selection under the mark-specific additive hazards model in the literature.

In this article, we conduct variable selection for a mark-specific additive hazards model via penalized estimating functions. Specifically, let $\lambda(t,\nu|z)$ be the conditional mark-specific hazard function, which is defined as

$$\lambda(t,\nu|z) = \lim_{h_1,h_2\to 0} P\{T\in[t,t+h_1), V\in[\nu,\nu+h_2)|T\geq t, Z=z\}/(h_1h_2),$$

where T is the failure time, V is a continuous mark variable standardized on the interval $[0,1]$, and Z is a p-dimensional covariate vector. We consider the following mark-specific additive hazards model:[6]

$$\lambda(t,\nu|z) = \lambda_0(t,\nu) + \boldsymbol{\beta_0}(\nu)^T z \tag{1}$$

where the baseline hazard function $\lambda_0(t,\nu)$ is an unknown function of t and ν, and $\boldsymbol{\beta_0}(\nu) = (\beta_{01}(\nu),\ldots,\beta_{0p}(\nu))^T$ is a p-dimensional vector, in which each element of $\boldsymbol{\beta_0}(\nu)$ is a one-dimensional unknown continuous function of ν. Our proposed method is based on an estimating equation with the first derivative of the ALASSO penalty function.

The rest of the article is organized as follows. Section 2.1 presents a variable selection procedure for model (1) via a penalized estimating function. The asymptotic properties of the proposed estimators are established in Section 2.2. A practical implementation of the procedure is discussed in Section 2.3. Section 3 reports results of simulation studies conducted for evaluating the proposed method. An application to a dataset from the first HIV vaccine efficacy trial is provided in Section 4, and some concluding remarks are made in Section 5. All proofs are given in the Appendix.

2 Estimation and inference

2.1 Estimation procedure

Suppose that the support of the mark variable V is taken to be $[0,1]$. Let C be the censoring time that is assumed to be conditionally independent of (T, V) given Z. Also let $X = \min(T,C)$ be the event time and $\delta = I(T \leq C)$ be the censoring indicator. The observations $(X_i, \delta_i, \delta_i V_i, Z_i)$ $(i = 1,\ldots,n)$ are assumed to be independent replicates of $(X, \delta, \delta V, Z)$. The mark V can be observed when $\delta = 1$, whereas it is undefined and is not meaningful when $\delta = 0$.

Let $Y_i(t) = I(X_i \ge t)$, $N_i(t, \nu) = I(X_i \le t, \delta_i = 1, V_i \le \nu)$, $S^{(0)}(t) = n^{-1}\sum_{i=1}^{n} Y_i(t)$, $S^{(1)}(t) = n^{-1}\sum_{i=1}^{n} Y_i(t)Z_i$, and $\overline{Z}(t) = S^{(1)}(t)/S^{(0)}(t)$. As discussed in Han et al.,[6] we can use the local constant method to construct the estimating equation for $\boldsymbol{\beta}$=$\boldsymbol{\beta}(\nu)$ at a fixed $\nu \in (0, 1)$:

$$U(\nu, \boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^1 \int_0^\tau K_h(u - \nu)\{Z_i - \overline{Z}(t)\} \times [N_i(dt, du) - Y_i(t)\boldsymbol{\beta}^T Z_i \mathrm{d}t\mathrm{d}u] \tag{2}$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a kernel function, $h = h_n$ is a bandwidth, and τ is the end of follow-up time. The kernel function $K(x)$ is a nonnegative real-valued integrable function used in nonparametric estimation techniques. For most applications, the kernel function is usually assumed to be a symmetric probability density function with a compact support. The bandwidth h_n is a sequence of positive numbers satisfying $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$.

We propose the following penalized estimating function[18] for variable selection:

$$U^P(\nu, \boldsymbol{\beta}) = U(\nu, \boldsymbol{\beta}) - n(q_{\lambda_\nu,\nu,1}(|\beta_1|)\mathrm{sgn}(\beta_1), \cdots, q_{\lambda_\nu,\nu,p}(|\beta_p|)\mathrm{sgn}(\beta_p))^T \tag{3}$$

where β_j is the jth element of $\boldsymbol{\beta}$, $q_{\lambda_\nu,\nu,j}(\theta) = dp_{\lambda_\nu,\nu,j}(\theta)/d\theta$, and $p_{\lambda_\nu,\nu,j}(\theta)$ is a penalty function, $j = 1, \ldots, p$. There are many possible choices for the penalty function $p_{\lambda_\nu,\nu,j}(\theta)$. The ALASSO penalty is defined as $p_{\lambda_\nu,\nu,j}(\theta) = \lambda_\nu \theta \omega_{\nu,j}$, where $\omega_{\nu,j}$ is a known data-driven weight and λ_ν is a tuning parameter. The LASSO penalty is given by $p_{\lambda_\nu,\nu,j}(\theta) = \lambda_\nu \theta$, which is a special case of the ALASSO penalty with $\omega_{\nu,j} = 1$. The LASSO may result in biased estimates for the large coefficients and inconsistent variable selection results,[9,19] while the ALASSO uses different weights for different coefficients and enjoys the oracle properties. In what follows, we will focus on the ALASSO penalty, where $q_{\lambda_\nu,\nu,j}(\theta) = \lambda_\nu \omega_{\nu,j}$. Since $q_{\lambda_\nu,\nu,j}(\theta)$ does not depend on θ, we denote it by $q_{\lambda_\nu,\nu,j}$ for simplicity. As discussed in Zou,[9] $\omega_{\nu,j}$ is usually taken as some function of a consistent estimator of $\boldsymbol{\beta_0}(\nu)$. For example, in our simulations and application, we set $\omega_{\nu,j} = 1/|\overline{\beta}_j(\nu)|$, where $\overline{\boldsymbol{\beta}}(\nu) = (\overline{\beta}_1(\nu), \ldots, \overline{\beta}_p(\nu))^T$ is the solution to the estimation equation $U(v, \boldsymbol{\beta}) = 0$ in (2). Since $\overline{\boldsymbol{\beta}}(\nu)$ is a consistent estimator of $\boldsymbol{\beta_0}(\nu)$ under some mild conditions,[6] the ALASSO penalty function depends on $\boldsymbol{\beta_0}(\nu)$ implicitly. In addition, the tuning parameter λ_ν is not user-specified and needs to be tuned by some commonly used criteria, such as the cross-validation criterion and the Bayesian information criterion (BIC)-type criterion. For any given ν and any tuning parameter λ_ν, one can estimate $\boldsymbol{\beta_0}(\nu)$ by $\hat{\boldsymbol{\beta}}(\nu)$ defined as the solution to the equation $U^P(\nu, \boldsymbol{\beta}) = 0$.

The number of significant variables (i.e. the corresponding regression coefficients are estimated to be not zero) as a function of ν, if not constant, will be discontinuous. This will lead to discontinuous estimates of coefficient functions and does not produce parsimonious and appealing models. In order to decide whether a covariate should be retained in the final model, we adopt a voting rule[11]: If a mark-specific covariate effect is estimated as zero over a certain percentage of grid points, then the corresponding covariate is regarded as unimportant and eliminated from model (1). For example, denote $\{\nu_k, k = 1, \ldots, 100\}$ as the equal grid points on $[0.1, 0.9]$, at which the jth covariate effect $\beta_{0j}(\nu)$ is estimated, and the percentage is taken as 50%. If at least 50 elements in the set $\{\beta_{0j}(\nu_k)|k = 1, \ldots, 100\}$ are estimated to be zero, then the corresponding jth covariate is eliminated from model (1). In our simulations and application, the voting rates are 40% and 50%.

2.2 Asymptotic properties

We present the asymptotic properties of the proposed estimators. Without loss of generality, assume that $\beta_{0j}(\nu) \neq 0$ for $1 \le j \le s$ and $\beta_{0j}(\nu) = 0$ for $s < j \le p$. Define the true value $\boldsymbol{\beta_0}(\nu)$=$(\boldsymbol{\beta_1}(\nu)^T, \boldsymbol{\beta_2}(\nu)^T)^T$, where $\boldsymbol{\beta_1}(\nu) = (\beta_{01}(\nu), \ldots, \beta_{0s}(\nu))^T$ and $\boldsymbol{\beta_2}(\nu)$=$(\beta_{0s+1}(\nu), \ldots, \beta_{0p}(\nu))^T$. Correspondingly, $\hat{\boldsymbol{\beta}}(\nu) = (\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\nu)^T, \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\nu)^T)^T$. To accommodate the discrete estimating function (3), we provide a formal definition of the solution to (3). An estimator $\hat{\boldsymbol{\beta}}(\nu) = (\hat{\beta}_1(\nu), \ldots, \hat{\beta}_p(\nu))^T$ is called an approximate zero-crossing[18] if for $j = 1, \ldots, p$,

$$\varlimsup_{n\to\infty} \varlimsup_{\eta\to 0+} n^{-1} h U_j^P(\nu, \hat{\boldsymbol{\beta}}(\nu) + \eta \boldsymbol{e}_j) U_j^P(\nu, \hat{\boldsymbol{\beta}}(\nu) - \eta \boldsymbol{e}_j) \le 0,$$

where $U_j^P(\nu, \cdot)$ is the jth component of $U^P(\nu, \cdot)$ and $\boldsymbol{e}_j$ is the jth canonical unit vector.

We summarize the asymptotic properties of $\hat{\boldsymbol{\beta}}(\nu)$ in the following theorems with the proof in the Appendix.

Theorem 1. *Under conditions (C1)–(C6) stated in the Appendix, (3) has an approximate zero-crossing solution* $\hat{\boldsymbol{\beta}}(\nu)$ *such that* $||\hat{\boldsymbol{\beta}}(\nu) - \boldsymbol{\beta_0}(\nu)|| = O_p(n^{-1/2}h^{-1/2})$ *for* $\nu \in [a, b] \subset (0, 1)$, *where* $|| \cdot ||$ *denotes the Euclidean norm.*

Theorem 2. *Under conditions (C1)–(C6) stated in the Appendix, for any root-nh-consistent approximate zero-crossing solution of* $U^P(\nu, \cdot)$, *denoted by* $\hat{\boldsymbol{\beta}}(\nu)$, *the following properties hold:*

(i) $\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\nu) = 0$ with probability tending to 1 for $\nu \in [a, b]$.

(ii) $(nh)^{1/2}(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\nu) - \boldsymbol{\beta_1}(\nu))$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mu_0 A_{11}^{-1}\Sigma_{11}(\nu)A_{11}^{-1}$ for $\nu \in [a, b]$, where A_{11} and $\Sigma_{11}(\nu)$ are the first $s \times s$ submatrices of A and $\Sigma(\nu)$, respectively, with

$$A = E\left[\int_0^\tau Y_i(t)\{Z_i - \bar{z}(t)\}^{\otimes 2}\mathrm{d}t\right],$$

$$\Sigma(\nu) = E\left[\int_0^\tau \{Z_i - \bar{z}(t)\}^{\otimes 2} Y_i(t)\{\lambda_0(t, \nu) + \boldsymbol{\beta_0}(\nu)^T Z_i\}\mathrm{d}t\right],$$

$\bar{z}(t)$ is the limit of $\overline{Z}(t)$ and $\mu_0 = \int K^2(u)du$.

The asymptotic variance of $(nh)^{1/2}\{\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\nu) - \boldsymbol{\beta_1}(\nu)\}$ can be consistently estimated by $\hat{A}_{11}^{-1}\hat{\Sigma}_{11}(\nu)\hat{A}_{11}^{-1}$, where $\hat{A}_{11}$ and $\hat{\Sigma}_{11}(\nu)$ are the first $s \times s$ submatrices of $\hat{A}$ and $\hat{\Sigma}(\nu)$, respectively, with

$$\hat{\Sigma}(\nu) = \frac{h}{n}\sum_{i=1}^n \int_0^1 \int_0^\tau (K_h(u - \nu))^2\{Z_i - \overline{Z}(t)\}^{\otimes 2} N_i(\mathrm{d}t, \mathrm{d}u),$$

$$\hat{A} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau Y_i(t)\{Z_i - \overline{Z}(t)\}^{\otimes 2}\mathrm{d}t.$$

2.3 Implementation

In this section, we discuss the computational issues. Johnson et al.[18] suggested local quadratic approximations to obtain the solution of (3). However, this algorithm does not give exact zeros for some coefficients. In what follows, we modified the shooting algorithm[20,21] to obtain the solution to (3). For any fixed $\nu \in [a, b]$, define

$$d_n = n^{-1}\sum_{i=1}^n \int_0^1 \int_0^\tau K_h(u - \nu)\{Z_i - \overline{Z}(t)\}N_i(\mathrm{d}t, \mathrm{d}u).$$

Denote $\boldsymbol{\beta}_{-j}$ as a $(p - 1)$-dimensional vector consisting of the $\boldsymbol{\beta}$'s other than β_j. Let $G_j(\beta_j, \boldsymbol{\beta}_{-j}) = d_{nj} - A_{nj}\boldsymbol{\beta}$, where d_{nj} is the jth component of d_n and A_{nj} is the jth row of $\hat{A}$. Our implementation is based on the following iterative algorithm:

Step 0. Set $\overline{\boldsymbol{\beta}}(\nu)$ as the initial estimator.

Step 1. Suppose that $\boldsymbol{\beta}^{(m-1)}$ has been obtained at the $(m - 1)$th iterative stage. At the mth iterative stage, for each j, let $G_0 = G_j(0, \boldsymbol{\beta}_{-j}^{(m-1)})$ and set

$$\beta_j^{(m)} = \begin{cases} \dfrac{\lambda_\nu\omega_{\nu,j} - G_0}{a_{jj}} & \text{if } G_0 > \lambda_\nu\omega_{\nu,j}, \\ \dfrac{-\lambda_\nu\omega_{\nu,j} - G_0}{a_{jj}} & \text{if } G_0 < -\lambda_\nu\omega_{\nu,j}, \\ 0 & \text{if } |G_0| \leq \lambda_\nu\omega_{\nu,j}, \end{cases}$$

where a_{jj} is the jth diagonal component of $\hat{A}$.

Step 2. Repeat Step 1 until the convergence criterion is met.

In our simulations and application, we set $\omega_{\nu,j} = 1/|\overline{\beta}_j(\nu)|, j = 1, \ldots, p$. To stress the dependence of the estimator on the tuning parameter λ_ν, we use $\hat{\boldsymbol{\beta}}_{\lambda_\nu}(\nu)$ to replace $\hat{\boldsymbol{\beta}}(\nu)$. For any fixed $\nu \in [a, b]$, we use the BIC-type criterion[22] to select λ_ν :

$$\mathrm{BIC}(\lambda_\nu) = (\hat{\boldsymbol{\beta}}_{\lambda_\nu}(\nu) - \overline{\boldsymbol{\beta}}(\nu))^T \hat{A} (\hat{\boldsymbol{\beta}}_{\lambda_\nu}(\nu) - \overline{\boldsymbol{\beta}}(\nu)) + df(\lambda_\nu) \times \frac{\log(nh)}{nh},$$

where $df(\lambda_\nu)$ is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_{\lambda_\nu}$. The final tuning parameter is chosen to minimize $\mathrm{BIC}(\lambda_\nu)$, that is,

$$\hat{\lambda}_\nu = \mathrm{argmin}_{\lambda_\nu}\{\mathrm{BIC}(\lambda_\nu)\}.$$

Several criteria can be used to check the convergence. In the simulation studies below, we used the absolute differences $\leq 10^{-3}$ between the iterative estimates of the parameters.

3 Simulation studies

In this section, we conducted simulation studies to examine the finite sample performance of the proposed method using the following mark-specific additive hazards model:

$$\lambda(t, \nu|z) = \lambda_0(t, \nu) + \sum_{i=1}^{6} z_i \beta_{0i}(\nu), \quad t \geq 0, \quad 0 \leq \nu \leq 1 \tag{4}$$

The covariates z_2, z_3, z_5, and z_6 were independently sampled from a uniform distribution on (0, 1), and z_1 and z_4 were independently sampled from a Bernoulli distribution with success probability 0.5. Under model (4), $z = (z_1, \ldots, z_6)^T$ and $\boldsymbol{\beta_0}(\nu) = (\beta_{01}(\nu), \ldots, \beta_{06}(\nu))^T$.

By some calculation, we can obtain that

$$F(t|z) = 1 - \exp\left\{ -\int_0^t \int_0^1 \lambda(s, \nu|z) \mathrm{d}s \mathrm{d}\nu \right\},$$

where $F(t|z)$ is the cumulative distribution function of T given $Z = z$. Hence, we can generate the failure time T by the inverse cumulative distribution function method. It is easy to obtain that $f(\nu|t, z) = \lambda(t, \nu|z)/\lambda(t|z)$, where $f(\nu|t, z)$ is the conditional probability density function of V given $T = t$ and $Z = z$. Then, given T, we can generate V by the inverse cumulative distribution function method. We considered the following two models:

$$\text{Case 1.} \quad \beta_{01}(\nu) = 2\exp(\nu),\ \beta_{02}(\nu) = 2\nu + 4,\ \beta_{0\mathrm{j}}(\mathrm{v}) = 0\ (3 \leq \mathrm{j} \leq 6)$$

$$\text{and } \lambda_0(t, \nu) = 0.3\nu^2;$$

$$\text{Case 2.} \quad \beta_{01}(\nu) = 5\cos\{\pi(\nu - 0.5)\},\ \beta_{02}(\nu) = 3\nu^2 + 3,\ \beta_{0\mathrm{j}}(\mathrm{v}) = 0\ (3 \leq \mathrm{j} \leq 6)$$

$$\text{and } \lambda_0(t, \nu) = \exp(0.3\nu).$$

The censoring time was generated from a uniform distribution on $(0, c)$, where c was selected to give a censoring rate of 30%. For the analysis, the interval for ν was set as $[a, b] = [0.1, 0.9]$. The kernel function was chosen as the Epanechnikov kernel $K(x) = 0.75(1 - x^2)I(|x| \leq 1)$. The bandwidth was chosen using the formula $h = 1.6\hat{\sigma}_\nu n_0^{-1/4}$,[6] where n_0 is the average number of observed failure times and $\hat{\sigma}_\nu$ is the estimated standard error (SE) of the observed marks for uncensored failure times for each simulation setting. The result presented

Table 1. Simulation results for Case 1: MMSE and the average numbers of correct (Corr) and incorrect (Incorr) zero coefficients.

		VR= 40%			VR= 50%		
n	Method	Corr	Incorr	MMSE	Corr	Incorr	MMSE
500	LASSO	3.84	0.43	4.45	3.67	0.22	4.23
	ALASSO	3.89	0.28	4.03	3.63	0.12	4.20
	Oracle	4.00	0.00	2.56	4.00	0.00	2.56
800	LASSO	3.95	0.38	3.56	3.83	0.16	3.44
	ALASSO	3.99	0.15	2.88	3.93	0.04	2.88
	Oracle	4.00	0.00	1.87	4.00	0.00	1.87
1200	LASSO	3.99	0.45	3.65	3.97	0.20	3.23
	ALASSO	4.00	0.08	2.24	3.99	0.01	2.23
	Oracle	4.00	0.00	1.35	4.00	0.00	1.35

VR: voting rate; MMSE: median of mean squared errors; ALASSO: adaptive LASSO.

Table 2. Simulation results for Case 2: MMSE and the average numbers of correct (Corr) and incorrect (Incorr) zero coefficients.

		VR= 40%			VR= 50%		
n	Method	Corr	Incorr	MMSE	Corr	Incorr	MMSE
500	LASSO	3.88	0.29	4.83	3.67	0.13	4.97
	ALASSO	3.92	0.23	4.26	3.64	0.08	4.56
	Oracle	4.00	0.00	2.56	4.00	0.00	2.56
800	LASSO	3.96	0.17	4.18	3.87	0.06	4.24
	ALASSO	3.98	0.11	3.21	3.89	0.04	3.29
	Oracle	4.00	0.00	2.13	4.00	0.00	2.13
1200	LASSO	4.00	0.15	3.96	3.98	0.03	3.96
	ALASSO	4.00	0.04	2.47	3.98	0.01	2.45
	Oracle	4.00	0.00	1.49	4.00	0.00	1.49

VR: voting rate; MMSE: median of mean squared errors; ALASSO: adaptive LASSO.

below is based on 500 replications with sample sizes $n = 500$, 800, and 1200. The voting rates (denoted by VR) are 40% and 50%. The performance of the estimator $\hat{\boldsymbol{\beta}}(\cdot)$ is measured by the mean square error (MSE):

$$\text{MSE} = \frac{1}{100}\sum_{i=1}^{100} ||\hat{\boldsymbol{\beta}}(\nu_i) - \boldsymbol{\beta}_0(\nu_i)||_2^2,$$

where $\{\nu_i, i = 1, \ldots, 100\}$ are the equal grid points on $[0.1, 0.9]$ at which $\boldsymbol{\beta_0}(\cdot)$ is estimated.

In our simulation, we also considered the LASSO penalty,[7] that is, set the weight $\omega_{\nu,j} = 1$. In addition, we compared the performance of the ALASSO and the oracle as well, where the oracle pertains to the situation in which we know a priori which coefficients are nonzero. Tables 1 and 2 give the average numbers of regression coefficients that are correctly or incorrectly shrunk to 0, along with the median of mean squared errors (MMSE) for Cases 1 and 2, respectively. From Tables 1 and 2, we see that all of the average numbers of correct zero coefficients are close to 4 and those of incorrect zero coefficients are close to 0. This implies that both the two penalty functions can discover the right sparse representation of model (4), and they perform comparably to the oracle. In addition, since the LASSO shrinks the coefficients excessively, the ALASSO method outperforms the LASSO method in terms of MMSE. As expected, the oracle performs better than the ALASSO in terms of MMSE. From Tables 1 and 2, it can also be seen that the average numbers of correct zero coefficients with VR= 50% are smaller than those with VR= 40%. Such phenomenon also occurs for the average numbers of incorrect zero coefficients. This is because that if a coefficient is estimated as zero in the case of VR= 50%, then it must be estimated as zero in the case of VR= 40%. However, the MMSE values are compared for both the situations. All results become better as the sample size increases.

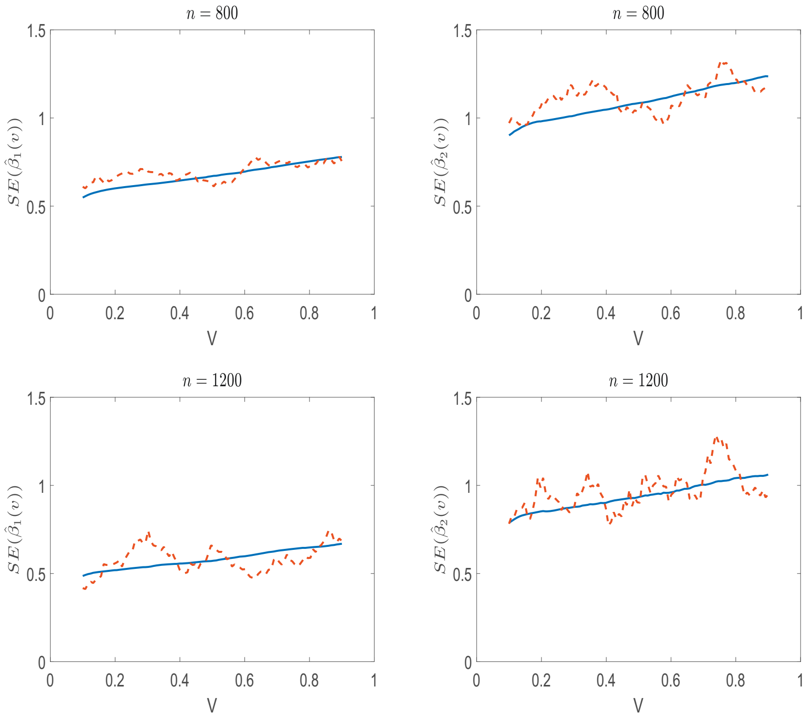

Figure 1. Plots of the estimated standard errors for Case 1 with $h = 1.6\hat{\sigma}_\nu n_0^{-1/4}$. The penalty function is the ALASSO penalty. The first row is for sample size $n = 800$, and the second row for sample size $n = 1200$. The dashed lines are the medians of the estimated standard errors of $\hat{\boldsymbol{\beta}}(\nu)$, while the solid lines are the sample median absolute deviations of $\hat{\boldsymbol{\beta}}(\nu)$ divided by 0.6745 based on 500 replications.
SE: standard error.

To test the accuracy of the proposed SE formula, we compared the median of the estimated SEs with the median absolute deviation of the estimated coefficients divided by 0.6745 among 500 simulations.[8] For Case 1, the results are given in Figures 1 and 2. Figure 3 depicts the results of the oracle estimator. Figures 1–3 suggest that the proposed SE formula performs well, especially when the sample size is large. In Figures 4 and 5, we further depict the estimated coefficient functions and their pointwise 95% confidence bands for the ALASSO and the oracle methods under Case 1. It can be seen that the estimated curves using the ALASSO are close to their true curves, the biases are negligible, and the confidence bands cover the entire true curves. Furthermore, all results in Figures 1–5 suggest that the performance of the ALASSO method is comparable to that of the oracle. The results for Case 2 are similar to those in Figures 1–5 and not reported.

Furthermore, we conducted simulation studies to examine the robustness of the proposed method to the choices of the kernel function, the bandwidth, and the voting rate. The results are reported in Tables A1–A4 and Figures A1–A8 of the supplement material. The simulation results show that the proposed method performs comparably well for the situations considered here, which suggests that our method is robust to the choices of the kernel function, the bandwidth, and the voting rate. Finally, we conducted simulation studies with unbalanced Bernoulli predictors when the baseline mortality hazard depends on time, and the covariates are dependent on each other. The results are also presented in Tables A1–A4 and Figures A1–A8 in the supplement material. Simulation results show that the proposed method still performs well in these settings.

4 Application

In this section, we applied the proposed method to a dataset from the HIV vaccine efficacy trial which was carried out in North America and The Netherlands. Sun et al.[3] and Han et al.[6] analyzed the same dataset. The vaccine was designed to protect subjects from HIV infection by stimulating high titer antibodies that neutralize exposing HIVs, and the HIV-gp120 region contains neutralizing epitopes that can prevent HIV infection by inducing anti-HIV antibody responses.[23] We defined the mark V as the percent mismatch of amino acids in the whole gp120

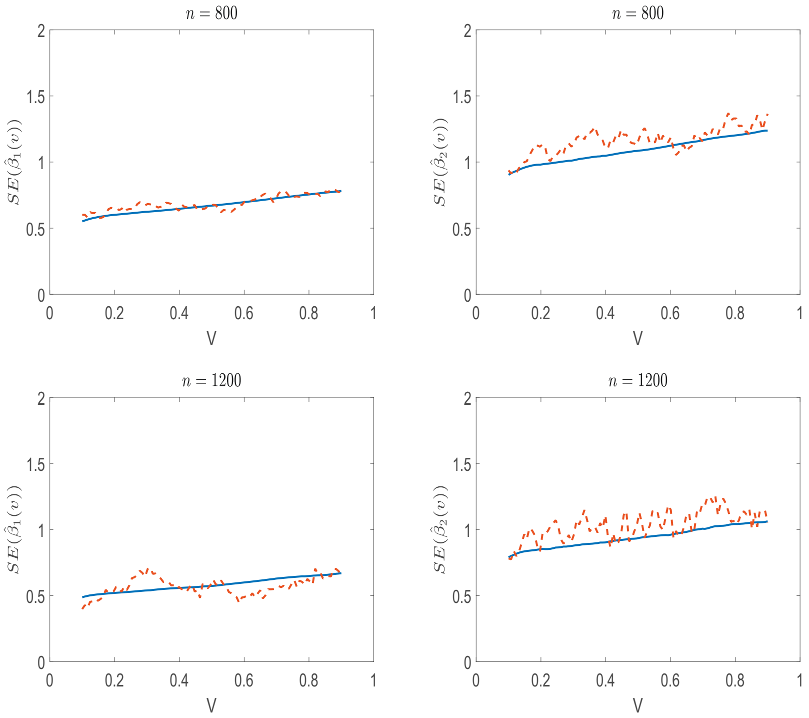

Figure 2. Plots of the estimated standard errors for Case I with $h = 1.6\hat{\sigma}_\nu n_0^{-1/4}$. The penalty function is the LASSO penalty. The first row is for sample size $n = 800$, and the second row for sample size $n = 1200$. The dashed lines are the medians of the estimated standard errors of $\hat{\boldsymbol{\beta}}(\nu)$, while the solid lines are the sample median absolute deviations of $\hat{\boldsymbol{\beta}}(\nu)$ divided by 0.6745 based on 500 replications.
SE: standard error.

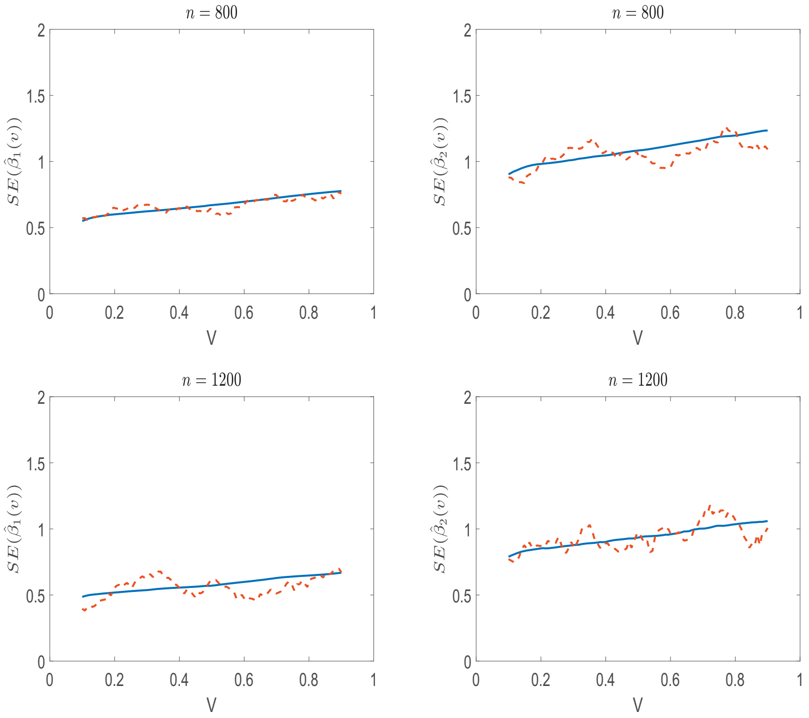

Figure 3. Plots of the estimated standard errors for the oracle under Case I with $h = 1.6\hat{\sigma}_\nu n_0^{-1/4}$. The first row is for sample size $n = 800$, and the second row for sample size $n = 1200$. The dashed lines are the medians of the estimated standard errors of $\hat{\boldsymbol{\beta}}(\nu)$, while the solid lines are the sample median absolute deviations of $\hat{\boldsymbol{\beta}}(\nu)$ divided by 0.6745 based on 500 replications.
SE: standard error.

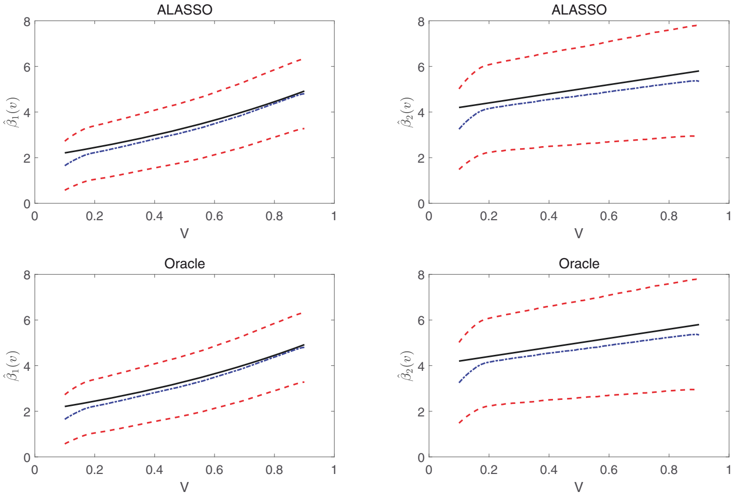

Figure 4. Plots of the estimated coefficients (dashed-dotted lines) and their 95% pointwise confidence bands (the dashed lines) under Case 1 with $n = 800$ and $h = 1.6\hat{\sigma}_\nu n_0^{-1/4}$. The solid lines are the true coefficient functions.
ALASSO: adaptive LASSO.

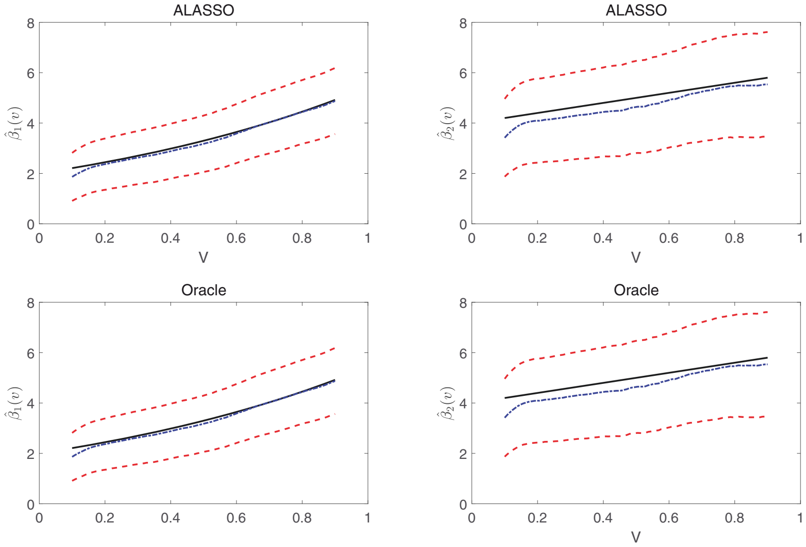

Figure 5. Plots of the estimated coefficients (dashed-dotted lines) and their 95% pointwise confidence bands (the dashed lines) under Case 1 with $n = 1200$ and $h = 1.6\hat{\sigma}_\nu n_0^{-1/4}$. The solid lines are the true coefficient functions.
ALASSO: adaptive LASSO.

region (581 amino acids long), where all possible mismatches of particular pairs of amino acids (e.g. A versus C) are weighted by the estimated probability of interchange.[24] The trial included 5403 HIV-negative volunteers who were at risk for acquiring HIV infection.[25] Volunteers were assigned randomly in a 2:1 ratio to receive a recombinant glycoprotein 120 vaccine (AIDSVAX) or placebo and were monitored for HIV infection at semiannual HIV testing visits for 36 months. Our objective was to select variables that are associated with the risk of infection under the mark-specific additive hazards model (1). During the trial, 368 individuals acquired HIV infection, but 32 individuals had missing marks. The analysis was based on the remaining 336 samples whose marks were unique (217 vaccine and 119 placebo samples).

In the dataset, eight covariates were included: treatment indicator, age at enrollment, sex, region, race, country, education, and behavioral risk score (taking values 0–7) as defined in Flynn et al.[25] Because the ALASSO method outperformed the LASSO method from the simulation results, we only used the ALASSO penalty in the analysis. We used the Epanechnikov kernel in the application. The estimated SE of the observed marks for uncensored

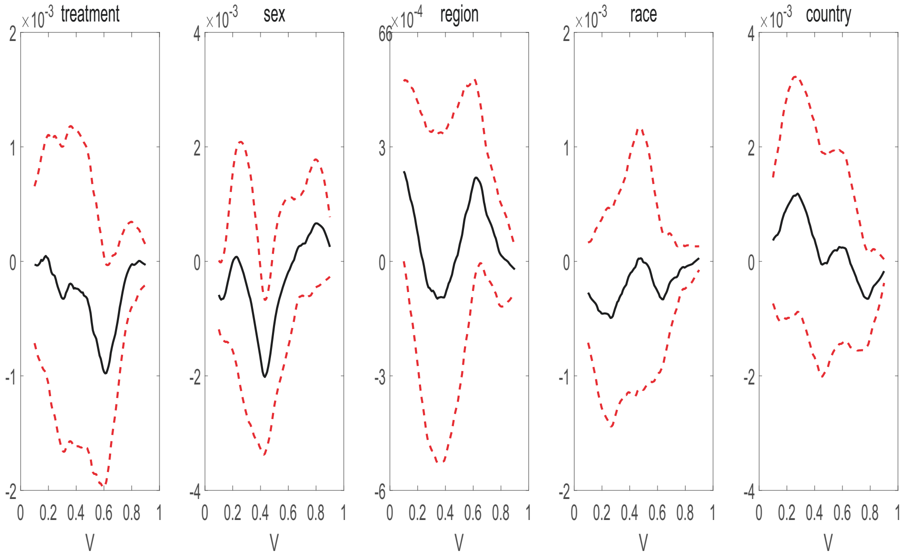

Figure 6. The unpenalized estimates of the irrelevant coefficients. The estimated functions (solid line) and their 95% pointwise confidence bands (dashed lines) are provided.

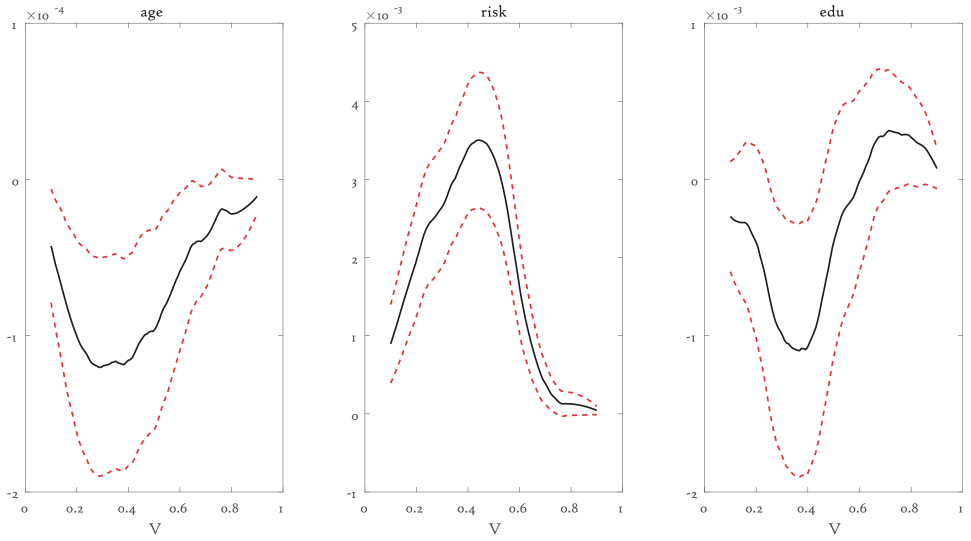

Figure 7. The estimates for the relevant coefficients after variable selection. The estimated functions (solid line) and their 95% pointwise confidence bands (dashed lines) are provided.

failure times is $\hat{\sigma}_V = 0.1591$. The bandwidth $h = 0.1491$ was estimated using the formula $h = 4\hat{\sigma}_V n_0^{-1/4}$, where n_0 is the number of observed failure times ($n_0 = 336$). The voting rates were taken as VR= 40% and 50%, and the results were identical using the proposed method. Specifically, the covariates age, risk score, and education are significantly relevant, whereas treatment indicator, sex, region, race, and country are not.

Figure 6 depicts the unpenalized estimates and their 95% pointwise confidence bands for the five unimportant variables.[6] It can be seen the pointwise confidence bands for the five coefficient functions cover zero in most of the range of the mark variable. This implies that the corresponding variables are indeed not important. Similarly, Figure 7 presents the estimates of the relevant factors and their 95% pointwise confidence bands after dropping the nonsignificant variables. Figure 7 shows that the pointwise confidence bands for the coefficient functions of age and behavioral risk score do not cover zero in most of the range of the mark variable. This indicates that the corresponding variables are significantly relevant. In addition, from Figure 7, we can obtain that there is a negative correlation between age and the risk of infection, whereas there is a positive correlation between the behavioral risk and the risk of infection. These results are consistent with the findings in Sun et al.[3] and Han et al.[6] Furthermore, our method also found that there is a negative correlation between education and the risk of infection in a subinterval of the mark variable. Since the variable treatment indicator is not significantly relevant, we conclude that the vaccine is not effective which is consistent with the finding in Sun et al.[3] and Han et al.[6]

5 Concluding remarks

In this article, we conducted variable selection for a mark-specific additive hazards model via penalized estimating functions. With the ALASSO penalty and proper choice of tuning parameters, our estimators are not only consistent but also enjoy the oracle properties. Simulation results demonstrated that the proposed method performed well, and an application to the first HIV vaccine efficacy trial was provided to illustrate our method.

Since our method is based on penalizing appropriate estimating functions to select variables, our method can be extended in a straightforward manner to accommodate other competing models, such as the mark-specific proportional hazards model with a univariate continuous mark and multivariate continuous marks.[3,4]

The proposed method cannot be extended in a straightforward manner to handle the mark-specific additive hazards model with missing marks, especially when some of the marks are missing not at random.[26] This merits future research. Note that the regression coefficients are not varying with time in model (1), which is a regular assumption for survival data with a continuous mark variable.[3–6] In some applications, however, treatment effects may vary with time. Although the additive hazards model is a convenient model to study time-varying effects, the proposed method cannot be directly extended to the case of time-varying coefficients, and substantial research efforts are required. Moreover, the proposed method cannot deal with the case when the number of explanatory variables is larger than the number of events. This is a challenging problem and requires further research efforts.

In addition, we only used the local constant fitting to construct the estimating equation for simplicity. But we must caution that the method proposed here requires a large sample size with moderate number of events to work well as demonstrated in the simulation studies. This does not cause a problem in our application to the first HIV vaccine efficacy trial, which has a sample size of 5403 with 336 events. The proposed method for the local constant fitting can be extended to deal with local linear fitting and general local polynomial fitting, but the resulting inference procedures would be much more complicated. We only considered the ALASSO penalty here. The proposed method can be extended to deal with some other penalty functions, such as the SCAD penalty[8] and the hard thresholding penalty.

Note that the corresponding penalized estimating function involves two tuning parameters (i.e. bandwidth and shrinkage parameter). Choosing both tuning parameters simultaneously may face substantial computational challenges. To tackle this issue, following Han et al.,[6] we set $h = \kappa\hat{\sigma}_\nu n_0^{-1/4}$, where κ is a prespecified constant. It would be worthwhile to develop some data-driven methods, such as the K-fold cross-validation method,[27] to select the optimal κ in the context of the mark-specific hazards models. In addition, methods with the tuning insensitivity property[28] would be explored to simplify the computation in future studies.

Acknowledgments

The authors thank the editor, Professor Evangelos Kontopantelis, and two reviewers for their constructive and insightful comments and suggestions that greatly improve the article. The authors thank Global Solutions for Infectious Diseases (GSID) and Dr Peter Gilbert for providing data from the phase 3 HIV vaccine trial VAX004.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dongxiao Han's research was supported by the Fundamental Research Funds for the Central Universities, Nankai University, 9920200110. Lianqiang Qu's research was partly supported by the National Natural Science Foundation of China

(Grant No. 12001219) and the Hubei Natural Science Foundation of China (Grant No. 2018CFB256). Liuquan Sun's research was partially supported by the National Natural Science Foundation of China (Grant Nos. 11771431 and 11690015) and Key Laboratory of RCSDS, CAS (No. 2008DP173182). Yanqing Sun's research was partly supported by the NIAID NIH award number R37AI054165 and the National Science Foundation grant DMS1915829.

ORCID iD

Liuquan Sun https://orcid.org/0000-0002-8816-942X

Supplemental Material

Supplemental material for this article is available online.

References

1. Gilbert PB, Mckeague IW and Sun Y. Tests for comparing mark-specific hazards and cumulative incidence functions. *Lifetime Data Anal* 2004; **10**: 5–28.
2. Gilbert PB, Mckeague IW and Sun Y. The two-sample problem for failure rates depending on a continuous mark: an application to vaccine efficacy. *Biostatistics* 2008; **9**: 263–276.
3. Sun Y, Gilbert PB and Mckeague IW. Proportional hazards models with continuous marks. *Ann Stat* 2009; **37**: 394–426.
4. Sun Y, Li M and Gilbert PB. Mark-specific proportional hazards model with multivariate continuous marks and its application to HIV vaccine efficacy trials. *Biostatistics* 2013; **14**: 60–74.
5. Gilbert PB and Sun Y. Inferences on relative failure rates in stratified mark-specific proportional hazards models with missing marks, with application to HIV vaccine efficacy trials. *J R Stat Soc C* 2015; **64**: 49–73.
6. Han D, Sun L, Sun Y, et al. Mark-specific additive hazards regression with continuous marks. *Lifetime Data Anal* 2017; **23**: 467–494.
7. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B* 1996; **58**: 267–288.
8. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
9. Zou H. The adaptive Lasso and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
10. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010; **38**: 894–942.
11. Fan J, Lin H and Zhou Y. Local partial likelihood estimation for lifetime data. *Ann Stat* 2006; **34**: 290–325.
12. Wang L, Li H and Huang J. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J Am Stat Assoc* 2008; **103**: 1556–1569.
13. Wang H and Xia Y. Shrinkage estimation of the varying coefficient model. *J Am Stat Assoc* 2009; **104**: 747–757.
14. Yan J and Huang J. Model selection for cox models with time-varying coefficients. *Biometrics* 2012; **68**: 419–428.
15. Wei X, Lu W and Hao H. Joint structure selection and estimation in the time-varying coefficient cox model. *Stat Sinica* 2016; **26**: 547–567.
16. Peng L, Xu J and Kutner N. Shrinkage estimation of varying covariate effects based on quantile regression. *Stat Comput* 2014; **24**: 853–869.
17. Johnson B, Long Q, Huang Y, et al. Model selection and inference for censored lifetime medical expenditures. *Biometrics* 2015; **72**: 731–741.
18. Johnson BA, Lin DY and Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. *J Am Stat Assoc* 2008; **103**: 672–680.
19. Meinshausen N and Bu̇̈hlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat* 2006; **34**: 1436–1462.
20. Fu W. Penalized regression: the bridge versus the LASSO. *J Comput Graph Stat* 1998; **7**: 397–416.
21. Zhang H and Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 2007; **94**: 691–703.
22. Wang H and Leng C. Unified LASSO estimation via least squares approximation. *J Am Stat Assoc* 2007; **102**: 1039–1048.
23. Wyatt R, Kwong PD, Desjardins E, et al. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* 1998; **393**: 705–711.
24. Nickle DC, Heath L, Jensen MA, et al. *Amino acid substitution matrices for HIV-1 subtype B*. Technical Report, University of Washington, May 2005.
25. Flynn NM, Forthal DN, Harro CD, et al. Placebo-controlled phase 3 trial of recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J Infect Dis* 2005; **191**: 654–665.
26. Little RJA and Rubin DB. *Statistical analysis with missing data*. New York: Wiley, 2002.
27. Tian L, Zucker D and Wei L. On the Cox model with time-varying regression coefficients. *J Am Stat Assoc* 2005; **100**: 172–183.
28. Belloni A, Chernozhukov V and Wang L. Square-root LASSO: pivotal recovery of sparse signals via conic programming. *Biometrika* 2011; **98**: 791–806.
29. Martinussen T and Scheike TH. *Dynamic regression models for survival data*. New York: Springer, 2006.

30. Aalen O and Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat* 1978; **5**: 141–150.

31. Pollard D. *Empirical processes: Theory and applications*. Hayward, CA: Institute of Mathematical Statistics, 1990.

Appendix

We assume that the following regularity conditions hold for Theorems 1 and 2:

(C1) $\boldsymbol{\beta}_{\boldsymbol{0}}(\nu)$ has componentwise continuous second derivatives on $[0,1]$. The second partial derivative of $\lambda_0(t,\nu)$ with respect to ν exists and is continuous on $[0,\tau]\times(0,1)$. The covariate vector Z is bounded.

(C2) $E[N_i(dt,d\nu)|\mathcal{F}_{t-}] = E[N_i(dt,d\nu)|Y_i(t),Z_i]$, where $\mathcal{F}_t = \sigma\{I(X_i \le s, \delta_i = 1), I(X_i \le s, \delta_i = 0), V_i I(X_i \le s, \delta_i = 1), Z_i; 0 \le s \le t, i = 1 \dots n\}$ is the right-continuous filtration generated by $\{N_i(s,\nu), Y_i(s), Z_i : 0 \le s \le t, 0 \le \nu \le 1, i = 1 \dots n\}$.

(C3) $P(X \ge \tau) > 0$, and the matrix A is nonsingular, where

$$A = E\left[\int_0^\tau Y_i(t)\{Z_i - \bar{z}(t)\}^{\otimes 2}\mathrm{d}t\right],$$

and $\bar{z}(t)$ is the limit of $\overline{Z}(t)$.

(C4) The kernel function $K(\cdot)$ is symmetric with support $[-1,1]$ and has bounded variation satisfying $\int K(u)du = 1$. The bandwidth satisfies $nh^2 \to \infty$ and $nh^5 \to 0$ as $n \to \infty$.

(C5) $(nh)^{1/2}\lambda_\nu \to 0$ and $nh\lambda_\nu \to \infty$.

(C6) $\omega_{\nu,j} = O_p(1)$ for $j = 1,\dots,s$ and $(nh)^{1/2}/\omega_{\nu,j} = O_p(1)$ for $j = s+1,\dots,p$.

Conditions (C1) and (C3) are standard assumptions in the context of survival analysis.[3] Condition (C4) is a standard assumption for kernel smoothing techniques. Condition (C2) implies that the mark-specific intensity of $N_i(t,\nu)$ with respect to $\mathcal{F}_t$ only depends on the failure status and the covariate Z_i. Thus, $E(N_i(dt,d\nu)|\mathcal{F}_{t-}) = Y_i(t)\lambda(t,\nu|Z_i)dtd\nu$, and $M_i(t,\nu) = \int_0^t\int_0^\nu [N_i(ds,du) - Y_i(s)\lambda(s,u|Z_i)dsdu]$ is a martingale with respect to $\mathcal{F}_t$ for each fixed v.[29] In addition, it can be checked that $M_i(\cdot,\nu_1)$ and $M_i(\cdot,\nu_2) - M_i(\cdot,\nu_1)$ are orthogonal square integrable martingales with respect to $\mathcal{F}_t$ for any $0 \le \nu_1 \le \nu_2 \le 1$.[30] A discussion of these conditions can be found in Sun et al.[3] Conditions (C5) and (C6) that pertain to the choices of tuning parameter and weight are the key to obtaining the oracle property. In order to avoid the boundary problems, we only study the asymptotic properties of $\hat{\boldsymbol{\beta}}(\nu)$ for $\nu \in [a,b] \subset (0,1)$.

Proof of Theorem 1. The functional central limit theorem[31] and condition (C1) imply that

$$\sup_{0\le t\le\tau} ||\overline{Z}(t) - \bar{z}(t)|| = O_p(n^{-1/2}) \tag{A.1}$$

By (A.1) and the uniform strong law of large numbers,[31] we have

$$\frac{1}{n}\frac{\partial U(\nu,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -A + o_p(1) \tag{A.2}$$

According to Theorem 1,[6] it can be checked that $(h/n)^{1/2}U(\nu,\boldsymbol{\beta}_{\boldsymbol{0}}(\nu))$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mu_0\Sigma(\nu)$. Hence, $(h/n)^{1/2}U(\nu,\boldsymbol{\beta}_{\boldsymbol{0}}(\nu)) = O_p(1)$. Next, we consider $\boldsymbol{b}_{\boldsymbol{1}}$ on the boundary of a ball around $\boldsymbol{\beta}_{\boldsymbol{1}}(\nu)$, that is, $\boldsymbol{b}_{\boldsymbol{1}} = \boldsymbol{\beta}_{\boldsymbol{1}}(\nu) + (nh)^{-1/2}u$ with $||u|| = r$ for some constant $r > 0$. Let $\boldsymbol{b} = (\boldsymbol{b}_{\boldsymbol{1}}{}^T, \boldsymbol{0}^{\mathrm{T}})^T$. Using the Taylor expansion, (A.2) and conditions (C5) and (C6), we obtain

$$(h/n)^{1/2}U^p_{1,s}(\nu,\boldsymbol{b}) = (h/n)^{1/2}U_{1,s}(\nu,\boldsymbol{\beta}_{\boldsymbol{0}}(\nu)) + (nh)^{1/2}A_{11}(\boldsymbol{b}_{\boldsymbol{1}} - \boldsymbol{\beta}_{\boldsymbol{1}}(\nu)) + o_p(1), \tag{A.3}$$

where $U^P_{1,s}(\nu,\cdot)$ and $U_{1,s}(\nu,\cdot)$ are the first s elements of $U^p(\nu,\cdot)$ and $U(\nu,\cdot)$, respectively. Since $(h/n)^{1/2}U_{1,s}(\nu,\boldsymbol{\beta_0}(\nu)) = O_p(1)$, it follows from (A.3) that

$$(h/n)^{1/2}(\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu))^T U^p_{1,s}(\nu,\boldsymbol{b}) = (nh)^{1/2}(\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu))^T A_{11}(\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu)) + O_p(||\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu)||).$$

Because A_{11} is nonsingular, the first term on the right side of the above equation is larger than $a_0 r^2(nh)^{-1/2}$, where a_0 is the smallest eigenvalue of A_{11}. The second term is of order $rO_p((nh)^{-1/2})$. Thus, for any $\epsilon > 0$, by choosing a sufficiently large r such that for large n, the probability that the absolute value of the first term is larger than that of the second term is less than ϵ. Hence, we get

$$P\left\{\min_{||\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu)||=r(nh)^{-1/2}}(\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu))^T U^p_{1,s}(\nu,(\boldsymbol{b_1}^T,\boldsymbol{0}^{\mathrm{T}})^T) > 0\right\} > 1-\epsilon.$$

By using the Brouwer fixed point theorem to the continuous function $U^p_{1,s}(\nu,(\boldsymbol{b_1}^T,\boldsymbol{0}^{\mathrm{T}})^T)$, it can be shown that $\min_{||\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu)||=(nh)^{-1/2}r}(\boldsymbol{b_1}-\boldsymbol{\beta_1}(\nu))^T U^p_{1,s}(\nu,(\boldsymbol{b_1}^T,\boldsymbol{0}^{\mathrm{T}})^T) > 0$ yields that $U^p_{1,s}(\nu,(\boldsymbol{b_1}^T,\boldsymbol{0}^{\mathrm{T}})^T)$ has a solution within this ball, denoted by $\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\nu)$. Let $\hat{\boldsymbol{\beta}}(\nu) = (\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\nu)^T,\boldsymbol{0}^{\mathrm{T}})^T$. Since $n^{-1}h$ goes to 0, $U^p_j(\nu,\hat{\boldsymbol{\beta}}(\nu)) = 0$ and $U^p_j(\nu,(\boldsymbol{b_1}^T,\boldsymbol{0}^{\mathrm{T}})^T)$ is a continuous function of $\boldsymbol{b_1}$ for $j = 1,\ldots,s$, we have

$$\lim_{n\to\infty}\lim_{\eta\to 0+}\frac{h}{n}U^P_j(\nu,\hat{\boldsymbol{\beta}}(\nu)+\eta\boldsymbol{e}_j)U^P_j(\nu,\hat{\boldsymbol{\beta}}(\nu)-\eta\boldsymbol{e}_j) = 0.$$

For $j = s+1,\ldots,p$, we have that for small $\eta > 0$,

$$(h/n)^{1/2}U^P_j(\nu,\hat{\boldsymbol{\beta}}(\nu)+\eta\boldsymbol{e}_j) = (h/n)^{1/2}U_j(\nu,\hat{\boldsymbol{\beta}}(\nu)+\eta\boldsymbol{e}_j) - (nh)^{1/2}q_{\lambda_\nu,\nu,j}.$$

Using the Taylor expansion and (A.2), we obtain that the first term on the right side of the above equation is of order $O_p(1)$ when η is small enough. Under conditions (C5) and (C6), $(nh)^{1/2}q_{\lambda_\nu,\nu,j}$ goes to infinity. As a result, $(h/n)^{1/2}U^P_j(\nu,\hat{\boldsymbol{\beta}}(\nu)+\eta\boldsymbol{e}_j)$ is dominated by $-(nh)^{1/2}q_{\lambda_\nu,\nu,j}$. Similarly, $(nh)^{1/2}q_{\lambda_\nu,\nu,j}$ dominates $(h/n)^{1/2}U^P_j(\nu,\hat{\boldsymbol{\beta}}(\nu)-\eta\boldsymbol{e}_j)$. Thus, $(h/n)^{1/2}U^P_j(\nu,\hat{\boldsymbol{\beta}}(\nu)+\eta\boldsymbol{e}_j)$ and $(h/n)^{1/2}U^P_j(\nu,\hat{\boldsymbol{\beta}}(\nu)-\eta\boldsymbol{e}_j)$ have opposite signs. Hence, $\hat{\boldsymbol{\beta}}(\nu)$ is an approximate zero-crossing by definition.

Proof of Theorem 2. Define $B_j = \{\hat{\beta}_j(\nu)\neq 0\}$, $j = s+1,\ldots,p$. To prove (*i*), it suffices to show that for any $\eta > 0$, and sufficiently large n, the probability $P(B_j) < \eta$. Since $\hat{\beta}_j(\nu) = O_p((nh)^{-1/2})$, there exists a positive constant M such that when n is sufficiently large,

$$P(B_j) < \eta/3 + P(\hat{\beta}_j(\nu)\neq 0, (nh)^{1/2}|\hat{\beta}_j(\nu)| < M) \tag{A.4}$$

Let

$$c_{j,n} = (h/n)^{1/2}U_j(\boldsymbol{\beta_0}(\nu)) + (nh)^{1/2}A_j(\hat{\boldsymbol{\beta}}(\nu)-\boldsymbol{\beta_0}(\nu)) - (nh)^{1/2}q_{\lambda_\nu,\nu,j}\mathrm{sgn}(\hat{\beta}_j(\nu)),$$

where A_j is the jth row of A. For any $\epsilon > 0$, using the definition of the approximate zero-crossing and the Taylor expansion, we have

$$\lim_{n\to\infty}P(|c_{j,n}| > \epsilon, \hat{\beta}_j(\nu)\neq 0, (nh)^{1/2}|\hat{\beta}_j(\nu)| < M) = 0. \tag{A.5}$$

Since the first two terms of $c_{j,n}$ are of order $O_p(1)$, according to (A.5), there exists some positive constant N such that for sufficiently large n,

$$P(\hat{\beta}_j(\nu)\neq 0, |(nh)^{1/2}\hat{\beta}_j(\nu)| < M, (nh)^{1/2}q_{\lambda_\nu,\nu,j} > N) < \eta/3.$$

For large n, using conditions (C5) and (C6), we can obtain

$$P((nh)^{1/2}q_{\lambda_\nu,\nu,j} < N) < \eta/3. \tag{A.6}$$

It then follows from (A.4)-(A.6) that $P(B_j) < \eta$.

For the part (*ii*), using conditions (C5)-(C6) and the Taylor expansion, we have

$$(h/n)^{1/2}U^p_{1,s}(\nu,\hat{\boldsymbol{\beta}}(\nu)) = (h/n)^{1/2}U_{1,s}(\nu,\boldsymbol{\beta_0}(\nu)) + (nh)^{1/2}A_{11}(\hat{\boldsymbol{\beta}}_{\boldsymbol{I}}(\nu) - \boldsymbol{\beta_I}(\nu)) + o_p(1).$$

According to the definition of the approximating zero-crossing, we obtain

$$(h/n)^{1/2}U^p_{1,s}(\nu,\hat{\boldsymbol{\beta}}(\nu)) = o_p(1).$$

As a result,

$$(h/n)^{1/2}U_{1,s}(\nu,\boldsymbol{\beta_0}(\nu)) + (nh)^{1/2}A_{11}(\hat{\boldsymbol{\beta}}_{\boldsymbol{I}}(\nu) - \boldsymbol{\beta_I}(\nu)) = o_p(1).$$

Since $(h/n)^{1/2}U(\nu,\boldsymbol{\beta_0}(\nu))$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mu_0\Sigma(\nu)$, it follows from the continuous mapping theorem and Slutsky's theorem that $(nh)^{1/2}(\hat{\boldsymbol{\beta}}_{\boldsymbol{I}}(\nu) - \boldsymbol{\beta_I}(\nu))$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mu_0 A_{11}^{-1}\Sigma_{11}(\nu)A_{11}^{-1}$.