Statistical Methods in Medical Research
XX(X):2–24

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Partly linear single-index cure models with a nonparametric incidence link function

Chun Yin Lee[1], Kin Yau Wong[1,2] and Dipankar Bandyopadhyay[3]

Abstract

In cancer studies, it is commonplace that a fraction of patients participating in the study are *cured*, such that not all of them will experience a recurrence, or death due to cancer. Also, it is plausible that some covariates, such as the treatment assigned to the patients or demographic characteristics, could affect both the patients' survival rates and cure/incidence rates. A common approach to accommodate these features in survival analysis is to consider a mixture cure survival model with the incidence rate modeled by a logistic regression model and latency part modeled by the Cox proportional hazards model. These modeling assumptions, though typical, restrict the structure of covariate effects on both the incidence and latency components. As a plausible recourse to attain flexibility, we study a class of semiparametric mixture cure models in this paper, which incorporates two single-index functions for modeling the two regression components. A hybrid nonparametric maximum likelihood estimation method is proposed, where the cumulative baseline hazard function for uncured subjects is estimated nonparametrically, and the two single-index functions are estimated via Bernstein polynomials. Parameter estimation is carried out via a curated EM algorithm. We also conducted a large-scale simulation study to assess the finite-sample performance of the estimator. The proposed methodology is illustrated via application to two cancer datasets.

Keywords

Bernstein polynomial, EM algorithm, mixture cure models, sieve estimation, survival analysis.

1 Introduction

In conventional survival analysis, a usual assumption is that all subjects will experience the event of interest in their lifetime. This assumption can be violated in some applications when there exists a fraction of the population that is non-susceptible to the event. For example, breast cancer patients [1;2] and melanoma patients [3;4] may be relapse-free for the rest of their lifetime after receiving a certain treatment at which all cancer cells are eliminated. In these studies, it is preferable to accommodate the existence of long-term survivors, or a cure fraction, in the population. The mixture cure model is the most popular class of cure models, which comprises two basic components, namely the incidence and latency components. The term 'incidence' pertains to whether an event would occur (i.e. whether a subject is uncured with finite failure time, or not), whereas the term 'latency' refers to when the event *actually* occurs, given the event will occur. In the pioneering work of Boag [5] and Berkson and Gage, [6] a subject is classified as cured or uncured according to a binary latent variable where the incidence parameter is an unknown constant, and the failure time of an uncured subject is assumed to follow a covariate-independent parametric distribution.

To introduce covariates to the incidence component, Farewell [7] and Ghitany et al. [8] assumed a logistic regression model for the probability of being uncured. Lam et al. [2] proposed the probit and complementary log-log models as alternatives to the logistic regression model. Although the popular logistic link function used in the incidence component is easy to interpret and implement, Amico et al. [9] have criticized its S-shaped functional form for being too restrictive, and a prespecified link function may not fit the data well in practice. They proposed a flexible nonparametric non-monotone link function for the incidence probability, which is more robust against different shapes of underlying functions compared to fully parametric models. In the same vein, Musta and Yuen [10] studied the use of a nonparametric monotone link function on the incidence probability model.

Farewell [11] and Ghitany et al. [8] assumed a Weibull and exponential regression model, respectively, for the latency component to capture the effects of covariates on the failure time distribution. Based on the semiparametric Cox proportional hazards (PH) model, [12] Kuk and Chen [13] proposed a maximum marginal likelihood approach with Monte Carlo approximations for parameter estimation, while Sy and Taylor [14] proposed

[1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong; [2]Shenzhen Research Institute, Hong Kong Polytechnic University; [3]Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia

Corresponding author:
Dipankar Bandyopadhyay, Department of Biostatistics, School of Population Health, Virginia Commonwealth University, 830 East Main St, PO Box 980032, Richmond, VA 23298-0032, USA.
Email: dbandyop@vcu.edu

Prepared using sagej.cls

the Expectation-Maximization (EM) algorithm where the parameters in the incidence and latency components can be updated separately in the M-step, and the profile likelihood[15] approach is still applicable in the estimation of the latency parameters. Lu[16] proposed a nonparametric maximum likelihood estimation approach, with the unknown cumulative baseline hazard function approximated via a step function, while Wang et al.[17] proposed a smoothing splines approach for the variance model for each of the two components, and adopted a penalized EM algorithm for estimation. A comprehensive summary of the development of the cure survival model was provided in Amico and Van Keilegom.[18]

The Cox PH model assumes that the covariates have linear effects on the log-hazard function. Nevertheless, there are practical situations where this assumption may not be satisfied. For example, in a variety of epidemiological and cancer studies, the effects of important covariates, such as age, tobacco/alcohol use, and socioeconomic status, can be nonlinear. In such situations, more flexible models are necessary to obtain reliable inferences.

In this paper, we develop a single-index mixture cure model that allows an unspecified link function for the incidence model, and a nonlinear effect of a single index in the latency model. In single-index models (SIMs), we assume that a linear combination of covariates, referred to as an index, affects the outcome variable nonparametrically. SIM[19–21] and partially linear single-index models (PLSIM)[22;23] have been studied extensively for non-censored data, serving as extensions to the generalized linear models that relaxes the linearity assumptions. For the analysis of right-censored data, the SIM and PLSIM are mainly constructed as generalizations of the Cox PH regression model.[24–28] A major motivation for considering the SIM/PLSIM over the fully nonparametric models[29;30] is to avoid the "curse of dimensionality" issue when the dimension of the covariates involved in the nonparametric function is large. Another popular alternative to the SIM for coping with the dimensionality issue is the additive model,[31] where covariate effects can be characterized nonparametrically. In comparison, the SIM is more computationally tractable in the presence of numerous covariates. Also, as remarked in Yu and Ruppert,[23] interactions among covariates are completely ignored in the additive models, but they are automatically captured in the SIM.

To fit SIM/PLSIM, a variety of estimation methods have been proposed, which include but are not limited to the kernel smoothing methods[9;21;24] and spline approximations;[25;27] piecewise linear functions[32;33] are special cases of splines. A particular choice of splines is the Bernstein polynomial (BP),[34] which has been adopted to approximate the baseline cumulative hazard functions in the bivariate transformation survival models,[35] to approximate nonlinear covariate effects in the additive Cox model for interval-censored data,[36] and to approximate the distribution function in the semiparametric transformation non-mixture cure models.[37] The BP is easy to implement, even if there is a monotonicity constraint on the underlying function, and requires no pre-specification of the interior knots. To the best of our knowledge, the BP has not been considered for the estimation of SIM/PLSIM in the analysis of survival data with a cure fraction. Utilizing the BP approach for SIM/PLSIM estimation is the central contribution of this paper.

The rest of the paper is organized as follows. Section 2 introduces the model, likelihood, associated sieve-nonparametric maximum likelihood (NPML) estimation

method, and related model identifiability. The computational details are presented in Section 3, where a carefully devised EM algorithm powers the sieve-NPML estimation. The standard error estimation for the Euclidean parameter estimates is also discussed. In Section 4, we conduct a simulation study to evaluate the finite-sample performance of the estimator. The proposed methodology is illustrated via application to two cancer datasets, on head and neck cancer and breast cancer. Finally, some concluding remarks appear in Section 5.

2 Methods

2.1 Model specification

Let U be a binary latent variable, which equals 0 if a subject is cured and 1 otherwise. Let T be the failure time, with $T = \infty$ if $U = 0$, and T follows a finite distribution if $U = 1$. Let $\boldsymbol{X}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ be p-, q- and r- dimensional vectors of covariates, respectively; they do not contain any constant elements. We assume that $\boldsymbol{W}$ and $\boldsymbol{Z}$ do not overlap, but both of them can overlap with $\boldsymbol{X}$. Furthermore, we assume a SIM for the incidence with

$$P(U = 1|\boldsymbol{X}) = \kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X})\} \equiv \pi(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}), \tag{1}$$

where $\kappa(x) = \{1 + \exp(-x)\}^{-1}$ is the standard logistic function, G is an unknown smooth monotone increasing link function, and $\boldsymbol{\alpha}$ is a vector of regression parameters. Model (1) includes the logistic regression model, with $G(x) = c + x$, and the complementary log-log model, with $G(x) = \log\left[\exp\{\exp(c + x)\} - 1\right]$, as special cases, where c is an unknown parameter. We propose a PLSIM for the latency component. The conditional hazard function of T given $U = 1$ takes the form

$$\lambda_u(t|\boldsymbol{W}, \boldsymbol{Z}) = \lambda(t) \exp\left\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{W} + H(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z})\right\}, \tag{2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression parameters, λ is an arbitrary baseline hazard function, and H is an unknown smooth link function. In model (2), we assume a linear effect of $\boldsymbol{W}$ on the log-hazard, but the effect of $\boldsymbol{Z}$ is potentially nonlinear. An appealing feature of the PLSIM is that the class of models includes both the partially linear survival model (with $\boldsymbol{Z}$ being univariate)[32;33] and the single-index survival models (in the absence of $\boldsymbol{W}$)[24;25] as special cases. Based on (1) and (2), the population survival function is

$$S(t|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z}) = 1 - \pi(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}) + \pi(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X})S_u(t|\boldsymbol{W}, \boldsymbol{Z}), \tag{3}$$

where $S_u(t|\boldsymbol{W}, \boldsymbol{Z}) = \exp\left[-\Lambda(t) \exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{W} + H(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z})\}\right]$, and $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$.

To ensure model identifiability, we note that any additive constants in H could be absorbed in λ; so, we impose the constraint $H(0) = 0$. Similarly, any additive constants in G could be absorbed, so $\boldsymbol{X}$ does not contain any constant elements. We also require $\|\boldsymbol{\alpha}\| = \|\boldsymbol{\gamma}\| = 1$, since the scaling factors of the parameters could be absorbed in G and H. In addition, the sign of $\boldsymbol{\gamma}$ cannot be identified, since we can find $\breve{H}(\cdot) = H(-\cdot)$ and $\breve{\boldsymbol{\gamma}} = -\boldsymbol{\gamma}$, such that $H(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}) = \breve{H}(\breve{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{Z})$. Therefore, we assume that the first component of $\boldsymbol{\gamma}$ is positive. By contrast, such a constraint is not needed for $\boldsymbol{\alpha}$, as G is monotone increasing.

2.2 Likelihood

Suppose that the failure time T is subject to right censoring, and let C be the random censoring time. We only observe $Y \equiv \min(T, C)$ and the event indicator $\Delta \equiv \mathrm{I}(T < C)$. We assume that T and C are independent given the covariates $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z})$. A random sample of size n comprises $\mathcal{O} = \{(Y_i, \Delta_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i), i = 1, \ldots, n\}$. Let $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda, G, H)$ be the set of all unknown parameters. We further let $\boldsymbol{\eta} \equiv (\boldsymbol{\beta}, H, \boldsymbol{\gamma})$ be the set of parameters in the survival model, and let $h_i(\boldsymbol{\eta}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{W}_i + H(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i)$. Then, the likelihood function for $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}|\mathcal{O}) = \prod_{i=1}^{n} \left[\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\lambda(Y_i)\exp\{h_i(\boldsymbol{\eta})\}\exp[-\Lambda(Y_i)\exp\{h_i(\boldsymbol{\eta})\}]\right]^{\Delta_i} \times \left[1 - \kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\} + \kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\exp[-\Lambda(Y_i)\exp\{h_i(\boldsymbol{\eta})\}]\right]^{1-\Delta_i}.$$

The maximization of this objective function is challenging because (i) it possesses three infinite-dimensional nonparametric components, namely Λ, G and H; and (ii) the partial likelihood approach is not applicable (despite a traditional right-censoring setup) since the failure time distribution depends on the cure status.

2.3 Sieve estimation

We propose a sieve-NPML estimation method to approximate the likelihood in a parameter space with finite dimensions. We approximate Λ by a step function with non-negative jump sizes at the unique observed event times. For G and H, we employ sieve methods and approximate the functions using BP. We define the sieve space for the approximation of G as

$$\mathcal{B}_G = \left\{G_{m_1}(x) = B(x; \boldsymbol{\psi}, m_1, u) : -M_g \leqslant \psi_0 \leqslant \cdots \leqslant \psi_{m_1} < M_g\right\},$$

where

$$B(x; \boldsymbol{\psi}, m_1, u) \equiv \sum_{j=0}^{m_1} \psi_j B_j(x, m_1, u) = \sum_{j=0}^{m_1} \psi_j \binom{m_1}{j} \left(\frac{x+u}{2u}\right)^j \left(1 - \frac{x+u}{2u}\right)^{m_1-j}$$

is a BP with degree m_1, M_g is some large constant, $\boldsymbol{\psi} \equiv (\psi_0, \ldots, \psi_{m_1})^{\mathrm{T}}$ is a vector of coefficients of the basis polynomials, and u is some large enough positive value such that $[-u, u]$ contains the support of $\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}$. In practice, we can set $u = \max_{i=1,\ldots,n} \|\boldsymbol{X}_i\|$. The order constraints on $\boldsymbol{\psi}$ ensure that the members of $\mathcal{B}_G$ are monotone increasing. Similarly, we define the sieve space

$$\mathcal{B}_H = \left\{H_{m_2}(z) = B(z; \boldsymbol{\phi}, m_2, v) : |\phi_i| < M_h, i = 0, \ldots, m_2\right\}$$

for the approximation of H, where M_h is a positive constant, $\boldsymbol{\phi} \equiv (\phi_0, \ldots, \phi_{m_2})^{\mathrm{T}}$ is a vector of coefficients of the basis polynomials, and v is such that $[-v, v]$ contains the support of $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{Z}$; we set $v = \max_{i=1,\ldots,n} \|\boldsymbol{Z}_i\|$. We denote the sieve maximum likelihood estimator by $\widehat{\boldsymbol{\theta}} \equiv (\widehat{\boldsymbol{\zeta}}, \widehat{\Lambda}, \widehat{G}, \widehat{H})$ where $\boldsymbol{\zeta} \equiv (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the set of Euclidean parameters.

2.4 Identifiability

Let $\boldsymbol{\zeta}_0 \equiv (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0), \Lambda_0, G_0$, and H_0 be the true parameter values. Let $\mathcal{S} \subset \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r$ be the support of $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z})$, and let $\mathcal{D}_1$ and $\mathcal{D}_2$ be the supports of $\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{Z}$, respectively. Let τ be the end-of-study time. We impose the following conditions for model identifiability.

(C1) The function G_0 is monotone increasing, differentiable, and non-constant on $\mathcal{D}_1$. The function H_0 is differentiable and non-constant on $\mathcal{D}_2$, with $H_0(0) = 0$. The function Λ_0 is strictly increasing and continuous on $[0, \tau]$ with $\Lambda_0(0) = 0$.

(C2) The parameters $\boldsymbol{\alpha}_0 \in \mathcal{A}_\alpha$ and $\boldsymbol{\gamma}_0 \in \mathcal{A}_\gamma$, where $\mathcal{A}_\alpha = \left\{(\alpha_1, \ldots, \alpha_p) \mid \sum_{j=1}^p \alpha_j^2 = 1\right\}$ and $\mathcal{A}_\gamma = \left\{(\gamma_1, \ldots, \gamma_r) \mid \sum_{j=1}^r \gamma_j^2 = 1, \gamma_1 \geqslant \epsilon\right\}$ for some positive constant ϵ.

(C3) The supports of $\boldsymbol{X}$ and $(\boldsymbol{W}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$ are not contained in any proper linear subspace of $\mathbb{R}^p$ and $\mathbb{R}^{q+r}$, respectively.

(C4) Both $\boldsymbol{X}$ and $\boldsymbol{Z}$ are continuous, and $P(\|\boldsymbol{X}\| + \|\boldsymbol{Z}\| + \|\boldsymbol{W}\| < M) = 1$ for some constant M. The set $\mathcal{S}$ includes the value $(\mathbf{0}, \mathbf{0}, \mathbf{0})$. For any $\boldsymbol{x} \neq 0$, there exists $(\boldsymbol{w}, \boldsymbol{z})$ such that $\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w} + H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z}) \neq 0$ and $(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}) \in \mathcal{S}$.

(C5) The censoring time C satisfies $P(C \geqslant \tau \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = P(C = \tau \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) > \epsilon$ almost surely, for some positive constant ϵ and $\tau < \infty$.

Remark 1. *Conditions (C1)–(C3) include regularity conditions for the identifiability of general single-index models, similar to those in Theorem 2.1 of Horowitz.*[38] *In particular, condition (C1) guarantees that there are no point masses in the event time distribution and no jumps in the transformation functions G and H. Condition (C2) guarantees that the scales of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, as well as the sign of $\boldsymbol{\gamma}$, are identifiable. Condition (C3) ensures that the covariates are not linearly dependent. Condition (C4) imposes mild conditions on the covariates that facilitate separate identification of the incidence and latency parts. In particular, it requires that the covariates have non-zero effects on the latency part. Condition (C5) guarantees that the event can be observed up to time τ.*

Proposition 1. *Under conditions (C1)–(C5), model (3) is identifiable.*

The proof of the proposition is provided in the Appendix.

Remark 2. *The aforementioned identifiability conditions are stated with $\boldsymbol{X}$ and $\boldsymbol{Z}$ being vectors of continuous covariates. According to Horowitz,*[38] *when $\boldsymbol{X}$ and $\boldsymbol{Z}$ contain a mixture of discrete and continuous variables, we additionally require*

(C6) The function H is nonperiodic.

(C7) When the values of the discrete component in $\boldsymbol{X}$ (and $\boldsymbol{Z}$) vary, the support $\mathcal{D}_1$ (and $\mathcal{D}_2$) must not be divided into disjoint subsets.

3 Computational details

3.1 EM algorithm

We devise an EM algorithm for the computation of the sieve NPML estimator. It suffices to write down the complete data log-likelihood function based on the augmented data $\mathcal{O}^{'} = \{(Y_i, \Delta_i, U_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i), i = 1, \ldots, n\}$, which takes the form

$$\ell(\boldsymbol{\theta} \mid \mathcal{O}^{'}) = \sum_{i=1}^{n} U_i \log\left[\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\right] + (1-U_i)\log\left[1-\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\right] \quad (4)$$
$$+ \sum_{i=1}^{n} \Delta_i \Big\{ \log\lambda(Y_i) + h_i(\boldsymbol{\eta}) \Big\} - U_i \Lambda(Y_i)\exp\{h_i(\boldsymbol{\eta})\}.$$

Clearly, the complete data log-likelihood is a sum of two terms. The first term only involves the parameters in (1), and the second term only involves the parameters in (2).

In the sequel, we use the superscript (d) to denote the parameter values in the dth step of the EM algorithm, $d = 0, \ldots$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_1 = (\boldsymbol{\alpha}, G)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \Lambda, H)$. In the E-step, we evaluate

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) \equiv \mathrm{E}\left\{\ell(\boldsymbol{\theta}) \mid \mathcal{O}, \boldsymbol{\theta}^{(d)}\right\} = Q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}^{(d)}) + Q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}^{(d)}) \quad (5)$$

where

$$Q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}^{(d)}) = \sum_{i=1}^{n} \widehat{\mathrm{E}}(U_i)\log\left[\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\right] + \Big\{1-\widehat{\mathrm{E}}(U_i)\Big\}\log\left[1-\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\right],$$
$$Q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}^{(d)}) = \sum_{i=1}^{n} \Delta_i \Big\{ \log\lambda(Y_i) + h_i(\boldsymbol{\eta}) \Big\} - \widehat{\mathrm{E}}(U_i)\Big[\Lambda(Y_i)\exp\{h_i(\boldsymbol{\eta})\}\Big],$$

and $\widehat{\mathrm{E}}$ denotes the conditional expectation given the observed data, evaluated at the parameter value $\boldsymbol{\theta}^{(d)}$. It suffices to compute

$$\widehat{\mathrm{E}}(U_i) = \Delta_i + \left.\frac{(1-\Delta_i)\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\exp\left[-\Lambda(Y_i)\exp\{h_i(\boldsymbol{\eta})\}\right]}{1-\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}+\kappa\{G(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\}\exp\left[-\Lambda(Y_i)\exp\{h_i(\boldsymbol{\eta})\}\right]}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(d)}}.$$

In the M-step, we propose a numerically stable two-stage optimization procedure for updating the pairs $(\boldsymbol{\psi}, \boldsymbol{\alpha})$ and $(\boldsymbol{\phi}, \boldsymbol{\gamma})$, where the BP coefficients and the single-index coefficients are updated sequentially.

We update $\boldsymbol{\theta}_1$ by maximizing $Q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}^{(d)})$, subject to the constraint $\|\boldsymbol{\alpha}\| = 1$. Note that the monotonicity constraints in $\mathcal{B}_G$ can be easily satisfied by the parameterization $\psi_0 = \psi_0^*$ and $\psi_q = \psi_0^* + \sum_{i=1}^{q} e^{\psi_i^*}$ for $q = 1, \ldots, m_1$, where $\psi_j^* \in (-\infty, \infty)$ for $j = 0, \ldots, m_1$. We first fix $\boldsymbol{\alpha}$ at $\boldsymbol{\alpha}^{(d)}$ and obtain $\boldsymbol{\psi}^{(d+1)}$ by the quasi-Newton method of Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm[39] for unconstrained nonlinear optimization problems. Then, by setting $\boldsymbol{\psi} = \boldsymbol{\psi}^{(d+1)}$, and for some real number c_1,

we update $(\boldsymbol{\alpha}^{(d)}, c_1^{(d)})$ by the standard Newton–Raphson iteration for the unconstrained Lagrange function, in which we have to compute

$$\begin{pmatrix} \boldsymbol{\alpha}_{(l+1)} \\ c_{1_{(l+1)}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_{(l)} \\ c_{1_{(l)}} \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 Q_1}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^{\mathrm{T}}} + 2c_1 \boldsymbol{I}_p & 2\boldsymbol{\alpha} \\ 2\boldsymbol{\alpha}^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial Q_1}{\partial \boldsymbol{\alpha}} + 2c_1 \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\alpha} - 1 \end{pmatrix} \Bigg|_{(\boldsymbol{\alpha}, c_1) = (\boldsymbol{\alpha}_{(l)}, c_{1_{(l)}})}$$

in an iterative manner for $l = 0, 1, \ldots$, where $(\boldsymbol{\alpha}_{(0)}, c_{1_{(0)}}) = (\boldsymbol{\alpha}^{(d)}, c_1^{(d)})$, and $\boldsymbol{I}_p$ denotes the identity matrix of size p. We obtain $(\boldsymbol{\alpha}^*, c_1^*)$ upon convergence and set $(\boldsymbol{\alpha}^{(d+1)}, c_1^{(d+1)}) = (\boldsymbol{\alpha}^*, c_1^*)$.

Next, we update $\boldsymbol{\theta}_2$ by maximizing $Q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}^{(d)})$, subject to the constraint $\|\boldsymbol{\gamma}\| = 1$. For fixed $\boldsymbol{\gamma}$, we can update $(\boldsymbol{\beta}^{(d+1)}, \boldsymbol{\phi}^{(d+1)})$ by first profiling out Λ in the expected complete-data log-likelihood. The "profile log-likelihood" for $(\boldsymbol{\beta}^{(d+1)}, \boldsymbol{\phi}^{(d+1)})$ takes the form

$$Q_3(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \Delta_i \left(h_i(\boldsymbol{\eta}) - \log \left[\sum_{j=1}^{n} R_j(Y_i) \widehat{\mathrm{E}}(U_j) \exp\{h_j(\boldsymbol{\eta})\} \right] \right), \tag{6}$$

where $R_j(t) \equiv \mathrm{I}(Y_j \geqslant t)$ is the risk-set indicator of the jth individual at time t. Hence, with $\boldsymbol{\gamma}$ fixed at $\boldsymbol{\gamma}^{(d)}$, we update $(\boldsymbol{\beta}, \boldsymbol{\phi})$ by maximizing Q_3. To avoid potential numerical instability under poorly assigned initial parameter values, we adopt a step-halving[40] Newton–Raphson algorithm that searches for candidates with current estimate plus 2^{-K} times $(K = 0, 1, \ldots)$ the usual updating term of the Newton–Raphson algorithm. Specifically, for a given K:

$$\begin{pmatrix} \boldsymbol{\beta}_{(K)} \\ \boldsymbol{\phi}_{(K)}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(d)} \\ \boldsymbol{\phi}^{(d)} \end{pmatrix} - 2^{-K} \begin{pmatrix} \frac{\partial^2 Q_3}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} & \frac{\partial^2 Q_3}{\partial \boldsymbol{\beta} \partial \boldsymbol{\phi}^{\mathrm{T}}} \\ \frac{\partial^2 Q_3}{\partial \boldsymbol{\phi} \partial \boldsymbol{\beta}^{\mathrm{T}}} & \frac{\partial^2 Q_3}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^{\mathrm{T}}} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial Q_3}{\partial \boldsymbol{\beta}} \\ \frac{\partial Q_3}{\partial \boldsymbol{\phi}} \end{pmatrix} \Bigg|_{(\boldsymbol{\beta}, \boldsymbol{\phi}) = (\boldsymbol{\beta}^{(d)}, \boldsymbol{\phi}^{(d)})}.$$

For identifiability, we set $\boldsymbol{\phi}_{(K)} = \boldsymbol{\phi}_{(K)}^* - H_{m_2}(0; \boldsymbol{\phi}_{(K)}^*)\mathbf{1}$ where $\mathbf{1}$ is a $(m_2 + 1)$-vector of ones, such that $H_{m_2}(0; \boldsymbol{\phi}_{(K)}) = 0$ is satisfied for every K. Then, we set $(\boldsymbol{\beta}^{(d+1)}, \boldsymbol{\phi}^{(d+1)}) = (\boldsymbol{\beta}_{(K*)}, \boldsymbol{\phi}_{(K*)})$, where $K*$ is chosen to be the smallest non-negative integer K such that $Q_3(\boldsymbol{\beta}_{(K)}, \boldsymbol{\phi}_{(K)}, \boldsymbol{\gamma}^{(d)}) \geqslant Q_3(\boldsymbol{\beta}^{(d)}, \boldsymbol{\phi}^{(d)}, \boldsymbol{\gamma}^{(d)})$. The Breslow-type estimator of Λ is given by

$$\Lambda^{(d+1)}(t) = \sum_{i=1}^{n} \frac{\Delta_i \mathrm{I}(Y_i \leqslant t)}{\sum_{j=1}^{n} R_j(Y_i) \widehat{\mathrm{E}}(U_j) \exp\{h_j(\boldsymbol{\eta})\}}.$$

Finally, we fix $(\boldsymbol{\beta}, \Lambda, \boldsymbol{\phi})$ at $(\boldsymbol{\beta}^{(d+1)}, \Lambda^{(d+1)}, \boldsymbol{\phi}^{(d+1)})$, and update $\boldsymbol{\gamma}$ by maximizing Q_2. Analogous to the above, we use the Lagrange multiplier method to resolve the constraint $\|\boldsymbol{\gamma}\| = 1$. For some real number c_2, we update $(\boldsymbol{\gamma}, c_2)$ iteratively, with

$$\begin{pmatrix} \boldsymbol{\gamma}_{(l+1)} \\ c_{2_{(l+1)}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}_{(l)} \\ c_{2_{(l)}} \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 Q_2}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{\mathrm{T}}} + 2c_2 \boldsymbol{I}_r & 2\boldsymbol{\gamma} \\ 2\boldsymbol{\gamma}^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial Q_2}{\partial \boldsymbol{\gamma}} + 2c_2 \boldsymbol{\gamma} \\ \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{\gamma} - 1 \end{pmatrix} \Bigg|_{(\boldsymbol{\gamma}, c_2) = (\boldsymbol{\gamma}_{(l)}, c_{2_{(l)}})}$$

for $l = 0, 1, \ldots$, where $(\boldsymbol{\gamma}_{(0)}, c_{2_{(0)}}) = (\boldsymbol{\gamma}^{(d)}, c_2^{(d)})$, and obtain $(\boldsymbol{\gamma}^*, c_2^*)$ upon convergence. We set $c_2^{(d+1)} = c_2^*$. If the first element of $\boldsymbol{\gamma}^*$ is positive, we set $\boldsymbol{\gamma}^{(d+1)} = \boldsymbol{\gamma}^*$. If it is negative, we set $\boldsymbol{\gamma}^{(d+1)} = -\boldsymbol{\gamma}^*$ and, by virtue of the symmetry of BPs, reverse the order of the elements in $\boldsymbol{\phi}^{(d+1)}$. We summarize the algorithm below.

Step 0 Start with $d = 0$ and initial value $\boldsymbol{\theta}^{(0)}$.

Step 1 With $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(d)}$ fixed, obtain $\boldsymbol{\psi}^{(d+1)}$ by maximizing Q_1 via the BFGS algorithm.

Step 2 With $\boldsymbol{\psi} = \boldsymbol{\psi}^{(d+1)}$ fixed, obtain $\boldsymbol{\alpha}^{(d+1)}$ by the Lagrange multiplier method.

Step 3 With $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(d)}$ fixed, obtain $(\boldsymbol{\beta}^{(d+1)}, \boldsymbol{\phi}^{(d+1)})$ by maximizing Q_3 with a step-halving Newton–Raphson algorithm, where $\boldsymbol{\phi}^{(d+1)}$ satisfies $H_{m_2}(0; \boldsymbol{\phi}^{(d+1)}) = 0$.

Step 4 Compute the Breslow-type estimator $\Lambda^{(d+1)}$.

Step 5 Obtain $\boldsymbol{\gamma}^{(d+1)}$ by the Lagrange multiplier method.

Step 6 Repeat Steps 1 to 5 for $d = 1, 2, \ldots$ until convergence at which the maximum absolute difference between two consecutive estimates of $\boldsymbol{\theta}$ is less than a small positive constant.

In the algorithm, we update $(\boldsymbol{\beta}, \boldsymbol{\phi}, \Lambda)$ by maximizing the expected complete-data log-likelihood $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ at fixed $\boldsymbol{\gamma}$ and then update $\boldsymbol{\gamma}$ by maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ at the updated $(\boldsymbol{\beta}, \boldsymbol{\phi}, \Lambda)$. Alternatively, we can update $(\boldsymbol{\beta}, \boldsymbol{\phi})$ by maximizing Q_3 at fixed $\boldsymbol{\gamma}$, update $\boldsymbol{\gamma}$ by maximizing Q_3 at the updated $(\boldsymbol{\beta}, \boldsymbol{\phi})$, and then update Λ using the Breslow-type estimator at the updated $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$. Both methods increase the value of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ in the M-step and should yield the sieve NPML estimator.

There are several merits of using the EM algorithm for maximizing the observed likelihood. First, the E-step involves only a simple closed-form expression concerning the conditional mean of the latent variable. Second, the objective function in (5) can be considered as two separate components that involve solely $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, which simplifies the maximization problem. Third, the algorithm is computationally efficient because (i) the commonly used Breslow-type estimator of Λ is preserved with an explicit solution; and (ii) the pairs of coefficients $(\boldsymbol{\alpha}, \boldsymbol{\psi})$ and $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ are updated sequentially, such that no inversion of high-dimensional matrices is involved in the estimation procedures.

Remarks: We present three remarks on the computational aspects of our algorithm. First, as proposed in Taylor,[41] a zero-tail constraint is typically required for the numerical stability of the EM algorithm. Hence, we set the conditional survival function to be 0 for those censored observations with observed failure times beyond the largest observed event time in the estimation procedures, that is, we set $\widehat{\mathrm{E}}(U_i) = 0$ for any subject i censored after the last observed event time. Second, the updating formula in (6) is concave in $(\boldsymbol{\beta}, \boldsymbol{\phi})$, but not necessarily concave in $\boldsymbol{\gamma}$. To avoid converging to local maxima, we suggest to begin with multiple sets of initial parameter values and choose the desired estimator to be the set of estimates that yields the largest observed likelihood. Third, in applying the proposed method, one has to decide the degrees of the BPs, namely

m_1 and m_2, used in the approximation of G and H respectively. We propose to set $m_1 = K_1\lfloor n^{1/4}\rfloor$ for a positive integer K_1, where, $\lfloor a\rfloor$ is the largest integer smaller than a, and use a data-adaptive approach to select m_2 based on the Akaike information criterion (AIC). Note that we adopt slightly different strategies to select m_1 and m_2. Since G is monotone, including many knots in the BP would not cause numerical instability in the estimation. Hence, we allow m_1 to increase with the sample size based on a simple rule. By contrast, H is generally nonlinear and can be very flexible. In our experience, using an information criterion, such as the AIC, to select m_2 yields an appropriate number of knots to capture the shape of the true function and gives a more stable estimation.

3.2 Confidence interval estimation

The computation of the inverse of the observed Fisher information matrix can be expensive due to the presence of the high-dimensional parameter Λ. To avoid heavy computations, we propose to use the profile likelihood approach[42] to approximate the covariance matrix of $\widehat{\boldsymbol{\zeta}}$. If α_1 is positive, then we reparameterize $\boldsymbol{\alpha} = ((1 - \|\boldsymbol{\rho}\|^2)^{1/2}, \boldsymbol{\rho}^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\rho} \equiv \boldsymbol{\alpha}_{-1} \equiv (\alpha_2, \ldots, \alpha_p)^{\mathrm{T}}$; otherwise, we reparametrize $\boldsymbol{\alpha} = (-(1 - \|\boldsymbol{\rho}\|^2)^{1/2}, \boldsymbol{\rho}^{\mathrm{T}})^{\mathrm{T}}$. Likewise, we write $\boldsymbol{\gamma}^{\mathrm{T}} = ((1 - \|\boldsymbol{\omega}\|^2)^{1/2}, \boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\omega} = \boldsymbol{\gamma}_{-1} \equiv (\gamma_2, \ldots, \gamma_r)^{\mathrm{T}}$. The inference for $\boldsymbol{\zeta}^* \equiv (\boldsymbol{\rho}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}}$ can be performed based on the profile likelihood function for $\boldsymbol{\zeta}^*$.

As noted by Zeng et al.[43], the covariance matrix estimator based on the sum of squares of individual score statistics is guaranteed to be positive semi-definite, and is more computationally efficient and stable than that based on the individual hessian matrix. Let

$$\mathrm{pl}(\boldsymbol{\zeta}^*) = \sup_{\Lambda, G, H} \log L(\Lambda, G, H; \boldsymbol{\zeta}^*)$$

be the profile-log-likelihood, which can be obtained through updating only Λ, G and H in the M-step of the EM-algorithm at fixed $\boldsymbol{\zeta}^*$. The estimated covariance matrix for $\widehat{\boldsymbol{\zeta}}^*$ is given by the inverse of

$$\boldsymbol{I}(\widehat{\boldsymbol{\zeta}}^*) \equiv \sum_{i=1}^{n} \left[\left\{ \frac{\partial}{\partial \boldsymbol{\zeta}^*} \mathrm{pl}_i(\boldsymbol{\zeta}^*) \Big|_{\boldsymbol{\zeta}^* = \widehat{\boldsymbol{\zeta}}^*} \right\}^{\otimes 2} \right],$$

where pl_i is the contribution to pl from the ith subject. We approximate $\partial \mathrm{pl}_i(\boldsymbol{\zeta}^*)/\partial \boldsymbol{\zeta}^*$ numerically by

$$\frac{\mathrm{pl}_i(\widehat{\boldsymbol{\zeta}}^* + h_n e_k) - \mathrm{pl}_i(\widehat{\boldsymbol{\zeta}}^*)}{h_n},$$

where, e_k is the kth canonical vector in $\mathbb{R}^{p+q+r-2}$, and h_n is a pre-specified perturbation constant that depends on n. One complication in applying this method to the single-index models is that the norm constraints may not be satisfied when a perturbation constant is added to an element in $\boldsymbol{\rho}$ or $\boldsymbol{\omega}$. For the perturbation of the jth element in $\boldsymbol{\rho}$, namely ρ_j for $j = 1, \ldots, (p-1)$, we propose to choose $h_n = -\mathrm{sign}(\widehat{\rho}_j) \min(|\widehat{\rho}_j|, K_2 n^{-1/2})$ for some positive integer K_2, and we choose h_n for the perturbation of elements in $\boldsymbol{\omega}$ in

the same manner. We simply take $h_n = K_2 n^{-1/2}$ for the perturbation of $\boldsymbol{\beta}$. Since $\widehat{\boldsymbol{\zeta}}^*$ is asymptotically normal, it is easy to observe that $\widehat{\boldsymbol{\zeta}}$ is also asymptotically normal, via the delta method. Specifically, the estimate for the covariance matrix of $\widehat{\boldsymbol{\zeta}}$ is given by

$$\boldsymbol{B}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\zeta}}^*)\boldsymbol{B}^{\mathrm{T}}.$$

where

$$\boldsymbol{B} = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{-1}^{\mathrm{T}}/\widehat{\alpha}_1 & \mathbf{0}_{1\times q} & \mathbf{0}_{1\times(r-1)} \\ \boldsymbol{I}_{p-1} & \mathbf{0}_{(p-1)\times q} & \mathbf{0}_{(p-1)\times(r-1)} \\ \mathbf{0}_{q\times(p-1)} & \boldsymbol{I}_q & \mathbf{0}_{q\times(r-1)} \\ \mathbf{0}_{1\times(p-1)} & \mathbf{0}_{1\times q} & \widehat{\boldsymbol{\gamma}}_{-1}^{\mathrm{T}}/\widehat{\gamma}_1 \\ \mathbf{0}_{(r-1)\times(p-1)} & \mathbf{0}_{(r-1)\times q} & \boldsymbol{I}_{r-1} \end{pmatrix}.$$

4 Simulation study

The finite-sample performance of the proposed estimator is studied via extensive simulation studies. We generate four independent variables where X_1 takes up the values -1 and 1 with equal probability, and X_2, X_3 and X_4 are independent standard normal random variables. The true parameter values are set to be $\boldsymbol{\alpha} = (0.5, -0.5, 0.5, -0.5)^{\mathrm{T}}$, $\boldsymbol{\beta} = (0.5, -0.5)^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (1/\sqrt{2}, -1/\sqrt{2})^{\mathrm{T}}$. Here, the covariates $\boldsymbol{W} = (X_1, X_2)^{\mathrm{T}}$ and $\boldsymbol{Z} = (X_3, X_4)^{\mathrm{T}}$. The baseline cumulative hazard function is set to be $\Lambda(t) = 0.8t^{1.2}$. Three scenarios are considered. In Scenario I, we have the classical logistic-Cox (LC) model with $\pi(x) = \kappa(x)$ and zero intercept value for the incidence component, and $H(z) = z$ for the latency component. Let Φ denote the cumulative distribution function of a standard normal random variable. For Scenario II, we have $\pi(x) = 0.4 + 0.6\Phi(2x - 1)$, a function that starts at level 0.4 and increases to 1, and $H(z) = \log(1 + z^2)$ is a v-shaped function. Finally, for Scenario III, we consider $\pi(x) = 0.5\Phi(2x + 3) + 0.5\Phi(2x - 3)$ and $H(z) = \sin(3z/2)$. The black solid lines in Figure 1 display π and H under Scenarios I–III. We set the censoring time C to follow an exponential distribution with a rate parameter 0.2. The response variables Y and Δ are generated as follows. For each individual, we first generate the binary random variable U according to (1). Then, we set $T = \infty$ if $U = 0$ and T follows the PLSIM described in (2) otherwise. Thus, we obtain Y and Δ by comparing the generated values of T and C. The average censoring rates are 63%, 48%, and 61%, and the average cure proportions are 50%, 40%, and 50% in Scenarios I, II, and III, respectively.

For the BP implementation, we fix $m_1 = K_1\lfloor n^{1/4}\rfloor$ $(K_1 = 1, 2, 3)$, and for each fixed m_1, choose m_2 from the candidates $\{1, \ldots, 5\}$ based on the AIC. For the implementation of the EM algorithm, we assign standard normal random variables re-scaled to norm 1 for $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$, respectively. We set $\boldsymbol{\psi}^{(0)}$ such that $\kappa\{G(\cdot)\}$ is roughly linear and increases from 0 to 1 over the support $[-u, u]$. We set $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, $\boldsymbol{\phi}^{(0)} = \mathbf{0}$, and set $\Lambda^{(0)}$ to be a step function that increases from 0 to 1 with equal step size at each observed event time. Five sets of random initial parameter values are generated for the parameter estimation for each given combination of m_1 and m_2. The convergence threshold of the EM algorithm is set to 10^{-3}. We set $K_2 = 1$, 5, and 10 in the perturbation constant h_n used in variance estimation of the Euclidean parameter estimators. For each scenario, we generate 1000

Table 1. Main simulation results. Table entries are the Bias, empirical standard deviation (ESD), average standard error (ASE), and coverage probability (CP) of the 95% nominal level, under Scenarios I–III, for $n = 500$ and 1000. Here, $K_1 = 3$ and $K_2 = 5$.

Scenario	Par.	$n = 500$				$n = 1000$			
		Bias	ESD	ASE	CP	Bias	ESD	ASE	CP
I	α_1	−0.019	0.112	0.108	0.94	−0.008	0.073	0.074	0.95
	α_2	0.014	0.114	0.108	0.93	0.001	0.077	0.075	0.94
	α_3	−0.012	0.116	0.109	0.94	−0.005	0.079	0.075	0.94
	α_4	0.008	0.114	0.109	0.94	0.009	0.077	0.075	0.94
	β_1	0.012	0.091	0.093	0.95	0.003	0.064	0.064	0.94
	β_2	−0.006	0.088	0.093	0.96	−0.002	0.063	0.064	0.95
	γ_1	−0.003	0.062	0.066	0.95	−0.001	0.042	0.044	0.97
	γ_2	0.003	0.064	0.066	0.95	0.001	0.041	0.044	0.96
II	α_1	−0.009	0.071	0.078	0.97	−0.007	0.048	0.052	0.96
	α_2	0.010	0.073	0.082	0.96	0.002	0.049	0.054	0.96
	α_3	0.001	0.073	0.080	0.96	−0.001	0.052	0.054	0.96
	α_4	0.003	0.075	0.081	0.94	0.000	0.049	0.054	0.96
	β_1	0.005	0.079	0.079	0.95	0.003	0.055	0.055	0.95
	β_2	−0.004	0.077	0.080	0.96	0.001	0.054	0.055	0.95
	γ_1	−0.001	0.070	0.074	0.93	−0.002	0.047	0.049	0.96
	γ_2	0.006	0.071	0.074	0.94	0.001	0.046	0.049	0.96
III	α_1	−0.017	0.134	0.140	0.94	−0.008	0.092	0.098	0.96
	α_2	0.009	0.117	0.127	0.94	0.008	0.080	0.084	0.95
	α_3	−0.015	0.120	0.126	0.94	−0.008	0.078	0.083	0.96
	α_4	0.020	0.120	0.126	0.94	0.003	0.080	0.084	0.95
	β_1	0.010	0.095	0.091	0.94	0.001	0.062	0.063	0.96
	β_2	−0.012	0.091	0.094	0.96	−0.002	0.062	0.064	0.96
	γ_1	−0.005	0.062	0.063	0.95	−0.001	0.042	0.042	0.94
	γ_2	0.001	0.062	0.062	0.95	0.002	0.042	0.042	0.94

replicates with sample sizes $n = 500$ and 1000. Based on the simulation results (see Table S.1 and Figure S.1 in the Supplementary Materials), we suggest to use $K_1 = 3$, which is sufficiently large for the estimated function to capture potential changes in G, while $K_2 = 5$ works well in rendering the nominal 95% empirical coverage probability level in all scenarios considered. Table 1 reports the estimation results for each Euclidean parameter with $K_1 = 3$ and $K_2 = 5$, including the bias, empirical standard deviation (ESD), average standard error (ASE) and coverage probability (CP) of the 95% nominal level. The coverage probability is computed based on the 95% confidence intervals constructed via the asymptotic normality of the estimators. It shows that the estimator is virtually unbiased in all scenarios and that the proposed method is robust to different underlying functional forms of π and H. The ASE aligns closely with the ESD, which illustrates that the profile likelihood approach is reliable for standard error estimation. In Figure 1, we plot the average estimates for π and H overlayed with the true values. The average estimates align closely with the true values implying that the proposed sieve-NPML estimation approach also provides good approximations of the nonparametric components.

We compare the performance of the proposed model to the classical LC model, and the Single-index/Cox (SIC) model proposed in Amico et al.[9] We follow the exact EM algorithm in Amico et al. for the estimation of the SIC model with their default parameter values. The performance is evaluated based on the average squared error (ASQE) for

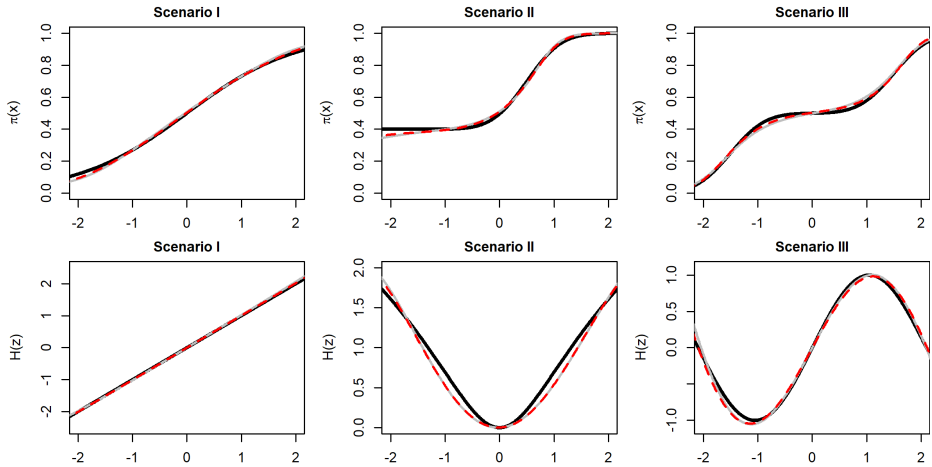

Figure 1. Average estimates of π and H in the simulation study. The black solid lines represent the true values. The grey solid line and red dashed line represent the average estimates under $n = 500$ and $n = 1000$, respectively.

the estimation of π and H in the incidence and latency components, respectively. Define a grid of realizations $(x_{i1}, x_{i2}, x_{i3}, x_{i4})$ of the variables (X_1, X_2, X_3, X_4) and a grid of realizations (z_{j1}, z_{j2}) of the variables (Z_1, Z_2), $i = 1, \ldots, B_1$ and $j = 1, \ldots, B_2$, where B_1 and B_2 denote the respective number of grid points. Here, X_1 takes values in $\{-1, 1\}$, whereas X_2, X_3, X_4 take the grid points on $[-1.5, 1.5]$ with a step size of 0.1. Then, for the proposed model and with estimates $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}, \widehat{G}, \widehat{H})$, we can compute $\mathrm{ASQE}_\pi = B_1^{-1} \sum_{i=1}^{B_1} \left[\kappa\{\widehat{G}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{x}_i)\} - \pi(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{x}_i)\right]^2$ and $\mathrm{ASQE}_H = B_2^{-1} \sum_{j=1}^{B_2} \left\{\widehat{H}(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{z}_j) - H(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{z}_j)\right\}^2$. Similarly, we can compute these two quantities for the LC model and SIC model, where the estimated uncured probability is used in ASQE_π, and $\overline{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{z}_j$ is used in ASQE_H, with $\overline{\boldsymbol{\gamma}}$ the estimated effect of $\boldsymbol{z}$. Figure 2 shows the results with $n = 500$ and 1000. For the incidence component, one can observe that the estimation performance of the three models is comparable when the logistic model is true, as in Scenario I (top-left figure). In particular, the LC model performs the best as expected. The proposed model is slightly superior to the SIC model because the logistic link function is indeed increasing. When the true model departs from the LC model but still maintains a monotone increasing incidence link function (i.e. Scenarios II and III), the proposed model outperforms the LC and SIC models. Next, as both LC and SIC models assumed a classical Cox PH model for the latency, the proposed model outperforms the LC and SIC models in the estimation of H under Scenarios II and III, where the true H is nonlinear. In Table 2, we present the simulation results for the estimation of β_1 and β_2 based on the LC and SIC models. The relative efficiency (RE) is defined as the ratio of the mean squared error of the proposed estimator to that

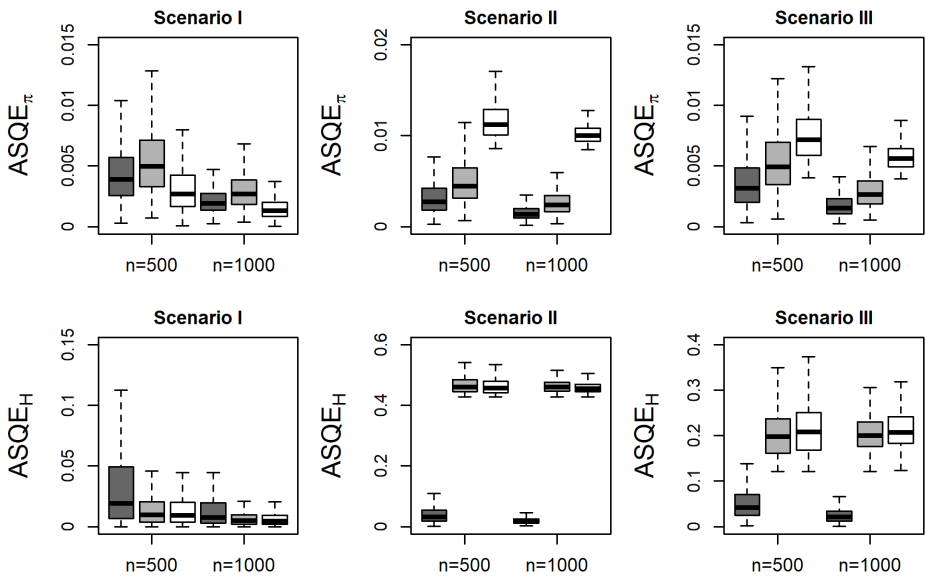

Figure 2. Average squared error (ASQE) for the estimation of π and H with $n = 500$ and 1000 in the simulation study. Dark grey, light grey and white box plots represent the ASQE computed based on the fitted proposed, SIC and LC models, respectively.

Table 2. Estimation results for parameters β_1 and β_2 based on the LC and SIC models where RE stands for relative efficiency compared to the estimator of the proposed model.

Scenario	Par.	$n = 500$			$n = 1000$		
		Bias	ESD	RE	Bias	ESD	RE
LC model							
I	β_1	0.008	0.089	1.052	0.002	0.064	1.003
	β_2	−0.003	0.086	1.041	−0.001	0.062	1.025
II	β_1	−0.075	0.077	0.539	−0.078	0.054	0.336
	β_2	0.070	0.076	0.557	0.075	0.053	0.345
III	β_1	−0.058	0.100	0.676	−0.068	0.067	0.420
	β_2	0.055	0.099	0.653	0.061	0.068	0.462
SIC model							
I	β_1	0.016	0.089	1.026	0.010	0.064	0.971
	β_2	−0.015	0.087	0.997	−0.011	0.063	0.982
II	β_1	−0.075	0.076	0.552	−0.076	0.054	0.350
	β_2	0.065	0.075	0.599	0.070	0.052	0.385
III	β_1	−0.056	0.100	0.700	−0.066	0.067	0.438
	β_2	0.044	0.098	0.733	0.054	0.068	0.516

of the LC and SIC models, respectively. As expected, the estimator for the LC model performs the best in Scenario I. However, we observe considerable efficiency gain by the proposed method in Scenarios II and III, compared to the estimators from the LC and SIC models. Hence, the results show that estimation accuracy and efficiency for $\boldsymbol{\beta}$ can be largely affected by the misspecification of the structures of π and H.

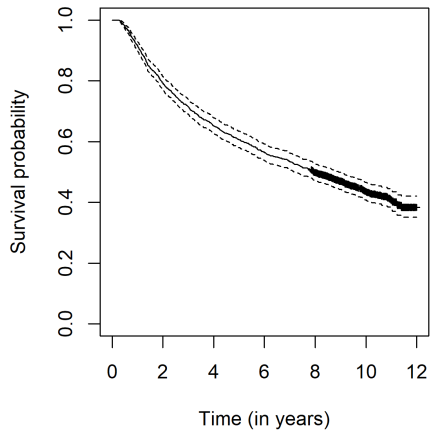

Figure 3. Kaplan-Meier estimator for the CHANCE data, with 95% pointwise confidence intervals.

5 Application

5.1 Head and neck cancer data

The first data set was collected from the Carolina Head and Neck Cancer Epidemiology (CHANCE) Study.[44;45] It was a retrospective case-control study that comprised the survival data of 1381 head and neck (HN) cancer patients and 1396 age-, sex- and race-matched controls. In this analysis, we focus on the HN cancer subjects, who were diagnosed with the first primary squamous cell carcinoma of the oral cavity, pharynx, or larynx, and resided in a 46-county region in central North Carolina between January 1, 2002, and February 28, 2006. The age-at-diagnosis of the patients ranges from 21 to 80 years. An administrative censoring occurred at the end of 2013, yielding a maximum follow-up period of 12 years. The primary objective is to identify potential risk factors that affect their survival since the initial tumor diagnosis. The plot of the Kaplan-Meier estimator of survival probability (see Figure 3) reveals a plateau behavior for the right tail at a high probability value, presumably due to the diminished impact of the initial tumor on survival.

We apply the proposed cure rate model to the CHANCE data. Various covariates are available to capture the heterogeneity between patients in terms of demographic background, disease severity, and oral health conditions, namely sex, race, smoking, alcohol use, tumor stage (T- and N- stages), number of dental visits in 10 years prior to study entry, and age-at-diagnosis in the analysis. Subjects with missing covariates are excluded. Subsequently, our analysis comprises 1255 HN cancer patients with 705 deceased during the 12-year observation period (i.e. censoring proportion is 44%). We standardize and include all the covariates in both the incidence and latency components. Referring to the PLSIM in (2), age-at-diagnosis is treated as Z since it is the only

Table 3. CHANCE data analysis results. Table entries are the estimate (Par), estimated standard error (ESE), and 95% confidence interval (95%CI) for the model parameters under the proposed model and LC model.

Covariates	Proposed model			LC model		
	Par	ESE	95%CI	Par	ESE	95%CI
Incidence part						
Intercept			–	0.547	0.092	(0.367, 0.726)
Sex (female = 1)	0.091	0.101	(−0.106, 0.288)	0.062	0.082	(−0.098, 0.222)
Race (white = 1)	−0.226	0.102	(−0.426, −0.027)	−0.134	0.085	(−0.301, 0.032)
Smoking (> 10 pack-year)	0.280	0.093	(0.097, 0.462)	0.246	0.082	(0.086, 0.407)
Alcohol (> 1 drink per week)	0.170	0.105	(−0.036, 0.377)	0.137	0.091	(−0.042, 0.316)
T-stage (T2–T4 vs T1)	0.214	0.099	(0.020, 0.408)	0.163	0.086	(−0.005, 0.332)
N-stage (N1–N3 vs N0)	0.231	0.095	(0.045, 0.417)	0.170	0.080	(0.013, 0.327)
Dental visits (⩾ 1 in 10 years)	−0.468	0.092	(−0.647, −0.288)	−0.391	0.083	(−0.554, −0.229)
Age-at-diagnosis	0.718	0.063	(0.594, 0.842)	0.581	0.088	(0.409, 0.754)
Latency part						
Sex (female = 1)	0.014	0.057	(−0.098, 0.126)	0.011	0.056	(−0.099, 0.121)
Race (white = 1)	−0.011	0.051	(−0.111, 0.089)	−0.004	0.054	(−0.109, 0.102)
Smoking (> 10 pack-year)	−0.007	0.059	(−0.122, 0.109)	−0.024	0.063	(−0.149, 0.100)
Alcohol (> 1 drink per week)	0.061	0.066	(−0.068, 0.190)	0.069	0.068	(−0.066, 0.203)
T-stage (T2–T4 vs T1)	0.160	0.060	(0.042, 0.277)	0.159	0.063	(0.034, 0.283)
N-stage (N1–N3 vs N0)	0.154	0.050	(0.057, 0.251)	0.159	0.051	(0.059, 0.259)
Dental visits (⩾ 1 in 10 years)	−0.156	0.056	(−0.267, −0.046)	−0.136	0.057	(−0.249, −0.024)
Age-at-diagnosis			–	−0.016	0.057	(−0.127, 0.095)

continuous variable in the dataset, and all other variables are treated as $\boldsymbol{W}$. As analogous to the simulation study, we set $K_1 = 3$ and $K_2 = 5$, and, based on the AIC, select $m_2 = 3$ from the candidates $\{1, \ldots, 8\}$. We consider 10 sets of different randomized initial parameter values under each candidate value of m_2, with the same randomization procedure as in the simulation studies. Then, we obtain the set of estimates with the highest likelihood. The convergence threshold of the EM algorithm is set to be 10^{-4}.

Table 3 summarizes the estimation results of the proposed model together with that of the LC model. Note, the estimates corresponding to the latency part pertain to effects on the log-hazard ratios. From the LC model, we observe that an increase in the incidence probability is associated with smoking, progressed N-stage, lack of dental visits, and ageing, while a decrease in survival rate is associated with both progressed T- and N-stages, and lack of dental visits. Specifically, sex, race, and alcohol consumption are not significant in both components of the LC model, and age-at-diagnosis is not significant in the latency component. Figure 4 illustrates the estimates of π and H in the proposed model. We observe that the uncured probability is capped at 0.9, and the shape of the fitted function deviates much from that of a logistic link function. This suggests that even though the follow-up time may not be long and the Kaplan-Meier curve in Figure 3 does not exhibit a plateau, we still identify a cure probability bounded away from zero; note that from Proposition 1, the cure probability is identifiable regardless of the end-of-study time τ. The estimated H function is nonlinear and non-monotone. On the contrary, the proposed model suggests that, in addition to the four identified risk factors in the LC model, race and progressed T-stage are significantly associated with the incidence probability. The conclusions drawn from the estimated linear effects of the covariates in

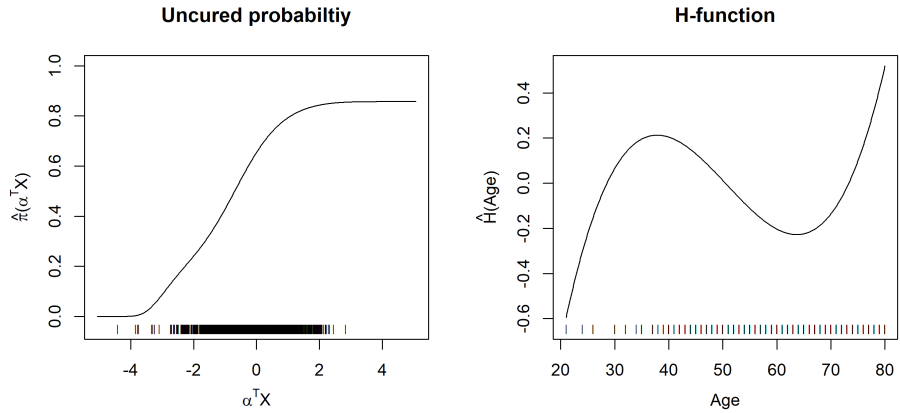

Figure 4. Estimates of π and H for the CHANCE data. The observed values of $\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{X}$ and the age variable are indicated at the bottom of the left and right panels, respectively.

the latency component are similar in both models, with comparable point and standard error estimates.

5.2 Rotterdam breast cancer data

The second motivating study is the Rotterdam breast cancer (RBC) data,[46] which includes the disease-free survival times of 2982 breast cancer patients, with a right censoring proportion of 42.6%. The Kaplan-Meier estimator presented in Figure 5 depicts a plateau well above zero at the right tail, suggesting that a certain fraction of breast cancer patients are cured. Nine prognostic baseline variables are available in the dataset, including hormonal and chemo- therapies indicators, tumor grade, menopausal status, tumor size, age, number of positive nodes, progesterone receptor, and estrogen receptor. The first five variables are discrete and the last four are continuous.

We apply the proposed methods to the RBC data. Since the continuous variables except age are right-skewed, we apply a logarithmic transformation $q(x) = \log(1 + x)$ to them. Then, we standardize and include all available covariates in both incidence and latency components. Specifically in the PLSIM in (2), the five discrete variables (i.e. hormonal and chemo-therapies indicators, tumor grade, menopausal status, tumor size) are treated as $\boldsymbol{W}$ and the four continuous variables (i.e. age, number of positive nodes, progesterone receptor, estrogen receptor) are treated as $\boldsymbol{Z}$. We employ the same configurations in the estimation procedure as in the analysis of CHANCE data (with $m_2 = 3$ selected via AIC). The estimation results of the proposed and LC models are summarized in Table 4, while the fitted nonparametric functions are provided in Figure 6. Once again, the estimates for the latency part pertain to effects on the log-hazard ratios. Both models suggest that an increase in incidence probability is associated with

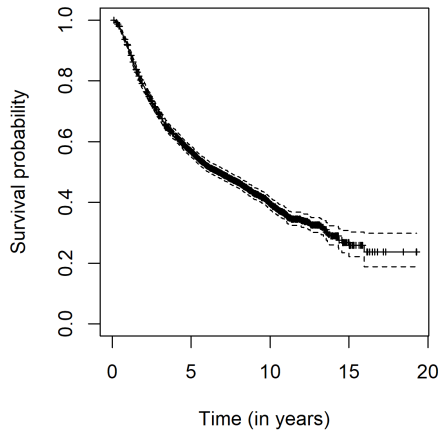

Figure 5. Kaplan-Meier estimator for the RBC data, with 95% pointwise confidence intervals.

Table 4. RBC data analysis results. Table entries are the estimate (Par), estimated standard error (ESE), and 95% confidence interval (95%CI) for the model parameters under the proposed model and LC model.

Covariates	Proposed model			LC model		
	Par	ESE	95%CI	Par	ESE	95%CI
Incidence part						
Intercept			–	1.351	0.148	(1.061, 1.642)
Hormonal therapy	−0.122	0.105	(−0.328, 0.084)	−0.114	0.110	(−0.330, 0.102)
Chemotherapy	−0.160	0.094	(−0.343, 0.023)	−0.174	0.090	(−0.351, 0.003)
Tumor grade	0.195	0.086	(0.026, 0.364)	0.139	0.089	(−0.035, 0.313)
Menopause	−0.108	0.129	(−0.361, 0.146)	−0.045	0.130	(−0.301, 0.211)
Tumor size	0.051	0.094	(−0.132, 0.234)	0.032	0.090	(−0.144, 0.208)
Age	0.247	0.120	(0.011, 0.482)	0.170	0.133	(−0.090, 0.430)
Number of positive nodes	0.812	0.064	(0.686, 0.938)	0.787	0.100	(0.592, 0.983)
Progesterone receptor	0.109	0.100	(−0.087, 0.306)	0.122	0.106	(−0.085, 0.329)
Estrogen receptor	0.419	0.102	(0.219, 0.619)	0.341	0.095	(0.155, 0.527)
Latency part						
Hormonal therapy	−0.130	0.030	(−0.189, −0.070)	−0.131	0.032	(−0.194, −0.068)
Chemotherapy	−0.200	0.035	(−0.269, −0.130)	−0.176	0.034	(−0.243, −0.109)
Tumor grade	0.135	0.037	(0.062, 0.207)	0.147	0.039	(0.070, 0.224)
Menopause	0.067	0.047	(−0.024, 0.159)	0.000	0.056	(−0.110, 0.110)
Tumor size	0.173	0.040	(0.095, 0.251)	0.181	0.042	(0.099, 0.264)
Age	0.406	0.069	(0.271, 0.542)	−0.107	0.053	(−0.210, −0.004)
Number of positive nodes	−0.814	0.031	(−0.876, −0.752)	0.441	0.031	(0.380, 0.501)
Progesterone receptor	0.341	0.063	(0.218, 0.464)	−0.184	0.041	(−0.264, −0.104)
Estrogen receptor	0.236	0.058	(0.123, 0.350)	−0.119	0.037	(−0.191, −0.048)

an increase in the number of positive nodes or estrogen receptors; chemotherapy is marginally significant, whereas tumor size is not significant. In contrast to the LC model, the proposed model additionally suggests that tumor grade (grade 2 versus grade 3) and

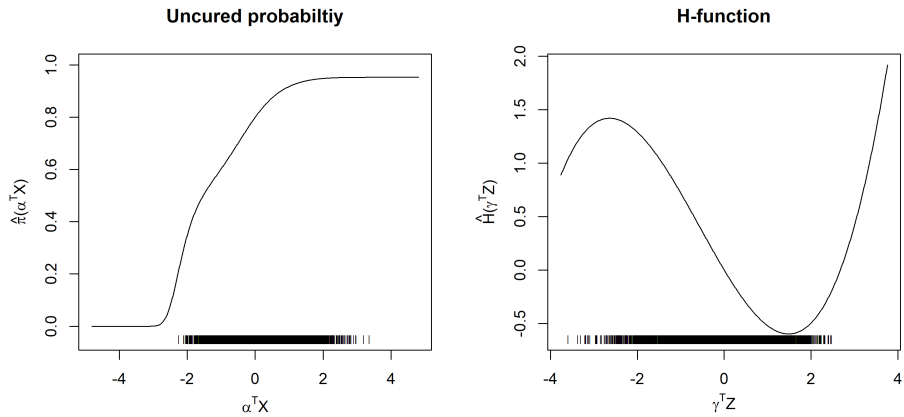

Figure 6. Estimates of π and H for the RBC data.

age at surgery are associated with the incidence probability. For the latency part, both models illustrate that the discrete variables except menopausal status are (significantly) predictive of the survival times of uncured subjects, where both therapies are shown to be effective in improving the survival rates of patients. The effects of all continuous variables are significant in the proposed model. Since the estimated H function is non-monotone, the interpretation of the parameter estimates of the continuous variables is not straightforward. However, from the magnitude of the estimated $\boldsymbol{\gamma}$, we observe that the number of positive nodes has the highest importance among the four variables, followed by age, and the two receptors.

6 Discussion

In this paper, we consider a flexible class of mixture cure models with a SIM for incidence probability and a PLSIM for the progression of the event time of the uncured subjects. A hybrid nonparametric approach is adopted for estimation, where the two link functions are approximated by the BPs, and the proposed estimator is computed via the EM algorithm. The simulation study demonstrates that the proposed model outperforms the LC model when the actual model deviates from the LC model, and it also outperforms the SIC model when the actual model has a monotone increasing incidence link function and a partially nonlinear structure in the latency component.

Massive and complex data arise naturally in the era of big data, providing unprecedented opportunities for developing more effective and predictive survival models. For instance, in the presence of numerous covariates $\boldsymbol{Z}$ and for dimension reduction, the single-index assumption in the latency component of the proposed model can be further generalized to the multiple-index model.[47–49] A partially linear multiple-index survival model may be specified as $\lambda_u(t|\boldsymbol{W},\boldsymbol{Z}) = \lambda(t)\exp\left\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{W} + V(\boldsymbol{\varsigma}_1^T\boldsymbol{Z},\ldots,\boldsymbol{\varsigma}_s^T\boldsymbol{Z})\right\}$ where V is an unknown s-variate function with

prespecified integer $s < r$. More recently, Zhong et al.[50] proposed the partially linear PH model with the nonlinear component estimated via the deep learning approach. They show that the rate of convergence for the linear effect estimate still achieves the semiparametric efficiency, even though the nonparametric component converges slower than $n^{1/2}$. Along this line, a natural extension of the proposed single-index cure model is to replace the single indexes with deep neural networks. In particular, the research work on single-index incidence probability modeling is scanty and deserves more attention with the implementation of some nonparametric methods. In this paper, we generalize the parametric link function with a nonparametric monotone single-index function, where the argument of the link function is assumed to be linear for simplicity. Future work may pertain to relaxing this assumption via deep learning methods, and exploring how this will affect the estimation of the latency components.

Acknowledgements

The authors would like to thank the Editor, Associate Editor, and four anonymous reviewers for their valuable feedback and insightful suggestions, which significantly enhanced the quality and presentation of this work.

Appendix

Proof of Proposition 1. In this proof, we consider a single observation and drop the subscript i. By the continuity of Λ_0 and condition (C5), we can set the survival time to be right censored at any time point within $[0, \tau]$ when establishing identifiability. Let $\boldsymbol{\theta}_0$ denote the set of true parameter values and $\widetilde{\boldsymbol{\theta}}$ denote an arbitrary set of parameter values. Suppose that the survival probabilities at $t \in [0, \tau]$ evaluated at the two sets of parameters are equal almost surely, i.e.,

$$\begin{aligned} &1 - \kappa\{G_0(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x})\} + \kappa\{G_0(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x})\} \exp\big\{ - \Lambda_0(t) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w} + H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z})} \big\} \\ =&1 - \kappa\{\widetilde{G}(\widetilde{\boldsymbol{\alpha}}^{T}\boldsymbol{x})\} + \kappa\{\widetilde{G}(\widetilde{\boldsymbol{\alpha}}^{T}\boldsymbol{x})\} \exp\big\{ - \widetilde{\Lambda}(t) e^{\widetilde{\boldsymbol{\beta}}^{T}\boldsymbol{w} + \widetilde{H}(\widetilde{\boldsymbol{\gamma}}^{T}\boldsymbol{z})} \big\} \end{aligned} \tag{7}$$

for all $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \in \mathcal{S}$ and $t \in [0, \tau]$. We wish to show that this implies $\boldsymbol{\alpha}_0 = \widetilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}_0 = \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0 = \widetilde{\boldsymbol{\gamma}}, \Lambda_0(t) = \widetilde{\Lambda}(t)$ for $t \in [0, \tau]$, $G_0(s) = \widetilde{G}(s)$ for $s \in \mathcal{D}_1$, and $H_0(s) = \widetilde{H}(s)$ for $s \in \mathcal{D}_2$. Define

$$\begin{aligned} L(s; \boldsymbol{x}) &= 1 - \kappa\{G_0(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x})\} + \kappa\{G_0(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x})\} \exp(-s) \\ S(t; \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}) &= L\{\Lambda_0(t) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w} + H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z})}; \boldsymbol{x}\}. \end{aligned}$$

For some small enough $\epsilon > 0$, $S(\cdot; \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}) \circ S^{-1}(\cdot; \mathbf{0}, \mathbf{0}, \mathbf{0})$ is well-defined on $(1 - \epsilon, 1]$, with

$$\begin{aligned} &S\{S^{-1}(s; \mathbf{0}, \mathbf{0}, \mathbf{0}); \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}\} \\ = \;& 1 - \kappa\{G_0(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x})\} + \kappa\{G_0(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x})\} \left[\frac{s - 1 + \kappa\{G_0(0)\}}{\kappa\{G_0(0)\}} \right]^{\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w} + H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z})\}}. \end{aligned}$$

Note that by condition (C4), $(\mathbf{0},\mathbf{0},\mathbf{0}) \in \mathcal{S}$. Let $\widetilde{S}(\cdot;\boldsymbol{x},\boldsymbol{w},\boldsymbol{z})$ denote $S(\cdot;\boldsymbol{x},\boldsymbol{w},\boldsymbol{z})$ with the parameter values set to be $\widetilde{\boldsymbol{\theta}}$. From (7), we have

$$S\{S^{-1}(s;\mathbf{0},\mathbf{0},\mathbf{0});\boldsymbol{x},\boldsymbol{w},\boldsymbol{z}\} = \widetilde{S}\{\widetilde{S}^{-1}(s;\mathbf{0},\mathbf{0},\mathbf{0});\boldsymbol{x},\boldsymbol{w},\boldsymbol{z}\} \tag{8}$$

for s in an open interval. Consider $\boldsymbol{x} \neq \mathbf{0}$ and $(\boldsymbol{w},\boldsymbol{z})$ such that $\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w} + H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z})$ is nonzero and $(\boldsymbol{x},\boldsymbol{w},\boldsymbol{z}) \in \mathcal{S}$. Clearly, (8) implies that $\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{w} + \widetilde{H}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{z}) \neq 0$. Differentiating both sides of (8) with respect to s, taking logarithm, and then differentiating again, we have

$$\frac{e^{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w}+H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z})}-1}{s-1+\kappa\{G_0(0)\}} = \frac{e^{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{w}+H(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{z})}-1}{s-1+\kappa\{\widetilde{G}(0)\}},$$

and therefore $\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w} + H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z}) = \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{w} + H(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{z})$ and $G_0(0) = \widetilde{G}(0)$. Plugging these results back into (8), we have

$$G_0(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x}) = \widetilde{G}(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{x}). \tag{9}$$

Note that this equality holds for all $\boldsymbol{x}$ in the support of $\boldsymbol{X}$. Differentiating both sides of (9) with respect to $\boldsymbol{x}$, we have

$$\boldsymbol{\alpha}_0 = \frac{\widetilde{G}'(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{x})}{G_0'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{x})}\widetilde{\boldsymbol{\alpha}}.$$

Because $\|\boldsymbol{\alpha}_0\| = \|\widetilde{\boldsymbol{\alpha}}\| = 1$ and G_0' and $\widetilde{G}'$ are positive, we conclude that $\boldsymbol{\alpha}_0 = \widetilde{\boldsymbol{\alpha}}$. From (9), we have $G_0(s) = \widetilde{G}(s)$ for $s \in \mathcal{D}_1$.

Similarly, by considering different values of $(\boldsymbol{w},\boldsymbol{z})$ (and arbitrary values of $\boldsymbol{x}$) in (8), we have

$$\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{w} + H_0(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{z}) = \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{w} + \widetilde{H}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{z}) \tag{10}$$

for all values of $(\boldsymbol{w},\boldsymbol{z})$ in the support of $(\boldsymbol{W},\boldsymbol{Z})$. By condition (C2) and a similar argument as the above, we conclude that $\boldsymbol{\beta}_0 = \widetilde{\boldsymbol{\beta}}$, $\boldsymbol{\gamma}_0 = \widetilde{\boldsymbol{\gamma}}$, and $H_0(s) = \widetilde{H}(s)$ for $s \in \mathcal{D}_2$. Finally, applying the established identifiability results to (7), we conclude that $\Lambda(t) = \widetilde{\Lambda}(t)$ for $t \in [0,\tau]$.

Declaration of conflicting interests

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

Funding

The authors acknowledge Prof. Andrew Olshan from the University of North Carolina for providing the motivating CHANCE data, and the context of this work. The CHANCE study was supported in part by grant R01CA90731 from the United States National Cancer Institute. This research was partially supported by the Central Guided Local Science and Technology Development Funds for Research Laboratories (No. 2021Szvup145). Bandyopadhyay acknowledges partial funding support from grants R21DE031879, R01DE031134 and P30CA016059 awarded by the United States National Institutes of Health.

Supplemental material

The R code associated with this article is available in the Github repository: `https://github.com/lcyjames/PLScure`, which includes the code for implementing the proposed methods and reproducible simulation results. The CHANCE dataset is not publicly available. The RBC dataset is publicly available from the R package `survival`. Supplemental material that contains additional simulation results are available online.

References

1. Peng Y and Dear KB. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; 56(1): 237–243.
2. Lam KF, Fong DY and Tang O. Estimating the proportion of cured patients in a censored sample. *Statistics in Medicine* 2005; 24(12): 1865–1879.
3. Chen MH, Ibrahim JG and Sinha D. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 1999; 94(447): 909–919.
4. Ibrahim JG, Chen MH and Sinha D. Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* 2001; 57(2): 383–388.
5. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society Series B (Methodological)* 1949; 11(1): 15–53.
6. Berkson J and Gage RP. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952; 47(259): 501–515.
7. Farewell VT. A model for a binary variable with time-censored observations. *Biometrika* 1977; 64(1): 43–46.
8. Ghitany M, Maller RA and Zhou S. Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis* 1994; 49(2): 218–241.
9. Amico M, Van Keilegom I and Legrand C. The single-index/Cox mixture cure model. *Biometrics* 2019; 75(2): 452–462.
10. Musta E and Yuen TP. Single-index mixture cure model under monotonicity constraints. *arXiv preprint arXiv:221109464* 2022; .
11. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; 38: 1041–1046.
12. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1972; 34(2): 187–220.
13. Kuk AYC and Chen CH. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* 1992; 79(3): 531–541.
14. Sy JP and Taylor JM. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; 56(1): 227–236.
15. Cox DR. Partial likelihood. *Biometrika* 1975; 62(2): 269–276.
16. Lu W. Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics* 2008; 60(3): 545–574.
17. Wang L, Du P and Liang H. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics* 2012; 68(3): 726–735.
18. Amico M and Van Keilegom I. Cure models in survival analysis. *Annual Review of Statistics and Its Application* 2018; 5: 311–342.

19. Friedman JH and Stuetzle W. Projection pursuit regression. *Journal of the American Statistical Association* 1981; 76(376): 817–823.
20. Härdle W and Stoker TM. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 1989; 84(408): 986–995.
21. Ichimura H. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics* 1993; 58(1-2): 71–120.
22. Carroll RJ, Fan J, Gijbels I et al. Generalized partially linear single-index models. *Journal of the American Statistical Association* 1997; 92(438): 477–489.
23. Yu Y and Ruppert D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 2002; 97(460): 1042–1054.
24. Wang W. Proportional hazards regression models with unknown link function and time-dependent covariates. *Statistica Sinica* 2004; 14: 885–905.
25. Huang JZ and Liu L. Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics* 2006; 62(3): 793–802.
26. Lu X, Chen G, Singh RS et al. A class of partially linear single-index survival models. *Canadian Journal of Statistics* 2006; 34(1): 97–112.
27. Sun J, Kopciuk KA and Lu X. Polynomial spline estimation of partially linear single-index proportional hazards regression models. *Computational Statistics and Data Analysis* 2008; 53(1): 176–188.
28. Shang S, Liu M, Zeleniuch-Jacquotte A et al. Partially linear single index Cox regression model in nested case-control studies. *Computational Statistics and Data Analysis* 2013; 67: 199–212.
29. Tibshirani R and Hastie T. Local likelihood estimation. *Journal of the American Statistical Association* 1987; 82(398): 559–567.
30. Dabrowska DM. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics* 1987; 14(3): 181–197.
31. Hastie T and Tibshirani R. Generalized additive models for medical research. *Statistical Methods in Medical Research* 1995; 4(3): 187–196.
32. Xue H, Lam KF and Li G. Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association* 2004; 99(466): 346–356.
33. Lee CY, Wong KY, Lam KF et al. Analysis of clustered interval-censored data using a class of semiparametric partly linear frailty transformation models. *Biometrics* 2022; 78(1): 165–178.
34. Lorentz GG. *Bernstein Polynomials*. American Mathematical Soc., 2013.
35. Zhou Q, Hu T and Sun J. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association* 2017; 112(518): 664–672.
36. Wu Q, Zhao H, Zhu L et al. Variable selection for high-dimensional partly linear additive Cox model with application to Alzheimer's disease. *Statistics in Medicine* 2020; 39(23): 3120–3134.
37. Lam KF, Lee CY, Wong KY et al. Marginal analysis of current status data with informative cluster size using a class of semiparametric transformation cure models. *Statistics in Medicine* 2021; 40(10): 2400–2412.
38. Horowitz JL. *Semiparametric and Nonparametric Methods in Econometrics*. Springer, 2009.

39. International Neural Network Society (INNS) tINNCCS, Battiti R and Masulli F. BFGS optimization for faster and automated supervised learning. In *International Neural Network Conference: July 9–13, 1990 Palais Des Congres—Paris—France*. Springer, pp. 757–760.
40. Jennrich RI and Sampson P. Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics* 1976; 18(1): 11–17.
41. Taylor JM. Semi-parametric estimation in failure time mixture models. *Biometrics* 1995; : 899–907.
42. Murphy SA and Van der Vaart AW. On profile likelihood. *Journal of the American Statistical Association* 2000; 95(450): 449–465.
43. Zeng D, Gao F and Lin DY. Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika* 2017; 104(3): 505–525.
44. Farquhar DR, Divaris K, Mazul AL et al. Poor oral health affects survival in head and neck cancer. *Oral Oncology* 2017; 73: 111–117.
45. Lenze NR, Farquhar D, Sheth S et al. Socioeconomic status drives racial disparities in HPV-negative head and neck cancer outcomes. *The Laryngoscope* 2021; 131(6): 1301–1309.
46. Foekens JA, Peters HA, Look MP et al. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer Research* 2000; 60(3): 636–643.
47. Cook RD and Li B. Dimension reduction for conditional mean in regression. *The Annals of Statistics* 2002; 30(2): 455–474.
48. Yin X and Cook RD. Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; 64(2): 159–175.
49. Chen D, Hall P and Müller HG. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics* 2011; 39(3): 1720–1747.
50. Zhong Q, Mueller J and Wang JL. Deep learning for the partially linear Cox model. *The Annals of Statistics* 2022; 50(3): 1348–1375.