Electronic Journal of Statistics
Vol. 18 (2024) 4292–4321
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2298

Deep learning for regression analysis of interval-censored data

Mingyue Du[*1], **Qiang Wu**[*2], **Xingwei Tong**[2] **and Xingqiu Zhao**[3]

[1] *School of Mathematics, Jilin University,*
e-mail: mingydu@jlu.edu.cn

[2] *School of Statistics, Beijing Normal University,*
e-mail: qiangwu@mail.bnu.edu.cn; xweitong@bnu.edu.cn

[3] *Department of Applied Mathematics, The Hong Kong Polytechnic University,*
e-mail: xingqiu.zhao@polyu.edu.hk

Abstract: This paper discusses regression analysis of interval-censored failure time data, and a new deep learning approach is proposed under the partially linear Cox model. For the analysis, we need to overcome theoretical and computational challenges arising from complex data structure where the partial likelihood function is no longer available. We propose to use a deep neural network and a B-spline function for approximating the nonlinear component and the baseline cumulative hazard function in the model, respectively. The proposed approach is flexible and able to circumvent the curse of dimensionality. At the same time, it facilitates the interpretability of covariate effects. The asymptotic properties of the resulting estimators are established. In particular, the finite-dimensional estimator of covariate effects is asymptotically normal and attains the semiparametric efficiency, while the deep nonparametric estimator achieves the minimax optimal rate of convergence. A simulation study is conducted to assess the finite-sample performance of the proposed approach and indicates that it works well in practical situations. Finally, the proposed method is applied to a set of real data that motivated this study.

MSC2020 subject classifications: Primary 62N02, 62G20; secondary 62G08.
Keywords and phrases: Deep neural network, interval-censored data, partially linear Cox model, semiparametric efficiency, spline function.

Received September 2023.

1. Introduction

Interval-censored failure time data commonly occur in many fields, including demographical studies, epidemiological studies, medical or public health studies and social studies [26, 27]. By interval-censored data, we usually mean that the failure time of interest is observed only to belong to an interval instead of being observed exactly and in other words, only incomplete information is available on the failure time of interest. Among others, one area that naturally and routinely yields interval-censored failure time data is longitudinal or periodic follow-up studies such as clinical trials.

[*] Mingyue Du and Qiang Wu contributed equally to this work and they are co-first authors.

An example of interval-censored data that motivated this study is given by the Atherosclerosis Risk in Communities study, a longitudinal epidemiological study that started in 1987 and in which the participants' health status was scheduled to be examined every three years on average [41]. Of course, as expected for such studies, real examination or visiting schedules differ from subject to subject and the occurrence of a disease such as diabetes was known only to belong to be within two consecutive examinations. In other words, only interval-censored data on the time to disease were available and more details will be given below. It is easy to see that there exist many studies similar to the study above and one such example is the Alzheimer's Disease Neuroimaging Initiative, a longitudinal study designed to collect various types of data for the early detection and tracking of Alzheimer's Disease as well as for the development of its treatments [30, 35].

Many authors have investigated regression analysis of interval-censored failure time data under various models, and among them, the most commonly used model is the Cox model given by $\Lambda(t \mid \boldsymbol{Z}) = \Lambda_0(t)\exp\left(\boldsymbol{\theta}_0'\boldsymbol{Z}\right)$ in terms of the cumulative hazard function of the failure time of interest [9, 29, 36, 37]. Here Λ_0 denotes an unknown baseline cumulative hazard function, $\boldsymbol{Z}$ a p-dimensional vector of covariates, and $\boldsymbol{\theta}_0$ a vector of regression parameters. For example, Huang [9] investigated the asymptotic properties of the maximum likelihood estimators of both the regression parameters and the baseline cumulative hazard function under the model above. Huang and Rossini [11] and Shen [22] considered the sieve maximum likelihood estimation for the proportional odds model, while Sun and Sun [29], Zhang et al. [38], Zhang and Zhao [39] and Zeng et al. [36] discussed the fitting of linear transformation models to interval-censored data. For more references on the analysis of interval-censored data, one can refer to the books by Sun [26] and Sun and Chen [27].

For all of methods mentioned above, one important limitation is that they all assume that covariate effects are linear and it is apparent that this may not be true in reality. To deal with this, several authors have considered the use of partially linear models that allow for nonlinear effects to be described by unknown smooth functions. For example, Huang [10] generalized the Cox model above to $\Lambda(t \mid \boldsymbol{Z}) = \Lambda_0(t)\exp\left(\boldsymbol{\theta}_0'\boldsymbol{Z} + \sum_{k=1}^{r} h_k(X_k)\right)$, where Λ_0, $\boldsymbol{Z}$ and $\boldsymbol{\theta}_0$ are defined as above, $\boldsymbol{X} = (X_1, \ldots, X_r)'$ denotes a r-dimensional vector of covariates that may have nonlinear effects, and the h_k's are unknown smooth functions. Also Ma and Kosorok [16] investigated the penalized semiparametric maximum likelihood estimation of partially linear transformation models, and Cheng and Wang [2] considered efficient estimation in a semiparametric additive transformation model. Note that all of the work described above focused on either right-censored data or current status data, a special case of interval-censored data where the failure time of interest is either left- or right-censored. Also all methods have some limitations on the type of covariate effects and to address them, we propose a deep learning-based method.

Deep learning has received increasing attention in failure time analysis and in particular, it provides an efficient and flexible approach to the estimation of unknown functions involved in regression models [12, 40]. Among others, Zhong et al. [40] considered the following partially linear Cox model

$$\Lambda(t \mid \boldsymbol{Z}, \boldsymbol{X}) = \Lambda_0(t) \exp\left(\boldsymbol{\theta}_0' \boldsymbol{Z} + g_0(\boldsymbol{X})\right) \tag{1}$$

where g_0 is an unknown function, and developed a deep neural network (DNN) estimation procedure. In particular, they pointed out that the model above not only inherits the simple interpretation of the finite-dimensional parameter $\boldsymbol{\theta}_0$ in the Cox model but also models more complex nonlinear effects of the covariate $\boldsymbol{X}$, thus more accurately capturing the properties of real data. Also they showed numerically that model (1) gives more stable results than the partially linear additive Cox model [10]. However, they only considered right-censored data, and as pointed out in the literature, the analysis of interval-censored data is quite different from and much more challenging than that of right-censored data.

Some researchers have also developed some DNN-based methods for regression analysis of interval-censored data. For example, Sun and Ding [28] and Meixide et al. [17] considered the nonparametric and partially linear Cox models. For the problem, they proposed LASSO-based penalized maximum likelihood estimation procedures and in particular, Meixide et al. [17] developed a LASSO optimization algorithm and demonstrated the usefulness of the DNN-based method to capture the nonlinear effects of covariates. Also Sun and Ding [28] used Bernstein polynomials for the estimation of the baseline cumulative hazard function. However, no theoretical justifications were provided for both methods. Note that as discussed in Huang [10] and Zhong et al. [40], sometimes there exist two types of covariates $\boldsymbol{Z}$ and $\boldsymbol{X}$ with $\boldsymbol{Z}$ representing the covariates that have linear effects and $\boldsymbol{X}$ the covariates that have nonlinear effects. In such situations, one may be mainly interested in making inferences about the linear effect and one common, important example is that $\boldsymbol{Z}$ denotes the treatment indicators with the treatment comparison being the focus. In this paper, we consider the situation where there exists both linear and nonlinear covariate effects with the focus on simultaneously estimating the linear effects such as the treatment effect and the nonlinear effects. Also instead of using Bernstein polynomials, we employ spline functions and provide the asymptotic theory of the proposed method. In particular, we show that the proposed nonparametric DNN estimator achieves the minimax optimal rate of convergence (up to a polylogarithmic factor). Furthermore, we derive that the resulting estimator of linear covariate effects is $\sqrt{n}$-consistent, asymptotically normal, and attains semiparametric efficiency.

To achieve the goal described above, we will consider case II interval-censored data under model (1) and generalize the method given in Zhong et al. [40]. The proposed method combines the statistical method with the DNN method. As mentioned above, one advantage of the DNN approach is that it can accommodate a rich class of nonlinear functions in the model to avoid the curse of dimensionality and yield faster convergence rates than usual nonparametric smoothing methods. A major difference between the proposed method and that given in Zhong et al. [40] is that we have to deal with much more complicated data structures, which make both computation and theoretical justification more difficult among others. In particular, to estimate covariate effects in the model above with right-censored data, a simple partial likelihood function that is independent of the unknown baseline cumulative hazard function is available and commonly used, while no such function exists for interval-censored data and a much more complicated full likelihood function has to be handled. In other words, we have to deal with an extra, unknown function in comparison to Zhong et al. [40], and one resulting major difficulty is that we have to perform the simultaneous estimation of the two unknown functions, one being the covariate-dependent function and the other being infinite-dimensional baseline cumulative hazard function of time. In contrast, Zhong et al. [40] only need to deal with the covariate-dependent function. To estimate the extra unknown baseline cumulative hazard function, we employ monotone B-spline functions to approximate it [21]. Due to these factors, in particular, more techniques such as those used in Wellner and Zhang [34] have to be employed to establish the asymptotic properties of the proposed estimators.

The remainder of the paper is organized as follows. In Section 2, we first introduce some notation, data structure, and the likelihood function, and then present spline, DNN-based sieve maximum likelihood estimation procedure. In Section 3, the theoretical properties of the resulting estimators are established. Section 4 presents some results obtained from a simulation study conducted to assess the finite-sample performance of the proposed method, and they suggest that the method works well in practical situations. The proposed method is applied to the Atherosclerosis Risk in Communities study discussed above in Section 5, and Section 6 concludes with some discussion and concluding remarks. The technical proofs are provided in the Supplementary Material.

2. Methodology

2.1. Likelihood function

Consider a failure time study and let T denote the failure time of interest. Also let $\boldsymbol{Z}$ and $\boldsymbol{X}$ be defined as above, representing p- and r-dimensional vectors of covariates that have linear and nonlinear effects on T, respectively. More comments on this will be given below. Assume that given $\boldsymbol{Z}$ and $\boldsymbol{X}$, the cumulative hazard function of T is given by model (1).

In the following, we will focus on the situation where for each study subject, there exist two observation times denoted by U and V with $U < V$ and one only observes the indicator functions $\delta_1 = 1_{[T \le U]}$, $\delta_2 = 1_{[U<T\le V]}$ and $\delta_3 = 1 - \delta_1 - \delta_2$. That is, only case II interval-censored are available [26]. Define $O = (\delta_1, \delta_2, \delta_3, U, V, \boldsymbol{Z}, \boldsymbol{X})$ and let $S(\cdot \mid \boldsymbol{z}, \boldsymbol{x})$ denote the survival function of T given $\boldsymbol{Z} = \boldsymbol{z}$ and $\boldsymbol{X} = \boldsymbol{x}$. Also assume that conditional on $\boldsymbol{Z}$ and $\boldsymbol{X}$, T is independent of (U, V). That is, we have independent interval censoring. Then the likelihood contribution of the observation O is proportional to

$$L(O) = \{1 - S(u \mid \boldsymbol{z}, \boldsymbol{x})\}^{\delta_1} \{S(u \mid \boldsymbol{z}, \boldsymbol{x}) - S(v \mid \boldsymbol{z}, \boldsymbol{x})\}^{\delta_2} S(v \mid \boldsymbol{z}, \boldsymbol{x})^{\delta_3}.$$

Suppose that the observed data consist of n i.i.d. samples of O, denoted by $\{O_i = (\delta_{1i}, \delta_{2i}, \delta_{3i}, U_i, V_i, \boldsymbol{Z}_i, \boldsymbol{X}_i)\,; i = 1, 2, \ldots, n\}$. Let $\phi = \log \Lambda$, then under model (1), the observed log likelihood function of $(\boldsymbol{\theta}, g, \phi)$ has the form

$$
\begin{aligned}
l_n(\boldsymbol{\theta}, g, \phi) = \sum_{i=1}^{n} \Big(& \delta_{1i} \log \Big[1 - \exp \Big\{ -\mathrm{e}^{\boldsymbol{\theta}' \boldsymbol{Z}_i + g(\boldsymbol{X}_i) + \phi(U_i)} \Big\} \Big] \\
& + \delta_{2i} \log \Big[\exp \Big\{ -\mathrm{e}^{\boldsymbol{\theta}' \boldsymbol{Z}_i + g(\boldsymbol{X}_i) + \phi(U_i)} \Big\} - \exp \Big\{ -\mathrm{e}^{\boldsymbol{\theta}' \boldsymbol{Z}_i + g(\boldsymbol{X}_i) + \phi(V_i)} \Big\} \Big] \\
& - \delta_{3i} \mathrm{e}^{\boldsymbol{\theta}' \boldsymbol{Z}_i + g(\boldsymbol{X}_i) + \phi(V_i)} \Big) .
\end{aligned} \tag{2}
$$

2.2. DNN-based estimation procedure

In this subsection, we discuss the estimation in model (1). For this, it is apparent that a natural approach is to maximizie (2) with respect to $\boldsymbol{\theta}$, g and ϕ. On the other hand, it is easy to see that the direct maximization would be very difficult or impossible since it involves the unknown functions g and ϕ. To deal with this, we propose first to approximate the function ϕ by a B-spline function and the function g by a DNN.

First we discuss the approximation of the function ϕ. For this, let a and b denote the lower and upper bounds of the observation times $\{(U_i, V_i) : i = 1, \ldots, n\}$ and $a = d_0 < d_1 < \cdots < d_{K_n} < d_{K_n+1} = b$ be a partition of $[a, b]$, where $K \equiv K_n \approx n^v$ is a positive integer such that $\max_{1 \le k \le K+1} |d_k - d_{k-1}| = O(n^{-v})$. Define $I_{Kt} = [d_t, d_{t+1}), t = 0, \ldots, K_n$, and $D_n = \{d_1, \ldots, d_{K_n}\}$. Following Schumaker [21] and Stone [25], we let $\mathcal{S}_n(D_n, K_n, m)$ be the space of polynomial splines of order $m \ge 1$ consisting of functions s satisfying: (i) the restriction of s to I_{Kt} is a polynomial of order m for $m \le K$; and (ii) for $m \ge 2$ and $0 \le m' \le m-2$, s is m' times continuously differentiable on $[a, b]$.

According to Corollary 4.10 in Schumaker [21], there exists a local basis $\mathcal{B}_n \equiv \{\mathbf{b}_t, 1 \le t \le q_n\}$, so-called B-spline, for $\mathcal{S}_n(D_n, K_n, m)$, where $q_n \equiv K_n + m$. These basis functions are non-negative and sum up to one at each point in $[a, b]$, and each $\mathbf{b}_t$ is zero outside the interval $[d_t, d_{t+m}]$. To approximate the function ϕ, define

$$
\mathcal{M}_n(D_n, K_n, m) = \left\{ \phi_n(t) = \sum_{j=1}^{q_n} \beta_j \mathbf{b}_j(t) \in \mathcal{S}_n(D_n, K_n, m) : \boldsymbol{\beta} \in B_n, t \in [a, b] \right\},
$$

where $B_n = \{\boldsymbol{\beta} : \beta_1 \le \beta_2 \le \cdots \le \beta_{q_n}\}$ since ϕ is non-decreasing. For the number q_n above, it is usually set to be a positive integer around n^v with $0 < v < 1/2$ [5, 15].

To approximate the covariate function g, we first briefly review the DNNs. A $(K+1)$-layer DNN with layer-width $\boldsymbol{p}$ is a composite function $g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{K+1}}$ recursively defined as

$$
\begin{aligned}
& g(x) = W_K g_K(x) + v_K, \\
& g_K(x) = \sigma(W_{K-1} g_{K-1}(x) + v_{K-1}), \ldots, g_1(x) = \sigma(W_0 x + v_0).
\end{aligned} \tag{3}
$$

Here $K \in \mathbb{N}_+$ denotes the depth of the network, $\boldsymbol{p} = (p_0, \ldots, p_K, p_{K+1}) \in \mathbb{N}_+^{K+2}$ lists the width of each layer, the matrices $W_k \in \mathbb{R}^{p_{k+1} \times p_k}$ and vectors $v_k \in \mathbb{R}^{p_{k+1}}$ (for $k = 0, \ldots, K$) are the parameters of the DNN. Chosen a priori, the activation functions $\sigma\left((x_1, \ldots, x_{p_k})'\right) = (\sigma(x_1), \ldots, \sigma(x_{p_k}))'$, which gives $g_k = (g_{k1}, \ldots, g_{kp_k})' : \mathbb{R}^{p_{k-1}} \to \mathbb{R}^{p_k}$ for $k = 1, \ldots, K$. The rectified linear unit (ReLU) [18], the most popular activation function, is given by $\sigma(x) = \max\{x, 0\}$. In the following, we will focus on the ReLU activation.

Note that in reality, a deep feedforward network with fully-connected layers can contain a huge number of parameters, which can lead to overfitting. This issue can be mitigated by pruning weights, which reduces the total number of nonzero parameters such that the network's layers are only sparsely connected [8, 20]. Following the similar methodology, for $s \in \mathbb{N}_+$ and $D > 0$, we focus on the following class of sparse neural networks

$$\mathcal{G}(K, s, \boldsymbol{p}, D) := \left\{ g \in \mathcal{G}(K, \boldsymbol{p}) : \sum_{k=1}^{K} \|W_k\|_0 + \|v_k\|_0 \leq s, \|g\|_\infty \leq D \right\},$$

where $\|\cdot\|_\infty$ is the sup-norm of a function, and $\|\cdot\|_0$ is the number of non-zero entries of a matrix or vector. Here $\mathcal{G}(K, \boldsymbol{p})$ is a class of DNN considered as

$$\mathcal{G}(K, \boldsymbol{p}) = \{g : g \text{ is a DNN with } (K+1) \text{ layers and width vector } \boldsymbol{p} \text{ such that } \max(\|W_k\|_\infty, \|v_k\|_\infty) \leq 1 \text{ for all } k = 0, \ldots, K\}.$$

Let $\Theta \in R^p$ denote the feasible domain for the regression parameter $\boldsymbol{\theta}$, and define $\mathcal{M}_n = \mathcal{M}_n(D_n, K_n, m)$ and $\mathcal{G} = \mathcal{G}(K, s, \boldsymbol{p}, \infty)$ for simplicity. We propose to estimate $\boldsymbol{\theta}$, g and ϕ by the estimator $\hat{\boldsymbol{\tau}} = \left(\hat{\boldsymbol{\theta}}, \hat{g}, \hat{\phi}\right)$ defined as the values of $\boldsymbol{\theta}$, g and ϕ that maximize the log likelihood function $l_n(\boldsymbol{\theta}, g, \phi)$ over the space $\Theta \times \mathcal{G} \times \mathcal{M}_n$ or $l_n(\boldsymbol{\theta}, g, \boldsymbol{\beta})$ over the space $\Theta \times \mathcal{G} \times B_n$. For the determination of $\hat{\boldsymbol{\tau}}$, it is apparent that one has to choose initial estimates and some parameters such as the degree of monotone splines m and the cardinality of the interior knot set K_n. Also one needs to choose the so-called hyperparameters including the number of hidden layers K, the number of neurons p_k in all K hidden layers, the dropout rate [24], the learning rate [7], and the number of epochs in deep learning. More details will be provided in Section 4. In the next section, we establish the asymptotic properties of $\hat{\boldsymbol{\tau}}$.

3. Asymptotic properties

To establish the asymptotic properties of the estimator $\hat{\boldsymbol{\tau}}$ defined above, we first describe some needed conditions. Assume that the function g belongs to a Hölder class of smooth functions defined as

$$\mathcal{H}_r^\alpha(\mathbb{D}, M) = \left\{ g : \mathbb{D} \to \mathbb{R} : \sum_{\boldsymbol{\beta}:|\boldsymbol{\beta}|<\alpha} \left\|\partial^{\boldsymbol{\beta}} g\right\|_\infty + \sum_{\boldsymbol{\beta}:|\boldsymbol{\beta}|=\lfloor\alpha\rfloor} \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{D},\boldsymbol{x}\neq\boldsymbol{y}} \frac{\left|\partial^{\boldsymbol{\beta}} g(\boldsymbol{x}) - \partial^{\boldsymbol{\beta}} g(\boldsymbol{y})\right|}{\|\boldsymbol{x}-\boldsymbol{y}\|_\infty^{\alpha-\lfloor\alpha\rfloor}} \leq M \right\}$$

with respect to the parameters $\alpha, M > 0$ and domain $\mathbb{D} \subset \mathbb{R}^r$ [20]. In the above, $\lfloor\alpha\rfloor$ is the largest integer strictly smaller than $\alpha, \partial^{\boldsymbol{\beta}} := \partial^{\beta_1} \dots \partial^{\beta_r}$ with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_r)$, and $|\boldsymbol{\beta}| = \sum_{k=1}^r \beta_k$. Let $q \in \mathbb{N}, M > 0, \boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_q) \in \mathbb{R}_+^{q+1}$, and $\boldsymbol{d} = (d_0, \dots, d_{q+1}) \in \mathbb{N}_+^{q+2}, \tilde{\boldsymbol{d}} = (\tilde{d}_0, \dots, \tilde{d}_q) \in \mathbb{N}_+^{q+1}$ with $\tilde{d}_j \leq d_j, j = 0, \dots, q$, where $\mathbb{R}_+$ is the set of all positive real numbers. Note that here $\boldsymbol{d}$ denotes the numbers of the functions in each layer, including the input layer and output layer, while $\tilde{\boldsymbol{d}}$ represents the numbers of the variables used for the function in each layer. Furthermore, assume that g_0 belongs to a composite smoothness function class:

$$\mathcal{H}(q, \boldsymbol{\alpha}, \boldsymbol{d}, \tilde{\boldsymbol{d}}, M) := \Big\{ g = g_q \circ \cdots \circ g_0 : g_i = (g_{i1}, \dots, g_{id_{i+1}})' \text{ and } g_{ij} \in \mathcal{H}_{\tilde{d}_t}^{\alpha_t}\left([a_i, b_i]^{\tilde{d}_i}, M \Big] \right), \text{ for some } |a_i|, |b_i| \leq M \Big\}.$$

Note that the functions in this class are characterized by two kind of dimensions $\boldsymbol{d}$ and $\tilde{\boldsymbol{d}}$ with the latter representing the intrinsic dimension of the function.

Denote $\tilde{\alpha}_i = \alpha_i \prod_{k=i+1}^q (\alpha_k \wedge 1)$ and $\gamma_n = \max_{i=0,\dots,q} n^{-\tilde{\alpha}_i/(2\tilde{\alpha}_i + \tilde{d}_i)}$, where $a \wedge b := \min\{a, b\}$. For any $\phi_1, \phi_2 \in \Phi$, define $\|\phi_1 - \phi_2\|_\Phi^2 = E\{\phi_1(U) - \phi_2(U)\}^2 + E\{\phi_1(V) - \phi_2(V)\}^2$, and for any $\boldsymbol{\tau}_1 = (\boldsymbol{\theta}_1, g_1, \phi_1)$ and $\boldsymbol{\tau}_2 = (\boldsymbol{\theta}_2, g_2, \phi_2)$ in the space of $\mathcal{T} = \Theta \times \mathcal{G} \times \Phi$, define an L_2-metric:

$$d(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = \|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\|_{\mathcal{T}} = \left\{ \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + \|g_1 - g_2\|_{L^2([0,1]^r)}^2 + \|\phi_1 - \phi_2\|_\Phi^2 \right\}^{1/2}.$$

Also define $\|\boldsymbol{v}\|_c^2 = (v_1^2, \dots, v_p^2)'$ for any vector $\boldsymbol{v} = (v_1, \dots, v_p)' \in \mathcal{R}^p$. The conditions referred in the theorems below are given in the Supplementary Material.

Theorem 3.1. *Assume that the conditions (C1)–(C7) hold. Then there exists an estimator $\hat{g}$ satisfying $\mathbb{E}\{\hat{g}(\boldsymbol{X})\} = 0$ such that $\|\hat{g} - g_0\|_{L^2([0,1]^r)} = O_p(\gamma_n \log^2 n)$.*

Theorem 3.2. *Assume that the conditions (C1)–(C7) hold. Then there exists a constant $0 < c < \infty$ such that*

$$\inf_{\hat{g}} \sup_{(\boldsymbol{\theta}_0, g_0, \phi_0) \in \Theta \times \mathcal{H}_0 \times \Phi} \mathbb{E}\{\hat{g}(\boldsymbol{X}) - g_0(\boldsymbol{X})\}^2 \geq c\gamma_n^2,$$

where the infimum is taken over all possible estimators $\hat{g}$ based on the observed data.

Theorem 3.3. *Assume that the conditions (C1)-(C7) hold. Then we have the efficient score for $\boldsymbol{\theta}$ as*

$$\ell_{\boldsymbol{\theta}}^*(\boldsymbol{\tau}) = \dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \dot{\ell}_\phi(\boldsymbol{\tau})[\mathbf{a}^*] - \dot{\ell}_g(\boldsymbol{\tau})[\mathbf{h}^*],$$

where $\mathbf{a}^* \in \mathcal{A}^p$ and $\mathbf{h}^* \in \mathcal{H}^p$ are the unique functions that minimize the distance $\|\dot{\ell}_\theta(\boldsymbol{\tau}) - \dot{\ell}_\phi(\boldsymbol{\tau})[\mathbf{a}] - \dot{\ell}_g(\boldsymbol{\tau})[\mathbf{h}]\|_c^2$, for $\mathbf{a} \in \mathcal{A}^p$ and $\mathbf{h} \in \mathcal{H}^p$. The information bound of $\boldsymbol{\theta}$ takes the form $\mathbf{I}(\boldsymbol{\theta}) = E\{\ell_{\boldsymbol{\theta}}^*(\boldsymbol{\tau})\}^{\otimes 2}$.

Theorem 3.4. *Assume that the conditions (C1)–(C7) hold and the information matrix* $\mathbf{I}(\boldsymbol{\theta}_0)$ *is nonsingular. Then if* $n\gamma_n^4 \to 0$ *as* $n \to \infty$*, we have* $n^{1/2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left\{0, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\right\}$, *as* $n \to \infty$.

The proof of the results above is sketched in the Supplementary Material. In particular, we show that the proposed nonparametric DNN estimator achieves the minimax optimal rate of convergence and the estimator of covariate effects attains semiparametric efficiency. To carry out the inference about $\boldsymbol{\theta}_0$ based on the results above, it is apparent that one needs to estimate the asymptotic covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)$. For this, note that

$$\boldsymbol{I}(\boldsymbol{\theta}_0) = \mathbb{E}\big\{[\boldsymbol{Z} - \boldsymbol{a}(U) - \boldsymbol{h}(\boldsymbol{X})]Q_4 + [\boldsymbol{Z} - \boldsymbol{a}(V) - \boldsymbol{h}(\boldsymbol{X})]Q_5\big\}^{\otimes 2},$$

and one may estimate $(\boldsymbol{a}^*, \boldsymbol{h}^*)$ by minimizing empirical objective function

$$(\boldsymbol{a}^*, \boldsymbol{h}^*) = \arg\min_{(\boldsymbol{a},\boldsymbol{h})} \frac{1}{n}\sum_{i=1}^{n} \|[\boldsymbol{Z}_i - \boldsymbol{a}(U_i) - \boldsymbol{h}(\boldsymbol{X}_i)]Q_{4i} + [\boldsymbol{Z}_i - \boldsymbol{a}(V_i) - \boldsymbol{h}(\boldsymbol{X}_i)]Q_{5i}\|_2^2,$$

where Q_4 and Q_5 are defined in the proof of Theorem 3.3. However, it would be difficult to obtain the convincing solution by utilizing the classical nonparametric methods due to the high dimensionality of function $\boldsymbol{h}$. As a result, we suggest to use a DNN-based method to approximate $(\boldsymbol{a}^*, \boldsymbol{h}^*)$ with the inputs and outputs being $(U, V, \boldsymbol{X})$ and $\{\boldsymbol{a}(U)+\boldsymbol{h}(\boldsymbol{X}), \boldsymbol{a}(V)+\boldsymbol{h}(\boldsymbol{X})\}$, respectively. Its implementation details are similar to the network used above for the maximization of the log likelihood function. Thus given the resulting estimate of $(\boldsymbol{a}^*, \boldsymbol{h}^*)$, the information bound can be estimated by

$$\hat{\boldsymbol{I}}(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \|(\boldsymbol{Z}_i - \boldsymbol{a}^*(U_i) - \boldsymbol{h}^*(\boldsymbol{X}_i))Q_{4i} + (\boldsymbol{Z}_i - \boldsymbol{a}^*(V_i) - \boldsymbol{h}^*(\boldsymbol{X}_i))Q_{5i}\|_2^2.$$

4. Simulation studies

In this section, we present some results obtained from a simulation study conducted to evaluate the finite-sample performance of the DNN-based estimation approach proposed in the previous sections. In the study, we considered the situation with one single covariate Z generated from the Bernoulli distribution with the success probability 0.5 and the covariate $\boldsymbol{X} = (X_1, \ldots, X_5)'$ generated from the multivariate normal distribution with mean zero, variance 1 and correlation 0.5 and truncated over the interval $[0, 2]$. With respect to the regression function $g_0(\boldsymbol{x})$ and the baseline cumulative hazard function $\Lambda_0(t)$, four different cases were considered and they are

- Case 1 (Linear): $g_0(\boldsymbol{x}) = \frac{x_1}{2} + \frac{x_2}{3} + \frac{x_3}{3} + \frac{x_4}{4} + \frac{x_5}{5} - 0.63$, $\Lambda_0(t) = \sqrt{t}/5$;
- Case 2 (Additive): $g_0(\boldsymbol{x}) = \frac{x_1^2}{3} + \frac{\log(x_2+1)}{2} + \frac{\sqrt{x_3}}{4} + \frac{e^{x_4}}{3} + \frac{x_5}{2} - 1.18$, $\Lambda_0(t) = \log\left(1 + \frac{t^{4/5}}{6}\right)$;

Table 1
Hyperparameter settings.

Number of layers	1	2	3	4	5
Number of neurons	30	50	100	150	200
Dropout rate	0.0	0.1	0.2	0.3	0.4
Learning rate	1e-5	5e-5	1e-4	1e-3	1e-2
Number of epochs	200	300	500	800	1000

- Case 3 (Deep 1): $g_0(\boldsymbol{x}) = \frac{x_1 x_2^2}{4} + \frac{\sqrt{x_3 x_4}}{5} + \frac{\log(x_4+1)}{4} + \frac{e^{x_5}}{2} - 1.11$, $\Lambda_0(t) = 2\sqrt{t}/9$;
- Case 4 (Deep 2): $g_0(\boldsymbol{x}) = \frac{1}{6}\{\frac{x_1 x_2^2}{4} + \frac{\sqrt{x_3 x_4}}{5} + \frac{\log(x_4+1)}{4} + \frac{e^{x_5}}{2}\}^2 - 0.36$, $\Lambda_0(t) = 4\sqrt{t}/17$.

Note that the various intercept terms 0.63, 1.18, 1.11 and 0.36 were added to g_0 to satisfy the assumption $\mathbb{E}g_0(\boldsymbol{X}) = 0$. Given the covariates $(Z, \boldsymbol{X})$, the event time T was generated under model (1). To generate interval-censored data, for each of the four cases above, the U and V were generated as follows:
(i) Case 1: $U \sim \text{Uniform}(0, \frac{\tau}{10})$ and $V \sim \min\{\frac{\tau}{5} + U + \frac{\tau}{2}\text{Exponetial}(1), \tau\}$;
(ii) Case 2: $U \sim \text{Uniform}(0, \frac{\tau}{6})$ and $V \sim \min\{\frac{\tau}{4} + U + \frac{\tau}{2}\text{Exponetial}(1), \tau\}$;
(iii) Case 3: $U \sim \text{Uniform}(0, \frac{\tau}{10})$ and $V \sim \min\{\frac{\tau}{4} + U + \frac{\tau}{2}\text{Exponetial}(1), \tau\}$;
(iv) Case 4: $U \sim \text{Uniform}(0, \frac{\tau}{9})$ and $V \sim \min\{\frac{\tau}{5} + U + \frac{\tau}{2}\text{Exponetial}(1), \tau\}$.

It was suggested to select the end time $\tau = 20$ to control the censoring rate with 30-35% left-censored observations and 30-35% right-censored ones on average in all simulation studies. We considered using cubic monotone B-splines (i.e., $m = 4$) to estimate $\Lambda_0(\cdot)$. We selected the cardinality of the interior knot set $K_n = [n^{1/3}]$ and divided the support set $[0, \tau]$ equally. Here the symbol $[\cdot]$ represents rounding. For the initial values of spline coefficients β_j's, we suggest setting them as a vector with all components being one. The initial value of θ was chose to be zero. Regarding the initialization of the neural network parameters (W_k's and v_k's), we used `Pytorch`'s default random initialization [6, 19]. Selecting good initial values for these matrices and vectors is not straightforward. The hyperparameters include the number of hidden layers K, the number of neurons p_k in each hidden layer, the number of epochs, the dropout rate, and the learning rate in deep learning. These hyperparameters need to be specified for implementing g_0. In our simulations, we consider the same number of neurons in each hidden layer (i.e., $p_k = p_j$ for $1 \leq k, j \leq K$). The dropout rate refers to randomly ignoring some neurons during training, and the learning rate determines the step size in the `Adam` optimization algorithm. The candidates of hyperparameters for DNNs in the simulations and data application can be found in Table 1.

In our simulations, the negative log-likelihood function was considered as the loss function, and the parameters were updated iteratively. The detailed estimation procedure is as follows:

- Step 1. Given the initial values of θ and spline parameters β_j's, the negative log-likelihood function was minimized by `Pytorch` to obtain the DNN parameters' estimates $\hat{W}_k$'s and $\hat{v}_k$'s.

Table 2

Simulation results of $\hat{\theta}$ by the DNN-based, CPH, and PLACM methods with 200 replications.

		DNN-based				CPH				PLACM			
	n	Bias	SSE	ESE	CP	Bias	SSE	ESE	CP	Bias	SSE	ESE	CP
Case 1	1000	0.030	0.088	0.088	0.950	0.014	0.086	0.086	0.955	0.043	0.092	0.089	0.910
	2000	0.015	0.067	0.065	0.930	0.002	0.066	0.064	0.945	0.017	0.067	0.065	0.930
Case 2	1000	0.020	0.089	0.088	0.960	−0.448	0.160	0.087	0.050	0.041	0.093	0.094	0.930
	2000	0.006	0.069	0.067	0.930	−0.464	0.156	0.061	0.045	0.014	0.070	0.067	0.930
Case 3	1000	0.005	0.091	0.089	0.935	−0.269	0.158	0.083	0.205	0.125	0.198	0.085	0.505
	2000	−0.000	0.067	0.065	0.935	−0.299	0.143	0.058	0.095	0.117	0.188	0.058	0.435
Case 4	1000	0.016	0.086	0.086	0.940	−0.324	0.232	0.082	0.220	0.055	0.194	0.060	0.420
	2000	0.008	0.065	0.064	0.930	−0.294	0.253	0.057	0.110	0.054	0.177	0.042	0.365

- Step 2. Plugging the obtained values of $\hat{W}_k$ and $\hat{v}_k$ into the loss function, the negative log-likelihood function was minimized again by the package `scipy.optimize.minimize` to obtain parameter estimates $\hat{\theta}$ and $\hat{\beta}_j$.
- Step 3. The iteration steps were repeated until convergence.

For comparison, we also include the numerical results from two alternative models: the Cox proportional hazards model (CPH) proposed by Cox [3, 4], as well as the partially linear additive Cox model (PLACM) introduced by Huang [10].

Table 2 shows the simulation results for $\hat{\theta}$ given by the DNN-based, CPH and PLACM estimation procedures, with the true value $\theta_0 = 1$. This table includes the estimated bias (Bias) is given by the mean of the estimates minus the true value, the sample standard error of the estimates (SSE), the mean of the estimated standard errors (ESE), and the 95% empirical coverage probability (CP). In all simulations, the performance of the proposed estimator improves as the sample size increases from 1000 to 2000. It is noteworthy that the DPLCM method significantly outperforms the CPH method and PLACM method in cases 3–4, where overly restrictive models may lead to large biases. However, in the simpler Case 1 where the CPH and PLACM assumptions are met, the DPLCM method remains very competitive with little efficiency loss. In addition, the coverage of the proposed DPLCM method's confidence intervals is generally close to the desired 95% level, especially as the sample size n increases. In contrast, the confidence intervals obtained using the CPH method have poor coverage in cases 2–4, while the PLACM confidence intervals have poor coverage in cases 3–4. This suggests that the uncertainty quantification provided by these alternative methods may not be reliable in practical applications.

Furthermore, we plot the estimates of the cumulative baseline hazard function obtained from the three different methods. As illustrated in Figure 1, the estimate produced by the DPLCM method converges to the true cumulative baseline function with only minimal bias. Conversely, the CPH estimates under cases 2–4 and the PLACM estimates under cases 3–4 display substantial biases.

For the evaluation of the performance of the proposed method on estimation

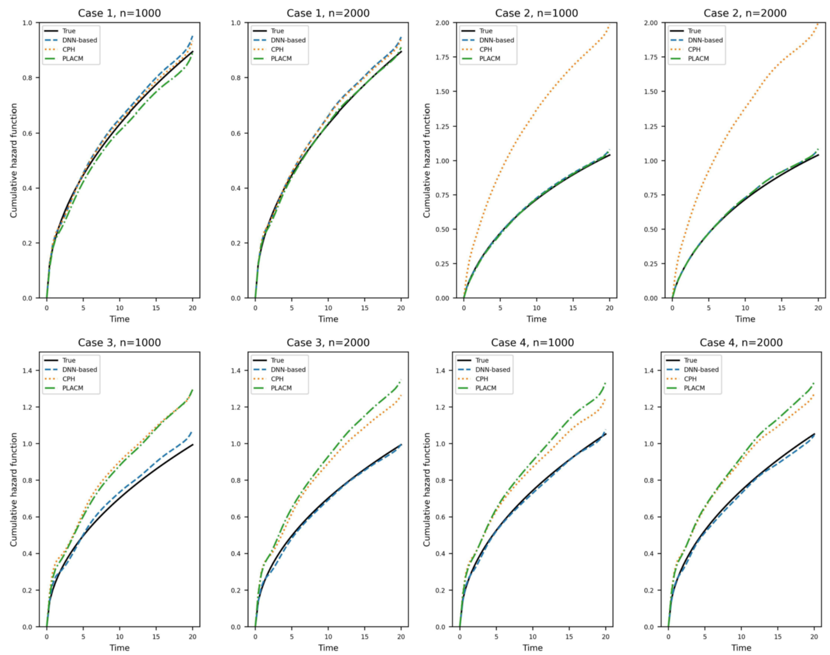

FIG 1. *Estimates of* $\Lambda_0(\cdot)$ *by the DNN-based (dashed line), CPH (dotted line), and PLACM (dash-dotted line) methods with 200 replications.*

of the covariate function g_0, we calculated the relative error given by

$$\mathrm{RE}(\hat{g}) = \left[\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\{(\hat{g}(\boldsymbol{X}_i) - \bar{\hat{g}}) - g_0(\boldsymbol{X}_i)\}^2}{\frac{1}{n_1}\sum_{i=1}^{n_1}\{g_0(\boldsymbol{X}_i)\}^2} \right]^{1/2},$$

where $\hat{g}$ and g_0 were evaluated on the covariates of the testing set $\{\boldsymbol{X}_i : i = 1, \ldots, n_1\}$ and $\bar{\hat{g}} = \sum_{i=1}^{n_1} \hat{g}(\boldsymbol{X}_i)/n_1$ with $n_1 = 200$. Note that since the maximizer of the log likelihood function is only unique up to a constant, we subtracted the mean of $\hat{g}$ on the testing set. Table 3 lists the RE given by the proposed DNN-based method, CPH method and PLACM method as well as the standard deviation (SD) of RE. They stated that, as expected, the proposed DNN-based method yielded better performance in terms of estimation. As the sample size n increases, the relative error of the DNN-based estimate generally decreases, a phenomenon that is theoretically guaranteed by Theorem 3.1.

Moreover, Figure 2 intuitively predicts the errors on the testing set between the function $\hat{g}$ estimated by the above three methods and the true function g_0, respectively, in four cases. The errors produced by the proposed DNN-based method for Case 1 are generally small across the whole testing set, with only a few outliers. Likewise, for more complex cases 2–4, the errors produced by the DPLCM method remain small except for a few outliers. However, the CPH method in case 2–4 and the PLACM method in case 3–4 produce larger errors.

TABLE 3
The relative error (RE) and standard deviation (SD) of $\hat{g}$ by the DNN-based, CPH, and PLACM methods on the testing data.

		DNN-based		CPH		PLACM	
	n	RE	SD	RE	SD	RE	SD
Case 1	1000	0.227	0.045	0.128	0.041	0.401	0.062
	2000	0.190	0.035	0.094	0.029	0.266	0.037
Case 2	1000	0.213	0.033	0.212	0.029	0.343	0.070
	2000	0.177	0.024	0.205	0.026	0.217	0.037
Case 3	1000	0.298	0.051	0.375	0.023	0.526	0.199
	2000	0.234	0.039	0.374	0.020	0.387	0.161
Case 4	1000	0.402	0.053	0.634	0.032	0.530	0.168
	2000	0.317	0.046	0.628	0.022	0.426	0.129

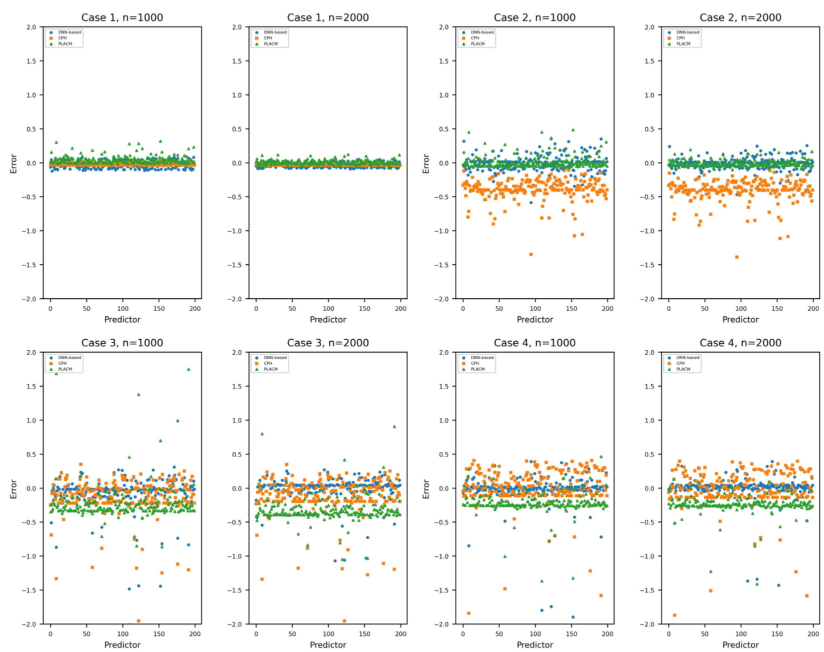

FIG 2. *Prediction errors between $\hat{g}$ estimated by the DDN-based (circle), CPH (square), and PLACM (triangle) methods and the true function g_0 on the testing data.*

It is worth noting that in real-world scenarios, CPH and PLACM methods often face limitations due to their specific model assumptions. Therefore, the DNN-based approach provides a more flexible alternative.

Our method has great flexibility, and when we do not need to interpret the effect sizes of certain covariates, we can include all covariates in a nonparametric form. A neural network on interval-censored data (DNN-IC) was developed by Sun and Ding [28] to handle the nonlinear effects of high-dimensional covariates within the Cox model. Their approach was designed to estimate both the covariate-dependent function and the infinite-dimensional baseline hazard function simultaneously. To tackle the problem of potential overfitting in neural networks, the authors employed an L_1 norm penalty on the parameters of the neural network. This penalty term was incorporated into the loss function and served to regulate the complexity of the model. We adopt the simulation setup from Sun and Ding [28] to compare with the DNN-IC method. Assume that given the covariate vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$, the cumulative hazard function of T has the form

$$\Lambda(t|\boldsymbol{X}) = \Lambda_0(t)\mathrm{e}^{g_0(\boldsymbol{X})}. \tag{4}$$

We consider the following scenarios:

Scenario 1: $g_0(\boldsymbol{X}) = \sum_{j=1}^{p} \beta_j X_j$,

Scenario 2: $g_0(\boldsymbol{X}) = \sum_{j=1}^{p} \beta_j X_j + X_1^2 + X_2^2$,

Scenario 3: $g_0(\boldsymbol{X}) = \sum_{j=1}^{p} \beta_j X_j + X_3X_4$,

Scenario 4: $g_0(\boldsymbol{X}) = \sum_{j=1}^{p} \beta_j X_j + X_1^2 + X_2^2 + X_3X_4$,

Scenario 5: $g_0(\boldsymbol{X}) = \sum_{j=1}^{p} \beta_j X_j + I(\{X_1 < -0.5\} \cup \{X_2 < -0.5\}) - I(\{X_1 \geq -0.5\} \cap \{X_2 \geq -0.5\} + X_3X_4)$.

We consider the Weibull cumulative hazard function $\Lambda_0(t) = (\lambda t)^k$ with $\lambda = 0.01$ and $k = 10$. We will generate the vector $\mathbf{X}$ from a multivariate normal distribution $MVN(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a covariance matrix defined as $\sigma_{jj'} = \exp(-|j - j'|),\ 1 \leq j,\ j' \leq p$.

Next, we will transform the components of $\mathbf{X}$ according to the following rules:

- The first 20% of X_j will remain continuous.
- The second 20% of X_j will be transformed into binary predictors using the indicator function $I(X_j > 0)$.
- The remaining 60% of X_j will be transformed into multinomial predictors using the indicator function $I(X_j > -0.5) + I(X_j > 0.5)$.

For the continuous and binary predictors, we set $\beta_j = 0.2$. However, for the multinomial predictors, we generate β_j from a multivariate normal distribution $MVN(\mu, 0.01 \times \mathbf{\Sigma}')$. Here, μ is a $[0.6p]$-dimensional vector with all elements equal to 0.2, and $\mathbf{\Sigma}'$ is a covariance matrix defined as $\sigma_{jj'} = \exp(-|j - j'|)$, $1 \leq j, j' \leq [0.6p]$.

To complete the data generation process, we set the end time $\tau' = 100$ and the sample size $n = 1000$. We aim to achieve the desired right censoring rate of 50%, and we generate two observation times under five scenarios using the following approach:

(i) Scenario 1: $U \sim \min\{\frac{\tau'}{5}\text{Exponetial}(1), \frac{5}{6}\tau'\}$ and $V \sim \min\{\frac{\tau'}{3} + U + \frac{\tau'}{4}\text{Exponetial}(1), \tau'\}$;

(ii) Scenario 2: $U \sim \min\{\frac{\tau'}{5}\text{Exponetial}(1), \frac{5}{6}\tau'\}$ and $V \sim \min\{\frac{\tau'}{5} + U + \frac{\tau'}{4}\text{Exponetial}(1), \tau'\}$;

(iii) Scenario 3: $U \sim \min\{\frac{\tau'}{6}\text{Exponetial}(1), \frac{5}{6}\tau'\}$ and $V \sim \min\{\frac{\tau'}{3} + U + \frac{\tau'}{4}\text{Exponetial}(1), \tau'\}$;

(iv) Scenario 4: $U \sim \min\{\frac{\tau'}{6}\text{Exponetial}(1), \frac{5}{6}\tau'\}$ and $V \sim \min\{\frac{\tau'}{3} + U + \frac{\tau'}{4}\text{Exponetial}(1), \tau'\}$;

TABLE 4
The mean squared prediction error (MSPE) averages, standard deviations (SD), and running times (RT) from 200 replications for the DNN-based method (our method) and DNN-IC method under five scenarios. The number of predictors is set as $p = 20, 50$.

		DNN-based Method			DNN-IC Method		
	p	MSPE	SD	RT (second)	MSPE	SD	RT (second)
Scenario 1	20	0.022	0.011	4340.301	0.039	0.006	51801.682
	50	0.029	0.017	3791.598	0.041	0.013	58893.634
Scenario 2	20	0.038	0.009	4640.448	0.051	0.006	56501.376
	50	0.044	0.013	3573.554	0.045	0.010	57141.779
Scenario 3	20	0.030	0.010	4324.937	0.046	0.010	48229.062
	50	0.035	0.016	3816.450	0.043	0.012	58575.708
Scenario 4	20	0.039	0.009	4538.649	0.055	0.007	56239.042
	50	0.047	0.012	3784.508	0.046	0.010	57241.471
Scenario 5	20	0.035	0.010	4816.534	0.049	0.009	55579.757
	50	0.039	0.016	3911.719	0.047	0.011	57410.442

(v) Scenario 5: $U \sim \min\{\frac{\tau'}{4}\text{Exponetial}(1), \frac{5}{6}\tau'\}$ and $V \sim \min\{\frac{\tau'}{3} + U + \frac{\tau'}{5}\text{Exponetial}(1), \tau'\}$.

The model will be trained on the training dataset and then evaluated on the testing dataset. To obtain reliable results, we will repeat this process 200 times. In simulation studies where the true survival function and event time T are known, the mean square prediction error (MSPE) is employed as a performance metric. The MSPE quantifies the average integrated L_2 distance between the true survival function and the estimated survival function, denoted as

$$L(\hat{S}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\int_0^{T_i}\{S_0(t|\boldsymbol{X}_i) - \hat{S}(t|\boldsymbol{X}_i)\}^2 dt,$$

where $S_0(t|\boldsymbol{X}_i) = \exp\{-\Lambda_0(t)e^{g_0(\boldsymbol{X}_i)}\}$ and $\hat{S}(t|\boldsymbol{X}_i) = \exp\{-\hat{\Lambda}_n(t)e^{\hat{g}_n(\boldsymbol{X}_i)}\}$. Obviously, smaller values of these indicators suggest better prediction performance. Additionally, we compared the computational efficiency of the two methods across various configurations and recorded their respective running times.

In the study conducted by Sun and Ding [28], they chose the following hyperparameters for five scenarios: two hidden layers, 50 nodes per hidden layer, activation function SeLU (scaled exponential linear unit), L_1 penalty 0.5, batch size 50, epoch size 1000, learning rate 0.01, uniformly distributed initial values, the degree of Bernstein polynomials $m_n = 3$. For the same dataset, our hyperparameters are selected as follows: three hidden layers, 50 ($p = 20$) and 75 ($p = 50$) nodes per hidden layer, activation function ReLU, epoch size 1000, learning rate 0.001, uniformly distributed initial values, the order of splines $m = 4$, the cardinality of the interior knot set $K_n = 8$. The specific comparison results are presented in Table 4. It is not difficult to see that in all cases, the prediction accuracy of our method is slightly higher than that of the DNN-IC method, and moreover, our method is obviously faster in computation.

Table 5
Summary of the subjects in the full dataset, dataset 1 and dataset 2.

	Complete ($n = 12204$)		Dataset 1 ($n = 11000$)		Dataset 2 ($n = 1204$)	
Censoring type	Number	Rate	Number	Rate	Number	Rate
Left-censored	210	0.017	189	0.017	21	0.018
Interval-censored	3294	0.270	2971	0.270	323	0.268
Right-censored	8700	0.713	7840	0.713	860	0.714

Table 6
Analysis results for the Atherosclerosis Risk in Communities study.

	DDN-based			CPH			PLACM		
Covariates	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
Body mass index	0.272	0.016	0.000	0.264	0.015	0.000	0.260	0.016	0.000
Glucose level	0.574	0.019	0.000	0.568	0.018	0.000	0.571	0.019	0.000
High-density Lipoprotein cholesterol	−0.297	0.021	0.000	−0.288	0.020	0.000	−0.296	0.021	0.000
Total cholesterol	0.042	0.018	0.020	0.046	0.017	0.007	0.045	0.018	0.012
Age	0.050	0.019	0.008	0.051	0.016	0.001	0.054	0.019	0.004

5. Application

Now we apply the methodology proposed in the previous sections to the epidemiological follow-up study, the Atherosclerosis Risk in Communities study. The data set is available at BioLINCC https://biolincc.nhlbi.nih.gov/studies/aric. It consists of the participants with ages between 45 to 64 at the beginning and recruited from four locations in the US, Forsyth County of North Carolina, Jackson of Mississippi, Minneapolis suburbs of Minnesota, and Washington County of Maryland. As mentioned before, the participants are examined only at discrete time times and the examination or visiting times differ from subject to subject. In consequence, only interval-censored data are available for the occurrence time of any event such as the onset of diabetes, the failure event of interest here. At each examination, the medical, social and demographic data are collected and in the following, we are interested in estimating the effects of various risk factors on the onset of diabetes.

For the analysis, we focus on the 12204 participants after excluding the subjects with prevalent diabetes, unknown status at baseline or missing values for risk factors or covariates. Among them, about 27% gives interval-censored observations and 71% right-censored observations. We randomly divide the full dataset into two parts: dataset 1 and dataset 2, where dataset 1 contains 90% of the full dataset and dataset 2 contains the remaining 10%. The characteristics of the subjects in the full dataset, dataset 1 and dataset 2 are summarized in Table 5.

For each subject, the data consist of the following risk factors, body mass index, glucose level, high-density lipoprotein cholesterol level, total cholesterol level, age, systolic blood pressure, diastolic blood pressure, race, gender, and location. Next, we consider fitting the dataset with model (1). We are mainly interested in estimating the effects of the first five risk factors (body mass index, glucose level, high-density lipoprotein cholesterol level, total cholesterol level, and age) on the risk of diabetes, we set $\boldsymbol{Z}$ to represent these five risk factors and $\boldsymbol{X}$ the remaining risk factors. We adopt a 5-fold cross-validation procedure.

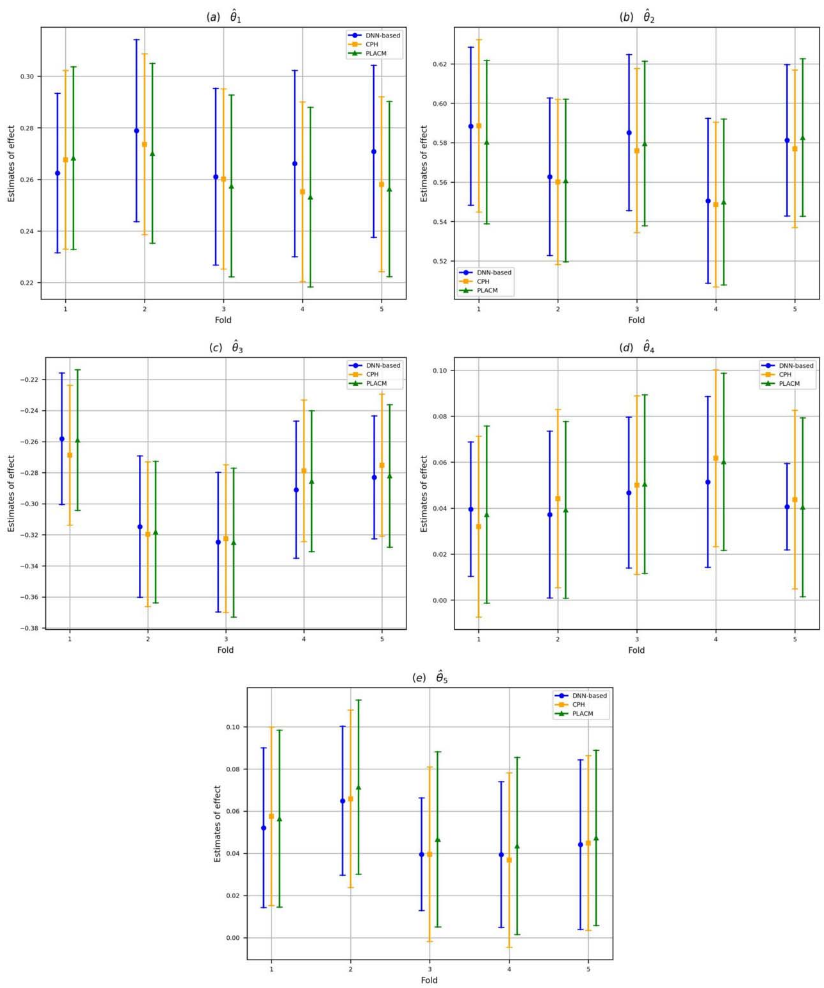

FIG 3. *Estimated effects $\hat{\theta}_1$ (body mass index), $\hat{\theta}_2$ (glucose level), $\hat{\theta}_3$ (high-density lipoprotein cholesterol), $\hat{\theta}_4$ (total cholesterol) and $\hat{\theta}_5$ (age) along with the 95% confidence intervals for five folds.*

In detail, we randomly split the dataset 1 into five folds and use four folds as the training set, and the remaining one fold as the validation set to select hyperparameters, conducting a total of five operations.

Table 6 presents the estimated effects of the five risk factors (body mass index, glucose level, high-density lipoprotein cholesterol level, total cholesterol level, and age), by the proposed DNN-based estimation procedure, the CPH method, and the PLACM method along with the estimated standard errors and p-values. They suggest that all of the five risk factors have significant effects on the risk of diabetes. Figure 3 presents the estimated values of γ_1, γ_2, γ_3,

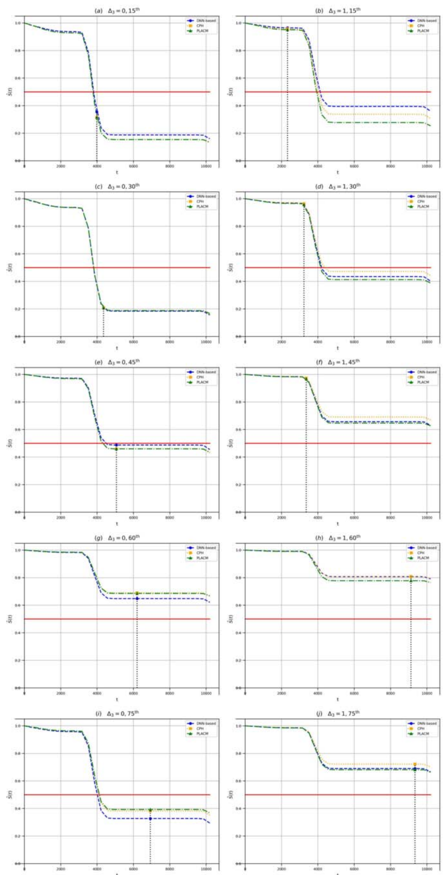

Fig 4. *Predictions of survival functions by the DDN-based (dashed line), CPH (dotted line), and PLACM (dash-dotted line) methods. The solid horizontal line in all plots represents a survival probability of 0.5. The five plots from top to bottom on the left (or right) side represent the survival curves of individuals, where the event occurred (not occurred) in the population of the test set with observation times at the* 15^{th}, 30^{th}, 45^{th}, 60^{th} *and* 75^{th} *quantiles, respectively.*

γ_4, and γ_5, which represent the effects of body mass index, blood glucose level, high-density lipoprotein cholesterol level, total cholesterol level, and age, respectively. The figure also shows the corresponding 95% confidence intervals across all five folds. Note that all confidence intervals obtained from the DNN-based method do not cover zero, while those obtained from the CPH and the PLACM methods include zero in a few cases. We also evaluate the performance of the above three methods in prediction. We use dataset 2 as the test set and classify the one into two categories: $\Delta_3 = 0$ (with 344 observations) and $\Delta_3 = 1$ (with 860 observations), which indicates that the event of interest occurred and did not occur, respectively. In each category, we consider five subjects with the second observation time at the 15^{th}, 30^{th}, 45^{th}, 60^{th}, and 75^{th} quantiles, respectively. Figure 4 depicts the predictions of survival functions of the ten representative subjects by the DNN-based, the CPH, and the PLACM methods. Note that if an event occurs, its survival probability is lower than 0.5, and vice versa. In Figure 4, the solid horizontal line in all plots represents a survival probability of 0.5. In the case of $\Delta_3 = 1$, the predicted values of the survival functions obtained by the three methods are above 0.5 at all five observation moments. In the case of $\Delta_3 = 0$, the predicted values for all five observation times are below 0.5 except for the 60^{th} observation time which is slightly above 0.5.

Overall, the results indicate that body mass index, glucose level, total cholesterol level, and age all exhibit positive effects on the risk of diabetes. Conversely, high-density lipoprotein cholesterol level has a negative effect on the risk of diabetes. Furthermore, we can conclude that the DNN-based method is more reliable and stable than the CPH and the PLACM methods when analyzing real data and estimating treatment effects.

6. Concluding remarks

In this article, we proposed a DNN-based sieve semiparametric maximum likelihood method for the partially linear Cox model with Case II interval-censored data. The method not only inherits the simple interpretation of the finite-dimensional parameters but also provides a powerful tool to remedy the curse of dimensionality with many covariates and capture complicated nonlinear effects. The estimators of the resulting regression parameters were shown to be asymptotically normal and efficient and the estimator of the unknown nonlinear covariate function g_0 to have the optimal convergence rate. The proposed method combines the statistical method with a deep learning approach and facilitates a practical and easy-to-implement inference for the partially linear Cox model based on interval-censored data.

As mentioned above, the proposed method can be seen as a generalization of that given in Zhong et al. [40] although both the situation and the employed objective function considered here are much more complicated than these in Zhong et al. [40]. As a consequence, we have to deal with much more difficult computation and theoretical issues. In addition, the proposed approach can also be regarded as a generalization of that proposed by Sun and Ding [28] in that we also considered the existence of linear covariate effects in addition to nonlinear covariate effects. The proposed method can be used not only for prediction but also for treatment comparison. Furthermore, we investigated the asymptotic properties of the proposed method and showed that the proposed nonparametric DNN estimator achieves the minimax optimal rate of convergence.

In the above, we have assumed that there exist two types of covariates Z and X that have linear and nonlinear effects, respectively, and there is no interaction between them. In practice, of course, this may not be known and in general, one usually chooses them based on the question of interest. For example, if the prediction is of main interest, one may include all covariates into X

or Z into the deep neural network. If the treatment comparison is the major goal of the study, one may set Z to be treatment indicator as discussed above and let X include all other covariates. For the situation where there may exist some interaction between Z and X, we can consider interaction terms into the unknown nonparametric effect form for a more comprehensive analysis.

There exist several directions for future research. One is that in the above, the failure time of interest has been assumed to be independent of the observation process and it is apparent that this may be not true in some situations. Therefore it would be useful to generalize the proposed method to the case of informatively interval-censored data under model (1). Another related direction for future research is to consider more generalized models or other models rather than model (1), which may not be proper sometimes. One such choice could be the partially linear transformation mode $\Lambda(t) = G(\Lambda_0(t)\exp(\boldsymbol{\theta}_0'\boldsymbol{Z} + g_0(\boldsymbol{X})))$, where G is a pre-specified link function. A third direction is to consider the situation where $\boldsymbol{Z}$ is a high-dimensional covariate vector and one is interested in identifying a small number of key or significant risk factors. For this, some penalized methods may be developed.

Appendix A: Additional simulations

We consider a model:

$$\Lambda(t \mid \boldsymbol{X}) = \Lambda_0(t)\mathrm{e}^{g_0(\boldsymbol{X})}.$$

Here we set $\Lambda_0(t) = \frac{\sqrt{t}}{7}$ and $g_0(\boldsymbol{X}) = \frac{1}{2}\log(X_1X_2+1) + \frac{1}{3}(X_3+X_4)^2 + \frac{\sqrt{X_5}}{4} + \frac{1}{5}\mathrm{e}^{X_6/2}$. The two observation times were generated as follows: $U \sim \mathrm{Uniform}(0, \frac{\tau}{6})$, $V \sim \min\{\frac{\tau}{3} + U + \frac{\tau}{2}\,\mathrm{Exponetial}(1), \tau\}$. When we continue to use our method, that is, $\Lambda(t \mid \boldsymbol{X}) = \Lambda_0(t)\mathrm{e}^{\theta_0 X_1 + h_0(X_2,X_3,X_4,X_5,X_6)}$, to estimate parameters, it is the misspecified model. The estimation and prediction results by our method are illustrated in Figure 5, which show that when the model structure assumption is incorrect, the estimated results are significantly biased and the predictions are inaccurate.

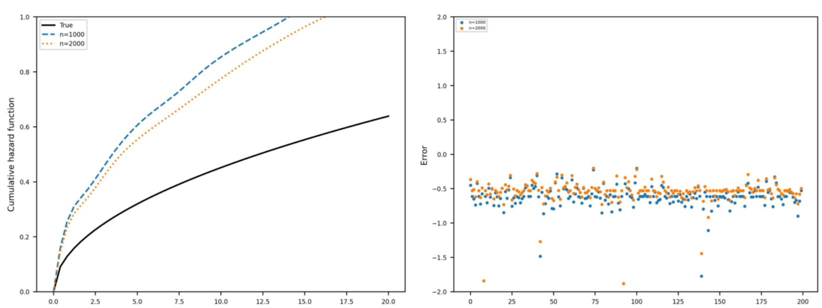

FIG 5. *Estimates of $\Lambda_0(\cdot)$ and prediction errors between $\hat{g}$ and the true function g_0 under two sample sizes 1000 and 2000.*

Supplementary material

(C1) $K = O(\log n)$, $s = O(n\gamma_n^2 \log n)$ and

$$n\gamma_n^2 \lesssim \min_{\{k=1,\dots,K\}} p_k \leq \max_{\{k=1,\dots,K\}} p_k \lesssim n.$$

(C2) The covariate $(\boldsymbol{Z}, \boldsymbol{X})$ takes value in a bounded subset of $\mathbb{R}^{p+r}$ with joint probability density function bounded away from zero. Without loss of generality, we assume that the domain of $\boldsymbol{X}$ is taken to be $[0,1]^r$. And Θ is a compact subset of $\mathbb{R}^p$.

(C3) (a) There exists a positive number η such that $P(V - U \geq \eta) = 1$; and (b) the union of the supports of U and V is contained in an interval $[a, b]$, where $0 < a < b < \infty$, and $0 < \Lambda_0(a) < \Lambda_0(b) < \infty$.

(C4) $\phi_0 = \log \Lambda_0$ belongs to Φ, a class of functions with bounded p th derivative in $[a, b]$ for $p \geq 1$ and the first derivative of ϕ_0 is strictly positive and continuous on $[a, b]$.

(C5) The conditional density $g(u, v \mid \boldsymbol{z}, \boldsymbol{x})$ of (U, V) given $\boldsymbol{Z}$ and $\boldsymbol{X}$ has bounded partial derivatives with respect to (u, v). The bounds of these partial derivatives do not depend on $(u, v, \boldsymbol{z}, \boldsymbol{x})$.

(C6) For some $\kappa \in (0,1)$, $\boldsymbol{a}^{\mathrm{T}} \operatorname{var}(\boldsymbol{Z} \mid U)\boldsymbol{a} \geq \kappa \boldsymbol{a}^{\mathrm{T}} \mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} \mid U\right)\boldsymbol{a}$ and $\boldsymbol{a}^{\mathrm{T}} \operatorname{var}(\boldsymbol{Z} \mid V)a \geq \kappa \boldsymbol{a}^{\mathrm{T}} \mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} \mid V\right)\boldsymbol{a}$ a.s. for all $\boldsymbol{a} \in R^p$.

(C7) The nonparametric function g_0 is an element of $\mathcal{H}_0 = \{g \in \mathcal{H}(q, \boldsymbol{\alpha}, \boldsymbol{d}, \tilde{\boldsymbol{d}}, M)$: $\mathbb{E}\{g(\boldsymbol{X})\} = 0\}$.

Condition (C1) determines the structure of a neural network family $\mathcal{G}(K, s, \boldsymbol{p}, D)$. Conditions (C2)-(C5) are common conditions in the context of survival analysis. Condition (C6) is a technical assumption which is similar to condition (C6) in Zhang and Hua [37]. Condition (C7) ensures the identifiability of the proposed model.

We first introduce some more notation. For any vector $\boldsymbol{v} = (v_1, \dots, v_p)' \in \mathcal{R}^p$, let $\|\boldsymbol{v}\|_c^2$ be defined as above, $\|\boldsymbol{v}\| = \left(\sum_{i=1}^p v_i^2\right)^{1/2}$ and $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$, and for any matrix $W = (w_{ij}) \in \mathbb{R}^{m \times n}$, $\|W\|_\infty = \max_{i,j} |w_{ij}|$. For any function h, $\|h\|_\infty$ and $\|h\|_{L^2}$ are the sup-norm and L^2-norm of h respectively, and for any vector function $h = (h_1, \dots, h_p)'$, $\|h\|_\infty = \max_i \|h_i\|_\infty$. Denote $a_n \lesssim b_n$ as $a_n \leq cb_n$ for some $c > 0$ and any n. And $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Denote $\boldsymbol{\tau} = (\boldsymbol{\theta}, g, \phi)$, the true value of $\boldsymbol{\tau}$, $\boldsymbol{\tau}_0 = (\boldsymbol{\theta}_0, g_0, \phi_0)$, $f(\boldsymbol{Z}, \boldsymbol{X}, t) = \exp\{\boldsymbol{\theta}'\boldsymbol{Z} + g(\boldsymbol{X}) + \phi(t)\}$ and $f_0(\boldsymbol{Z}, \boldsymbol{X}, t) = \exp\{\boldsymbol{\theta}_0'\boldsymbol{Z} + g_0(\boldsymbol{X}) + \phi_0(t)\}$. Define

$$\begin{aligned} m(\boldsymbol{\tau}) = {} & \delta_1 \log\{1 - \exp[-f(\boldsymbol{Z}, \boldsymbol{X}, U)]\} \\ & + \delta_2 \log[\exp(-f(\boldsymbol{Z}, \boldsymbol{X}, U)) - \exp(-f(\boldsymbol{Z}, \boldsymbol{X}, V))] \\ & - \delta_3 f(\boldsymbol{Z}, \boldsymbol{X}, V) \end{aligned}$$

and then $l_n(\boldsymbol{\tau}) = \mathbb{P}_n m(\boldsymbol{\tau})$ and $l(\boldsymbol{\tau}) = \mathbb{P} m(\boldsymbol{\tau})$.

Proof of Theorem 3.1. Before deriving the convergence rate, we need to show that $d(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}_0) \xrightarrow{p} 0$ as $n \to \infty$. This can be accomplished by verifying the conditions of Theorem 5.7 in van der Vaart [32].

For some $D > 0$, let $\mathbb{R}_D^p = \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta}\|_\infty < D\}$, $\mathcal{G}_D := \mathcal{G}(K, s, \boldsymbol{p}, D)$, and $\mathcal{M}_D = \left\{\phi_n : \phi_n(t) = \sum_{j=1}^{q_n} \beta_j b_j(t), \beta_1 \le \beta_2 \le \cdots \le \beta_{q_n} \le D, t \in [a,b]\right\}$. Define

$$\hat{\boldsymbol{\tau}}_D = \left(\hat{\boldsymbol{\theta}}_D, \hat{g}_D, \hat{\phi}_D\right) = \underset{(\boldsymbol{\theta},g,\phi)\in\mathbb{R}_D^p\times\mathcal{G}_D\times\mathcal{M}_D}{\arg\max} l_n(\boldsymbol{\theta}, g, \phi). \tag{5}$$

Note that $\mathbb{P}\left(d\left(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}_0\right) < \infty\right) = 1$. Thus, it suffices to show that $d\left(\hat{\boldsymbol{\tau}}_D, \boldsymbol{\tau}_0\right) \xrightarrow{p} 0$ as $n \to \infty$ for some large enough D.

According to the bracketing number calculations developed by Shen and Wong [23] and Lu et al. [14], for any $\eta > 0$ and $0 < \varepsilon < \eta$, the logarithm of the bracketing number of $\mathcal{M}_D$, computed with $L_2(P)$, is bounded by $Mq_n \log(\eta/\varepsilon)$. Based on the Lemma 6 in Zhong et al. [40], the logarithm of the bracketing number of $\mathcal{G}_D$ with $L_2(P)$ is bounded by $s\log(U/\varepsilon)$, where $U = K\prod_{k=0}^{K}(p_k + 1)\sum_{k=0}^{K} p_k p_{k+1}$. It is known that the neighborhood $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \eta\}$ can be covered by $M(\eta/\varepsilon)^p$ balls with radius ε. In view of Theorem 9.23 of Kosorok [13], the bracketing number of $\left\{\boldsymbol{\theta}' Z : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \eta\right\}$ is bounded by $M(\eta/\varepsilon)^p$. It is shown that the class $\{l(\boldsymbol{\tau}) : \boldsymbol{\tau} \in \mathbb{R}_D^p \times \mathcal{G}_D \times \mathcal{M}_D\}$ is Glivenko-Cantelli. Therefore,

$$\sup_{\boldsymbol{\tau}\in\mathbb{R}_D^p\times\mathcal{G}_D\times\mathcal{M}_D} |l_n(\boldsymbol{\tau}) - l(\boldsymbol{\tau})| \xrightarrow{p} 0. \tag{6}$$

Some algebra yields that

$$\begin{aligned}
l\left(\boldsymbol{\tau}_0\right) - l(\boldsymbol{\tau}) =& \mathbb{E}\Bigg([1 - \exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, U)\}] \log \frac{1 - \exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, U)\}}{1 - \exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\}} \\
&+ [\exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, U)\} - \exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, V)\}] \\
&\times \log \frac{\exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, U)\} - \exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, V)\}}{\exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\} - \exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\}} \\
&+ \exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, V)\} \log \frac{\exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, V)\}}{\exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\}}\Bigg) \\
=& \mathbb{E}\Bigg([1 - \exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\}] h\left[\frac{1 - \exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, U)\}}{1 - \exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\}}\right] \\
&+ [\exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\} - \exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\}] \\
&\times h\left[\frac{\exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, U)\} - \exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, V)\}}{\exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\} - \exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\}}\right] \\
&+ \exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\} h\left[\frac{\exp\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, V)\}}{\exp\{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\}}\right]\Bigg),
\end{aligned}$$

where $h(x) = x\log x - x + 1 \ge (x-1)^2/4$ for $0 \le x \le 5$. Further analysis by using Taylor expansion and conditions (C2)-(C4) leads to

$$
\begin{aligned}
& l\left(\boldsymbol{\tau}_0\right)-l(\boldsymbol{\tau}) \\
\geq & C \mathbb{E}\left(\frac{1}{1-\exp \{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\}}\left[\exp \left\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, U)\right\}-\exp \{-f(\boldsymbol{Z}, \boldsymbol{X}, U)\}\right]^2\right. \\
& \left.+\frac{1}{\exp \{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\}}\left[\exp \left\{-f_0(\boldsymbol{Z}, \boldsymbol{X}, V)\right\}-\exp \{-f(\boldsymbol{Z}, \boldsymbol{X}, V)\}\right]^2\right) \\
\geq & C \mathbb{E}\left[\left\{\left(\boldsymbol{\theta}_0-\boldsymbol{\theta}\right)^{\prime} Z+\left(g_0-g\right)(\boldsymbol{X})+\left(\phi_0-\phi\right)(U)\right\}^2\right. \\
& \left.+\left\{\left(\boldsymbol{\theta}_0-\boldsymbol{\theta}\right)^{\prime} \boldsymbol{Z}+\left(g_0-g\right)(\boldsymbol{X})+\left(\phi_0-\phi\right)(V)\right\}^2\right]
\end{aligned}
$$

Let $w_1(\boldsymbol{Z})=\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)^{\prime} \boldsymbol{Z}$ and $w_2(U, V, \boldsymbol{X})=\left(\phi-\phi_0\right)(U)+\left(\phi-\phi_0\right)(V)+(g-g_0)(\boldsymbol{X})$. The Cauchy-Schwarz inequality and the law of total expectation yield

$$
\begin{aligned}
& \left[\mathbb{E}\left\{w_1(\boldsymbol{Z}) w_2(U, V, \boldsymbol{X})\right\}\right]^2 \\
\leq & \mathbb{E}_{U, V, \boldsymbol{X}}\left\{w_2^2(U, V, \boldsymbol{X})\right\} \mathbb{E}_{U, V, \boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{Z} \mid U, V, \boldsymbol{X}}\left\{w_1(\boldsymbol{Z}) \mid U, V, \boldsymbol{X}\right\}\right]^2 .
\end{aligned}
$$

Using the similar arguments as those in Wellner and Zhang [34] and Lemma 25.86 of van der Vaart [32] yields

$$
l\left(\boldsymbol{\tau}_0\right)-l(\boldsymbol{\tau}) \geq C\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\right\|^2+\left\|g-g_0\right\|_{L^2\left([0,1]^r\right)}+\left\|\phi-\phi_0\right\|_{\Phi}^2\right)=C d^2\left(\boldsymbol{\tau}_0, \boldsymbol{\tau}\right) .
$$

Then, it implies that

$$
\sup _{\boldsymbol{\tau}: d\left(\boldsymbol{\tau}, \boldsymbol{\tau}_0\right) \geq \epsilon} l(\boldsymbol{\tau}) \leq l\left(\boldsymbol{\tau}_0\right)-C \epsilon^2<l\left(\boldsymbol{\tau}_0\right) . \tag{7}
$$

.

For $\phi_0 \in \Phi$, Lu [14] has shown that there exists a $\phi_{0,n} \in \mathcal{M}_n$ of order $m \geq p+2$ such that

$$
\left\|\phi_{0, n}-\phi_0\right\|_{\infty} \leq C q_n^{-p}=O\left(n^{-p v}\right) .
$$

This also implies that $\left\|\phi_{0, n}-\phi_0\right\|_{\Phi} \leq C q_n^{-p}=O\left(n^{-p v}\right)$. By the proof of Theorem 1 in Schmidt-Hieber [20], we know $\left\|g_{0, n}-g_0\right\|_{L^2}=O(\gamma_n \log^2 n)$. Using the fact that the function $h(x)=x \log x-x+1 \leq(x-1)^2$ in the neighborhood of $x=1$, it can be easily be argued that $l(\boldsymbol{\tau}_0)-l(\boldsymbol{\tau}) \leq C d^2\left(\boldsymbol{\tau}_0, \boldsymbol{\tau}\right)$. Thus, it follows that $l(\tau_0)-l(\tau) \asymp d^2\left(\boldsymbol{\tau}_0, \boldsymbol{\tau}\right)$. Then, by this, (6) and the law of large numbers, we have

$$
\begin{aligned}
& \left|l_n\left(\boldsymbol{\theta}_0, g_{0, n}, \phi_{0, n}\right)-l_n\left(\boldsymbol{\theta}_0, g_0, \phi_0\right)\right| \\
\leq & \left|l_n\left(\boldsymbol{\theta}_0, g_{0, n}, \phi_{0, n}\right)-l\left(\boldsymbol{\theta}_0, g_{0, n}, \phi_{0, n}\right)\right|+\left|l\left(\boldsymbol{\theta}_0, g_{0, n}, \phi_{0, n}\right)-l\left(\boldsymbol{\theta}_0, g_{0, n}, \phi_0\right)\right| \\
& +\left|l\left(\boldsymbol{\theta}_0, g_{0, n}, \phi_0\right)-l\left(\boldsymbol{\theta}_0, g_0, \phi_0\right)\right|+\left|l\left(\boldsymbol{\theta}_0, g_0, \phi_0\right)-l_n\left(\boldsymbol{\theta}_0, g_0, \phi_0\right)\right| \\
= & o_p(1) .
\end{aligned}
$$

Since $\hat{\boldsymbol{\tau}}_D$ is the maximizer of (5), we obtain

$$
l_n\left(\hat{\boldsymbol{\theta}}_D, \hat{g}_D, \hat{\phi}_D\right) \geq l_n\left(\boldsymbol{\theta}_0, g_{0, n}, \phi_{0, n}\right)=l_n\left(\boldsymbol{\theta}_0, g_0, \phi_0\right)-o_p(1),
$$

which gives

$$l_n\left(\hat{\boldsymbol{\tau}}_D\right) \geq l_n\left(\boldsymbol{\tau}_0\right) - o_p(1). \tag{8}$$

Therefore, the conditions in Theorem 5.7 of van der Vaart [32] follows from (6), (7) and (8), and this implies that $d\left(\hat{\boldsymbol{\tau}}_D, \boldsymbol{\tau}_0\right) \to 0$ as $n \to \infty$.

Next, we show the convergence rates $d\left(\hat{\boldsymbol{\tau}}_D, \boldsymbol{\tau}_0\right) = O_p\left(\gamma_n \log^2 n\right)$. We need to verify the conditions of Theorem 3.4.1 of van der Vaart and Wellner [33]. Let $\mathcal{A}_\eta = \{\boldsymbol{\tau} = (\boldsymbol{\theta}, g, \phi) \in \mathbb{R}_D^p \times \mathcal{G}_D \times \mathcal{M}_D : \eta/2 \leq d\left(\boldsymbol{\tau}, \boldsymbol{\tau}_0\right) \leq \eta\}$. We first need to show that

$$\mathbb{E}^* \sup_{\boldsymbol{\tau} \in \mathcal{A}_\eta} \sqrt{n} \left|\left(l_n - l\right)(\boldsymbol{\tau}) - \left(l_n - l\right)\left(\boldsymbol{\tau}_0\right)\right| \lesssim \varphi_n(\eta).$$

Let $\mathcal{L}_1(\eta) = \{m(\boldsymbol{\tau}) - m\left(\boldsymbol{\tau}_0\right) : \boldsymbol{\tau} \in \mathcal{A}_\eta\}$. Note that, for any $\boldsymbol{\tau}, \boldsymbol{\tau}_1 \in \mathcal{A}_\eta$,

$$\mathbb{E}\left[m(\boldsymbol{\tau}) - m(\boldsymbol{\tau}_1)\right]^2 \lesssim d^2\left(\boldsymbol{\tau}, \boldsymbol{\tau}_1\right).$$

Then, by Lemma 6 in Zhong et al. [40], it follows, if $p, q_n \leq s$ and $\eta \leq U$,

$$\log \mathcal{N}_{[]}\left(\epsilon, \mathcal{L}_1(\eta), L_2(\mathbb{P})\right) \lesssim p \log \frac{\eta}{\epsilon} + q_n \log \frac{\eta}{\epsilon} + s \log \frac{U}{\epsilon} \lesssim s \log \frac{U}{\epsilon}.$$

This leads to

$$\begin{aligned}
J_{[]}\left(\eta, \mathcal{L}_1(\eta), L_2(\mathbb{P})\right) &:= \int_0^\eta \sqrt{1 + \log \mathcal{N}_{[]}\left(\epsilon, \mathcal{L}_1(\eta), L_2(P)\right)} d\epsilon \\
&\lesssim \int_0^\eta \sqrt{1 + s \log\left(\frac{U}{\epsilon}\right)} d\epsilon \\
&= \sqrt{\frac{s}{2}} U \int_{\sqrt{2 \log \frac{U}{\eta}}}^\infty v^2 e^{-v^2/2} dv \\
&\asymp \eta \sqrt{s \log \frac{U}{\eta}}
\end{aligned}$$

Based on Lemma 3.4.2 in van der Vaart and Wellner [33], we have

$$\begin{aligned}
\mathbb{E}^* \left\|\mathbb{G}_n\right\|_{\mathcal{L}_1(\eta)} &\lesssim J_{[]}\left(\eta, \mathcal{L}_1(\eta), L_2(\mathbb{P})\right) \left\{1 + \frac{J_{[]}\left(\eta, \mathcal{L}_1(\eta), L_2(\mathbb{P})\right)}{\eta^2 \sqrt{n}}\right\} \\
&\lesssim \eta \sqrt{s \log \frac{U}{\eta}} + \frac{s}{\sqrt{n}} \log \frac{U}{\eta}.
\end{aligned}$$

Thus, the key function $\varphi_n(\eta)$ is given by

$$\varphi_n(\eta) = \eta \sqrt{s \log \frac{U}{\eta}} + \frac{s}{\sqrt{n}} \log \frac{U}{\eta}.$$

Furthermore, we have shown that $l(\boldsymbol{\tau}_0) - l(\boldsymbol{\tau}) \asymp d^2\left(\boldsymbol{\tau}_0, \tau\right)$, which leads to

$$\sup_{\boldsymbol{\tau} \in A_\eta} \left[l(\boldsymbol{\tau}) - l\left(\boldsymbol{\tau}_0\right)\right] \lesssim -\eta^2.$$

Denote $\delta_n = \gamma_n \log^2 n$. By assumption (C1), it is clear that

$$\delta_n^{-2} \varphi_n(\delta_n) \leq \sqrt{n}.$$

On the other hand,

$$\begin{aligned}
&|l_n(\boldsymbol{\theta}_0, g_{0,n}, \phi_{0,n}) - l_n(\boldsymbol{\theta}_0, g_0, \phi_0)| \\
&\lesssim O_p\left(n^{-1/2}\varphi_n(\delta_n)\right) + |l(\boldsymbol{\theta}_0, g_{0,n}, \phi_{0,n}) - l(\boldsymbol{\theta}_0, g_0, \phi_0)| \\
&\lesssim O_p\left(n^{-1/2}\varphi_n(\delta_n)\right) + \|g_{0,n} - g_0\|^2_{L^2([0,1]^r)} + \|\phi_{0,n} - \phi_0\|^2_\Phi \\
&= O_p\left(\delta_n^2\right)
\end{aligned}$$

Since, by the definition of $\hat{\boldsymbol{\tau}}_D$ in (5), we have

$$l_n\left(\hat{\boldsymbol{\theta}}_D, \hat{g}_D, \hat{\phi}_D\right) \geq l_n(\boldsymbol{\theta}_0, g_{0,n}, \phi_{0,n}) = l_n(\boldsymbol{\theta}_0, g_0, \phi_0) - O_p\left(\delta_n^2\right).$$

Thus, by Theorem 3.4.1 in van der Vaart and Wellner [33], we have

$$d(\hat{\boldsymbol{\tau}}_D, \boldsymbol{\tau}_0) = O_p(\delta_n).$$

This gives $d(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}_0) = O_p(\delta_n)$. Furthermore, we have $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\| = O_p(\delta_n)$, $\left\|\hat{\phi} - \phi_0\right\|_\Phi = O_p(\delta_n)$, and $\|\hat{g} - g_0\|_{L^2} = O_p(\delta_n)$. ☐

Proof of Theorem 3.2. Denote $P_{(\boldsymbol{\theta}_0, g_0, \phi_0)}$ be the probability distribution with respect to the parameter $\boldsymbol{\theta}_0$, ϕ_0, and nonparametric function g_0. For $(\boldsymbol{\theta}_0, \phi_0) \in \Theta \times \Phi$ and $g^{(0)}, g^{(1)} \in \mathcal{H}_0$, let P_0 and P_1 be the joint probability distribution of the observed data $\{(\delta_{1i}, \delta_{2i}, \delta_{3i}, U_i, V_i, \boldsymbol{Z}, \boldsymbol{X}), i = 1, \ldots, n\}$ under $P_{(\boldsymbol{\theta}_0, g^{(0)}, \phi_0)}$ and $P_{(\boldsymbol{\theta}_0, g^{(1)}, \phi_{0,})}$, respectively.

The Kullback-Leibler distance between P_1 and P_0 is

$$\begin{aligned}
KL(P_1, P_0) &= \mathbb{E}_{P_1} \log \frac{P_1}{P_0} \\
&= n\mathbb{E}_{P_1}\Bigg([1 - \exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, U_i)\}] h\left[\frac{1 - \exp\left\{-f_{(1)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, U_i)\right\}}{1 - \exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, U_i)\}}\right] \\
&\quad + [\exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, U_i)\} - \exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, V_i)\}] \\
&\quad \times h\left[\frac{\exp\left\{-f_{(1)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, U_i)\right\} - \exp\left\{-f_{(1)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, V_i)\right\}}{\exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, U_i)\} - \exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, V_i)\}}\right] \\
&\quad + \exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, V_i)\} h\left[\frac{\exp\left\{-f_{(1)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, V_i)\right\}}{\exp\{-f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, V_i)\}}\right]\Bigg),
\end{aligned}$$

where $f_{(0)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, t) = \exp\{\boldsymbol{\theta}_0' \boldsymbol{Z}_i + g^{(0)}(\boldsymbol{X}_i) + \phi_0(t)\}$ and $f_{(1)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, t) = \exp\{\boldsymbol{\theta}_0' \boldsymbol{Z}_i + g^{(1)}(\boldsymbol{X}_i) + \phi_0(t)\}$. Using the fact that the function $h(x) = x\log x - x + 1 \leq (x-1)^2$ in the neighborhood of $x = 1$ and Taylor expansion, we have that $KL(P_1, P_0) \leq cn \left\|g^{(1)} - g^{(0)}\right\|_{L^2}^2$. Therefore, there exist a constant $c > 0$, such that

$$KL(P_1, P_0) \leq cn \left\|g^{(1)} - g^{(0)}\right\|_{L^2}^2, \tag{9}$$

By the proof of Theorem 3 of Schmidt-Hieber [20], there exist $g^{(0)}, \ldots, g^{(N)} \in \mathcal{H}_0$ and constant $c_1, c_2 > 0$, such that

$$\left\|g^{(j)} - g^{(k)}\right\|_{L^2} \geq 2c_1\gamma_n > 0 \tag{10}$$

and

$$\frac{cn}{N}\sum_{j=1}^{N}\left\|g^{(j)} - g^{(0)}\right\|_{L^2}^2 \leq c_2 \log N. \tag{11}$$

Then with (9)-(11), Theorem 2.5 in Tsybakov [31] implies that

$$\inf_{\hat{g}} \sup_{g_0\in\mathcal{H}_0} \mathbb{P}\left(\|\hat{g} - g_0\|_{L^2} \geq c_1\gamma_n\right) \geq \frac{\sqrt{N}}{1+\sqrt{N}}\left(1 - 2c_2 - \sqrt{\frac{2c_2}{\log N}}\right).$$

This shows that

$$\inf_{\hat{g}} \sup_{(\boldsymbol{\theta}_0, g_0, \phi_0)\in\Theta\times\mathcal{H}_0\times\Phi} \mathbb{E}_{P_{(\boldsymbol{\theta}_0,g_0,\phi_0)}}\left\{\hat{g}(\boldsymbol{X}) - g_0(\boldsymbol{X})\right\}^2 \geq c_3\gamma_n^2,$$

for some constant $0 < c_3 < \infty$. Therefore, the proof is completed. □

Proof of Theorem 3.3. Let

$$\begin{aligned}\ell(\boldsymbol{\tau}) =& \delta_1 \log\{1 - \exp[-f(\boldsymbol{Z}, \boldsymbol{X}, U)]\} \\ &+ \delta_2 \log\{\exp[-f(\boldsymbol{Z}, \boldsymbol{X}, U)] - \exp[-f(\boldsymbol{Z}, \boldsymbol{X}, V)]\} - \delta_3 f(\boldsymbol{Z}, \boldsymbol{X}, V).\end{aligned}$$

be the log-likelihood for a sample of size one. We define functions Q_i, i=1, 2, 3, by

$$Q_1(U, V, \boldsymbol{Z}, \boldsymbol{X}) = \frac{\exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, U)\right) f(\boldsymbol{Z}, \boldsymbol{X}, U)}{1 - \exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, U)\right)},$$

$$Q_2(U, V, \boldsymbol{Z}, \boldsymbol{X}) = \frac{\exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, U)\right) f(\boldsymbol{Z}, \boldsymbol{X}, U)}{\exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, U)\right) - \exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, V)\right)},$$

and

$$Q_3(U, V, \boldsymbol{Z}, \boldsymbol{X}) = \frac{\exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, V)\right) f(\boldsymbol{Z}, \boldsymbol{X}, V)}{\exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, U)\right) - \exp\left(-f(\boldsymbol{Z}, \boldsymbol{X}, V)\right)}.$$

The score function for $\boldsymbol{\theta}$ is

$$\begin{aligned}\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) = \frac{\partial \ell(\boldsymbol{\tau})}{\partial \boldsymbol{\theta}} =& Z\{\delta_1 Q_1 - \delta_2 Q_2 + \delta_2 Q_3 - \delta_3 f(\boldsymbol{Z}, \boldsymbol{X}, V)\} \\ =& ZQ_4(U, V, \boldsymbol{Z}, \boldsymbol{X}) + ZQ_5(U, V, \boldsymbol{Z}, \boldsymbol{X}),\end{aligned}$$

where $Q_4(U, V, \boldsymbol{Z}, \boldsymbol{X}) = \delta_1 Q_1 - \delta_2 Q_2$ and $Q_5(U, V, \boldsymbol{Z}, \boldsymbol{X}) = \delta_2 Q_3 - \delta_3 f(\boldsymbol{Z}, \boldsymbol{X}, V)$.

Consider a parametric smooth submodel (ϕ_t, g_s), such that $\phi_t|_{t=0} = \phi$ and $g_s|_{s=0} = g$, with $\frac{\partial \phi_t}{\partial t}|_{t=0} = a$ and $\frac{\partial g_s}{\partial s}|_{s=0} = h$. The score operators for ϕ and g are defined as

$$\begin{aligned}\dot{\ell}_\phi(\boldsymbol{\tau})[a] = & \left.\frac{\partial \ell\left(\boldsymbol{\theta}, \phi_t, g_s\right)}{\partial t}\right|_{t=s=0} \\ = & \delta_1 a(U) Q_1 - \delta_2[a(U)Q_2 - a(V)Q_3] - \delta_3 f(\boldsymbol{Z}, \boldsymbol{X}, V) a(V) \\ = & a(U) Q_4(U, V, \boldsymbol{Z}, \boldsymbol{X}) + a(V) Q_5(U, V, \boldsymbol{Z}, \boldsymbol{X}).\end{aligned}$$

and

$$\begin{aligned}\dot{\ell}_g(\boldsymbol{\tau})[h] = & \left.\frac{\partial \ell\left(\boldsymbol{\theta}, \phi_t, g_s\right)}{\partial s}\right|_{t=s=0} \\ = & h(\boldsymbol{X})\{\delta_1 Q_1 - \delta_2[Q_2 - Q_3] - \delta_3 f(\boldsymbol{Z}, \boldsymbol{X}, V)\} \\ = & h(\boldsymbol{X}) Q_4(U, V, \boldsymbol{Z}, \boldsymbol{X}) + h(\boldsymbol{X}) Q_5(U, V, \boldsymbol{Z}, \boldsymbol{X}).\end{aligned}$$

Define $\mathcal{A} = \left\{a : Ea^2 < \infty\right\}$ and $\mathcal{H} = \left\{h : E\{h(\boldsymbol{X})\} = 0, E\left\{h^2(\boldsymbol{X})\right\} < \infty\right\}$. For $\mathbf{h} = (h_1, \ldots, h_p)' \in \mathcal{H}^p$, define $\dot{\ell}_g(\boldsymbol{\tau})[\mathbf{h}] = \left(\dot{\ell}_g(\boldsymbol{\tau})\left[h_1\right], \ldots, \dot{\ell}_g(\boldsymbol{\tau})\left[h_p\right]\right)'$, and similarly for $\mathbf{a} = (a_1, \ldots, a_p)' \in \mathcal{A}^p$ define $\dot{\ell}_\phi(\boldsymbol{\tau})[\mathbf{a}] = \left(\dot{\ell}_\phi(\boldsymbol{\tau})\left[a_1\right], \ldots, \dot{\ell}_\phi(\boldsymbol{\tau})\left[a_p\right]\right)'$. According to Bickel et al. [1], the efficient score vector for θ is

$$\ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}) = \dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \dot{\ell}_\phi(\boldsymbol{\tau})[\mathbf{a}^*] - \dot{\ell}_g(\tau)[\mathbf{h}^*],$$

where $a^* \in \mathcal{A}^p$ and $h^* \in \mathcal{H}^p$ satisfies that

$$\mathbb{E}\{(\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \dot{\ell}_\phi(\boldsymbol{\tau})[\mathbf{a}^*] - \dot{\ell}_g(\boldsymbol{\tau})[\mathbf{h}^*])\dot{\ell}_\phi(\boldsymbol{\tau})[a]\} = 0$$

and

$$\mathbb{E}\{(\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \dot{\ell}_\phi(\boldsymbol{\tau})\left[\mathbf{a}^*\right] - \dot{\ell}_g(\boldsymbol{\tau})[\mathbf{h}^*])\dot{\ell}_g(\boldsymbol{\tau})[h]\} = 0$$

for any $a \in \mathcal{A}$ and $h \in \mathcal{H}$. The information bound of $\boldsymbol{\theta}$ takes the form

$$\mathbf{I}(\boldsymbol{\theta}) = E\left\{\ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau})\right\}^{\otimes 2}. \qquad \square$$

Lemma A.1. *Assume that*

(i) $\mathbb{P}_n\{\dot{\ell}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\tau}})\} = o_p(n^{-1/2})$ *and* $\mathbb{P}_n\{\dot{\ell}_\phi(\widehat{\boldsymbol{\tau}})\left[\mathbf{a}^*\right]\} = \mathbb{P}_n\left\{\dot{\ell}_g(\hat{\boldsymbol{\tau}})\left[\mathbf{h}^*\right]\right\} = o_p(n^{-1/2})$,

(ii) $(\mathbb{P}_n - P)\left\{\ell^*_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\tau}}) - \ell^*_{\boldsymbol{\theta}}\left(\boldsymbol{\tau}_0\right)\right\} = o_p\left(n^{-1/2}\right)$,

(iii) $P\left\{\ell^*_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\tau}}) - \ell^*_{\boldsymbol{\theta}}\left(\boldsymbol{\tau}_0\right)\right\} = -\mathbf{I}\left(\boldsymbol{\theta}_0\right)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|\right) + o_p\left(n^{-1/2}\right)$ *and* $\mathbf{I}\left(\boldsymbol{\theta}_0\right)$ *is nonsingular. Then*

$$n^{1/2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = n^{1/2}\mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)\mathbb{P}_n\left\{\ell^*_{\boldsymbol{\theta}}\left(\boldsymbol{\tau}_0\right)\right\} + o_p(1) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)\right), n \to \infty.$$

Proof of Lemma A.1. Combining (ii) and (iii), we have

$$\mathbb{P}_n\{\ell^*_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\tau}}) - \ell^*_{\boldsymbol{\theta}}\left(\boldsymbol{\tau}_0\right)\} = -\mathbf{I}\left(\boldsymbol{\theta}_0\right) + o_p\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|\right) + o_p\left(n^{-1/2}\right).$$

By Condition (i), it follows that

$$\mathbb{P}_n\ell^*_{\boldsymbol{\theta}}\left(\boldsymbol{\tau}_0\right) = \mathbf{I}\left(\boldsymbol{\theta}_0\right)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|\right) + o_p\left(n^{-1/2}\right).$$

Because $\mathbf{I}(\boldsymbol{\theta}_0)$ is nonsingular, and $\mathbb{P}_n \ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0) = O_p\left(n^{-1/2}\right)$ owing to the ordinary large sample theory, one has $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\| = O_p\left(n^{-1/2}\right)$. Thus, $o_p\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|\right) = o_p\left(n^{-1/2}\right)$ and therefore

$$\mathbb{P}_n \ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0) = \mathbf{I}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p\left(n^{-1/2}\right).$$

The result follows. □

Proof of Theorem 3.4. Now we verify the conditions in Lemma A.1 to show the asymptotic normality of $\hat{\boldsymbol{\theta}}$. By the definition of $\hat{\boldsymbol{\tau}}$, $\mathbb{P}_n\left\{\dot{\ell}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\tau}})\right\} = 0$. It can be easily shown that there exists a $a_n \in \mathcal{M}_n$ such that $\|a_n - a^*\|_\Phi = 0(q_n^{-1}) = O(n^{-\nu})$, and $\mathbb{P}_n\left\{\dot{\ell}_\phi(\widehat{\boldsymbol{\tau}})[a_n]\right\} = 0$. Therefore, we can write $\mathbb{P}_n\left\{\dot{\ell}_\phi(\widehat{\boldsymbol{\tau}})[a^*]\right\} = I_{1,n} + I_{2,n}$, where $I_{1,n} = (\mathbb{P}_n - P)\left\{\dot{\ell}_\phi(\widehat{\boldsymbol{\tau}})[a^* - a_n]\right\}$ and $I_{2,n} = P\{\dot{\ell}_\phi(\widehat{\boldsymbol{\tau}})[a^* - a_n] - \dot{\ell}_\phi(\boldsymbol{\tau}_0)[a^* - a_n]\}$. Let

$$\mathcal{L}_2 = \left\{\dot{\ell}_\phi(\boldsymbol{\tau})[a^* - a] : \phi, a \in \mathcal{M}_n, g \in \mathcal{G}_D, d(\boldsymbol{\tau}, \boldsymbol{\tau}_0) < \eta, \|a^* - a\|_\Phi \le \eta\right\}.$$

It can be similarly argued that the ε-bracketing number associated with $L_2(P)$-norm is bounded by $C(\eta/\varepsilon)^p(\eta/\varepsilon)^{Cq_n}(U/\varepsilon)^s$, which leads to $\mathcal{L}_2$ being Donsker. Furthermore, for any $r(\boldsymbol{\tau}) \in \mathcal{L}_2$, $\mathrm{Pr}^2 \to 0$ as $n \to \infty$. Hence $I_{3,n} = o_p\left(n^{-1/2}\right)$ by corollary 2.3.12 of van der Vaart and Wellner [33]. The Cauchy-Schwatz inequality yields $I_{2,n} = Cd(\hat{\boldsymbol{\tau}}, \tau_0)\|a^* - a_n\|_\Phi = o_p(n^{-1/2})$. Thus, we have $\mathbb{P}_n\left\{\dot{\ell}_\phi(\widehat{\boldsymbol{\tau}})[\mathbf{a}^*]\right\} = o_p(n^{-1/2})$. The proof of $\mathbb{P}_n\left\{\dot{\ell}_g(\hat{\boldsymbol{\tau}})[\mathbf{h}^*]\right\} = o_p\left(n^{-1/2}\right)$ is similar. According to Schmidt-Hieber [20], there exist $h_{n,j} \in \mathcal{G}_D$ such that $\|h_j^* - h_{n,j}\|_{L^2} = O(\gamma_n \log^2 n)$. By the definition of $\hat{\boldsymbol{\tau}}$, $\mathbb{P}_n\left\{\dot{\ell}_g(\hat{\boldsymbol{\tau}})[h]\right\} = 0$ for any $h \in \mathcal{G}_D$. Therefore, it suffices to show $\mathbb{P}_n\left\{\dot{\ell}_g(\widehat{\boldsymbol{\tau}})\left[h_j^* - h_{n,j}\right]\right\} = o_p\left(n^{-1/2}\right)$. The term $\mathbb{P}_n\left\{\dot{\ell}_g(\widehat{\boldsymbol{\tau}})\left[h_j^* - h_{n,j}\right]\right\}$ can be written as $I_{3,n} + I_{4,n}$, where $I_{3,n} = (\mathbb{P}_n - P)\left\{\dot{\ell}_g(\boldsymbol{\tau})\left[h_j^* - h_{n,j}\right]\right\}$ and $I_{4,n} = P\left\{\dot{\ell}_g(\boldsymbol{\tau})\left[h_j^* - h_{n,j}\right]\right\}$. The proofs of $I_{3,n} = o_p(n^{-1/2})$ and $I_{4,n} = o_p(n^{-1/2})$ is similar to the proofs of $I_{1,n}$ and $I_{2,n}$.

For $0 < \varepsilon < \eta$, it can be shown that $\mathcal{L}_3 = \{\ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0) : \phi \in \mathcal{M}_n, g \in \mathcal{G}_D, d(\boldsymbol{\tau}, \boldsymbol{\tau}_0) \le \eta\}$ is P-Donsker and for $r(\boldsymbol{\tau}) \in \mathcal{L}_3$, $Pr^2 \to 0$ as $\eta \to 0$. Using Taylor expansion, the convergence rate and assumption $n^{1/2}\gamma_n^2 \to 0$ as $n \to \infty$, $P\{\ell^*_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\tau}}) - \ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0)\} = -\mathbf{I}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p\left(n^{-1/2}\right)$ can be easily established. □

Acknowledgments

The authors would like to thank the Editor, the Associate Editor and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper.

Funding

This research was supported in part by the National Natural Science Foundation of China (12271459, 12101522, 12371622), the CAS AMSS-PolyU Joint Laboratory of Applied Mathematics, and The Hong Kong Polytechnic University (P0038663, P0043955, P0045385).

References

[1] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University. Press, Baltimore, MD. MR1245941

[2] Cheng, G. and Wang, X. (2011). Semiparametric additive transformation model under current status data. *Electronic Journal of Statistics*, **5**:1735–1764. MR2870149

[3] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, **34**(2):187–202. MR0341758

[4] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**(2):269–276. MR0400509

[5] Deng, S., Liu, L. and Zhao, X. (2015). Monotone spline-based least squares estimation for panel count data with informative observation times. *Biometrical Journal*, **57**(5): 743–765. MR3394808

[6] Glorot, X. and Bengio, Y.(2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

[7] Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning Series)*. MIT Press, Cambridge, MA. MR3617773

[8] Han, S., Pool, J., Tran, J. and Dally, W. (2015). Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, 1135–1143.

[9] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, **24**(2):540–568. MR1394975

[10] Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *The Annals of Statistics*, **27**(5):1536–1563. MR1742499

[11] Huang, J. and Rossini, A. J. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, **92**(439):960–967. MR1482126

[12] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, **18**:1–12.

[13] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Science and Business Media. MR2724368

[14] Lu, M. (2007). *Monotone Spline Estimations for Panel Count Data*. PhD Dissertation, Department of Biostatistics, University of Iowa.

[15] Lu, M., Zhang, Y. and Huang, J. (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *Journal of the American Statistical Association*, **104**(487):1060–1070. MR2750237

[16] Ma, S. and Kosorok, M. (2005). Penalized log-likelihood estimation for

partly linear transformation models with current status data. *The Annals of Statistics*, **33**(5):2256–2290. MR2211086

[17] Meixide, C. G., Matabuena, M. and Kosorok, M. R. (2022). Neural interval-censored Cox regression with feature selection. *arXiv preprint arXiv:2206.06885*.

[18] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 807–814.

[19] Saxe, A. M., McClelland, J. L. and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

[20] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, **48**(4):1875–1897. MR4134774

[21] Schumaker, L. (1981). *Spline Function: Basic Theory*. John Wiley, New York. MR0606200

[22] Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*, **85**(1):165–177. MR1627289

[23] Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, **22**(2):580–615. MR1292531

[24] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, **15**(1):1929–1958. MR3231592

[25] Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**(2):689–705. MR0790566

[26] Sun, J. (2006). *Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer. MR2287318

[27] Sun, J. and Chen, D. (2022). *Emerging Topics in Modeling Interval-Censored Survival Data*. Springer Nature.

[28] Sun, T. and Ding, Y. (2023). Neural network on interval-censored data with application to the prediction of Alzheimer's disease. *Biometrics*, **79**(3):2677–2690. MR4644024

[29] Sun, J. and Sun, L. (2005). Semiparametric linear transformation models for current status data. *The Canadian Journal of Statistics*, **33**(1):85–96. MR2155000

[30] Tian, T. and Sun, J. (2023). Variable selection for nonparametric additive Cox model with interval-censored data. *Biometrical Journal*, **65**(1):2100310. MR4534397

[31] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York. MR2724359

[32] van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University. Press, Cambridge. MR1652247

[33] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in

Statistics. Springer, New York. MR1385671

[34] Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics*, **35**(5):2106–2142. MR2363965

[35] Wu, Q., Zhao, H., Zhu, L. and Sun, J. (2020). Variable selection for high-dimensional partly linear additive Cox model with application to Alzheimer's disease. *Statistics in Medicine*, **39**(23):3120–3134. MR4151923

[36] Zeng, D., Mao, L. and Lin, D. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, **103**(2):253–271. MR3509885

[37] Zhang, Y. and Hua, L. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics*, **37**(2):338–354. MR2682304

[38] Zhang, Z., Sun, L., Zhao, X. and Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, **33**(1):61–70. MR2154998

[39] Zhang, Z. and Zhao, Y. (2013). Empirical likelihood for linear transformation models with interval-censored failure time data. *Journal of Multivariate Analysis*, **116**:398–409. MR3049912

[40] Zhong, Q., Mueller, J. and Wang, J. (2022). Deep learning for the partially linear Cox model. *The Annals of Statistics*, **50**(3):1348–1375. MR4441123

[41] Zhou, Q., Zhou, H. and Cai, J. (2017). Case-cohort studies with interval-censored failure time data. *Biometrika*, **104**(1):17–29. MR3626480