Lifetime Data Analysis (2026) 32:14
https://doi.org/10.1007/s10985-026-09689-y

Deep learning for the change-point Cox model with current status data

Qiyue Huang[1] · Anyin Feng[2] · Qiang Wu[2] · Xingwei Tong[1,3]

Received: 10 July 2025 / Accepted: 2 January 2026

Abstract
This study develops estimation methods for a deep partially linear Cox proportional hazards model with a change point under current status data, aiming to accommodate complex change-point effects. Prior work has largely relied on linear models, which may inadequately capture relationships among multivariate covariates and thus hinder accurate change-point detection. To address this, we use a deep neural network to model covariate effects within the Cox framework and propose a maximum likelihood estimation procedure for the model. We establish asymptotic properties of the resulting estimators, including consistency, asymptotic independence, and semiparametric efficiency. Simulation studies indicate that the proposed inference procedure performs well in finite samples. An analysis of a breast cancer dataset is provided to illustrate the methodology.

Keywords Change point · Current status data · Deep neural network · Semiparametric efficiency

1 Introduction

A change-point model is a statistical framework for detecting points or periods at which the properties of the data shift abruptly (Pons 2003; Deng et al. 2017, 2022). Such models are vital across many domains, including economics (Li et al. 2024), environmental science, signal processing, and medical settings (Deng et al. 2022),

✉ Qiang Wu
qiangwu@mail.bnu.edu.cn

[1] School of Statistics, Beijing Normal University, Beijing, China

[2] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

[3] Department of Statistics, Faculty of Arts and Science, Beijing Normal University at Zhuhai, Zhuhai, China

Published online: 09 February 2026

where identifying when changes occur is crucial for decision-making and subsequent analysis.

Many studies have examined change-point models in survival analysis (Loader 1991; Pons 2003; Dupuy 2006; Kosorok and Song 2007; Torben and Scheike 2007; Oueslati and Lopez 2013; Deng et al. 2022). Among these, the Cox proportional hazards model has attracted the most attention. Luo and Boyett (1997) first considered the model $\Lambda(t) = \Lambda_0(t)\exp(\beta_0 I_{\{X \le \zeta_0\}} + \alpha_0 Z)$, where the covariates X and Z are one-dimensional and a constant term β_0 is added when the threshold ζ_0 is reached. Pons (2003) considered the Cox model $\Lambda(t) = \Lambda_0(t)\exp\{\boldsymbol{\alpha}_0^\top \boldsymbol{Z}_1 + \boldsymbol{\beta}_0^\top \boldsymbol{Z}_2 I_{\{Z_3 \le \zeta_0\}} + \boldsymbol{\gamma}_0^\top \boldsymbol{Z}_2 I_{\{Z_3 > \zeta_0\}}\}$, where the covariate jumps at a threshold ζ_0, and showed that the change-point estimator is n-consistent. Kosorok and Song (2007) and Song et al. (2009) extended the asymptotic properties of the Cox model to a class of transformation models and introduced a maximal score-based test for the existence of a change point using random shuffling. Lee et al. (2020) further studied tests for the change-point effect with right-censored data and proposed three maximum Wald or score-based procedures that are important for model identification. More recently, for right-censored data, Deng et al. (2022) extended the Cox model with a change point to one with a change hyperplane. They developed a computational method for this model, proved the convergence of the change hyperplane, and derived the asymptotic distribution of the estimators.

Most existing methods focus on right-censored data. In this type of data, some individuals have fully observed failure times, while others are right-censored, known only after the time of censoring. However, in some clinical settings, individuals scheduled for observation to detect clinically significant changes in their health or disease status may miss some scheduled appointments and return with altered status (Ma et al. 2014), such as in studies of traumatic brain injury (Vanier et al. 2020). In such cases, change-point models developed for right-censored data may fail because the event time of interest can be left-censored or right-censored. To overcome this limitation, we study a change-point model for "case-I" interval-censored data, also known as current status data.

Current status data arise when the exact event time is unobserved and only the occurrence status at a single examination is recorded (Huang 1996; Huang and Wellner 1997; Sun 2006). Compared with right-censored data, the precise value of the survival time T is unknown, making statistical inference with current status data more challenging. Huang (1996) first established the Cox model framework for current status data and derived its asymptotic properties. For regression analysis, a partially linear proportional hazards model was later proposed (Lu and McMahan 2018).

However, in clinical studies, some covariates can exert complex nonlinear interaction effects on the logarithm of the cumulative hazard of the failure time T. When interaction effects are present, simple linear or purely additive models may fail to capture them, leading to substantial bias. To address this, Wu et al. (2024) proposed a deep learning approach that flexibly models the nonparametric effects of multivariate covariates and their potential relationships. The partially linear Cox model is then formulated as

$$\Lambda(t|\boldsymbol{X},\boldsymbol{Z}) = \Lambda_0(t)\exp\{\boldsymbol{\beta}_0^\top \boldsymbol{X} + g_0(\boldsymbol{Z})\}, \tag{1}$$

where Λ_0 denotes the baseline cumulative hazard function, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown parameter vector, and $g_0 : \mathbb{R}^r \to \mathbb{R}$ represents an unknown smooth function.

Although many methods have been developed for failure-time data with change points, most are designed for right-censored data and are not directly applicable to interval-censored settings. We therefore study a partially linear Cox model with a change point for current status data.

Our main contributions are summarized below:

(i) We introduce a deep partially linear Cox model with a change point for current status data and provide a practical estimation algorithm;
(ii) We establish consistency and convergence rates for the maximum likelihood estimators of all unknown components, including both parametric and nonparametric parts, derive the asymptotic normality of the parametric estimators, and attain semiparametric efficiency;
(iii) We develop a score-based testing procedure to assess the existence of a change point.

The remainder of the paper is organized as follows. Section 2 presents the estimation procedure for a change point in a partially linear Cox model with current status data. Section 3 provides theoretical results for the estimators, including consistency, convergence rates for all components, and the asymptotic normality of the parametric estimates. Section 4 introduces a score-based testing procedure to assess the presence of a change point. Section 5 reports numerical experiments, and Sect. 6 illustrates the method with the Rotterdam breast cancer data. Section 7 offers a summary and further discussion. All technical lemmas, necessary proofs, and computational details are provided in the supplementary material.

2 Model specification and methodology

2.1 Framework of deep neural network

Before constructing the model, we first outline the deep neural network framework. A DNN of depth $K \in \mathbb{N}^+$ and width $\boldsymbol{p} = (p_0, \ldots, p_{K+1})^\top$ is a composite function $g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{K+1}}$, expressed as

$$\begin{aligned} g(\mathrm{z}) &= W_K g_K(\mathrm{z}) + \boldsymbol{\nu}_K, \\ g_K(\mathrm{z}) &= \boldsymbol{\sigma}(W_{K-1} g_{K-1}(\mathrm{z}) + \boldsymbol{\nu}_{K-1}), \ldots, g_1(\mathrm{z}) = \boldsymbol{\sigma}(W_0 \mathrm{z} + \boldsymbol{\nu}_0), \end{aligned}$$

where the matrices $W_k \in \mathbb{R}^{p_{k+1}\times p_k}$ and the vectors $\boldsymbol{\nu}_k \in \mathbb{R}^{p_{k+1}}$ $(k = 0, \ldots, K)$ are the parameters to be estimated, and the activation function $\boldsymbol{\sigma}$ acts component-wise on vectors, i.e., for $\mathrm{z} = (z_1, \ldots, z_m)^\top$ we have $\boldsymbol{\sigma}(\mathrm{z}) = (\sigma(z_1), \ldots, \sigma(z_m))^\top$. This study adopts the rectified linear unit (ReLU) activation function (Nair and Hinton 2010), defined as $\sigma(z) = \max\{0, z\}$. Compared with other activation functions,

ReLU is computationally efficient, alleviates the vanishing gradient problem, and promotes stable gradient propagation in deep neural networks (Schmidt-Hieber 2020). For notational simplicity, let $\tilde{W}_k = (W_k, \boldsymbol{\nu}_k) \in \mathbb{R}^{p_{k+1}\times(p_k+1)}$. It follows from the definition of $g(\mathrm{z})$ that g is also the function of the $\tilde{W}_k$ for $k = 0, \ldots, K$, so we can rewrite $g(\mathrm{z})$ as $g(\mathrm{z}; \tilde{W}_k' s)$.

The total number of parameters is $\sum_{k=0}^{K} p_{k+1}(p_k + 1)$. Typically, an excessive number of parameters can induce overfitting. Weight pruning addresses this by reducing the total number of nonzero parameters, thereby yielding sparsely connected network layers (Schmidt-Hieber 2020). Given $D > 0$ and $s \in \mathbb{N}^+$, we consider a sparsely connected neural network class

$$\begin{aligned}\mathcal{G} &:= \mathcal{G}(K, \boldsymbol{p}, s, D)\\ &= \{g : g \text{ is a DNN with } (K+1) \text{ layers and width vector } \boldsymbol{p}\\ &\qquad |\tilde{W}_k \in \mathbb{R}^{p_{k+1}\times(p_k+1)}, \|\tilde{W}_k\|_\infty \le 1, \text{for } k = 0, \ldots, K, \sum_{k=0}^{K} \|\tilde{W}_k\|_0 \le s, \|g\|_\infty \le D\Big\},\end{aligned}$$

where $\|\cdot\|_\infty$ and $\|\cdot\|_0$ represent the L^∞ norm and L^0 norm, respectively.

2.2 Model and estimation

Consider a survival study involving *n* individuals with current status data. For each subject *i*, the event time of interest T_i is not directly observed; instead, we observe a single examination time U_i and a censoring indicator $\Delta_i = I_{\{T_i \le U_i\}}$, where $I_{\{\cdot\}}$ denotes the indicator function. Let X_i be a *p*-dimensional covariate vector representing the treatments, which contains categorical explanatory variables and binary variables, Z_i be a vector of *r*-dimensional covariate vector that affect the response in a nonparametric way, and $E_i \in \mathbb{R}$ denote the random variable of interest that exhibits a change point. Therefore, the observed data contains $\{(\Delta_i, U_i, \mathrm{X}_i, \mathrm{Z}_i, E_i), i = 1, \ldots, n\}$.

Given a covariate vector $(\mathrm{X}, \mathrm{Z}, E) \in \mathbb{R}^{p+r+1}$, assume that the cumulative hazard function for the failure time *T* satisfies the following partial linear Cox model:

$$\Lambda(t|\mathrm{X}, \mathrm{Z}, E) = \Lambda_0(t) \exp\{\boldsymbol{\beta}_0^\top \mathrm{X} + g_0(\mathrm{Z}) + \{\boldsymbol{\gamma}_0^\top \mathrm{X} + h_0(\mathrm{Z})\} I_{\{E>\zeta_0\}}\}, \qquad (2)$$

where $\Lambda_0(\cdot) : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}$ is the true baseline cumulative hazard function, $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0 \in \mathbb{R}^p$ are true values of treatment effect vectors, $g_0(\cdot) : \mathbb{R}^r \to \mathbb{R}$ and $h_0(\cdot) : \mathbb{R}^r \to \mathbb{R}$ are true smooth functions, $\zeta_0 \in \mathbb{R}$ is the true value of the change point. In this paper, we assume conditional independent censoring: $T \perp U \mid (\mathrm{X}, \mathrm{Z}, E)$. In addition, the distribution of $(U, \mathrm{X}, \mathrm{Z}, E)$ does not depend on $(\Lambda_0, \boldsymbol{\beta}_0, g_0, \boldsymbol{\gamma}_0, h_0, \zeta_0)$.

Under the Model (2), the cumulative distribution function $F(t|\mathrm{X}, \mathrm{Z}, E)$ can be expressed as:

$$F(t|\mathrm{X}, \mathrm{Z}, E) = 1 - \exp\{-\Lambda_0(t) \exp\{\boldsymbol{\beta}_0^\top \mathrm{X} + g_0(\mathrm{Z}) + \{\boldsymbol{\gamma}_0^\top \mathrm{X} + h_0(\mathrm{Z})\} I_{\{E>\zeta_0\}}\}\}.$$

To guarantee that $F(t|\mathrm{X}, \mathrm{Z}, E)$ is well defined, we require $\Lambda_0(\cdot)$ to be a monotone nondecreasing function satisfying $\Lambda_0(0) = 0$; it can be approximated by a mono-

tone spline function $\Lambda_{0,n}(\cdot) = \sum_{j=0}^{q_n} c_j M_j(\cdot|\ell)$, where $c_j \geq 0$ for $j = 1, \ldots, q_n$ and q_n indicates the number of basis functions (Ramsay 1988). In this formula, $\Lambda_{0,n}(\cdot)$ is the projection of $\Lambda_0(\cdot)$ onto the finite-dimensional sieve space, where $M_j(\cdot|\ell)$ denotes the integral spline basis function, and c_j represents the corresponding spline coefficient. Assuming that the support set of the observation time U lies within the interval $[L_U, R_U]$, where $0 < L_U < R_U < \tau$ and τ is the end time of observation, and all $M_j(\cdot|\ell)$ are nondecreasing piecewise polynomials of degree ℓ defined over $[L_U, R_U]$. Let the interval $[L_U, R_U]$ be partitioned as $L_U = t_0 < t_1 < \cdots < t_{m_n+1} = R_U$, where $m_n = O(n^\nu)$ represents the cardinality of the interior nodes with $\nu \in (0, 1/2)$. Given the polynomial degree ℓ and the set of interior nodes $\mathcal{S}_n = \{t_1, \ldots, t_{m_n}\}$, the total number of basis functions q_n can be determined as $q_n = m_n + \ell$.

In this paper, we use the integrated piecewise polynomial spline function $M_j(\cdot|\ell) = \int_0^t B_j(s|\ell - 1)ds$ to estimate the monotonic function Λ_0. Define

$$\mathcal{F} := \mathcal{F}(\mathcal{S}_n, m_n, \ell) = \left\{ \sum_{j=0}^{q_n} c_j M_j(t|\ell) : \ell \in \mathbb{N}, c_j \geq 0 \text{ for } j = 0, \ldots, q_n, t \in [L_U, R_U] \right\}.$$

Let $\boldsymbol{\eta} := (\Lambda, \boldsymbol{\beta}, g, \boldsymbol{\gamma}, h, \zeta)$ and denote its true value by $\boldsymbol{\eta}_0 := (\Lambda_0, \boldsymbol{\beta}_0, g_0, \boldsymbol{\gamma}_0, h_0, \zeta_0)$. Specifically, we estimate $\boldsymbol{\eta}_0$ by

$$\hat{\boldsymbol{\eta}}_n = (\hat{\Lambda}_n, \hat{\boldsymbol{\beta}}_n, \hat{g}_n, \hat{\boldsymbol{\gamma}}_n, \hat{h}_n, \hat{\zeta}_n) = \arg\max_{\boldsymbol{\eta} \in \mathcal{N}} \ell_n(\boldsymbol{\eta}), \tag{3}$$

where

$$\begin{aligned} \ell_n(\boldsymbol{\eta}) := \frac{1}{n} \sum_{i=1}^{n} \Delta_i \log \Big\{ 1 - \exp \Big\{ -\Lambda(U_i) \exp \Big\{ \boldsymbol{\beta}^\top \mathrm{X}_i + g(\mathrm{Z}_i) \\ + \{\boldsymbol{\gamma}^\top \mathrm{X}_i + h(\mathrm{Z}_i)\} I_{\{E_i > \zeta\}} \Big\} \Big\} \Big\} \\ - (1 - \Delta_i) \Lambda(U_i) \exp \Big\{ \boldsymbol{\beta}^\top \mathrm{X}_i + g(\mathrm{Z}_i) + \{\boldsymbol{\gamma}^\top \mathrm{X}_i + h(\mathrm{Z}_i)\} I_{\{E_i > \zeta\}} \Big\}, \end{aligned}$$

and $\mathcal{N} = \mathcal{F} \times \mathbb{R}^p \times \mathcal{G} \times \mathbb{R}^p \times \mathcal{G} \times \mathbb{R}$.

3 Asymptotic properties

This section establishes the consistency and asymptotic properties of the estimates in (3). Before discussing them, we first introduce some notation. The estimates of the change point ζ_0 and the remaining parameters $\boldsymbol{\xi}_0 = (\Lambda_0, \boldsymbol{\beta}_0, g_0, \boldsymbol{\gamma}_0, h_0)$ are denoted by $\hat{\zeta}_n$ and $\hat{\boldsymbol{\xi}}_n = (\hat{\Lambda}_n, \hat{\boldsymbol{\beta}}_n, \hat{g}_n, \hat{\boldsymbol{\gamma}}_n, \hat{h}_n)$, respectively. Note that $\boldsymbol{\eta} = (\boldsymbol{\xi}, \zeta) = (\Lambda, \boldsymbol{\beta}, g, \boldsymbol{\gamma}, h, \zeta)$. For any q-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_q)^\top \in \mathbb{R}^q$, $\|\boldsymbol{a}\| = \sqrt{\sum_{j=1}^q a_j^2}$ and $\|\boldsymbol{a}\|_\infty = \max_j |a_j|$ denote

the L^2 norm and L^∞ norm, respectively. For any matrix $W = (w_{ij}) \in \mathbb{R}^{m\times n}$, let $\|W\|_\infty = \max_{i,j} |w_{ij}|$. For any function $h(x)$, $\|h\|_\infty = \sup_x |h(x)|$. Denote $a_n \lesssim b_n$ as $a_n \le cb_n$ and $a_n \gtrsim b_n$ as $a_n \ge cb_n$ for some constant $c > 0$ respectively, and $a_n \asymp b_n$ indicates $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. Define the distance of $\boldsymbol{\eta}$ as

$$d(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \Big(\|\Lambda_1 - \Lambda_2\|^2_{L^2([0,\tau])} + \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + \|g_1 - g_2\|^2_{L^2([0,1]^r)} + \|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|^2 + \|h_1 - h_2\|^2_{L^2([0,1]^r)} + |\zeta_1 - \zeta_2| \Big)^{1/2},$$

where $|\cdot|$ represents the absolute value,

$$\|g_1 - g_2\|_{L^2([0,1]^r)} = \Big(\mathbb{E}\left\{ g_1(\mathrm{Z}) - g_2(\mathrm{Z}) \right\}^2 \Big)^{1/2},$$

and

$$\|\Lambda_1 - \Lambda_2\|_{L^2([0,\tau])} = \Big(\mathbb{E}\left\{ \Lambda_1(U) - \Lambda_2(U) \right\}^2 \Big)^{1/2},$$

respectively. Denote $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n^\top, \hat{\boldsymbol{\gamma}}_n^\top)^\top$, $\tilde{\mathrm{X}}(\zeta) = (\mathrm{X}^\top, \mathrm{X}^\top I_{\{E>\zeta\}})^\top$, and $\tilde{K}(\mathrm{Z};\zeta) = g(\mathrm{Z}) + h(\mathrm{Z}) I_{\{E>\zeta\}}$. Write

$$\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V}) = \exp\left\{ \log \Lambda(U) + \boldsymbol{\beta}^\top \mathrm{X} + g(\mathrm{Z}) + \{\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})\} I_{\{E>\zeta\}} \right\},$$

where $\boldsymbol{V} = (U, \mathrm{X}, \mathrm{Z}, E)$, and define

$$m_{\boldsymbol{\eta}}(\boldsymbol{V}) = \Delta \log(1 - \exp\{-\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V})\}) - (1-\Delta)\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V}),$$

then $\ell_n(\boldsymbol{\eta}) = \mathbb{P}_n m_{\boldsymbol{\eta}}(\boldsymbol{V}) := \frac{1}{n}\sum_{i=1}^n m_{\boldsymbol{\eta}}(\boldsymbol{V}_i)$ and define $\ell(\boldsymbol{\eta}) = \mathbb{P} m_{\boldsymbol{\eta}}(\boldsymbol{V}) := \int m_{\boldsymbol{\eta}}(\boldsymbol{V}) dP$.

To establish the theoretical results, the following regularity conditions are required:

(C1) $K = O(\log n)$, $n\alpha_n^2 \lesssim \min(p_k)_{k=1,\ldots,K} \le \max(p_k)_{k=1,\ldots,K} \lesssim n$ and $s = O(n\alpha_n^2 \log n)$.

(C2) For $\nu \in (0, \frac{1}{2})$, assume the maximum spacing of the knots is $O(n^{-\nu})$, i.e.

$$\max_{1\le k\le m_n+1} |t_k - t_{k-1}| = O(n^{-\nu}).$$

(C3) The true parameter vector $\boldsymbol{\theta}_0$ lies in the interior of a compact subset of $\mathbb{R}^{2p}$.

(C4) The uth derivative of Λ_0 satisfies the Lipschitz conditions with $u \ge 1$ and Λ_0 increases in $[0, \tau]$.

(C5) The support of U is an interval within $[L_U, R_U]$, with $0 < L_U < R_U < \tau$.

(C6) Assume the covariates (X, Z) lie in a bounded subset of $\mathbb{R}^{p+r}$. Without loss of generality, we further assume $\mathrm{Z} \in [0,1]^r$.

(C7) $\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V})$ is almost surely bounded away from 0 for all $\boldsymbol{\eta} \in \{\boldsymbol{\eta} : d(\boldsymbol{\eta}, \boldsymbol{\eta}_0) < \epsilon\}$ with some small $\epsilon > 0$.

(C8) For any $\mathrm{Z} \in [0,1]^r$, assume that $\mathbb{E}g_0(\mathrm{Z}) = 0$ and $\mathbb{E}h_0(\mathrm{Z}) = 0$.

(C9) We assume the existence of a change point $\zeta_0 \in \mathbb{R}$, i.e. $\mathbb{P}(\boldsymbol{\gamma}_0^\top \mathrm{X} + h_0(\mathrm{Z}) \neq 0) > 0$. Also, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and fixed $\zeta \in \mathbb{R}$, $\mathbb{P}(\boldsymbol{\theta}^\top \tilde{\mathrm{X}}(\zeta) \neq \boldsymbol{\theta}_0^\top \tilde{\mathrm{X}}(\zeta)) > 0$.

(C10) The variable E admits a density $h_3(\cdot)$ that is strictly positive, bounded, and continuous on a neighborhood of ζ_0.

(C11) The information matrix $I(\boldsymbol{\theta}_0; \zeta_0)$ defined in Theorem 4 is nonsingular.

Condition (C1) determines the structure and complexity of the deep neural network. Condition (C2) is a mild regularity assumption on the knot placement and is similar to the conditions in Ramsay (1988). Condition (C3) is a standard assumption in semi-parametric estimation The smoothness and monotonicity conditions on $\Lambda_0(\cdot)$ in (C4) guarantee that it can be well approximated by a monotone spline basis. Conditions (C5)-(C7) are needed for the calculation in the proofs of Theorems 1–5. Conditions (C8) and (C9) are required to ensure that the model parameters are identifiable. Condition (C10) indicates that E is continuously distributed in a bounded neighborhood of ζ_0, which is crucial for establishing the asymptotic independence and asymptotic properties at ζ_0. (C11) ensures the asymptotic efficiency of the regression parameter $\boldsymbol{\theta}_0$.

Theorem 1 *(Identifiability) Under the conditions (C1)-(C9), the parameter* $\boldsymbol{\eta} = (\boldsymbol{\xi}, \zeta)$ *is identifiable.*

Theorem 1 limits the range of the change point, illustrating that this methodology is identifiable only when $\zeta_0 \in (\min_i(E_i), \max_i(E_i))$, i.e. $\mathbb{P}(\{E > \zeta_0\}) \in (0,1)$. Another characteristic of identifiability is the event that $\boldsymbol{\gamma}_0 \neq \mathbf{0}$ or $g_0(\cdot) \neq 0$ occurs.

Before deriving the overall convergence rate, we introduce a Hölder class that contains both g_0 and h_0. Let κ and B be two positive constants, and $\lfloor \kappa \rfloor$ is the largest integer smaller than κ. A (κ, B)-Hölder class of smooth functions is denoted as

$$\mathcal{H}_r^\kappa(\mathbb{D}, B) = \Bigg\{ g : \mathbb{D} \subset \mathbb{R}^r \to \mathbb{R} : \sum_{\boldsymbol{\upsilon}:|\boldsymbol{\upsilon}|<\kappa} \|\partial^{\boldsymbol{\upsilon}} g\|_\infty + \sum_{\boldsymbol{\upsilon}:|\boldsymbol{\upsilon}|=\lfloor\kappa\rfloor} \sup_{x,y\in\mathbb{D},x\neq y} \frac{|\partial^{\boldsymbol{\upsilon}} g(x) - \partial^{\boldsymbol{\upsilon}} g(y)|}{\|x-y\|_\infty^{\kappa-\lfloor\kappa\rfloor}} \leq B \Bigg\},$$

where $\partial^{\boldsymbol{\upsilon}} = \partial^{\upsilon_1} \cdots \partial^{\upsilon_r}$ with $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_r)^\top$ and $|\boldsymbol{\upsilon}| = \sum_{j=1}^r \upsilon_j$. Denote $K > 0, R \in \mathbb{N}^+, \boldsymbol{\kappa} = (\kappa_0, \ldots, \kappa_K)^\top \in \mathbb{R}_+^{K+1}$, and let $\boldsymbol{d} = (d_0, \ldots, d_{K+1}) \in \mathbb{N}_+^{K+2}$, $\tilde{\boldsymbol{d}} = (\tilde{d}_0, \ldots, \tilde{d}_K)^\top \in \mathbb{N}_+^{K+1}$ with $\tilde{d}_j \leq d_j$ where $j = 0, \ldots, L$. Further, assume that the functions g_0 and h_0 belong to a composite smoothness function class:

$$\mathcal{H}(K,\boldsymbol{\kappa},\boldsymbol{d},\tilde{\boldsymbol{d}},B) := \big\{h = h_K \circ \cdots \circ h_0 : h_i = (h_{i1},\ldots,h_{id_{i+1}})^\top \text{ and } h_{ij} \in \mathcal{H}_{\tilde{d}_i}^{\kappa_i}\left([a_i,b_i]^{\tilde{d}_i},B\right) \text{ for some } |a_i|,|b_i| \le B\big\},$$

where $\tilde{\boldsymbol{d}}$ denotes the intrinsic dimension of the function in this class. Furthermore, denote $\tilde{\upsilon}_i = \upsilon_i \prod_{j=i+1}^{K}(\upsilon_j \wedge 1)$ and $\alpha_n = \max_{i=0,\ldots,K} n^{-\tilde{\upsilon}_i/(2\tilde{\upsilon}_i+\tilde{d}_i)}$, where $a \wedge b = \min\{a,b\}$.

Theorem 2 *(Consistency and convergence rate) Let* $(2u+1)^{-1} < \nu < (2u)^{-1}$ *with* $u \ge 1$ *and* $q_n = O(n^\nu)$. *Suppose conditions (C1)-(C9) hold, we have* $d(\hat{\boldsymbol{\eta}}_n, \boldsymbol{\eta}_0) = O_P(\alpha_n \log^2 n + n^{-u\nu})$.

Theorem 2 illustrates that the rate is determined by the smoothness α_n, the intrinsic dimension $\tilde{\boldsymbol{d}}$ of the DNN, the smoothness ℓ, and the maximum number of nodes m_n of the spline function. As a result, when the intrinsic dimension $\tilde{\boldsymbol{d}}$ is relatively small, the suggested strategy can mitigate the curse of dimensionality and has a faster convergence rate. It is well known that when nonparametric smoothing techniques (such as splines or kernel functions) are used to estimate g_0, the model suffers from a severe "curse of dimensionality". A key advantage of the DNN estimator over traditional smoothing methods is its ability to alleviate the curse of dimensionality by accurately projecting data into a lower-dimensional representation space (Anthony and Bartlett 1999).

We next establish the semiparametric efficiency. Unlike the classical regression model, the unknown change point affects the procedure of estimation and projection, which makes it difficult to obtain the score and information functions (Pons 2003; Deng et al. 2022). Therefore, before establishing the semiparametric efficiency, we first show that the effect of the change point is asymptotically negligible; in other words, when n is large enough, $\hat{\zeta}_n$ is asymptotically independent of $\hat{\boldsymbol{\xi}}_n$. In particular, denote two counting processes

$$N_{1i}(t) = (1-\Delta_i)I(U_i \le t),$$
$$N_{2i}(t) = \Delta_i I(U_i \le t).$$

For simplicity, let

$$\boldsymbol{m}(t;\boldsymbol{\xi},\zeta) = -\Lambda(t)\exp\{\boldsymbol{\beta}^\top \mathrm{X} + g(\mathrm{Z}) + \{\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})\}I_{\{E>\zeta\}}\},$$
$$\mathcal{M}(t;\boldsymbol{\xi}) = \Lambda(t)\exp\{\boldsymbol{\beta}^\top \mathrm{X} + g(\mathrm{Z})\}\left[\exp\{\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})\} - 1\right],$$

$$\mathcal{M}^+(t;\boldsymbol{\xi}) = \log\left(\frac{1-\exp\left\{-\Lambda(t)\exp\{\boldsymbol{\beta}^\top \mathrm{X} + g(\mathrm{Z}) + \{\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})\}(I_{\{E>\zeta_0\}} - 1)\}\right\}}{1-\exp\left\{-\Lambda(t)\exp\{\boldsymbol{\beta}^\top \mathrm{X} + g(\mathrm{Z}) + \{\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})\}I_{\{E>\zeta_0\}}\}\right\}}\right),$$

$$\mathcal{M}^-(t;\boldsymbol{\xi}) = \log\left(\frac{1-\exp\left\{-\Lambda(t)\exp\{\boldsymbol{\beta}^\top \mathrm{X} + g(\mathrm{Z}) + \{\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})\}(I_{\{E>\zeta_0\}} + 1)\}\right\}}{1-\exp\left\{-\Lambda(t)\exp\{\boldsymbol{\beta}^\top \mathrm{X} + g(\mathrm{Z}) + \{\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})\}I_{\{E>\zeta_0\}}\}\right\}}\right),$$

and

$$Q_n(\boldsymbol{\xi}_0, \hat{\zeta}_n) = \sum_{i=1}^{n} (1-\Delta_i)\mathcal{M}(U_i;\boldsymbol{\xi}_0)\tilde{I}(\hat{\zeta}_n) + \Delta_i \left\{ \mathcal{M}^+(U_i;\boldsymbol{\xi}_0) I_{\{\zeta_0 < E \le \hat{\zeta}_n\}} - \mathcal{M}^-(U_i;\boldsymbol{\xi}_0) I_{\{\hat{\zeta}_n < E \le \zeta_0\}} \right\}.$$

Theorem 3 *(Asymptotic independence) Under (C1)-(C10), there exists*

$$\ell_n(\hat{\boldsymbol{\eta}}_n) = n^{-1} Q_n(\boldsymbol{\xi}_0, \hat{\zeta}_n) + \ell_n(\hat{\boldsymbol{\xi}}_n, \zeta_0) + o_P(1),$$

where

$$\ell_n(\hat{\boldsymbol{\xi}}_n, \zeta_0) := \frac{1}{n}\sum_{i=1}^{n} \Delta_i \log\left\{1 - \exp\left\{-\Lambda(U_i)\exp\left\{\boldsymbol{\beta}^\top \mathrm{X}_i + g(\mathrm{Z}_i) + \{\boldsymbol{\gamma}^\top \mathrm{X}_i + h(\mathrm{Z}_i)\} I_{\{E_i > \zeta_0\}}\right\}\right\}\right\} - (1-\Delta_i)\Lambda(U_i)\exp\left\{\boldsymbol{\beta}^\top \mathrm{X}_i + g(\mathrm{Z}_i) + \{\boldsymbol{\gamma}^\top \mathrm{X}_i + h(\mathrm{Z}_i)\} I_{\{E_i > \zeta_0\}}\right\}.$$

Furthermore, $\hat{\boldsymbol{\xi}}_n$ is asymptotically independent of $\hat{\zeta}_n$.

The unknown change point causes inconvenience in statistical inference, especially in deriving the asymptotic distribution. Based on asymptotic independence, we establish the effective score and information bound for θ_0.

Let $\Omega_0 = \left\{ \begin{matrix} \Lambda \in L^2([0,\tau]) \\ : \Lambda(\tau) < \infty \end{matrix} \right\}$, and let $\begin{matrix} \mathcal{H}_0 = \{g \in \mathcal{H}\{K, \kappa, d, \tilde{d}, B\} \\ : \mathbb{E}\{g(Z)\} = 0\} \end{matrix}$. Denote Ω_{Λ_0} be the collection of all subfamilies $\{\log\Lambda_t : t \in (-1,1)\} \subset \{\log\Lambda : \Lambda \in \Omega_0\}$ such that $\lim_{t\to\infty} \|t^{-1}(\log\Lambda_t - \log\Lambda_0) - a\|_{L^2([0,\tau])} = 0$ where $a \in L^2([0,\tau])$, and let

$$\mathbb{T}_{\Lambda_0} = \Big\{ a \in L^2([0,\tau]) : \lim_{t\to\infty} \|t^{-1}(\log\Lambda_t - \log\Lambda_0) - a\|_{L^2([0,\tau])} = 0 \text{ for some subfamily } \{\log\Lambda_t : t \in (-1,1)\} \subset \Omega_0 \Big\}.$$

Similarly, define

$$\mathbb{T}_{\tilde{K}_0} = \Big\{ b \in L^2([0,1]^r) : \lim_{s\to\infty} \|s^{-1}(\tilde{K}_s - \tilde{K}_0) - b\|_{L^2([0,1]^r)} = 0 \text{ for some subfamily } \{\tilde{K}_s : s \in (-1,1)\} \subset \mathcal{H}_{\tilde{K}_0}, \text{ and } \mathbb{E}\{b(\mathrm{Z})\} = 0 \Big\},$$

where $\mathcal{H}_{\tilde{K}_0}$ denotes the collection of all subfamilies of $\{\tilde{K}_s(\cdot;\zeta_0) \in L^2([0,1]^r) : s \in (-1,1)\} \subset \mathcal{H}_0$ such that $\lim_{s\to 0} \|u^{-1}(\tilde{K}_s - K_0) - b\| = 0$ and $b \in L^2([0,1]^r)$. Let $\bar{\mathbb{T}}_{\Lambda_0}$ and $\bar{\mathbb{T}}_{\tilde{K}_0}$ be the linear span of $\mathbb{T}_{\Lambda_0}$ and $\mathbb{T}_{\tilde{K}_0}$, respectively. Denote the single observation of the log-likelihood $\ell(\boldsymbol{\eta}|\mathrm{Y}) = \Delta\log(1 - \exp\{-\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V})\}) - (1-\Delta)\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V})$, where $\mathrm{Y} = (\Delta, \boldsymbol{V})$. The score function w.r.t. $\boldsymbol{\theta}$ is

$$\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\eta}_0) = \tilde{\mathrm{X}}(\zeta_0)\mathcal{Q}(\mathrm{Y};\boldsymbol{\xi}_0,\zeta_0),$$

where

$$\mathcal{Q}(\mathrm{Y};\boldsymbol{\xi},\zeta) = \boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V})\left\{\Delta\frac{\exp\{-\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V})\}}{1-\exp\{-\boldsymbol{h}_{\boldsymbol{\eta}}(\boldsymbol{V})\}} - (1-\Delta)\right\}.$$

Let a parametric smooth model $(\Lambda_t(U), \tilde{K}_s(\mathrm{Z};\zeta_0))$ satisfy $\Lambda_t(U)|_{t=0} = \Lambda_0(U)$ and $\tilde{K}_s(\mathrm{Z};\zeta_0)|_{s=0} = \tilde{K}_0(\mathrm{Z};\zeta_0)$, with $\frac{\partial}{\partial t}\Lambda_t(U)\big|_{t=0} = a(U)$ and $\frac{\partial}{\partial s}\tilde{K}_s(\mathrm{Z};\zeta_0)\big|_{s=0} = b(\mathrm{Z})$, the score functions for Λ_0 and $\tilde{K}_0(\cdot;\zeta_0)$ are

$$\dot{\ell}_{\Lambda}(\boldsymbol{\eta}_0)[a] = \frac{\partial\ell(\boldsymbol{\theta}_0,\Lambda_t,\tilde{K}_0 \mid \mathrm{Y})}{\partial t}\Big|_{t=0} = a(U)\mathcal{Q}(\mathrm{Y};\boldsymbol{\xi}_0,\zeta_0)$$

and

$$\dot{\ell}_{\tilde{K}}(\boldsymbol{\eta}_0)[b] = \frac{\partial\ell(\boldsymbol{\theta}_0,\Lambda_0,\tilde{K}_s \mid \mathrm{Y})}{\partial s}\Big|_{s=0} = b(\mathrm{Z})\mathcal{Q}(\mathrm{Y};\boldsymbol{\xi}_0,\zeta_0),$$

respectively. For $\boldsymbol{a} = (a_1,\ldots,a_{2p})^\top \in \bar{\mathbb{T}}_{\Lambda_0}$, let $\dot{\ell}_{\Lambda}(\boldsymbol{\eta}_0)[\boldsymbol{a}] = (\dot{\ell}_{\Lambda}(\boldsymbol{\eta}_0)[a_1],\ldots,\dot{\ell}_{\Lambda}(\boldsymbol{\eta}_0)[a_{2p}])^\top$, and for $\boldsymbol{b} = (b_1,\ldots,b_{2p})^\top \in \bar{\mathbb{T}}_{\tilde{K}_0}$, define $\dot{\ell}_{\tilde{K}}(\boldsymbol{\eta}_0)[\boldsymbol{b}] = (\dot{\ell}_{\tilde{K}}(\boldsymbol{\eta}_0)[b_1],\ldots,\dot{\ell}_{\tilde{K}}(\boldsymbol{\eta}_0)[b_{2p}])^\top$. The score and information bound are given in the following theorem.

Theorem 4 *Under (C1)–(C11),*

if $(\boldsymbol{a}_*^\top, \boldsymbol{b}_*^\top)^\top \in \bar{\mathbb{T}}_{\Lambda_0} \times \bar{\mathbb{T}}_{\tilde{K}_0}$ *minimizes the distance*

$$\mathbb{E}\left\{\mathcal{Q}^2(\mathrm{Y};\boldsymbol{\xi}_0,\zeta_0)[(\tilde{\mathrm{X}}(\zeta) - \boldsymbol{a}(U) - \boldsymbol{b}(\mathrm{Z})) \odot (\tilde{\mathrm{X}}(\zeta) - \boldsymbol{a}(U) - \boldsymbol{b}(\mathrm{Z}))]\right\},$$

where $\boldsymbol{y}\odot\boldsymbol{y} = (y_1^2,\ldots,y_q^2)^\top$ *for any q-dimensional vector* $\boldsymbol{y} = (y_1,\ldots,y_q)^\top \in \mathbb{R}^q$ *and the minimization operates component-wise on the vector, the efficient score function for* $\boldsymbol{\theta}_0$ *is*

$$U(\boldsymbol{\theta}_0;\zeta_0) = \ell^*_{\boldsymbol{\theta}}(\mathrm{Y},\boldsymbol{\eta}_0) = \{\tilde{\mathrm{X}}(\zeta_0) - \boldsymbol{a}_*(U) - \boldsymbol{b}_*(\mathrm{Z})\}\mathcal{Q}(\mathrm{Y};\boldsymbol{\xi}_0,\zeta_0),$$

and the information matrix

$$I(\boldsymbol{\theta}_0;\zeta_0) = \mathbb{E}(U(\boldsymbol{\theta}_0;\zeta_0)^{\otimes 2}) = \mathbb{E}\left\{\mathcal{Q}^2(\mathrm{Y};\boldsymbol{\xi}_0,\zeta_0)\{\tilde{\mathrm{X}}(\zeta_0) - \boldsymbol{a}_*(U) - \boldsymbol{b}_*(\mathrm{Z})\}^{\otimes 2}\right\}.$$

After determining the score function and the information bound of $\boldsymbol{\theta}$, we present the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ and $\hat{\zeta}_n$. Denote two variables

$$\xi^+ = \xi^+(U,\Delta,\mathrm{X},\mathrm{Z},E) = (1-\Delta)\mathcal{M}(t,\boldsymbol{\xi}_0) + \Delta\mathcal{M}^+(t,\boldsymbol{\xi}_0),$$

$$\xi^- = \xi^-(U, \Delta, \mathrm{X}, \mathrm{Z}, E) = (1-\Delta)\mathcal{M}(t, \boldsymbol{\xi}_0) + \Delta\mathcal{M}^-(t, \boldsymbol{\xi}_0),$$

respectively, and define

$$\tilde{Q}_n^+(v) = I_{\{v>0\}} \sum_{i=1}^{n} \xi_i^+ I_{\{\zeta_0 < E_i \le \zeta_0 + n^{-1}v\}},$$

$$\tilde{Q}_n^-(v) = I_{\{v<0\}} \sum_{i=1}^{n} \xi_i^- I_{\{\zeta_0 + n^{-1}v < E_i \le \zeta_0\}},$$

where $\xi_i^+ = \xi^+(U_i, \Delta_i, \mathrm{X}_i, \mathrm{Z}_i, E_i)$ and $\xi_i^- = \xi^-(U_i, \Delta_i, \mathrm{X}_i, \mathrm{Z}_i, E_i)$. Furthermore, for $v \in \mathbb{R}^+$, define

$$\tilde{Q}^+(v) = I_{\{v>0\}} \sum_{j \ge 1} \frac{e^{-vh_3(\zeta_0)}(vh_3(\zeta_0))^j}{(j-1)!} \mathbb{E}(\xi^+ | E = \zeta_0^+) \ ,$$

where

$$\mathbb{E}(\xi^+ \mid E = \zeta_0^+) = \mathbb{E}\left\{(1-\Delta)\mathcal{M}(U, \boldsymbol{\xi}_0) + \Delta \log\left(\frac{1 - \exp\{-\Lambda_0(U)\exp\{\boldsymbol{\beta}_0^\top \mathrm{X} + g_0(\mathrm{Z})\}}{1 - \exp\{-\Lambda_0(U)\exp\{(\boldsymbol{\beta}_0 + \boldsymbol{\gamma}_0)^\top \mathrm{X} + g_0(\mathrm{Z}) + h_0(\mathrm{Z})\}}\right)\right\},$$

and for $v \in \mathbb{R}^-$, define

$$\tilde{Q}^-(v) = I_{\{v<0\}} \sum_{j \ge 1} \frac{e^{vh_3(\zeta_0)}(-vh_3(\zeta_0))^j}{(j-1)!} \mathbb{E}(\xi^- | E = \zeta_0) \ ,$$

where

$$\mathbb{E}(\xi^- \mid E = \zeta_0) = \mathbb{E}\left\{(1-\Delta)\mathcal{M}(U, \boldsymbol{\xi}_0) - \Delta \log\left(\frac{1 - \exp\{-\Lambda_0(U)\exp\{\boldsymbol{\beta}_0^\top \mathrm{X} + g_0(\mathrm{Z})\}}{1 - \exp\{-\Lambda_0(U)\exp\{(\boldsymbol{\beta}_0 + \boldsymbol{\gamma}_0)^\top \mathrm{X} + g_0(\mathrm{Z}) + h_0(\mathrm{Z})\}}\right)\right\}.$$

Theorem 5 *(Semiparametric efficiency) Under (C1)-(C11),*

if $n\alpha_n^4 \to 0$ *as* $n \to \infty$ *and* $(2u+1)^{-1} < \nu < (2u)^{-1}$ *for* $u \ge 1$, *the parametric regression parameter estimator* $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n^\top, \hat{\boldsymbol{\gamma}}_n^\top)^\top$ *satisfies* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0; \zeta_0))$

where $I(\boldsymbol{\theta}_0; \zeta_0)$ *is defined in Theorem 4. The change point estimator* $\hat{\zeta_n}$ *satisfies*

$$n(\hat{\zeta}_n - \zeta_0) \xrightarrow{d} \inf\left\{v : v = \arg\max_u \tilde{Q}(v)\right\},$$

where $\tilde{Q}(v) := \tilde{Q}^+(v) - \tilde{Q}^-(v)$.

Theorem 5 presents the $\sqrt{n}$-consistency and asymptotic normality of the estimate of the regression parameter and shows that the estimate of the change point converges distributionally to a distribution of the maximum value of a composite Poisson-type process with rate n^{-1}.

4 Model diagnostic

In practice, the detection of change points is a crucial problem. Accurate identification improves understanding of the data structure, enhances predictive models, and supports better decision-making. We aim to test whether a change point exists, i.e., the null hypothesis $H_0 : \boldsymbol{\gamma}_0 = \mathbf{0},\ h_0(\cdot) = 0$. Note that if $\boldsymbol{\gamma}^\top \mathrm{X} + h(\mathrm{Z})$ is degenerate (identically zero) or the threshold diverges ($\zeta_0 = \pm\infty$), the change-point model fails to be identifiable. If H_0 holds, model (2) reduces to the deep partially linear Cox model (1) and one can estimate $\boldsymbol{\beta}_0$, $\Lambda_0(\cdot)$ and $g_0(\cdot)$ by the estimation procedure proposed by Wu et al. (2024). To test H_0, we propose a functional SUP statistic inspired by the supremum test (SUP) proposed by Kosorok and Song (2007).

Denote $\boldsymbol{L}_n(\boldsymbol{\gamma}, h, \Lambda, \boldsymbol{\beta}, g, \zeta) = \sum_{i=1}^n \Delta_i \log(1 - \exp\{-\boldsymbol{h_\eta}(\boldsymbol{V}_i)\}) - (1 - \Delta_i)\boldsymbol{h_\eta}(\boldsymbol{V}_i)$. The score functional evaluated at direction $f \in \mathcal{G}$ is defined as

$$\begin{aligned}
\boldsymbol{U}_n(\zeta, f, \boldsymbol{\xi}) &= \frac{\partial}{\partial(\boldsymbol{\gamma}^\top, \varepsilon)^\top} \boldsymbol{L}_n(\boldsymbol{\gamma}, h + \varepsilon f, \Lambda, \boldsymbol{\beta}, g, \zeta)\Big|_{\varepsilon=0} \\
&= \sum_{i=1}^n \mathcal{Q}(\mathrm{Y}_i; \boldsymbol{\xi}, \zeta) \frac{\partial}{\partial(\boldsymbol{\gamma}^\top, \varepsilon)^\top} \left\{\log \Lambda(U_i) + \boldsymbol{\beta}^\top \mathrm{X}_i + g(\mathrm{Z}_i) \right. \\
&\quad \left. + \{\boldsymbol{\gamma}^\top \mathrm{X}_i + h(\mathrm{Z}_i) + \varepsilon f(\mathrm{Z}_i)\} I_{\{E_i > \zeta\}}\right\}\Big|_{\varepsilon=0} \\
&= \sum_{i=1}^n I_{\{E_i > \zeta\}} \mathcal{Q}(\mathrm{Y}_i; \boldsymbol{\xi}, \zeta)(\mathrm{X}_i^\top, f(\mathrm{Z}_i))^\top,
\end{aligned}$$

and the covariance of $\boldsymbol{U}_n(\zeta, f, \boldsymbol{\xi})$ is

$$\boldsymbol{\Sigma}_n(\zeta, f, \boldsymbol{\xi}) = \sum_{i=1}^n I_{\{E_i > \zeta\}} \mathcal{Q}^2(\mathrm{Y}_i; \boldsymbol{\xi}, \zeta)(\mathrm{X}_i^\top, f(\mathrm{Z}_i))^\top (\mathrm{X}_i^\top, f(\mathrm{Z}_i)).$$

Then the SUP statistic is defined as

$$\mathrm{SUP} = \sup_{\zeta \in [\zeta_1, \zeta_2]} \sup_{f \in \mathcal{G}} \boldsymbol{U}_n(\zeta, f, \tilde{\boldsymbol{\Omega}}_n)^\top \boldsymbol{\Sigma}_n(\zeta, f, \tilde{\boldsymbol{\Omega}}_n)^{-1} \boldsymbol{U}_n(\zeta, f, \tilde{\boldsymbol{\Omega}}_n),$$

where $\tilde{\boldsymbol{\Omega}}_n = (\tilde{\Lambda}_n, \tilde{\boldsymbol{\beta}}_n, \tilde{g}_n, \mathbf{0}, 0)$, and where $\tilde{\Lambda}_n$, $\tilde{\boldsymbol{\beta}}_n$, and $\tilde{g}_n$ are the restricted MLE of $\boldsymbol{\beta}_0$, Λ_0 and g_0 under H_0, respectively.

For application, the change point ζ can be selected by grid search, and the function f can be chosen from the DNN function space $\mathcal{G}$ to match the potentially complex nonlinear forms of the direction. However, directly selecting $f \in \mathcal{G}$ may be infeasible due to the high dimensionality and complexity of the DNN function space. In practice, one can instead work with a simpler space that contains G, such as a finite-dimensional linear span $\mathcal{T}$, and choose *f* from $\mathcal{T}$. To reduce the computational burden of evaluating *f*(*Z*), the selection of direction functions can be further decomposed into a sequence of one-dimensional problems. Specifically, for each dimension $j = 1, \ldots, r$ of Z, we can select *f* from $\mathcal{T}_j$, the projection on the *j*th dimension of $\mathcal{T}$, and calculate the corresponding score-based statistics. Ultimately, we take the maximum of the statistics across all dimensions, which substantially reduces computational complexity while retaining flexibility. For example, if we take $\mathcal{T} = \text{span}\{I_{\{Z_1 < z_1, \ldots, Z_r < z_r\}}\}$ to replace $\mathcal{G}$, the projection on its *j*th dimension $\mathcal{T}_j = \text{span}\{I_{\{Z_j < z_j\}}\}$ can be used to choose *f*; we can take $\mathrm{Z} = {}_{\{Z_j < z_j\}} \in \mathcal{T}_j$ for $j = 1, \ldots, r$ and then take the supremum over *j*, rather than using $\mathrm{Z} = I_{\{Z_1 < z_1, \ldots, Z_r < z_r\}} \in \mathcal{T}$ to compute the SUP statistic directly. In summary, the SUP_k statistic is defined as

$$\mathrm{SUP}_k = \sup_{\zeta \in \{\zeta_1^*, \ldots, \zeta_k^*\}} \sup_{j \in \{1, \ldots, r\}} \sup_{f \in \mathcal{T}_j} \boldsymbol{U}_n(\zeta, f, \tilde{\boldsymbol{\Omega}}_n)^\top \boldsymbol{\Sigma}_n(\zeta, f, \tilde{\boldsymbol{\Omega}}_n)^{-1} \boldsymbol{U}_n(\zeta, f, \tilde{\boldsymbol{\Omega}}_n),$$

where $\{\zeta_1^*, \ldots, \zeta_k^*\} \subset [\zeta_1, \zeta_2]$. We construct the permutation distribution of SUP_k by repeatedly permuting E and recomputing the statistic. At a significance level α, reject H_0 if SUP_k exceeds the upper α-quantile of its permutation distribution; otherwise, do not reject H_0. The algorithm below summarizes the testing procedure.

1. Obtain the estimate $(\tilde{\Lambda}_n, \tilde{\boldsymbol{\beta}}_n, \tilde{g}_n)$ based on model (2) under the restriction of $\boldsymbol{\gamma} = \mathbf{0}$ and $h = 0$.
2. Calculate the test statistics SUP_k for a given $k \geq 2$.
3. Obtain the permutation distribution of the SUP_k by shuffling the covariate E at random for adequate times.
4. Reject H_0 if SUP_k is greater than the upper α-quantile of the permutation distribution.

Algorithm 1: Test procedure for H_0

5 Simulation studies

In this section, we first present the computational procedure. The loss is the negative log-likelihood function, which is nonconcave in the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, W_j, $\boldsymbol{\nu}_j$, and c_j; consequently, conventional optimization methods may perform poorly. To overcome this difficulty, we propose an iterative profiling estimation procedure, detailed in the algorithm below.

1. Let $\breve{\zeta} \in (\min_i(E_i), \max_i(E_i))$, $\breve{\Lambda}(\cdot) = \sum_{j=0}^{q_n} \breve{c}_j M_j(\cdot|\ell) = \breve{\boldsymbol{c}}^\top \boldsymbol{M}(\cdot)$ with $\breve{c}_j \geq 0$ for $\breve{\boldsymbol{c}} = (\breve{c}_0, \ldots, \breve{c}_{q_n})^\top$, and $\breve{\boldsymbol{\theta}}$ be the initial values, where $\boldsymbol{M}(\cdot) = (M_0(\cdot|\ell), \ldots, M_{q_n}(\cdot|\ell))^\top$. Define a grid $(\zeta_1, \ldots, \zeta_{N_R})$ for estimate ζ_0, where $\min(E_i) \leq \zeta_1 < \cdots < \zeta_{N_R} \leq \max(E_i)$ and $N_R \geq 2$. At the beginning, let $\hat{\Lambda}_n = \breve{\Lambda}$, $\hat{\boldsymbol{\theta}}_n = \breve{\boldsymbol{\theta}}$ and $\hat{\zeta}_n = \breve{\zeta}$.

2. Use DNN to estimate the effects of $\mathbf{Z}$ by maximizing $\ell_n(\boldsymbol{\eta})$ given $\Lambda = \hat{\Lambda}_n$, $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n$ and $\zeta = \hat{\zeta}_n$, and denote the estimates as $\hat{g}_n$ and $\hat{h}_n$.

3. Obtain $\hat{\Lambda}_n$ by maximizing $\ell_n(\boldsymbol{\eta})$ for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n$, $\zeta = \hat{\zeta}_n$, $g(\cdot) = \hat{g}_n(\cdot)$ and $h(\cdot) = \hat{h}_n(\cdot)$.

4. Estimate $\hat{\boldsymbol{\theta}}_n$ by optimizing $\ell_n(\boldsymbol{\eta})$ given $\Lambda = \hat{\Lambda}_n$, $g = \hat{g}_n$, $h = \hat{h}_n$ and $\zeta = \tilde{\zeta}_n$.

5. Given $\hat{\Lambda}_n$, $\hat{\boldsymbol{\theta}}_n$, $\hat{g}_n$ and $\hat{h}_n$, find $\hat{\zeta}_n$ in grid $(\zeta_1, \ldots, \zeta_{N_R})$ by which maximizes $\ell_n(\boldsymbol{\eta})$.

6. Iterate Steps 2-5 until $\hat{\boldsymbol{\theta}}_n$ converge.

Algorithm 2: Profiling estimation procedure for parameter $\boldsymbol{\eta}$.

Step 1 gives the initial values, and it is recommended to let $\breve{\boldsymbol{\theta}}$ be a zero vector with dimension $2p$ to keep the proportional term from influencing the initial estimation and resulting in overfitting. In Step 2, given the initial values of $\boldsymbol{\theta}$ and spline parameters c_j's, the $\hat{W}_j$'s and $\hat{\boldsymbol{\nu}}_j$'s can be obtained by minimizing the loss function using the `Pytorch`. We estimate g_0 and h_0 jointly by choosing $p_{K+1} = 2$, producing the two-dimensional output $(\hat{g}_n(\mathrm{Z}), \hat{h}_n(\mathrm{Z}))$ for each Z. Then, in Step 3, fixing the remaining parameters, we estimate $\hat{c}_j$ for $\hat{\Lambda}_n$ and set $\hat{\Lambda}_n(\cdot) = \sum_{i=1}^{q_n} \hat{c}_j M_j(\cdot|\ell)$. In Step 4, we apply the sequential least squares programming to estimate $\boldsymbol{\theta}_0$ while keeping the remaining parameters fixed. While a larger N_R in Step 5 enhances the precision of ζ_0, it also raises computational burden. In practice, one should balance accuracy against cost. Finally, in Step 6, if we define $\hat{\boldsymbol{\theta}}_n^{(j)}$ as the estimation of regression parameters in the *j* iteration and $\|\cdot\|_\infty$ as the L^∞ norm, our iteration procedures stop at the *j*th step when $\|\hat{\boldsymbol{\theta}}_n^{(j)} - \hat{\boldsymbol{\theta}}_n^{(j-1)}\|_\infty < \epsilon$ for some given $\epsilon > 0$.

Some details of the neural network optimization are discussed here. The `Adam` optimizer estimates the W_j and $\boldsymbol{\nu}_j$ of the deep neural network, which is reliable and practical in various models (Kingma and Ba 2014). The `Adam` optimization algorithm utilizes first-order moment estimates and second-order moment estimates of the gradients to dynamically adjust the learning rates for each parameter. The tuning parameters of the DNN contain the number of hidden layers K, the number of neurons p_k in all hidden layers, the number of epochs, the dropout rate and the learning rate. The learning rate is the step size of gradient descent, while the dropout rate refers to the random disregard of partial neurons during training.

In all simulations, based on Model (2), X is generated from the Bernoulli distribution with a successful probability of 0.5. The covariate Z is generated from a multivariate *t*-distribution $t_5(\mathbf{0}, \boldsymbol{\Sigma})$ with truncation on [0, 2], where $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5$ for $i \neq j$. The variable E is distributed as $N(2, 1)$ truncated on the interval [1.5, 2.5]. We set the true values of the regression parameters $\boldsymbol{\theta}_0 = (\beta_0, \gamma_0)^\top = (-1, 2)^\top$, the change point $\zeta_0 = 2$ and the baseline cumulative hazard function $\Lambda_0(t) = \sqrt{t}/5$. For covariate Z, we design the following 3 cases with $\mathrm{z} = (z_1, \ldots, z_5)^\top \in [0, 2]^5$:

Case 1 (Linear): $g_0(\mathrm{z}) = -z_1/2 - z_2/3 - z_3/4 - z_4/5 - z_5/6 + 0.94$, $h_0(\mathrm{z}) = -z_1/3 - z_2/4 - z_3/5 - z_4/6 - z_5/7 + 0.71$,

Case 2 (Additive): $g_0(\mathrm{z}) = z_1^2/2 + 2\log(z_2+1)/5 + 3\sqrt{z_3}/10 + e^{z_4}/5 + z_5^3/10 - 1.62$, $h_0(\mathrm{z}) = \sin(2\pi z_1) + e^{z_2}/5 + 3\sqrt{z_3}/5 + \log(z_4+1)/3 + z_5^2/3 - 1.38$,

Case 3 (Deep): $g_0(\mathrm{z}) = \sqrt{z_1 z_2}/5 + z_3^2 z_4/4 + \log(z_4+1)/3 + e^{z_5}/2 - 1.91$, $h_0(\mathrm{z}) = \{\sqrt{z_1 z_2}/5 + z_3^2 z_4/4 + \log(z_4+1)/3 + e^{z_5}/2\}^2/5 - 1.36$.

All intercept terms are added to satisfy the identifiable conditions $\mathbb{E}g_0(\mathrm{Z}) = 0$ and $\mathbb{E}h_0(\mathrm{Z}) = 0$. In cases 1 and 2, the effects of all covariates on the log cumulative hazard function are linear and additive, respectively, while the effects in Case 3 are more complicated.

To compare with our proposed method (CPDPLCM), we also include an approach that estimates the effect of Z using a linear model. This method assumes the form $\Lambda(t|\mathrm{X},\mathrm{Z},E) = \Lambda_0(t)\exp\{\boldsymbol{\beta}_0{}^\top \mathrm{X} + \boldsymbol{\varsigma}_0^\top \mathrm{Z} + (\boldsymbol{\gamma}_0^\top \mathrm{X} + \boldsymbol{\vartheta}_0^\top \mathrm{Z} + \vartheta_0) I_{\{E>\zeta_0\}}\}$, where $\boldsymbol{\varsigma}_0, \boldsymbol{\vartheta}_0 \in \mathbb{R}^r$ and $\vartheta_0 \in \mathbb{R}$. This approach, referred to as CPCPH, is a direct extension of methods of Huang (1996) and Pons (2003). For both the two methods, we chose $\breve{\zeta} = \frac{1}{n}\sum_{i=1}^n E_i$, $\breve{\boldsymbol{\theta}} = (\breve{\beta}, \breve{\gamma})^\top = (0,0)^\top$ and $\breve{\boldsymbol{c}} = (\breve{c}_0, \ldots, \breve{c}_{q_n})^\top$ with $\breve{c}_j = 0.1$ for $j = 0, \ldots, q_n$ to be the initial values. The sample size for each case was $n = 1000, 2000$, or 4000, respectively, and 200 replications were performed.

We first describe the evaluation criteria for the estimates. The estimates of regression parameters are evaluated by the estimated biases (Bias), sampling standard errors (SSE), sampling means of the estimated standard error estimates (ESE), and the 95% empirical coverage probabilities (CP). To evaluate the performance of DNN functions, we consider the relative error (RE), defined as

$$RE(\hat{g}) = \left\{ \frac{\frac{1}{n_t}\sum_{i=1}^{n_t}\{\{\hat{g}(\mathrm{Z}_i) - \frac{1}{n_t}\sum_{i=1}^{n_t}\hat{g}(\mathrm{Z}_i)\} - g_0(\mathrm{Z}_i)\}^2}{\frac{1}{n_t}\sum_{i=1}^{n_t}\{g_0(\mathrm{Z}_i)\}^2} \right\}^{1/2}$$

with its SSE, where $\hat{g}$ and g_0 are evaluated at the test set's covariates $\{\mathrm{Z}_i, i = 1, \ldots, n_t\}$, with $n_t = 200$. We also consider the prediction error for the ith test observation, defined as $\mathrm{Error}_i(\hat{g}) = \hat{g}(\mathrm{Z}_i) - g_0(\mathrm{Z}_i)$. Obtain the change point estimate $\hat{\zeta}_n$ by searching a grid with a given gap; in this study, we choose the gap of 0.01 to balance the accuracy and the computational cost. The simulation results include the estimated bias, the length of the 95% empirical confidence interval (95% CI Length), and the SSE of the estimation. Specifically, for M replications, denote $\boldsymbol{\zeta}_M = (\hat{\zeta}_n^{(1)}, \ldots, \hat{\zeta}_n^{(M)})$ be the estimation collection of ζ_0, the 95% CI Length is defined as the difference of the upper 2.5% quantile and the lower 2.5% quantile of $\boldsymbol{\zeta}_M$.

As reported in Table 1, the estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$ exhibit improved performance as the sample size increases in all cases. For the proposed CPDPLCM method, the results indicate that $\hat{\beta}_n$ and $\hat{\gamma}_n$ are unbiased and asymptotically normal, in accordance with Theorem 5. Moreover, the procedure in Theorem 4 for estimating the asymptotic variance is validated as reliable. The 95% coverage probabilities for both $\hat{\beta}_n$ and $\hat{\gamma}_n$ under the proposed method are generally close to 0.95 as the sample size increases. Note that $\hat{\beta}_n$ is estimated more accurately than $\hat{\gamma}_n$ because the change-point effect

does not occur for every observation, reducing the information available for estimating $\hat{\gamma}_n$. In addition, in Cases 2 and 3, the CPCPH method exhibits substantial bias due to model misspecification. Although Case 1 is much simpler and satisfies the assumptions of CPCPH, our method remains highly competitive, with only a slight loss of efficiency. Overall, the proposed approach is flexible for estimating regression parameters and outperforms CPCPH.

Table 2 summarizes the relative errors (REs) and their standard deviations on the test set, and Fig. 1 visualizes the corresponding prediction errors. Both the RE and its SSE decrease as the sample size increases. The relative error of $\hat{h}$ exceeds that of $\hat{g}$, mirroring the behavior observed for $\hat{\boldsymbol{\beta}}$ versus $\hat{\boldsymbol{\gamma}}$. When the true functions correspond to Case 1, the proposed method is slightly less efficient than the perfectly specified CPCPH in estimating g0 and h0. In contrast, for Cases 2 and 3, as sample size grows, our method outperforms CPCPH, yielding smaller REs and prediction errors and exhibiting greater flexibility.

Table 3 presents the simulation results for the change-point estimates. When the identifiability conditions are satisfied, both methods produce valid estimates, corroborating the asymptotic independence established in Theorem 3.3. For both methods, performance improves with increasing sample size. The bias of the change-point estimates is smaller than that of the regression parameters, reflecting the faster convergence rate predicted by Theorem 3.5. Across settings, the accuracy of the change-point estimates degrades slightly as the covariate effects become more complex.

Overall, the CPDPLCM method performs well across all scenarios considered and demonstrates strong flexibility to varying model structures.

Table 1 Estimation of the regression parameters $\boldsymbol{\theta}_0 = (\beta_0, \gamma_0)^\top$

Model	n	Para	CPDPLCM				CPCPH			
			Bias	SSE	ESE	CP	Bias	SSE	ESE	CP
Case 1	1000	β	− 0.018	0.193	0.197	0.965	− 0.020	0.197	0.189	0.945
		γ	0.088	0.256	0.268	0.935	0.061	0.251	0.260	0.955
	2000	β	− 0.003	0.139	0.127	0.935	− 0.006	0.139	0.132	0.935
		γ	0.061	0.189	0.182	0.940	0.026	0.186	0.179	0.950
	4000	β	0.002	0.094	0.096	0.960	− 0.001	0.093	0.093	0.945
		γ	0.029	0.132	0.126	0.950	0.013	0.129	0.125	0.970
Case 2	1000	β	0.003	0.237	0.247	0.950	0.011	0.216	0.221	0.965
		γ	0.059	0.329	0.324	0.940	− 0.109	0.315	0.316	0.920
	2000	β	0.024	0.160	0.155	0.935	0.029	0.156	0.151	0.935
		γ	− 0.041	0.224	0.219	0.930	− 0.135	0.212	0.213	0.875
	4000	β	− 0.012	0.104	0.115	0.950	0.054	0.095	0.104	0.945
		γ	− 0.008	0.153	0.154	0.945	− 0.189	0.140	0.146	0.745
Case 3	1000	β	− 0.035	0.255	0.252	0.930	0.061	0.238	0.239	0.950
		γ	0.061	0.346	0.367	0.960	− 0.325	0.339	0.368	0.845
	2000	β	− 0.033	0.181	0.187	0.950	0.095	0.152	0.162	0.940
		γ	0.043	0.273	0.290	0.955	− 0.400	0.215	0.246	0.610
	4000	β	− 0.022	0.117	0.113	0.940	0.121	0.101	0.111	0.845
		γ	0.003	0.182	0.176	0.955	− 0.467	0.161	0.169	0.195

Table 2 Evaluation of relative errors for $\hat{g}$ and $\hat{h}$ on the test data

Scenario	n	$\hat{g}$				$\hat{h}$			
		CPDPLCM		CPCPH		CPDPLCM		CPCPH	
		RE	SSE	RE	SSE	RE	SSE	RE	SSE
Case 1	1000	0.240	0.044	0.295	0.086	0.298	0.106	0.484	0.150
	2000	0.227	0.037	0.222	0.049	0.272	0.086	0.348	0.101
	4000	0.186	0.026	0.181	0.033	0.218	0.056	0.275	0.073
Case 2	1000	0.270	0.038	0.246	0.038	0.556	0.053	0.561	0.082
	2000	0.229	0.024	0.217	0.017	0.487	0.056	0.482	0.078
	4000	0.194	0.019	0.205	0.009	0.414	0.051	0.474	0.062
Case 3	1000	0.221	0.036	0.303	0.025	0.603	0.068	0.774	0.092
	2000	0.185	0.018	0.288	0.014	0.585	0.064	0.762	0.068
	4000	0.162	0.013	0.283	0.008	0.542	0.055	0.756	0.046

6 Applications to Rotterdam breast cancer data

6.1 Data and model

We applied our proposed methods to a breast cancer dataset comprising records for 2,982 primary breast cancer patients from the Rotterdam Tumor Bank, as used by Royston and Altman (2013). They also recommended using the German Breast Cancer Study Group (GBSG) data, which contain 686 complete patient records, as the test set. Both datasets are included in the R package `survival` and are publicly available at https://cran.r-project.org/web/packages/survival/index.html.

In our study, we use the Rotterdam data to train the model and aim to detect the change-point effect of chemotherapy across ages in preventing breast cancer relapse. Prior work (Keogh and Morris 2018; Chen et al., 2023) suggests that the Cox model is appropriate, and the censoring mechanism is typically treated as right censoring. However, event times may be imprecise and subject to bias (e.g., delayed detection), so treating them as interval-censored is more general and often more robust. A summary of the variables, with minor preprocessing for computational convenience, is provided in the following table.

In this study, we use the binary variable chemo as X, and define the event of interest as breast cancer relapse, with observation time U and censoring indicator Δ taken from rtime and recur, respectively. As noted by Lee et al. (2020), in cancer research the primary outcomes are overall survival or disease-free survival, for which age is often a key prognostic factor; accordingly, we take the change-point variable E to be age. The 7-dimensional covariate Z comprises the remaining variables. According to Table 4, patients who received chemotherapy are a minority (about 20%). Patient ages range from 24 to 90, with middle-aged and older patients predominating. Approximately half of the patients experience relapse, yielding a censoring rate of about 50%.

The model relating relapse time to chemotherapy is specified as

$$\Lambda(t|X, \mathrm{Z}, E) = \Lambda_0(t) \exp\left\{\beta_0 X + g_0(\mathrm{Z}) + \{\gamma_0 X + h_0(\mathrm{Z})\} I_{\{E>\zeta_0\}}\right\}.$$

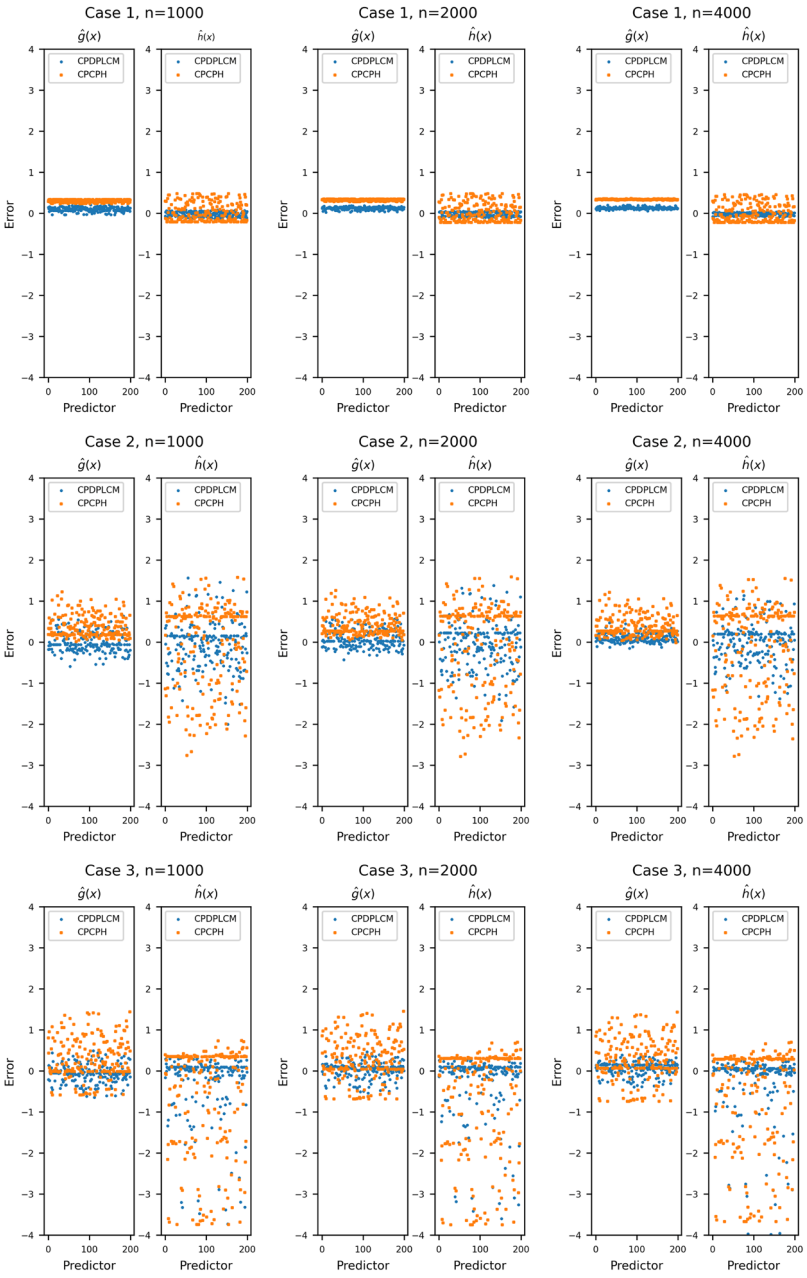

Fig. 1 Prediction errors of the DNN function g and h estimated by the proposed CPDPLCM (circle) and CPCPH (square) on test data (sample size = 200) with sample sizes $n = 1000, 2000$ and 4000 for three different cases

Table 3 Evaluation of the estimation of the change point ζ_0

Scenario	n	CPDPLCM			CPCPH		
		Bias	95% CI length	SSE	Bias	95% CI length	SSE
Case 1	1000	− 0.0003	0.0800	0.0114	− 0.0001	0.0800	0.0110
	2000	− 0.0005	0.0400	0.0050	− 0.0002	0.0400	0.0047
	4000	− 0.0003	0.0200	0.0018	− 0.0005	0.0200	0.0017
Case 2	1000	0.0043	0.2000	0.0270	0.0001	0.1600	0.0205
	2000	0.0028	0.1000	0.0142	0.0020	0.0700	0.0114
	4000	− 0.0007	0.0500	0.0058	− 0.0007	0.0500	0.0058
Case 3	1000	0.0213	0.3400	0.0545	0.0018	0.2900	0.0471
	2000	0.0114	0.1900	0.0262	0.0009	0.1500	0.0208
	4000	0.0058	0.1500	0.0190	− 0.0029	0.1000	0.0149

Table 4 A summary of the Rotterdam data

Variable	Description	Min	Mean	Max
Rtime	Time to relapse or last follow-up (year)	0.0986	5.7477	19.2959
Recur	0 = no relapse, 1 = relapse	0.0000	0.5091	1.0000
Chemo	0 = no chemotherapy, 1 = experienced	0.0000	0.1945	1.0000
Age	Age at surgery (year)	24.0000	55.0600	90.0000
Meno	0 = premenopausal, 1 = postmenopausal	0.0000	0.5600	1.0000
Size	Tumor size (0 = "$\leq$ 20", 1 ="20–50", 2 ="$\geq$50")	0.0000	0.6368	2.0000
Grade	Differentiation grade (take values 2 and 3)	2.0000	2.7340	3.0000
Nodes	Number of positive lymph nodes	0.0000	2.7120	34.0000
Pgr	Progesterone receptors (10^2 fmol/L)	0.0000	1.6180	50.0400
Er	Estrogen receptors (10^2 fmol/L)	0.0000	1.6660	32.7500
Hormon	Hormonal treatment (0 = no, 1 = yes)	0.0000	0.1137	1.0000

In this model, β_0 is the baseline effect of chemotherapy on relapse prevention, while γ_0 captures the additional effect that manifests beyond the age threshold ζ_0. The covariate vector Z represents other control variables, whose effect is modeled by $g_0(\cdot)$; analogous to the treatment effect, the impact of Z may vary with age, and this differential is captured by $h_0(\cdot)$. For comparison, we also apply the CPCPH method.

6.2 Results

This section presents the empirical results based on the real data described above. Before conducting the regression analysis, we first implement the testing procedure in Sect. 4. For computational convenience, we randomly permute the change-point

Table 5 Estimation results for regression parameter (β, γ)

Methods	Para	Value	SE	p value	95% CI
CPDPLCM	β	−0.5832	0.1247	$< 10^{-5}$	[− 0.8275, − 0.3388]
	γ	0.3508	0.1508	0.0100	[0.0552, 0.6464]
CPCPH	β	−0.0068	0.1279	0.4788	[− 0.2575, 0.2439]
	γ	−0.0201	0.1504	0.4469	[− 0.3150, 0.2748]

covariate E 1,000 times, set $k = 5$ in the statistic SUP_k , and obtain a p-value of 0.007, indicating the presence of a change point and yielding an estimate of $\hat{\zeta}_n = 48$.

As shown in Table 5, the p-values obtained by our method are below 0.05, whereas those from the CPCPH method exceed 0.05, indicating that our estimates are statistically significant while CPCPH's are not. Moreover, the 95% confidence interval produced by our method excludes 0, whereas that of CPCPH includes 0, underscoring the greater reliability of our interval estimates. Furthermore, the hazard ratio of patients older than 48 versus those younger than 48 is $\exp(0.3508) \approx 142.02\%$ on average.

The reduced immunological recovery and the menopause are two potential explanations for the change-point effect. First, it is well-known that chemotherapy is harmful both to cancer cells and to normal cells, so it is more difficult for elderly patients to recover from the damage caused by chemotherapy. In addition, according to Ossewaarde et al. (2005), menopause is strongly associated with an increased risk of breast cancer. They discovered that the average age at menopause is 49.0 ± 4.5 years, suggesting that the effect of the change point may come from menopause.

In conclusion, our method effectively detects a shift in chemotherapy effects before and after age 48 and yields reliable 95% confidence interval estimates. We find that chemotherapy substantially reduces recurrence risk across all age groups; however, its preventive effectiveness diminishes among patients older than 48, indicating a change-point effect in efficacy.

7 Concluding remarks

This study focuses on the Cox model with a change point at the covariate for the current status data, proposes a DNN-based estimation method, gives the semiparametric validity of the maximum likelihood estimation, and provides a computational scheme. This method is an effective tool for analyzing interval-censored survival data and identifying significant changes in risk factors, especially when the effects of covariates or nuisance parameters are complex.

This method can be directly extended to a class of generalized partial linear models, such as the transformation model (Kosorok and Song 2007) and the Cox model with right censoring (Pons 2003). Furthermore, the assumption that all covariates are affected by the change-point effect of a single univariate variable is restrictive. Thus, a multivariate $\boldsymbol{E} = (E_1, \ldots, E_q)^\top$ with threshold $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_q)^\top$ could be considered, and variable selection techniques may help detect these more complex change-point effects. In addition, heterogeneity across observations, common in precision medicine, suggests that the change point may be individualized; it may there-

fore be valuable to consider models with individualized change-point effects and to design personalized therapies.

Another challenging issue is how to effectively verify model identifiability. Identifiability is crucial for ensuring estimator consistency; however, current approaches (e.g., Kosorok and Song (2007)) rely heavily on random shuffling and sampling, which may be ineffective when training a DNN and consume substantial computing resources. An optimal approach is to avoid resampling and repeated DNN training to save time. We will explore advances on this issue in future work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10985-026-09689-y.

Funding Open access funding provided by The Hong Kong Polytechnic University. This paper was supported by the National Natural Science Foundation of China (No. 12171374, 12371262, 12501387) and the Ministry of Education of China for Humanities and Social Sciences Research (No. 25YJC910004).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Anthony M, Bartlett PL (1999) Neural network learning: theoretical foundations. Cambridge University Press, Cambridge

Chen R, Cai N, Luo Z, Wang H, Liu X, Li J (2023) Multi-task banded regression model: a novel individual survival analysis model for breast cancer. Comput Biol Med 162:107080

Deng Y, Zeng D, Zhao J, Cai J (2017) Proportional hazards model with a change point for clustered event data. Biometrics 73(3):835–845

Deng Y, Cai J, Zeng D (2022) Maximum likelihood estimation for Cox proportional hazards model with a change hyperplane. Stat Sin 32(2):983–1006

Dupuy JF (2006) Estimation in a change-point hazard regression model. Stat Probab Lett 76(2):182–190

Huang J (1996) Efficient estimation for the proportional hazards model with interval censoring. Ann Stat 24(2):540–568

Huang J, Wellner JA (1997) Interval censored survival data: a review of recent progress. In: Proceedings of the first Seattle symposium in biostatistics: survival analysis. Springer, pp 123–169

Keogh RH, Morris TP (2018) Multiple imputation in Cox regression when there are time-varying effects of covariates. Stat Med 37(25):3661–3678

Kingma DP, Ba JL (2014) ADAM: a method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kosorok MR, Song R (2007) Inference under right censoring for transformation models with a change-point based on a covariate threshold. Ann Stat 35(3):957–989

Lee CY, Chen X, Lam KF (2020) Testing for change-point in the covariate effects based on the Cox regression model. Stat Med 39(10):1473–1488

Li M, Gao L, Lv G, Tong X (2024) Random change point model with an application to the China Household Finance Survey. Sci China Math 67(10):2373–2386
Loader CR (1991) Inference for a hazard rate change point. Biometrika 78(4):749–757
Lu M, McMahan CS (2018) A partially linear proportional hazards model for current status data. Biometrics 74(4):1240–1249
Luo X, Boyett JM (1997) Estimations of a threshold parameter in Cox regression. Commun Stat Theory Methods 26(10):2329–2346
Ma L, Feng Y, Chen D-GD, Sun J (2014) Interval-censored time-to-event data and their applications in clinical trials. Clin Trial Biostat Biopharm Appl 307
Nair V, Hinton, GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
Ossewaarde ME, Bots ML, Verbeek AL, Peeters PH, van der Graaf Y, Grobbee DE, van der Schouw YT (2005) Age at menopause, cause-specific mortality and total life expectancy. Epidemiology 16(4):556–562
Oueslati A, Lopez O (2013) A proportional hazards regression model with change-points in the baseline function. Lifetime Data Anal 19(1):59–78
Pons O (2003) Estimation in a Cox regression model with a change-point according to a threshold in a covariate. Ann Stat 31(2):442–463
Ramsay JO (1988) Monotone regression splines in action. Stat Sci 3:425–441
Royston P, Altman DG (2013) External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol 13(1):33
Schmidt-Hieber J (2020) Nonparametric regression using deep neural networks with ReLU activation function. Ann Stat 48(4):1875–1897
Song R, Kosorok MR, Fine JP (2009) On asymptotically optimal tests under loss of identifiability in semiparametric models. Ann Stat 37(5A):2409–2444
Sun J (2006) The statistical analysis of interval-censored failure time data. Springer series in statistics for biology and health. Springer, New York
Torben M, Scheike TH (2007) Aalen additive hazards change-point model. Biometrika 94(4):861–872
Vanier C, Pandey T, Parikh S, Rodriguez A, Knoblauch T, Peralta J, Hertzler A, Ma L, Nam R, Musallam S et al (2020) Interval-censored survival analysis of mild traumatic brain injury with outcome based neuroimaging clinical applications. J Concuss 4:1–13
Wu Q, Tong X, Zhao X (2024) Deep partially linear Cox model for current status data. Biometrics 80(2):ujae024