Journal of Data Science 21 (4), 681–695
October 2023
DOI: 10.6339/22-JDS1061
Statistical Data Science

Maximum Likelihood Estimation for Shape-restricted Single-index Hazard Models

Jing Qin[1,*], Yifei Sun[2], Ao Yuan[3], and Chiung-Yu Huang[4]

[1]*Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Maryland, U.S.A.*
[2]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, U.S.A.*
[3]*Department of Biostatistics, Bioinformatics & Biomathematics, Georgetown University, Washington D.C., U.S.A.*
[4]*Department of Epidemiology & Biostatistics, University of California San Francisco, California, U.S.A.*

Abstract

Single-index models are becoming increasingly popular in many scientific applications as they offer the advantages of flexibility in regression modeling as well as interpretable covariate effects. In the context of survival analysis, the single-index hazards models are natural extensions of the Cox proportional hazards models. In this paper, we propose a novel estimation procedure for single-index hazard models under a monotone constraint of the index. We apply the profile likelihood method to obtain the semiparametric maximum likelihood estimator, where the novelty of the estimation procedure lies in estimating the unknown monotone link function by embedding the problem in isotonic regression with exponentially distributed random variables. The consistency of the proposed semiparametric maximum likelihood estimator is established under suitable regularity conditions. Numerical simulations are conducted to examine the finite-sample performance of the proposed method. An analysis of breast cancer data is presented for illustration.

Keywords *isotonic regression; pool-adjacent-violators algorithm; profile likelihood; semiparametric estimation*

1 Introduction

Single-index models have received much attention in many fields, including medicine, economics, finance, and environmental science. The single-index models can be viewed as a natural extension of the generalized linear models, where the link function is left unspecified and the covariate effects are summarized using a one-dimensional variable, referred to as the index (Powell et al., 1989; Härdle and Stoker, 1989; Ichimura, 1993; Härdle et al., 1993). In this paper, we consider single-index models for right-censored survival data. Let T denote the failure time of interest and let X be a $p \times 1$ vector of covariates. The popular Cox proportional hazards model assumes that, given $X = x$, the conditional hazard function $\lambda(t \mid x)$ satisfies

$$\lambda(t \mid x) = \lambda(t) \exp(x^\top \beta),$$

where β is a p-dimensional vector of regression parameters and the baseline hazard $\lambda(t)$ is left unspecified. The Cox model imposes an exponential functional form for the covariate effects on

*Corresponding author. Email: jingqin@niaid.nih.gov.

Received June 2, 2022; Accepted July 5, 2022

the hazard of failure, so that the covariates have a linear effect on the log hazard function. As pointed out by many authors, including Prentice (Prentice and Self, 1983), such assumption can be easily violated and it is desirable to consider a more flexible class of regression models

$$\lambda(t \mid x) = \lambda(t) \exp\{\phi(x^\top \beta)\}, \tag{1}$$

where ϕ is the link function. This model allows characterization of covariate effects on the risk of experiencing the failure event in a parsimonious way via a single index $x^\top \beta$. When ϕ is known, the partial likelihood method (Cox, 1975) can be applied directly to estimate the regression parameters β with right-censored survival data.

When the link function ϕ is unspecified, local partial likelihood methods, derived by employing either spline or local polynomial smoothing approximations for the unknown link function, have been proposed for estimating the link function in the case where X is an univariate continuous variable (Tibshirani and Hastie, 1987; Fan et al., 1997; Chen and Zhou, 2007). The local partial likelihood method can be inefficient because only data from individuals with covariate values in the neighborhood of x_0 are used to estimate $\phi(x_0)$. Gentleman and Crowley (Gentleman and Crowley, 1991) considered the local version of the full likelihood and developed an estimation procedure that alternates between estimating λ and estimating ϕ. Lately, Chen et al (Chen et al., 2010) proposed a global partial likelihood method that use all observations to estimate the value of the link function at any x_0. Statistical methods for Model (1) with multi-dimensional X have been developed in the same spirit as that for the univariate case. In particular, Wang et al (Wang et al., 2009) described an algorithm that iterates between maximizing the local partial likelihood function with respect to the smoothed approximation of ϕ for a given β and maximizing the global partial likelihood with respect to β with the estimated ϕ. Huang and Liu (Huang and Liu, 2006), on the other hand, proposed to approximate ϕ with cubic splines, thus reduces to a parametric model for ϕ which can be estimated directly using standard partial likelihood methods. However, large sample properties of the spline based approach are not well studied as, in theory, an infinite number of spline bases may be needed to span the unknown link function.

In many applications, it is desirable to impose shape restrictions, such as monotonicity and concavity, on the form of the covariate effects. Incorporating such a constraint can lead to improved efficiency and reduction in model complexity while allowing for more straightforward interpretation. For example, in dose finding trials for combination therapies, the marginal dose-response curve is often believed to be monotonic. While estimation procedures have been proposed for estimating the usual single-index models under shape-constraints with complete data (Foster et al., 2013; Groeneboom and Hendrickx, 2019), less attention has been paid to single-index hazards models under shape-constraints with right-censored data. Recently, Chung et al (Chung et al., 2018) considered a Cox model with shape constraints on the covariate effects, where, conditioning on covariates $X = x$ and $Z = z$, the hazard function is assumed to take the form

$$\lambda(t \mid x, z) = \lambda(t) \exp\{\psi(x) + z\beta\},$$

with $\psi(x)$ being an unspecified monotone function of the univariate variable X. The authors modified the iterative convex minorant algorithm of Jongbloed (Jongbloed, 1998) and proposed a pseudo-iterative convex minorant algorithm to maximize the partial likelihood. Specifically, the partial likelihood is sequentially approximated by quadratic functions and, as a result, the pool adjacent violators algorithm (PAVA) (Ayer et al., 1955) can be readily applied. In this paper, we fill in the gap by studying the single index model (1) with an unspecified, monotonic link

function ϕ. Instead of maximizing the partial likelihood, we embed the full likelihood function in the isotonic regression problem with exponentially distributed random variables and develop an iterative convex maximization algorithm. Our method provides a computationally stable estimation and, as demonstrated by the simulation studies, offers substantial efficiency gains over the Cox proportional hazards model when the link function is misspecified.

2 An Iterative Convex Maximization Algorithm

For ease of discussion, we write $r(z) = \exp\{-\phi(z)\}$ and assume that $r(z)$ is an unspecified non-decreasing function; or, equivalently, $\phi(z)$ is non-increasing in z. We adopt the convention with upper case letters, such as (Y, Δ, X), for random variables, and lower case letters, such as (y, δ, x) for observed values of (Y, Δ, X). With a minor modification, the estimation procedure discussed below can be applied to deal with the case where ϕ is non-decreasing or unimodal. Under the single-index model (1), the conditional hazard function of the survival time T is given by

$$\lambda(t \mid x) = \lambda(t)/r(x^\top \beta), \quad t \in [0, \tau],$$

where one is interested in making inference about the survival time distribution on a prespecified time interval $[0, \tau]$. Define the baseline cumulative hazard function $\Lambda(t) = \int_0^t \lambda(u)du$. It is easy to see that, for any constant $k > 0$, the pair $r^*(z) = kr(z)$ and $\lambda^*(t) = k\lambda(t)$ gives the same model as $r(z)$ and $\lambda(t)$, owing to the semiparametric nature of the Cox model. Moreover, the pair $r^*(z) = r(kz)$ and $\beta^* = k^{-1}\beta$ also yields the same model as that from $r(z)$ and β. In this paper, we impose $\Lambda(\tau) = 1$ and $\|\beta\| = 1$ to ensure model identifiability and note that the hazard function for the reference group $(X = 0)$ is given by $\Lambda(t)/r(0)$. A proof of the model identifiability is given in the Appendix.

In practice, the observation of the survival time T is usually subject to right censoring due to study end or premature dropout. Thus, instead of observing the actual value of T, we observe the possibly censored survival time $Y = \min(T, C)$, where C is the time of censoring. In many applications, it is reasonable to assume that C is independent of T given the observed covariates X. Denote by $\Delta = I(T \leqslant C)$ the indicator function of a failure event. Assume that the observed data $\{(y_i, \delta_i, x_i), i = 1, \ldots, n\}$ are independent realizations of (Y, Δ, X). Then the likelihood function based on the observed data is

$$\mathcal{L}(\beta, r, \Lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda(y_i)}{r(x_i^\top \beta)} \right\}^{\delta_i} \exp\left\{ -\frac{\Lambda(y_i)}{r(x_i^\top \beta)} \right\}.$$

Given β and $r(\cdot)$, the full likelihood $\mathcal{L}$ is maximized by the Breslow-type estimator

$$\widehat{\Lambda}^*(t) = \sum_{i=1}^{n} \frac{\delta_i I(y_i \leqslant t)}{\sum_{j=1}^{n} I(y_j \geqslant y_i)/r(x_j^\top \beta)}.$$

It is easy to see that replacing Λ with $\widehat{\Lambda}^*$ in the full likelihood $\mathcal{L}$ yields the partial likelihood

$$\prod_{i=1}^{n} \left\{ \frac{1/r(x_i^\top \beta)}{\sum_{j=1}^{n} I(y_j \geqslant y_i)/r(x_j^\top \beta)} \right\}^{\delta_i}.$$

However, direct maximization of the partial likelihood under the monotonicity constraint of $r(\cdot)$ is challenging and the conventional isotonic regression methods are not directly applicable.

In what follows, we consider semiparametric maximum likelihood estimation for Model (1) under the monotonicity constraint of $r(\cdot)$. We note that the maximum likelihood estimator derived without imposing the monotonicity constraint on the link function can be inconsistent. It is known that the nonparametric maximum likelihood estimate of a function concentrates its masses only on (some or all of) the observed data points. Without constraints, the masses may take any values and make the likelihood arbitrarily large. Thus we impose the monotone constraint on the link function $r(\cdot)$. Instead of directly maximizing the partial likelihood, we embed the full likelihood into the isotonic regression problems with exponential random variables and apply the pool adjacent violators algorithm (PAVA) to obtain the constrained maximizer for the link function. The proposed algorithm uses an inner loop to calculate $\{\Lambda(\cdot), r(\cdot)\}$ that maximize the full likelihood function $\mathcal{L}(\beta, r, \Lambda)$ at a given value of β and an outer loop to maximize the profile likelihood with respect to β.

Specifically, given β and $r(\cdot)$, we derive the maiximiser $\widehat{\Lambda}^*$ of the full likelihood and estimate Λ by $\widehat{\Lambda}(t) = \widehat{\Lambda}^*(t)/\widehat{\Lambda}^*(\tau)$, where rescaling is carried out to ensure the identifiability condition $\Lambda(\tau) = 1$. Next, for given β and $\Lambda(\cdot)$, we define $Z_i = X_i^\top \beta$ and sort the observed data $\{(y_i, \delta_i, x_i), i = 1, \ldots, n\}$ according to the value of z_i's, so that $x_1^\top \beta \leqslant \cdots \leqslant x_n^\top \beta$. Write $\eta_i = \Lambda(y_i)$ and $r_i = \exp\{-\phi(z_i)\} = \exp\{-\phi(x_i^\top \beta)\}$. Then, for fixed β and Λ, the full likelihood is proportional to

$$\mathcal{L}_r = \prod_{i=1}^{n} \left(\frac{1}{r_i}\right)^{\delta_i} \exp\left(-\frac{\eta_i}{r_i}\right)$$

We maximize $\mathcal{L}_r$ with respect to r_i's under the monotone constraint $r_1 \leqslant r_2 \leqslant \cdots \leqslant r_n$ by embedding into the isotonic regression problem with exponential random variables. Let $z_1^*, \ldots, z_L^*$ be the ordered, distinct values of the z_i's from *uncensored* observations. We further define the subintervals $I_1 = (-\infty, z_1^*], \ldots, I_L = (z_{L-1}^*, z_L^*]$, and $I_{L+1} = (z_L^*, \infty)$. It's easy to see that, if the event time of the ith subject is censored, the contribution of the observation to the likelihood is given by $\exp(-\eta_i/r_i)$, so that the likelihood increases with r_i. If z_i falls into the subinterval I_l, then, under the monotone constraint, maximization is achieved by setting $r_i = r(z_l^*)$. In other words, $\mathcal{L}_r$ is maximized by setting $r_i = r_l$ if $z_i \in I_l$, $l \leqslant L$, and $r_i = \infty$ if $z_i > z_L^*$. To perform isotonic regression, we exclude data from individuals whose z value is greater than z_L^* and consider the following likelihood based on the reduced dataset

$$\mathcal{L}_r^* = \prod_{l=1}^{L} \left(\frac{1}{r_l^*}\right)^{m_l^d} \exp\left(-\frac{m_l^d \bar{\eta}_l}{r_l^*}\right),$$

where r_l^* is the value of $r(z)$ evaluated at $z = z_l^*$, $m_l^d = \sum_{i=1}^n \delta_i I(z_i = z_l^*)$ is the number of uncensored individuals whose Z value is z_l^*, and $\bar{\eta}_l = \frac{1}{m_l^d} \sum_{i=1}^n \Lambda(y_i) I(z_i \in I_l)$, $l = 1, \ldots, L$. Note that the numerator of $\bar{\eta}_l$ include data from all individual, either censored or uncensored, but the denominator only include data from uncensored individuals. Maximisation of $\mathcal{L}_r^*$ subject to the monotone constraint can be viewed as an isotonic regression problem because $\mathcal{L}_r^*$ is mathematically equivalent to the likelihood of a sequence of L independent trials, where the outcomes are exponentially distributed with means satisfying $r_1^* \leqslant r_2^* \leqslant \cdots \leqslant r_L^*$ and the sample size of the lth trial is m_l^d. As pointed out in Chapter 1 of Robertson et al (Robertson et al., 1988), the pool-adjacent-violators algorithm (PAVA) (Ayer et al., 1955) can be used to solve exponential family isotonic regression problems. Hence, given β and $\Lambda(\cdot)$, we propose to maximize $\mathcal{L}_r^*$ with

the PAVA estimator,

$$\widehat{r}_l^* = \max_{k\leqslant l}\min_{l\leqslant q}\frac{\sum_{k\leqslant l\leqslant q} w_l\bar{\eta}_l}{\sum_{k\leqslant l\leqslant q} w_l}.$$

Thus we obtain $\widehat{r}(z_i)=\widehat{r}_l^*$ for $z_i\in I_l$, $l\leqslant L$, and $\widehat{r}(z_i)=\widehat{r}_L^*$ for $z_i\in I_{L+1}$.

For a given β, the estimation algorithm for $\{\Lambda(\cdot), r(\cdot)\}$ is summarized below. We alternate between (M1) and (M2) below repeatedly until some convergence criteria are met. Specifically, suppose the value of parameters in the bth step is $(\Lambda^{(b)}, r^{(b)})$. Then in the $(b+1)$th step,

(M1) Calculate $\bar{\eta}_l^{(b+1)}=\frac{1}{m_l^d}\sum_{i=1}^n \Lambda^{(b)}(y_i)I(z_i\in I_l)$. Apply PAVA to obtain

$$r_l^{(b+1)} = \max_{k\leqslant l}\min_{l\leqslant q}\frac{\sum_{k\leqslant l\leqslant q} m_l^d\ \bar{\eta}_l^{(b+1)}}{\sum_{k\leqslant l\leqslant q} m_l^d}.$$

Set $r^{(b+1)}(z_i)=r_l^{(b+1)}$ for $z_i\in I_l$, $l\leqslant L$, and $r^{(b+1)}(z_i)=r_L^{(b+1)}$ for $z_i\in I_{L+1}$.

(M2) Update Λ with the Breslow-Type estimator

$$\Lambda^{(b+1)*}(t)=\sum_{i=1}^n \frac{\delta_i I(y_i\leqslant t)}{\sum_{j=1}^n I(y_j\geqslant y_i)/r^{(b+1)}(z_j)}$$

to obtain $\Lambda^{(b+1)}(t)=\Lambda^{(b+1)*}(t)/\Lambda^{(b+1)*}(\tau)$.

For fixed β, we iterate between (M1) and (M2) until convergence. Denote the limit by $(\widehat{\Lambda}(\cdot;\beta),\widehat{r}(\cdot;\beta))$.

Finally, we plug in $\widehat{\Lambda}(\cdot;\beta)$ and $\widehat{r}(\cdot;\beta)$ back to the full likelihood function $\mathcal{L}(\beta,r,\Lambda)$ to obtain the profile likelihood function $\mathcal{L}_p(\beta)=\mathcal{L}(\beta,\widehat{r}(\cdot;\beta),\widehat{\Lambda}(\cdot;\beta))$. For a given β, $\widehat{r}(\cdot;\beta)$ is only uniquely defined at the ordered, distinct values of $x_i^\top\beta$ from uncensored observations. With a finite sample, the maximizer of $\mathcal{L}_p(\beta)$ is not unique, as $\mathcal{L}_p(\beta)$ only depends on the ordering of $x_i^\top\beta$ from uncensored observations induced by β. As shown in Theorem 1 below, the maximizer converges to the true parameters β_0 as the sample size goes to infinity. To account for the constraint $\|\beta\|=1$, we use the spherical coordinate system to represent β on the unit sphere $\mathcal{B}=\{\beta:\|\beta\|=1,\beta\in\mathbb{R}^p\}$. Following Balabdaoui et al (Balabdaoui et al., 2019), we use the following map to reduce the parameters to a $(p-1)$-dimensional vector, $\mathbb{S}:[0,\pi]^{(p-2)}\times[0,2\pi]\mapsto\mathcal{B};\theta\mapsto\beta$, where $\theta=(\theta_1,\theta_2,\ldots,\theta_{p-1})$, and

$$\begin{aligned}\beta \;=\; & (\cos(\theta_1),\sin(\theta_1)\cos(\theta_2),\ldots,\sin(\theta_1)\cdots\sin(\theta_{p-2})\cos(\theta_{p-1}),\\ & \sin(\theta_1)\cdots\sin(\theta_{p-2})\sin(\theta_{p-1})).\end{aligned}$$

Maximization of $\mathcal{L}_p(\beta)$ can be implemented using Nelder-Mead's downhill simplex method (Nelder and Mead, 1965) with respect to $(\theta_1,\theta_2,\ldots,\theta_{p-1})$. Different initial values can be used in the optimization for improved performance.

Let $\mathcal{R}$ be the collection of monotone increasing functions on $\mathbb{R}$, and $\mathcal{A}$ be a collection of monotone increasing functions on $\mathbb{R}^+$ such that the function takes value 1 at τ. Denote by $(\widehat{\beta},\widehat{r},\widehat{\Lambda})$ the maximum likelihood estimator of the true parameters (β_0,r_0,Λ_0), that is,

$$(\widehat{\beta},\widehat{r},\widehat{\Lambda})=\arg\max_{(\beta,r,\Lambda)\in(\mathcal{B},\mathcal{R},\mathcal{A})}\mathcal{L}(\beta,r,\Lambda).$$

The consistency of the maximum likelihood estimator $(\widehat{\beta},\widehat{r},\widehat{\Lambda})$ is stated in Theorem 1, with proof given in the Appendix.

Theorem 1. *Let $[z_1, z_2]$ be a bounded interval in the support of $X^\top\beta_0$. Under conditions (C1)∼(C4) in the Appendix, as $n \to \infty$, we have*

$$\widehat{\beta} \xrightarrow{a.s.} \beta_0 \quad , \qquad \sup_{z\in[z_1,z_2]} |\, \widehat{r}(z) - r_0(z) \,| \xrightarrow{a.s.} 0 \quad , \qquad \sup_{t\in[0,\tau]} |\, \widehat{\Lambda}(t) - \Lambda_0(t) \,| \xrightarrow{a.s.} 0.$$

For single index models where the nonparametric component is estimated by nonparametric maximum likelihood estimation under shape constraints, the $\sqrt{n}$-convergence rate and asymptotic normality of the estimator for regression parameters is an open question (Huang and Wellner, 1997; Murphy et al., 1999; Groeneboom and Hendrickx, 2018). Our estimation procedure encounters similar technical challenges. The main reason is that $\widehat{r}$ is a step function which is not smooth, and the regression part is bundled inside it. As a result, the asymptotic distribution of $(\widehat{\beta}, \widehat{r}, \widehat{\Lambda})$ requires further investigation. Let $[z_1, z_2]$ be a bounded interval in the support of $X^\top\beta_0$. Define $\|\widehat{r} - r_0\| = [\int_{z_1}^{z_2}\{\widehat{r}(z) - r_0(z)\}^2 dz]^{1/2}$ and $\|\widehat{\Lambda} - \Lambda_0\| = [\int_0^\tau\{\widehat{\Lambda}(t) - \Lambda_0(t)\}^2 dt]^{1/2}$. In Theorem 2, we show that the convergence rate of $(\widehat{\beta}, \widehat{r}, \widehat{\Lambda})$ is at least $n^{1/3}$. The proof of Theorem 2 is given in the Supplementary Material.

Theorem 2. *Under conditions (C1)∼(C5) in the Appendix, we have*

$$\|\widehat{\beta} - \beta_0\| + \|\widehat{r} - r_0\| + \|\widehat{\Lambda} - \Lambda_0\| = O_p(n^{-1/3}).$$

It is worthwhile to point out that when the covariates have elliptically symmetric distribution, fitting a Cox model to the data yields consistent estimate of the direction of β_0. The result is summarized in Proposition 1 and the proof is given in the Appendix.

Proposition 1. *Let $\widehat{\beta}_P$ be the maximum partial likelihood estimator under the usual Cox model with a (potentially misspecified) exponential link function, $\lambda(t \mid x) = \lambda(t)\exp(x^\top\beta)$. If X has an elliptically symmetric distribution and the censoring is completely random, then β_0 can be consistently estimated by $-\widehat{\beta}_P/\|\widehat{\beta}_P\|$.*

3 Simulation Studies

We conduct simulation studies to evaluate the performance of the proposed method. Given covariates X, we generated survival times from the Weibull distribution with shape parameter 2 and scale parameter $\sqrt{r(X^\top\beta^*)}$, where $r(X^\top\beta^*) = \exp\{|X^\top\beta^*|^a \mathrm{sign}(X^\top\beta^*)\}$. The hazard function is $\lambda(t \mid X) = 2t\exp\{-|X^\top\beta^*|^a \mathrm{sign}(X^\top\beta^*)\}$. We included p covariates and set $\beta^* = (\beta_1^*, \ldots, \beta_p^*)$, where $\beta_{2m-1}^* = -1$ and $\beta_{2m}^* = 1$ for $m \geqslant 1$, thus the true value is $\beta_0 = \beta^*/\sqrt{p}$. We considered the following scenarios: (I) $p = 2$, X_j were generated from the exponential distribution with rate parameter 1, denoted by exp(1), for $j = 1, 2$. (II) $p = 2$, X_j were generated from the standard normal distribution $N(0, 1)$ for $j = 1, 2$. (III) $p = 5$, X_1, X_2 were generated from exp(1), X_4, X_5 were generated from $N(0, 1)$, and X_3 was generated from the Bernoulli distribution with success probability 1/2. (IV) $p = 5$, $X_1, \ldots, X_5$ were generated from $N(0, 1)$. We also considered three cases under each scenario, that is, (A) $a = 1/5$, (B) $a = 1/3$, and (C) $a = 1$. The censoring time was set as $C = \min(C^*, \tau)$, where C^* was generated from exponential distributions with rate parameters λ_C, and λ_C and τ were chosen to yield approximately 25% and 50% censoring rates. In each simulation, we generated 1000 datasets with sample sizes of 200 and 800. We compared the proposed method with the negative normalized coefficients from maximum partial likelihood estimator (MPLE) assuming an exponential link function.

In the Supplementary Material, we also included the simulation results of an estimator without imposing the monotone constraint. Specifically, the estimator replaces the PAVA estimator of the link function with a kernel smoothing estimator. More details can be found in the Supplementary Material.

The non-smoothness of the profile likelihood function precludes the use of methods that utilize derivative because its derivative does not exist. Thus we considered applying Nelder-Mead's method, which only requires the value of likelihood functions. Moreover, we used multiple initial values to improve the search for the maximizer. The first initial value of the Nelder-Mead algorithm is chosen as the negative normalized MPLE transformed to a $(p-1)$-dimensional vector via the map $\mathbb{S}^{-1}$. When the solution is obtained, we further add a small perturbation (e.g., a random vector whose elements are generated from the uniform distribution on $[-0.5, 0.5]$) to the solution. We then run the Nelder-Mead's algorithm with the perturbed solution as the initial value. If the likelihood function value is larger than that of the previous step, we replace the estimated parameters with the solution from the current step. We repeat this procedure twenty times and use the parameter value that yields the largest profile likelihood. In our current implementation, we applied the R function `optim` (R Core Team, 2020) for the Nelder-Mead algorithm to maximize $\mathcal{L}_p(\beta)$. To obtain $\{r, \Lambda\}$ that maximize $\mathcal{L}(\beta, r, \Lambda)$ at each value of β, we applied the R function `squarem` in the package `SQUAREM` (Du and Varadhan, 2020), which is used to accelerate the convergence of general fixed-point iterations. We used the stopping criteria from the default setting in each function.

The results are reported in Table 1 and 2. In Scenario I, the covariates were generated from the exponential distribution, and the survival data was not generated from the Cox model. It can be observed that, when the link function is misspecified (i.e., $a = 1/3$ and $a = 1/5$, MPLE has substantially larger bias and variance compared to the proposed approach, and the bias does not decrease as the sample size increases; when the link function is correctly specified, the MPLE has smaller variances. In Scenarios II and IV, the covariates were generated from the normal distribution. Both methods yield small biases, and the variance decreases as the sample size increases. This is consistent with Proposition 1, that is, when the covariates have elliptically symmetric distribution, the negative normalized MPLE is consistent for β_0 even if the proportional hazards model assumption is violated. When the link function is misspecified, the proposed method has smaller variances; when the link function is correctly specified, the MPLE has smaller variances. In Scenario III, we include more covariates generated from different types of distributions. The biases and variances of the proposed estimator decrease as the sample size increases. However, when the link function is misspecified, the biases of MPLE remain large when $n = 800$, and the variances of MPLE are larger compared to the proposed method. In summary, the proposed method performs well and outperforms MPLE when the assumption on the link function does not hold.

4 Breast Cancer Data Example

The proposed method is applied to a multicenter randomized clinical trial conducted by the German Breast Cancer Study Group (Schumacher et al., 1994). The aim of the trial was to compare the time-to-event outcomes between different treatment modalities. The data used in this paper to illustrate our findings are available in the R package `mfp` on the Comprehensive R Archive Network (Ambler and Benner, 2015). The primary outcome is the recurrence-free survival time, which is a composite endpoint of breast cancer recurrence and death. The median

Table 1: Summary of simulation studies ($n = 200$).

		a = 1/5, cen = 25%				a = 1/3, cen = 25%				a = 1, cen = 25%			
		Proposed		MPLE		Proposed		MPLE		Proposed		MPLE	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
I	$\widehat{\beta}_1$	2	60	64	107	2	69	60	96	2	88	−1	51
	$\widehat{\beta}_2$	−3	61	45	93	−5	71	45	83	−9	89	−4	52
II	$\widehat{\beta}_1$	3	53	10	78	2	61	9	70	4	63	4	43
	$\widehat{\beta}_2$	−1	53	1	77	−3	62	2	69	−2	63	1	43
III	$\widehat{\beta}_1$	6	68	−5	92	4	71	−6	80	−3	52	−0.2	36
	$\widehat{\beta}_2$	−6	71	11	95	−5	73	10	82	−5	52	−0.1	37
	$\widehat{\beta}_3$	17	119	4	162	16	119	4	140	11	90	7	65
	$\widehat{\beta}_4$	−7	76	5	100	−10	77	10	88	−6	61	−4	43
	$\widehat{\beta}_5$	8	94	100	102	10	92	83	89	4	54	0.1	37
IV	$\widehat{\beta}_1$	−1	70	8	88	1	71	7	76	0.3	50	3	36
	$\widehat{\beta}_2$	−10	70	−14	87	−13	70	−11	76	−7	50	−4	36
	$\widehat{\beta}_3$	4	67	2	89	2	68	0.2	76	−1	49	−2	36
	$\widehat{\beta}_4$	−7	68	−12	88	−5	70	−8	75	−2	50	−1	35
	$\widehat{\beta}_5$	7	69	9	90	8	71	7	79	6	51	1	36
		a = 1/5, cen = 50%				a = 1/3, cen = 50%				a = 1, cen = 50%			
		Proposed		MPLE		Proposed		MPLE		Proposed		MPLE	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
I	$\widehat{\beta}_1$	3	84	95	131	5	97	85	118	5	107	0.1	64
	$\widehat{\beta}_2$	−7	87	65	108	−8	97	61	99	−11	111	−6	65
II	$\widehat{\beta}_1$	−0	75	3	91	4	83	3	82	3	77	−0.4	51
	$\widehat{\beta}_2$	−8	76	−8	90	−6	83	−7	82	−6	77	−4	52
III	$\widehat{\beta}_1$	−2	100	1	115	−1	98	1	102	1	63	3	45
	$\widehat{\beta}_2$	−13	100	−6	116	−13	96	−4	103	−5	64	−3	45
	$\widehat{\beta}_3$	25	152	11	187	27	153	8	162	8	106	3	73
	$\widehat{\beta}_4$	−15	107	10	124	−15	105	13	110	−12	75	−7	54
	$\widehat{\beta}_5$	30	127	108	122	22	120	88	107	6	60	−0.2	42
IV	$\widehat{\beta}_1$	5	94	17	107	4	93	12	94	0.3	60	3	42
	$\widehat{\beta}_2$	−9	90	−14	108	−8	91	−11	94	−6	62	−3	45
	$\widehat{\beta}_3$	4	95	8	108	6	91	6	94	0.2	62	1	44
	$\widehat{\beta}_4$	−19	91	−16	105	−19	88	−12	91	−7	58	−2	41
	$\widehat{\beta}_5$	11	93	9	105	9	90	7	93	8	63	2	43

Note: MPLE stands for the maximum partial likelihood estimator and cen stands for the censoring rate. Bias and SE are the empirical bias (×1000) and empirical standard deviation (×1000) of 1000 simulated datasets, respectively.

follow-up was 56 months. During the study period, 299 of the 686 patients had disease recurrence or died. The covariates included in the model are hormonal treatment (yes/no), tumor size, tumor grade (1/2/3), and the number of positive lymph nodes. To ensure stable numerical performance, we standardize the covariates taking numeric values to have zero mean and unit variance.

Table 2: Summary of simulation studies ($n = 800$).

		a = 1/5, cen = 25%				a = 1/3, cen = 25%				a = 1, cen = 25%			
		Proposed		MPLE		Proposed		MPLE		Proposed		MPLE	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
I	$\widehat{\beta}_1$	0.1	19	77	61	1	26	74	55	1	44	2	24
	$\widehat{\beta}_2$	−0.4	19	65	50	−0.1	26	64	45	−2	44	1	24
II	$\widehat{\beta}_1$	−1	20	−0.02	36	−1	26	−0.3	33	0.2	31	−0.1	20
	$\widehat{\beta}_2$	−1	20	−2	37	−2	26	−2	33	−1	31	−1	20
III	$\widehat{\beta}_1$	−0.03	24	−14	47	1	28	−11	41	0.4	26	1	18
	$\widehat{\beta}_2$	−1	25	16	47	−1	29	13	41	−2	26	−0.2	18
	$\widehat{\beta}_3$	4	39	−21	78	3	49	−17	68	4	45	2	30
	$\widehat{\beta}_4$	0.01	27	22	48	0.3	30	24	43	1	30	1	21
	$\widehat{\beta}_5$	0.4	31	104	53	3	38	88	46	1	26	1	18
IV	$\widehat{\beta}_1$	0.5	23	1	44	1	27	0.3	38	0.3	24	−0.1	17
	$\widehat{\beta}_2$	0.2	23	1	43	1	27	1	37	1	25	1	18
	$\widehat{\beta}_3$	0.5	24	5	44	1	28	3	38	1	25	1	17
	$\widehat{\beta}_4$	−1	23	−2	45	−1	27	−2	39	−1	25	−1	18
	$\widehat{\beta}_5$	2	23	4	44	2	27	3	38	2	24	1	17
		a = 1/5, cen = 50%				a = 1/3, cen = 50%				a = 1, cen = 50%			
		Proposed		MPLE		Proposed		MPLE		Proposed		MPLE	
		Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
I	$\widehat{\beta}_1$	1	26	104	70	1	35	94	63	0.02	52	−1	30
	$\widehat{\beta}_2$	0.05	26	86	53	−1	35	79	49	−4	53	−3	30
II	$\widehat{\beta}_1$	0.4	25	3	46	2	33	2	42	2	38	1	25
	$\widehat{\beta}_2$	−0.5	25	−0.3	46	0.05	33	−0.3	41	−0.2	38	0.3	25
III	$\widehat{\beta}_1$	−1	35	−10	58	−1	39	−8	51	1	29	1	21
	$\widehat{\beta}_2$	−0.1	33	11	58	−1	39	10	51	−2	30	−1	21
	$\widehat{\beta}_3$	4	55	−14	99	7	62	−12	85	3	52	1	37
	$\widehat{\beta}_4$	−1	37	34	66	−1	40	32	58	−1	38	−1	27
	$\widehat{\beta}_5$	5	44	113	64	5	51	95	56	2	29	−0.2	20
IV	$\widehat{\beta}_1$	−1	32	2	52	−2	36	1	45	0.2	28	0.1	20
	$\widehat{\beta}_2$	−2	32	−3	55	−1	36	−2	48	0.3	30	−0.2	21
	$\widehat{\beta}_3$	1	31	3	56	2	36	2	49	1	30	0.3	21
	$\widehat{\beta}_4$	−1	31	−4	55	−3	36	−3	47	0.02	28	−0.1	20
	$\widehat{\beta}_5$	3	33	6	56	4	37	5	48	4	30	2	21

Note: MPLE stands for the maximum partial likelihood estimator and cen stands for the censoring rate. Bias and SE are the empirical bias (×1000) and empirical standard deviation (×1000) of 1000 simulated datasets, respectively.

The conventional method to analyze this data is the standard Cox regression, which uses a pre-specified monotonic link function. By using an unspecified monotonic link function, we allow greater flexibility than the standard Cox regression. Moreover, compared to models with non-monotonic link functions, the use of monotone link leads to an easier interpretation in

Table 3: Estimated coefficients for the German Breast Cancer Study.

	Proposed model[1]	Cox model[2]
Hormonal therapy	−0.325 (−0.519, −0.047)	−0.248
Tumor size	0.165 (−0.052, 0.344)	0.073
Tumor grade 2	0.444 (0.124, 0.642)	0.567
Tumor grade 3	0.474 (0.166, 0.843)	0.754
Number of positive nodes	0.667 (0.280, 0.862)	0.207

[1] For the proposed model, we reported $-\widehat{\beta}$ so that the parameters can be compared with the Cox model. We also reported the 2.5th and 97.5th percentiles from 500 Bootstrap replicates. Note that the coverage probability of this interval may not be close to 95%.
[2] For the Cox model, we reported the normalized regression parameters $\widehat{\beta}_P/\|\widehat{\beta}_P\|$.

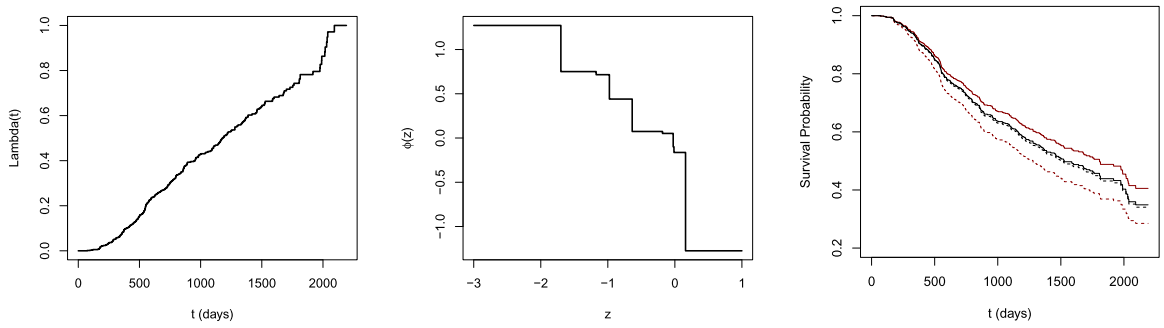

Figure 1: Estimated functions for the German Breast Cancer Study. The left and middle panel shows the estimated baseline cumulative hazard function Λ and link function ϕ, respectively. In the right panel, black lines are the predicted survival probabilities using the proposed model; the red lines are the predicted survival probabilities using the Cox model; the dashed lines are patients without hormonal therapy; the solid lines are patients with hormonal therapy.

practical applications. Table 3 reports the coefficients that are normalized to have a unit norm. We reported $-\widehat{\beta}$ from the proposed method and $\widehat{\beta}_P/\|\widehat{\beta}_P\|$ from the Cox model, and thus a positive coefficient indicates a larger value of the covariate is associated with a higher risk. To ensure identifiability, we assume the coefficient of hormonal therapy is negative. In both Cox model and monotone single index model, hormonal therapy is associated with a lower survival risk; larger tumors, higher tumor grades, and more positive lymph nodes are associated with higher risks. The Cox model estimates a much smaller effect for the positive lymph node count relative to other covariates.

The estimated hazard (normalized to have $\Lambda_0(\tau) = 1$, $\tau = 6$ years) and estimated link function are reported in Figure 1. The shape of the estimated link function $\widehat{\phi}(z) = -\log \widehat{r}(z)$ provides evidence against the assumption of the exponential link function in the Cox model. In this case, the proposed method may yield less biased estimates. To provide more insight into the difference between the proposed method with the Cox model, we plot the predicted survival probability of two hypothetical patients in Figure 1. The two patients have tumors grade 3, and the other covariates are set to be the median values among grade 3 patients; one of them undergoes hormonal treatment while the other does not. The proposed method yields a smaller difference in survival curves compared to the Cox proportional hazards model. When predicting

the survival probability, the proposed method is expected to be more robust than the Cox model under model misspecification. The difference in the predicted survival functions using the two methods may suggest a potential violation of the Cox model assumption.

5 Remarks

This paper focuses on semiparametric maximum likelihood estimation of the single-index hazards model under a shape-constraint. In our method, the link function is obtained from the PAVA algorithm and is not smooth. One may also consider the monotone splines (see, for example, Wang and Yan, 2021) to estimate the link function; this will be studied in our future work. Extensions of the proposed estimation procedure to the partially linear single-index hazards model under the same shape-constraint are straightforward, and their asymptotic properties will be studied elsewhere.

Supplementary Material

The Supplementary Material includes the proof of Theorem 2, additonal simulation results, and the R code to implement the proposed method.

Appendix

Regularity Conditions

To establish the large sample property of the maximum likelihood estimator, we impose the following conditions:

(C1) $\Pr(\Delta = 1) > 0$ and $\Pr(Y > \tau) > 0$ for a prespecified constant τ.

(C2) Define $p(Y, \Delta \mid X; \beta, r, \Lambda) = \{\lambda(Y)/r(\beta^\top X)\}^\Delta \exp\{-\Lambda(Y)/r(\beta^\top X)\}$. For any $(\beta, r, \Lambda) \in (\mathcal{B}, \mathcal{R}, \mathcal{A})$, the Kullback-Leilber divergence,

$$E \log \left\{ \frac{p(Y, \Delta \mid X; \beta_0, r_0, \Lambda_0)}{p(Y, \Delta \mid X; \beta, r, \Lambda)} \right\},$$

is strictly larger than zero if $(\beta, r, \Lambda) \neq (\beta_0, r_0, \Lambda_0)$.

(C3) There exit constants $0 < a < b < \infty$ so that $a \leqslant \inf_{z\in\mathbb{R}} r(z) \leqslant \sup_{z\in\mathbb{R}} r(z) \leqslant b$, $\forall r(\cdot) \in \mathcal{R}$.

(C4) The support of X is a bounded convex set of $\mathbb{R}^d$. For any $\beta \in \mathcal{B}$, the density of $\beta^\top X$ is bounded from above and below by some constants.

(C5) The quantity $E\big[\Delta r_0^2(X^\top\beta_0)\{r^{-1}(X^\top\beta) - r_0^{-1}(X^\top\beta_0)\}^2\big] - E\,\Delta\{S(Y, r, \beta) - S(Y, r_0, \beta_0)\}^2 S^{-2}(Y, r_0, \beta_0)$ is bounded from below and above up to some constants by $E\{r(X^\top\beta) - r_0(X^\top\beta_0)\}^2$, where $S(y, r, \beta) = E\{I(Y \geqslant y)/r(X^\top\beta)\}$.

Proof of Model Identifiability

Assume that there exists two sets of parameters $\{\beta_1, \lambda_1(t), r_1(z))\}$ and $\{\beta_2, \lambda_2(t), r_2(z))\}$, with $\int_0^\tau \lambda_1(t)dt = \int_0^\tau \lambda_2(t)dt = 1$, $r_1(z)$ and $r_2(z)$ being non-decreasing, and $\|\beta_1\| = \|\beta_2\| = 1$, that give the same conditional hazard function, that is,

$$\frac{\lambda_1(t)}{r_1(x^\top\beta_1)} = \frac{\lambda_2(t)}{r_2(x^\top\beta_2)} \quad \text{for all } t, x,$$

or, equivalently,

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{r_1(x^\top \beta_1)}{r_2(x^\top \beta_2)} \quad \text{for all } t, x.$$

We want to show that $\lambda_1 = \lambda_2$, $r_1 = r_2$, and $\beta_1 = \beta_2$. Since the left-hand-side of the equality depends only on t while the right-hand-side depends only on x, the equality holds if and only if there exists a constant k such that

$$\lambda_1(t) = k\, \lambda_2(t) \quad \text{and} \quad r_1(x^\top \beta_1) = k\, r_2(x^\top \beta_2) \text{ for all } t, x.$$

The identifiability condition $\int_0^\tau \lambda_1(t)dt = \int_0^\tau \lambda_2(t)dt = 1$ implies $k = 1$, and hence we have $\lambda_1(t) = \lambda_2(t)$ for all t and $r_1(x\beta_1) = r_2(x\beta_2)$ for all x. It is clear that $r_1 \equiv r_2$ under the monotone constraint if $\beta_1 = \beta_2$, hence it suffices to show $\beta_1 = \beta_2$.

Since r is continuous and nonconstant, there exists $B(x_0, \kappa) = \{x_0 + \gamma u, \|\gamma\| = 1, u \in [-\kappa, \kappa]\}$ for some x_0 such that r is nonconstant on $B(x_0, \kappa)$. For any $w \in [-\kappa, \kappa]$, it follows from the identifiability condition $\|\beta_1\| = \beta_1^\top \beta_1 = 1$ that

$$r_1(x_0^\top \beta_1 + u) = r_1((x_0 + u\beta_1)^\top \beta_1) = r_2((x_0 + u\beta_1)^\top \beta_2) = r_2(x_0^\top \beta_2 + \beta_1^\top \beta_2 u),$$

and, similarly,

$$r_2(x_0^\top \beta_2 + u) = r_2((x_0 + u\beta_2)^\top \beta_2) = r_1((x_0 + u\beta_2)^\top \beta_1) = r_1(x_0^\top \beta_1 + \beta_2^\top \beta_1 u).$$

Without loss of generalizability, we assume that the first element of β is positive. So if $\beta_1 \neq \beta_2$, we have $|\beta_1^\top \beta_2| < 1$ by Cauchy-Schwartz inequality. As a result,

$$r_1(x_0^\top \beta_1 + u) \;=\; r_2(x_0^\top \beta_2 + \beta_1^\top \beta_2 u) = r_1(x_0^\top \beta_1 + (\beta_1^\top \beta_2)^2 u) = \cdots = r_1(x_0^\top \beta_1).$$

For any $x \in B(x_0, \kappa)$, we can express it as $x = x_0 + \gamma w$ for some unit vector γ and $w \in [-\kappa, \kappa]$. Thus we have $r_1(x^\top \beta_1) = r_1(x_0^\top \beta_1 + \gamma^\top \beta_1 w) = r_1(x_0^\top \beta_1)$. This implies r_1 is constant on $B(x_0, \kappa)$, which is a contradiction. Hence we show that $\beta_1 = \beta_2$, and therefore $r_1 \equiv r_2$.

Proof of Theorem 1

We denote expectation with respect to the empirical distribution of the data by P_n and denote expectation with respect to the true underlying distribution of the data by P. Define

$$\ell(\beta, r, \Lambda) = -\delta \log r(x^\top \beta) + \delta \log d\Lambda(y) - \Lambda(y)/r(x^\top \beta),$$

where $d\Lambda(y) = \Lambda(y) - \Lambda(y-)$. Let $(\widehat{\beta}, \widehat{r}, \widehat{\Lambda})$ be the maximizer of the likelihood under the constraints, that is, for any (β, r, Λ),

$$P_n \ell(\widehat{\beta}, \widehat{r}, \widehat{\Lambda}) \geqslant P_n \ell(\beta, r, \Lambda),$$

where $\widehat{\Lambda}$, Λ are monotonically nondecreasing functions satisfying $\widehat{\Lambda}(0) = \Lambda(0) = 0$ and $\widehat{\Lambda}(\tau) = \Lambda(\tau) = 1$, $\widehat{r}$, r are monotonically nondecreasing and bounded functions, and $\|\widehat{\beta}\| = \|\beta\| = 1$. Define $N(t) = \Delta I(Y \leqslant t)$ and $R(t) = I(Y \geqslant t)$. Let

$$\widehat{\Lambda}_1(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n R_i(u)/\widehat{r}(x_i^\top \widehat{\beta})}$$

be the Breslow type estimator. The constrained MLE for Λ is the normalized version

$$\widehat{\Lambda}(t) = \widehat{\Lambda}_1(t)/\widehat{\Lambda}_1(\tau).$$

Since $\widehat{r}(\cdot)$ is a bounded monotonic function and has bounded variation, by Helly's selection theorem, it has a convergence subsequence to a function $r^*(\cdot)$. Moreover, since $\widehat{\beta}$ falls in a compact subset of $\mathbb{R}^p$, it also has a convergence subsequence to a limiting value β^*. Then there exists a further subsequence $\{n_k\}$ along which $\widehat{r}(z) \to r^*(z)$ and $\|\widehat{\beta} - \beta^*\| \to 0$, and thus along $\{n_k\}$, $\widehat{\Lambda}(t)$ converges almost surely to $\Lambda^*(t)$, where $\Lambda^*(t) = \Lambda_1^*(t)/\Lambda_1^*(\tau)$ and $\Lambda_1^*(t) = \int_0^t \frac{P\{dN(u)\}}{P\{R(u)/r^*(x^\top\beta^*)\}}$.

Define

$$\widetilde{\Lambda}_1(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n R_i(u)/r_0(x_i^\top\beta_0)}, \ \widetilde{\Lambda}(t) = \widetilde{\Lambda}_1(t)/\widetilde{\Lambda}_1(\tau).$$

By Lemma 9.10 and Corollary 9.27 in Kosorok (Kosorok, 2008), it can be seen that $\{N(t) : t \geqslant 0\}$ and $\{R(t)/r_0(x^\top\beta_0) : t \geqslant 0\}$ are Glivenko-Cantelli classes. Then $\widetilde{\Lambda}(t)$ converges almost surely to $\Lambda_0(t)$, where $\Lambda_1^0(t) = \int_0^t \frac{P\{dN(u)\}}{P\{R(u)/r_0(x^\top\beta_0)\}}$ and $\Lambda_0(t) = \Lambda_1^0(t)/\Lambda_1^0(\tau)$.

Note that $\{\delta \log r(x^\top\beta) - \Lambda(y)/r(x^\top\beta) : (\beta, r, \Lambda) \in (\mathcal{B}, \mathcal{R}, \mathcal{A})\}$ is a Glivenko-Cantelli class since it is indexed by monotonic functions Λ, r and parameters β and it has an integrable envelop function (Theorem 3, van der Vaart and Wellner (van der Vaart and Wellner, 2000)). Then we have

$$\begin{aligned}
& P_{n_k}\{\ell(\widehat{\beta}, \widehat{r}, \widehat{\Lambda}) - \ell(\beta_0, r_0, \widetilde{\Lambda})\} \\
= \ & -P_{n_k}\delta \log \frac{\widehat{r}(x^\top\widehat{\beta})}{r_0(x^\top\beta_0)} + P_{n_k}\delta \log \frac{d\widehat{\Lambda}(y)}{d\widetilde{\Lambda}(y)} - P_{n_k}\{\widehat{\Lambda}(y)/\widehat{r}(x^\top\widehat{\beta}) - \widetilde{\Lambda}(y)/r_0(x^\top\beta_0)\} \\
= \ & -P_{n_k}\delta \log \frac{\widehat{r}(x^\top\widehat{\beta})}{r_0(x^\top\beta_0)} + P_{n_k}\delta \log \frac{\widetilde{\Lambda}_1(\tau)P_{n_k}\{R(y)/r_0(x^\top\beta_0)\}}{\widehat{\Lambda}_1(\tau)P_{n_k}\{R(y)/\widehat{r}(x^\top\widehat{\beta})\}} - P_{n_k}\{\widehat{\Lambda}(y)/\widehat{r}(x^\top\widehat{\beta}) - \widetilde{\Lambda}(y)/r_0(x^\top\beta_0)\} \\
\xrightarrow{a.s.} \ & -P\delta \log \frac{r^*(x^\top\beta^*)}{r_0(x^\top\beta_0)} + P\delta \log \frac{\Lambda_1^0(\tau)P\{R(y)/r_0(x^\top\beta_0)\}}{\Lambda_1^*(\tau)P\{R(y)/r^*(x^\top\beta^*)\}} - P\{\Lambda^*(y)/r^*(x^\top\beta^*) - \Lambda_0(y)/r_0(x^\top\beta_0)\} \\
= \ & -P\delta \log \frac{r^*(x^\top\beta^*)}{r_0(x^\top\beta_0)} + P\delta \log \frac{d\Lambda^*(y)}{d\Lambda_0(y)} - P\{\Lambda^*(y)/r^*(x^\top\beta^*) - \Lambda_0(y)/r_0(x^\top\beta_0)\}.
\end{aligned}$$

Therefore, we obtain $0 \leqslant P_{n_k}\{\ell(\widehat{\beta}, \widehat{r}, \widehat{\Lambda}) - \ell(\beta_0, r_0, \widetilde{\Lambda})\} \to P \log \frac{dP_{\beta^*, r^*, \Lambda^*}}{dP_{\beta_0, r_0, \Lambda_0}}$, where $P_{\beta, r, \Lambda}$ is the probability measure of a single observation on the specified model at parameter value (β, r, Λ). By identifiability of the model, we have

$$\Lambda^* = \Lambda_0, r^* = r_0, \beta^* = \beta_0.$$

Therefore, we have shown that any convergence subsequence has a limiting to the true underlying parameters. Since every subsequence of n contains a further subsequence for which $(\widehat{\beta}, \widehat{r}, \widehat{\Lambda})$ converges uniformly to $(\beta_0, r_0, \Lambda_0)$, we have convergence for the entire sequence.

Proof of Proposition 1

We need to show that there exists a constant $c^* < 0$ such that $\beta = c^*\beta_0$ is the solution to the limiting value of the partial score equation,

$$U(\beta) = E\{XN(\tau)\} - \int_0^\tau \frac{E\{X \exp(X^\top\beta) I(Y \geqslant t)\}}{E\{\exp(X^\top\beta) I(Y \geqslant t)\}} E\{dN(t)\} = 0, \tag{2}$$

where $N(t) = \Delta I(Y \leqslant t)$. Without loss of generality, we assume $E(X) = 0$. Define $W = X^\top \beta_0$. Using the property of elliptically symmetric random variables, we have $E\{X \mid W\} = Wb$, where b is a $p \times 1$ deterministic vector such that $b = \Sigma\beta_0(\beta_0^\top \Sigma\beta_0)^{-1}$ and $\Sigma = var(X)$. Under random censoring, it can be shown that for any $c \in \mathbb{R}$,

$$\begin{aligned} E\{X \exp(cX^\top \beta_0) I(Y \geqslant t)\} &= E\{\exp(cW) I(Y \geqslant t) W\} b, \\ E\{\exp(cX^\top \beta_0) I(Y \geqslant t)\} &= E\{\exp(cW) I(Y \geqslant t)\}, \\ E\{X dN(u)\} &= E\{W dN(u)\} b. \end{aligned}$$

By plugging the above quantities into (2), we have

$$U(c\beta_0) = b\left[E\{WN(\tau)\} - \int_0^\tau \frac{E\{W \exp(cW) I(Y \geqslant t)\}}{E\{\exp(cW) I(Y \geqslant t)\}} E\{dN(t)\}\right].$$

Le c^* be the limiting value of estimated c as $n \to \infty$ from fitting the Cox model $\lambda(t \mid W) = \lambda(t)\exp\{cW\}$ under the true model $\lambda(t \mid W) = \lambda(t)\exp\{-\phi(W)\}$. Then we have $U(c^*\beta_0) = 0$. Moreover, the monotonicity of ϕ will result in a negative value of c^*.

References

Ambler G, Benner A (2015). mfp: Multivariable fractional polynomials. R package version 1.5.2.

Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26: 641–647.

Balabdaoui F, Groeneboom P, Hendrickx K (2019). Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics*, 46(2): 517–544.

Chen K, Guo S, Sun L, Wang JL (2010). Global partial likelihood for nonparametric proportional hazards models. *Journal of the American Statistical Association*, 105(490): 750–760.

Chen S, Zhou L (2007). Local partial likelihood estimation in proportional hazards regression. *The Annals of Statistics*, 35(2): 888–916.

Chung Y, Ivanova A, Hudgens MG, Fine JP (2018). Partial likelihood estimation of isotonic proportional hazards models. *Biometrika*, 105: 133–148.

Cox DR (1975). Partial likelihood. *Biometrika*, 62(2): 269–276.

Du Y, Varadhan R (2020). SQUAREM: An R package for off-the-shelf acceleration of EM, MM and other EM-like monotone algorithms. *Journal of Statistical Software*, 92(7): 1–41.

Fan J, Gijbels I, King M, et al. (1997). Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25(4): 1661–1690.

Foster JC, Taylor JMG, Nan B (2013). Variable selection in monotone single-index models via the adaptive lasso. *Statistics in Medicine*, 32(22): 3944–3954.

Gentleman R, Crowley J (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics*, 47: 1283–1296.

Groeneboom P, Hendrickx K (2018). Current status linear regression. *The Annals of Statistics*, 46: 1415–1444.

Groeneboom P, Hendrickx K (2019). Estimation in monotone single-index models. *Statistica Neerlandica*, 73: 78–99.

Härdle W, Hall P, Ichimura H (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21: 157–178.

Härdle W, Stoker TM (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408): 986–995.

Huang J, Wellner J (1997). Interval censored survival data: A review of recent progress. In: *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*. 123–169.

Huang JZ, Liu L (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, 62(3): 793–802.

Ichimura H (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1–2): 71–120.

Jongbloed G (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7(3): 310–321.

Kosorok MR (2008). *Introduction to empirical processes and semiparametric inference Springer Series in Statistics*. Springer, New York.

Murphy S, van der Vaart AW, Wellner J (1999). Current status regression. *Mathematical Methods of Statistics*, 8(3): 407–425.

Nelder JA, Mead R (1965). A simplex method for function minimization. *The Computer Journal*, 7(4): 308–313.

Powell JL, Stock JH, Stoker TM (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57: 1403–1430.

Prentice RL, Self SG (1983). Asymptotic distribution theory for cox-type regression models with general relative risk form. *The Annals of Statistics*, 11: 804–813.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robertson T, Wright FT, Dykstra R (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester.

Schumacher M, Bastert G, Bojar H, Huebner K, Olschewski M, Sauerbrei W, et al. (1994). Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German breast cancer study group. *Journal of Clinical Oncology*, 12(10): 2086–2093.

Tibshirani R, Hastie T (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82(398): 559–567.

van der Vaart A, Wellner JA (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. 115–133.

Wang W, Wang JL, Wang Q (2009). Proportional hazards regression with unknown link function. *IMS Lecture Notes-Monograph Series*, 57: 47–66.

Wang W, Yan J (2021). Shape-restricted regression splines with R package splines2. *Journal of Data Science*, 19(3): 498–517.