The Canadian Journal of Statistics

Vol. xx, No. yy, 2020, Pages 1–43

La revue canadienne de statistique

Continuous threshold models with two-way interactions in survival analysis

Shuo Shuo Liu[1] and Bingshu E. Chen[2]*

[1]*Department of Statistics, The Pennsylvania State University, PA, U.S.A*

[2]*Canadian Cancer Trials Group and Department of Public Health Sciences, Queen's University, Ontario, Canada*

Key words and phrases: Asymptotic statistics; change-point; rectified linear unit; single-index model; subset two-way interaction effect; survival analysis ; threshold model.

MSC 2010: Primary 62N02; secondary 62P10

Abstract: Proportional hazards model with the biomarker-treatment interaction plays an important role in the survival analysis of the subset treatment effect. A threshold parameter for a continuous biomarker variable defines the subset of patients who can benefit or lose from a certain new treatment. In this paper, we focus on a continuous threshold effect using the rectified linear unit (ReLU) and propose a gradient descent method to obtain the maximum likelihood estimation of the regression coefficients and the threshold parameter simultaneously. Under certain regularity conditions, we prove the consistency, asymptotic normality and provide a robust estimate of the covariance matrix when the model is misspecified. To illustrate the finite sample properties of the proposed methods, we simulate data to evaluate the empirical biases, the standard errors, and the coverage probabilities for both the correctly specified models and misspecified models. The proposed continuous threshold model is applied to a prostate cancer data with serum prostatic acid phosphatase as a biomarker.

The Canadian Journal of Statistics xx: 1–43; 2020

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–43; 2020 © 2020 Société statistique du Canada

1. INTRODUCTION

In biomedical research, a threshold parameter (cut-point) is often used to identify a subset of patients who may benefit or lose more from the new treatment than other patients based on certain biomarker measurement. Liu, Jiang & Chen (2015) propose a local partial linear bootstrap test for the nonlinear treatment-biomarker interaction for Cox proportional hazards model. When the null hypothesis of a linear interaction is rejected, one strategy of dealing with the nonlinear interaction is through semi-parametric models with a specific interaction form. Several methods based on biomarker threshold models have been developed. For example, Subtil & Rabilloud (2014) propose a Bayesian method to estimate the optimal threshold of a quantitative biomarker for the diagnostic test. They utilize the Bayesian method to estimate the threshold of a biomarker which is not normally distributed but follows a Student-t distribution and a mixture of Dirichlet processes. Spencer et al. (2016) create a design which can learn from the biomarker and make adjustments on the threshold optimally using the com-

* *Author to whom correspondence may be addressed.*

E-mail: bechen@ctg.queensu.ca

bination of Bayesian prediction and the generalized linear model. This design aims to figure out the subpopulation for which the treatment is effective and estimate the threshold parameter. Biomarker threshold models by Chen, Jiang & Tu (2014) and Fang et al. (2017) are widely used in identifying the optimal cut-off point for a continuous biomarker, which is often dichotomized using an indicator function. Li & Jin (2018) propose a two-stage procedure for simultaneously detecting multiple thresholds in the accelerated failure time (AFT) model by the indicator function. Chen, Li & Chong (2017) propose a method of using an indicator covariate effect for the competing risks model.

Disadvantages of the dichotomization are information loss and a non-differentiable likelihood function. Because the classical regularity conditions do not hold for the discontinuous likelihood of a dichotomized threshold function, it is a challenge to conduct statistical inference on such a model (Pons, 2003). Furthermore, the derivative of an indicator function is either 0 or undefined, so a smooth approximation to the indicator function is often required to maximize the likelihood function (Seo & Linton, 2007; Lin et al. 2011; He, Lin & Tu, 2018). In this paper, we investigate the biomarker threshold models for survival data using the Rectified Linear Unit (ReLU) function instead of the indicator function. A rectifier is defined by the positive part of its argument, which has the advantages of having more biological plausibility and computational efficiency for modelling the effects of biomarker variables than an indicator (Hahnloser et al. 2000). A rectifier can also optimally balance the linear or piece-wise linear models by the cut-off point. One of the benefits of the ReLU function compared to the indicator function is that we can quantitatively find the unit-increase effect of the biomarker above the cut-off point, which could be biologically meaningful in certain situations. The purpose of this paper is not to show the ReLU function is better than the indicator function, but to propose an alternative way to fit the biomarker model in cases when the ReLU function is more appropriate.

In statistics, a rectifier linear unit is also known as a piece-wise linear function or continuous threshold function. It is widely used in the literature, such as the least square regression by Hinkley (1971) and Hansen (2017), and the logistic regression models by Fong et al. (2016). Several researchers have also worked on a smooth version of this function called bent cable regression (Chiu, Lockhart & Routledge, 2006). Using the rectifier biomarker effect results in a continuous and piece-wise differentiable likelihood function, therefore, a gradient descent method can be used directly to obtain the maximum likelihood estimate without any smooth approximation. Fong et al. (2016) propose two continuous threshold logistic regression models with the change-point covariate and study their asymptotic behaviours when the model is misspecified. Since the likelihood function for the logistic model is well defined, asymptotic results from van der Vaart (2000) can be applied to their situation directly. Jensen and Lütkebohmert (2008) propose a Cox model with the ReLU function but with the main effect only. However, in this paper, we propose methods which can deal with the main effects as well as two-way interaction effects.

When the covariate effect does not have a traditional linear combination form, Prentice & Self (1983) prove the asymptotic results for survival data with the Cox partial likelihood function and general relative risk form. However, in their approach, the relative risk function is required to have a smooth second order derivative. In our model, the gradient descent algorithm can be used to obtain the maximum likelihood estimator of the continuous biomarker threshold model. For the gradient descent algorithm, we make use of the results from the profile likelihood method and the *coxph* function in R (R, 2020). Our method estimates the threshold parameter and regression coefficients simultaneously. We further study the asymptotic properties of the estimation of the proposed model under a set of new regularity conditions. We demonstrate the limiting distribution (van der Vaart, 2000) via a local approximation to the partial likelihood function (Luo & Boyett, 1997) and the counting process formulation (Andersen & Gill, 1982). When the model is misspecified, we develop the robust estimate for the threshold parameter and regression coefficients and study their joint asymptotic distribution.

2. MODEL SPECIFICATION

Let T_i and C_i be the failure time and censoring time for a patient i in a study, respectively. Define $\delta_i = I(T_i < C_i)$ as an indicator of the patient's status. Let

$Y_i(t) = I(\tilde{T}_i > t)$ be the at-risk status for subject i, where $\tilde{T}_i = \min(T_i, C_i)$ is the observed failure or censoring time. Let $\boldsymbol{W}_i = (w_{i1}, \ldots, w_{iq})^T$ be a $q \times 1$ biomarker covariate and $\boldsymbol{Z}_{1i} = (z_{1i}, \ldots, z_{pi})^T$ be the $(p \times 1)$ vector for the rest of covariates. Here we assume that the biomarker $\boldsymbol{W}_i$ is a q-dimensional continuous variable with the probability density function $f_{\boldsymbol{W}}(\boldsymbol{w}) > 0$ in the parameter space $(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_q, b_q) \in R^q$. It is not necessary to have a threshold model if w_i has a discrete distribution (e.g. a binary variable). Here $\boldsymbol{Z}_{1i}$ is either a time-dependent covariate (e.g. $\boldsymbol{Z}_{1i}(t)$) or a time-fixed covariate. To simplify the notation, we omit the notation t and simply use $\boldsymbol{Z}_{1i}$. We assume that given covariates $\boldsymbol{W}_i$ and $\boldsymbol{Z}_{1i}$, the censoring time C_i and the failure time T_i are independent. For $i = 1, 2, \ldots, n$, $(T_i, C_i, \boldsymbol{W}_i, \boldsymbol{Z}_{1i})$ are independent, identically distributed random variables.

For the covariate-biomarker interaction, let $\boldsymbol{Z}_{2ij}$, a $(p_j \times 1)$ covariate vector, be the subset of $\boldsymbol{Z}_{1i}$ plus the intercept term that interacts with the biomarker variable w_{ij}, for $j = 1, \ldots, q$. Each $\boldsymbol{Z}_{2ij}$ can be either as simple as just having an intercept term or containing all the covariates in $\boldsymbol{Z}_{1i}$. They are not necessarily the same from one to another. Technically, we can choose the covariates that potentially interact with w_{ij} to be included in $\boldsymbol{Z}_{2ij}$ for each $j = 1, \ldots, q$ based on the study objectives and the prior knowledge about the interactions. Model selection techniques can also be used to identify the important interaction covariates. We denote $p^* = (p_1 + \cdots + p_q)$ as the total number of interaction terms. To model the covariate biomarker interaction using a continuous threshold model, for any threshold parameter $a_j < c_j < b_j$, we define a $q \times 1$ vector

$$\varphi(\boldsymbol{W}_i, \boldsymbol{c}) = (\boldsymbol{W}_i - \boldsymbol{c})^+ = \{(w_{i1} - c_1)^+, \ldots, (w_{iq} - c_q)^+\}^T,$$

where $\boldsymbol{c} = (c_1, \ldots, c_q)$ and $(w_{ij} - c_j)^+ = \max(w_{ij} - c_j, 0)$ is a piece-wise linear function with the slope change from 0 to 1 at point $c_j \in (a_j, b_j)$. This function is also known as the Rectified Linear Unit (ReLU) activation function in the field of artificial neural network and machine learning (Hahnloser, 2000). The covariate-biomarker interaction between $\boldsymbol{Z}_{2ij}$ and w_{ij} is given by $\boldsymbol{Z}_{2ij}(w_{ij} - c_j)^+$ for $j = 1, \ldots, q$. Here the main effect of biomarker $(w_{ij} - c_j)^+$ is expressed as the interaction between the intercept term and the biomarker variable. Let $\boldsymbol{Z}_{\boldsymbol{c}i} = \{\boldsymbol{Z}_{1i}^T, \boldsymbol{Z}_{2i1}^T(w_{i1} - c_1)^+, \ldots, \boldsymbol{Z}_{2iq}^T(w_{iq} - c_q)^+\}^T$ be a $(p + p^*) \times 1$ covariate vector and $\boldsymbol{\beta} = (\boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T)^T$ be the regression coefficient, where $\boldsymbol{\alpha}$ is a $p \times 1$ coefficient for $\boldsymbol{Z}_{1i}$ and where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_q^T)^T$ is a $p^* \times 1$ coefficient for interaction terms $\{\boldsymbol{Z}_{2ij}^T(w_{ij} - c_j)^+\}_{j=1}^q$, respectively.

To simplify the notation, we define the following $p^* \times q$ matrix

$$\boldsymbol{Z}_{2i} = \begin{bmatrix} \boldsymbol{Z}_{2i1} & 0_{p_1\times 1} & \ldots & 0_{p_1\times 1} \\ 0_{p_2\times 1} & \boldsymbol{Z}_{2i2} & \ldots & 0_{p_2\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{p_q\times 1} & 0_{p_q\times 1} & \ldots & \boldsymbol{Z}_{2iq} \end{bmatrix}.$$

Then $\boldsymbol{Z}_{\boldsymbol{c}i}$ can be written as $\{\boldsymbol{Z}_{1i}^T, (\boldsymbol{Z}_{2i}\varphi(\boldsymbol{W}_i, \boldsymbol{c}))^T\}^T$. An alternative way to model the multiple biomarkers $\boldsymbol{W}_i$ is the single-index model (He, Lin & Tu, 2018; Carroll et al. 1997). With a $q \times 1$ parameter $\boldsymbol{\eta}$ and a single cut-point c, we model $\varphi(\boldsymbol{W}_i, c) = (\boldsymbol{\eta}^T\boldsymbol{W}_i - \boldsymbol{c})^+$. For the single-index model to be identifiable, we need a restriction on parameter $\boldsymbol{\eta}$ such as $||\boldsymbol{\eta}||_2 = 1$, or $\eta_1 = 1$. With the interaction between the $p_1 \times 1$ covariate $\boldsymbol{Z}_{2i1}$ and $(\boldsymbol{\eta}^T\boldsymbol{W}_i - \boldsymbol{c})^+$, the single-index model is much simpler (with $p + p_1 + q + 1$ parameters) compared to the multi-threshold models (with $p + p^* + q$ parameters). We show later that the single-index model can be incorporated into our methodological framework.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{c}^T)^T \in R^{(p+p^*+q)}$ (or $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T, \boldsymbol{\eta}^T, c)^T \in R^{(p+p_1+q+1)}$ for the single-index model) denote the parameter vector to be estimated and $\boldsymbol{\theta}_0$ to be the true value of $\boldsymbol{\theta}$. The counting process for survival time T_i is defined as $N_i(t) = \delta_i I(T_i \leq t)$. The intensity function of a counting process is defined as

$$\lambda(t; \boldsymbol{\theta}) = \lim_{\Delta t \to 0} \frac{P(N_i(t + \Delta t) - N_i(t) = 1 \mid \mathcal{H}_{t-})}{\Delta t},$$

where $\mathcal{H}_{t-}$ is the history of the counting process and covariate process up to but not including time t. The cumulative hazard function is defined as $\Lambda(t; \boldsymbol{\theta}) = \int_0^t \lambda(t; \boldsymbol{\theta})$. Under the assumption of the Cox proportional hazards, we have the following threshold model for two-way interaction:

$$\begin{aligned} \lambda(t; \boldsymbol{\theta}) &= \lambda_0(t)e^{\boldsymbol{\beta}^T\boldsymbol{Z}_{\boldsymbol{c}i}} \\ &= \lambda_0(t)e^{\boldsymbol{\alpha}^T\boldsymbol{Z}_{1i} + \boldsymbol{\gamma}^T\boldsymbol{Z}_{2i}\varphi(\boldsymbol{W}_i, \boldsymbol{c})}, \end{aligned} \tag{1}$$

where $\lambda_0(t)$ is the baseline hazards function and $\boldsymbol{c}$ is a vector of unknown threshold parameters. For example, we can consider a simple situation with interaction between treatment z_{1i} and biomarker w_i, while adjusting for the other covariates $(z_{2i}, ..., z_{pi})$. Let z_{1i} be a treatment variable with 1 if patient i is assigned to the treatment group and 0 if to the control group. Therefore, we have $\boldsymbol{Z}_{1i} = (z_{1i}, \ldots, z_{pi})^T$ and $\boldsymbol{Z}_{2i} = (1, z_{1i})^T$ (i.e. $q^* = 2$ and $q = 1$) and the regression parameter $\boldsymbol{\beta} = (\boldsymbol{\alpha},^T \boldsymbol{\gamma}^T)^T = (\alpha_1, \ldots, \alpha_p, \gamma_1, \gamma_2)^T$ is a vector of the treatment effect α_1, the covariate effects $(\alpha_2, \ldots, \alpha_p)$, the biomarker main effect γ_1 and the treatment-biomarker interaction effect γ_2, respectively.

For model (1), let $(\tilde{T}_i, \delta_i, \boldsymbol{W}_i, \boldsymbol{Z}_{1i})$, $i = 1, 2, \ldots, n$, be the independent, identically distributed observed data during the study period $\tilde{T}_i \in (0, \tau]$, where τ is the end of the study time. The set of individuals who are "at risk" for failure at time t is denoted as $R(t) = \{k : \tilde{T}_k \geq t\}$. The partial likelihood function is defined as:

$$\mathcal{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{Z}_{\boldsymbol{c}i})}{\sum_{k \in R(\tilde{T}_i)} \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_{\boldsymbol{c}k})} \right\}^{\delta_i}, \tag{2}$$

and the corresponding log-likelihood is $\ell_n(\boldsymbol{\theta}) = \log(\mathcal{L}_n(\boldsymbol{\theta}))$. The profile likelihood method is often used to obtain the maximum likelihood estimate for a threshold model with indicator function biomarker effect $I(w_{ij} > c_j)$ because the likelihood function is not continuous (Pons, 2003; Chen, Jiang & Tu, 2014). The profile likelihood method may not be feasible when the dimension of the biomarker variable is high. However, the utilization of ReLU function guarantees the continuity of the likelihood function of the proposed model. The property of continuity provides the possibility to find the maximum likelihood estimators of the threshold parameter c and regression coefficient $\boldsymbol{\beta}$ simultaneously.

Here we propose a gradient descent method to estimate $\boldsymbol{\theta}$. We first initialize the values for c and $\boldsymbol{\beta}$, then the estimators are obtained when the gradients converge. To obtain a faster convergence and more accurate estimators, we initialize $\boldsymbol{\theta}$ generated from the profile likelihood method with a refined line search of c. In detail, we conduct a refined line search of the threshold which is within the support of the biomarker in the dataset. For example, when $q = 1$ with a one-dimension biomarker variable, for each given c, we can obtain maximum likelihood estimator $\hat{\boldsymbol{\beta}}_c$ using the *coxph* function in the R survival package. The maximum likelihood estimator of the threshold parameter can be estimated by $\hat{c}$ which maximizes the profile log-likelihood function $\ell_p(c) = \ell_n(\hat{\boldsymbol{\beta}}_c, c)$. The partial derivative of $\ell_n(\boldsymbol{\theta})$ with respect to c is a step function and may not be 0 numerically for a given finite sample size n. We stop the gradient descent algorithm when the gradient is small enough (on the order of $O(n^{-1})$). The variance-covariance matrix for the estimated parameter from the gradient descent methods can be calculated using the asymptotic results from Section 3.

3. ASYMPTOTIC RESULTS

The typical asymptotic distribution for a maximum likelihood estimate based on a score function with Lipschitz continuity can be found in van der Vaart (2000).

It shows that the estimator $\hat{\boldsymbol{\theta}}$ has the following asymptotic normal distribution

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(\mathbf{0}, \boldsymbol{V}^{-1}\boldsymbol{\Gamma}\boldsymbol{V}^{-1}),$$

when the expected score $\dot{\ell}(\boldsymbol{\theta}) = E(\partial\ell/\partial\boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}_0$ and satisfies the Lipschitz condition. Here $\boldsymbol{V}^{-1}$ is the inverse of the expected Hessian matrix at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and $\boldsymbol{\Gamma}$ is a function of the expected first order derivative with $\boldsymbol{\Gamma} = \dot{\ell}(\boldsymbol{\theta}_0)\dot{\ell}(\boldsymbol{\theta}_0)^T$. Fong et al. (2016) use this result to prove the asymptotic distribution of logistic regression models with continuous threshold covariate effects. However, the proof cannot be extended to the Cox proportional hazards model directly because the partial likelihood function from each individual i includes contributions from other subjects who are still "at risk" at time $\tilde{T}_i$.

In this section, we prove the consistency and asymptotic normality of the maximum likelihood estimate for continuous threshold functions in a Cox model. Methods from Andersen (1982) and Pons (2003) are adopted to show the asymptotic consistency and normality of the maximum likelihood estimate. The main proof is based on a Taylor-type expansion of the log likelihood function in Lemma 3 below.

We denote some additional notations and regularity conditions for the proof of the asymptotic results. The martingale in Andersen et al. (2012) is defined as

$$M_i(t) = N_i(t) - \Lambda_i(t) = N_i(t) - \int_0^t Y_i(s)\exp(\boldsymbol{\beta}^T \boldsymbol{Z}_{\boldsymbol{ci}})\lambda_0(s)ds.$$

To obtain the partial derivative of the linear predictor $r_i(\boldsymbol{\theta}) = \boldsymbol{\beta}^T \boldsymbol{Z}_{\boldsymbol{c}i}$ with respect to $\boldsymbol{\theta}$, let $\boldsymbol{Z}_i(\boldsymbol{\theta}) = \left\{\boldsymbol{Z}_{\boldsymbol{c}i}^T, -\boldsymbol{\gamma}_1^T \boldsymbol{Z}_{2i1} I(w_{i1} > c_1), \ldots, -\boldsymbol{\gamma}_q^T \boldsymbol{Z}_{2iq} I(w_{iq} > c_q)\right\}^T$ be a $(p + p^* + q)$ vector with the first $(p + p^*)$ entries as covariate $\boldsymbol{Z}_{\boldsymbol{c}i}$ and the $(p + p^* + 1)$th to the $(p + p^* + q)$th entries as the partial derivative of $r_i(\boldsymbol{\theta})$ with respect to $\boldsymbol{c}$, which are well defined when $c_j \neq w_{ij}$ for $j = 1, \ldots, q$. This is different from the indicator function biomarker model where the partial derivative with respect to c_j is either 0 or undefined, and here we are able to use the gradient descent method without any smoothing approximation. Let $\boldsymbol{I}(\boldsymbol{W}_i > \boldsymbol{c}) = \operatorname{diag}\{I(W_{i1} > c_1), \ldots, I(W_{iq} > c_q)\}$ be a $q \times q$ diagonal matrix, then we have $\boldsymbol{Z}_i(\boldsymbol{\theta}) = \left\{\boldsymbol{Z}_{1i}^T, (\boldsymbol{Z}_{2i}(\boldsymbol{W}_i - \boldsymbol{c})^+)^T, (-\boldsymbol{\gamma}^T \boldsymbol{Z}_{2i} \boldsymbol{I}(\boldsymbol{W}_i > \boldsymbol{c}))^T\right\}^T$. Since $-\boldsymbol{\gamma}^T \boldsymbol{Z}_{2i} \boldsymbol{I}(\boldsymbol{W}_i > \boldsymbol{c}) = \partial r_i(\boldsymbol{\theta})/\partial \boldsymbol{c}$, we have $\boldsymbol{Z}_i(\boldsymbol{\theta}) = \partial r_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$. If the single-index model $\varphi(\boldsymbol{W}_i, c) = (\boldsymbol{\eta}^T \boldsymbol{W}_i - \boldsymbol{c})^+$ is used, we can define $\boldsymbol{Z}_i(\boldsymbol{\theta}) = \partial r_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \left\{\boldsymbol{Z}_{1i}^T, \boldsymbol{Z}_{2i}^T (\boldsymbol{\eta}^T \boldsymbol{W}_i - c)^+, \boldsymbol{\gamma}^T \boldsymbol{Z}_{2i} \boldsymbol{W}_i^T I(\boldsymbol{\eta}^T \boldsymbol{W}_i > c), -\boldsymbol{\gamma}^T \boldsymbol{Z}_{2i} I(\boldsymbol{\eta}^T \boldsymbol{W}_i > c)\right\}^T$.

Let the notations $Y(t), \boldsymbol{Z}_1, \boldsymbol{Z}_{2j}, \boldsymbol{Z}_{\boldsymbol{c}}, \boldsymbol{Z}(\boldsymbol{\theta}), r(\boldsymbol{\theta})$, and T without the subscript i be the random variables that generate the corresponding observations. The following regularity conditions are similar to those by Pons (2003) to establish the asymptotic results for Cox model with change points:

- C1. There exists a $\tau > 0$ such that $\sup_{t \in [0,\tau]} \lambda_0(t) < \infty$ and $P(T \geq \tau) > 0$.
- C2. For each $j = 1, \ldots, q$ and any given covariate $\boldsymbol{Z}_{2j}$, the biomarker variable $\boldsymbol{W}_j$ has a continuous conditional density function $f_{\boldsymbol{W}_j|\boldsymbol{Z}_{2j}}(\cdot)$. Furthermore, for

$\boldsymbol{w}_j \in v(c_{0j})$, a neighbourhood of c_{0j}, there exist positive numbers $0 < m_1 < m_2$, such that $m_1 < f_{\boldsymbol{W}_j|\boldsymbol{Z}_{2j}}(\boldsymbol{w}_j) < m_2$, for all $\boldsymbol{Z}_{2j}$. The conditional density function will be replaced by $f_{\boldsymbol{\eta}_0^T \boldsymbol{W}|\boldsymbol{Z}_{21}}(\cdot)$ for the single index model.

- C3. For each $j = 1, \ldots, q$, the regression coefficient $\gamma_{0j} \neq 0$ to ensure that the threshold model is identifiable.
- C4. There exists a convex and bounded neighbourhood $v(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$, such that $\boldsymbol{\gamma} \neq 0$ for all $\boldsymbol{\theta} \in v(\boldsymbol{\theta}_0)$ and $P(\boldsymbol{a}^T \boldsymbol{Z}(\boldsymbol{\theta}) = 0) = 1$ if and only if $\boldsymbol{a} = \boldsymbol{0}$ for vector $\boldsymbol{a}$ with length $(p + p^* + q)$. That is, the elements of $\boldsymbol{Z}(\boldsymbol{\theta})$ are linearly independent and $E(\boldsymbol{Z}(\boldsymbol{\theta})\boldsymbol{Z}(\boldsymbol{\theta})^T)$ is strictly positive definite.
- C5. There exists a convex and bounded neighbourhood of $\boldsymbol{\theta}_0$, $v(\boldsymbol{\theta}_0)$, such that

$$E\left[\sup_{\theta \in v(\theta_0)} \left\{(||\boldsymbol{Z}_1||^2 + \sum_{j=1}^{q} ||\boldsymbol{Z}_{2j}||^2) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_{\boldsymbol{c}})\right\}^2\right] < \infty.$$

We derive the asymptotic results based on conditions C1–C5 above. It is easy to check condition C4 as long as elements of covariate vector $\boldsymbol{Z}_1$ are not linearly dependent. Because $\boldsymbol{\gamma} \neq \boldsymbol{0}$ for $\boldsymbol{\theta} \in v(\boldsymbol{\theta}_0)$, we have that elements of $\boldsymbol{Z}(\boldsymbol{\theta})$ defined above are not linearly dependent. Here linearly dependent means that at least one of the covariate can be expressed as a linear combination of the other covariates. The same statement is also true for a single-index model.

For $k = 0, 1, 2$, and $t \in [0, \tau]$, let

$$S^{(k)}(t; \boldsymbol{\theta}) = \sum_{i=1}^{n} Y_i(t) \boldsymbol{Z}_i(\boldsymbol{\theta})^{\otimes k} e^{r_i(\boldsymbol{\theta})},$$

where $\boldsymbol{x}^{\otimes 0} = 1, \boldsymbol{x}^{\otimes 1} = \boldsymbol{x}$ and $\boldsymbol{x}^{\otimes 2} = \boldsymbol{x}\boldsymbol{x}^T$ for any column vector $\boldsymbol{x}$. For the corresponding expected value of each $S^{(k)}(t;\boldsymbol{\theta})$ scaled by n, we let

$$s^{(k)}(t;\boldsymbol{\theta}) = E\{Y(t)\boldsymbol{Z}(\boldsymbol{\theta})^{\otimes k} e^{r(\boldsymbol{\theta})}\}, \text{ for } k = 0, 1, 2,$$

where $r(\boldsymbol{\theta}) = \boldsymbol{\beta}^T \boldsymbol{Z}_c$. By condition C1, we have $s^{(0)}(t;\boldsymbol{\theta}) > 0$ for all $t \in [0,\tau]$ and $\boldsymbol{\theta} \in v(\boldsymbol{\theta}_0)$. We divide the vector $s^{(1)}(t;\boldsymbol{\theta})$ into three components: $s^{(1)}_{\boldsymbol{\alpha}}(t;\boldsymbol{\theta}), s^{(1)}_{\boldsymbol{\gamma}}(t;\boldsymbol{\theta})$, and $s^{(1)}_{\boldsymbol{c}}(t;\boldsymbol{\theta})$, which are the first p, the $(p+1)$th to $(p+p^*)$th, and the $(p+p^*+1)$th to $(p+p^*+q)$th elements of the vector, respectively. For a single-index model, we have these three components to be the first p, the $(p+1)$th to $(p+p_1)$th, and the $(p+p_1+1)$th to $(p+p_1+q+1)$th elements of the vector, respectively. Let $s^{(1)}_{\boldsymbol{\beta}}(t;\boldsymbol{\theta}) = (s^{(1)}_{\boldsymbol{\alpha}}(t;\boldsymbol{\theta}), s^{(1)}_{\boldsymbol{\gamma}}(t;\boldsymbol{\theta}))$. We further divide the matrix $s^{(2)}(t;\boldsymbol{\theta})$ into the following blocks:

$$s^{(2)}(t;\boldsymbol{\theta}) = \begin{bmatrix} s^{(2)}_{\boldsymbol{\alpha},\boldsymbol{\alpha}} & s^{(2)}_{\boldsymbol{\alpha},\boldsymbol{\gamma}} & s^{(2)}_{\boldsymbol{\alpha},\boldsymbol{c}} \\ s^{(2)}_{\boldsymbol{\gamma},\boldsymbol{\alpha}} & s^{(2)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} & s^{(2)}_{\boldsymbol{\gamma},\boldsymbol{c}} \\ s^{(2)}_{\boldsymbol{c},\boldsymbol{\alpha}} & s^{(2)}_{\boldsymbol{c},\boldsymbol{\gamma}} & s^{(2)}_{\boldsymbol{c},\boldsymbol{c}} \end{bmatrix} = \begin{bmatrix} s^{(2)}_{\boldsymbol{\beta},\boldsymbol{\beta}} & s^{(2)}_{\boldsymbol{\beta},\boldsymbol{c}} \\ s^{(2)}_{\boldsymbol{c},\boldsymbol{\beta}} & s^{(2)}_{\boldsymbol{c},\boldsymbol{c}} \end{bmatrix}.$$

By the Glivenko-Cantellli theorem (Tucker, 1959), we have the uniform convergence of $S^{(k)}(t;\boldsymbol{\theta})$ below.

Lemma 1. *For $k = 0, 1, 2$, and $\boldsymbol{\theta}$ in a neighbourhood of $v(\boldsymbol{\theta}_0)$,*

$$\sup_{t\in[0,\tau]} \sup_{\boldsymbol{\theta}\in v(\boldsymbol{\theta}_0)} \left|\left|\frac{1}{n}S^{(k)}(t;\boldsymbol{\theta}) - s^{(k)}(t;\boldsymbol{\theta})\right|\right| \to 0 \text{ a.e.}$$

We work on $X_n(\boldsymbol{\theta}) = 1/n\,\{\ell_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}_0)\}$ instead of $\ell_n(\boldsymbol{\theta})$ because maximizing $\ell_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is equivalent to maximizing

$$X_n(\boldsymbol{\theta}) = 1/n \sum_i \delta_i \left\{ (r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0)) - \log \frac{S^{(0)}(T_i;\boldsymbol{\theta})}{S^{(0)}(T_i;\boldsymbol{\theta}_0)} \right\},$$

where $r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0)$ is defined by

$$(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \boldsymbol{Z}_{1i} + \sum_{j=1}^{q} \{(\boldsymbol{\gamma}_j (w_{ij} - c_j)^+ - \boldsymbol{\gamma}_{0j}(w_{ij} - c_{0j})^+)^T \boldsymbol{Z}_{2ij}\}.$$

The second term above will be replaced by $(\boldsymbol{\gamma}(\boldsymbol{\eta}^T \boldsymbol{W}_i - c)^+ - \boldsymbol{\gamma}_0(\boldsymbol{\eta}_0^T \boldsymbol{W}_i - c_0)^+)^T \boldsymbol{Z}_{2i1}$ in a single-index model. Using the method in Lemma 2 of Pons (2003), we have that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |X_n(\boldsymbol{\theta}) - X(\boldsymbol{\theta})|$ converges to zero uniformly in probability as $n \to \infty$, where $X(\boldsymbol{\theta})$ is defined as

$$X(\boldsymbol{\theta}) = E\left[\int_0^\tau Y(t) \left\{ r(\boldsymbol{\theta}) - r(\boldsymbol{\theta}_0) - \log \frac{s^{(0)}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta}_0)} \right\} d\Lambda(t;\boldsymbol{\theta}_0)\right].$$

When the proportional hazards model is correctly specified, $d\Lambda(t;\boldsymbol{\theta}) = e^{r(\boldsymbol{\theta}_0)} d\Lambda_0(t;\boldsymbol{\theta})$. In the appendix, we derive the first order partial derivatives of $X(\boldsymbol{\theta})$ and $X_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$\begin{aligned} \dot{X}(\boldsymbol{\theta}) &= \partial X(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = E\left[\int_0^\tau Y(t)\left\{ \boldsymbol{Z}(\boldsymbol{\theta}) - \frac{s^{(1)}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta})} \right\} d\Lambda(t;\boldsymbol{\theta}_0)\right], \\ \dot{X}_n(\boldsymbol{\theta}) &= \partial X_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = n^{-1} \sum_{i=1}^n \delta_i \left\{ \boldsymbol{Z}_i(\boldsymbol{\theta}) - \frac{S^{(1)}(T_i,\boldsymbol{\theta})}{S^{(0)}(T_i,\boldsymbol{\theta})} \right\}. \end{aligned} \tag{3}$$

Theorem 1. *Under regularity conditions C1–C5, $X(\boldsymbol{\theta})$ is a locally strictly concave function within a neighbourhood of $\boldsymbol{\theta}_0$. If $\boldsymbol{\theta}_0$ is unique and $\sup_{\boldsymbol{\theta}\in v(\boldsymbol{\theta}_0)} X(\boldsymbol{\theta}) < \infty$, for $\hat{\boldsymbol{\theta}}$ that maximizes $X_n(\boldsymbol{\theta})$, then $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ in probability.*

Proof. In the appendix, we show that for any $\boldsymbol{\theta}$ in a small neighbourhood of $\boldsymbol{\theta}_0$, $X(\boldsymbol{\theta})$ is a concave function of $\boldsymbol{\theta}$ under the similar regularity conditions of Andersen (1982). When the proportional hazards model is correctly specified, it is obvious that the function $X(\boldsymbol{\theta})$ is maximized at $X(\boldsymbol{\theta}_0) = 0$. Since $X_n(\boldsymbol{\theta})$ is a continuous function with respect to $\boldsymbol{\theta}$ and converges uniformly to $X(\boldsymbol{\theta})$, for all $\hat{\boldsymbol{\theta}}$ that maximizes $X_n(\boldsymbol{\theta})$, we have that $X(\hat{\boldsymbol{\theta}}) \geq X(\boldsymbol{\theta}_0) - o_p(1)$ and that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ in probability based on Theorem 5.7 of van der Vaart (2000). ■

To prove the asymptotic normality of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, we notice that while $X_n(\boldsymbol{\theta})$ is a continuous function, it does not have a continuous partial derivative with respect to parameter c. However, we show below that $X_n(\boldsymbol{\theta})$ has a Taylor-type expansion similar to a differentiable function. Let $f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{c}_0)$ define the $q \times q$ diagonal matrix of the conditional probability density function of $\boldsymbol{w}_j$ given $\boldsymbol{Z}_{2j}$, for $j = 1, \ldots, q$, with proofs in the appendix, we have Lemma 2 and Lemma 3.

Lemma 2. *Under conditions C1–C3, for any $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}_0$ and linear predictor $r_i(\boldsymbol{\theta}) = \boldsymbol{\beta}^T \boldsymbol{Z}_{\boldsymbol{c}i}$, as $n \to \infty$, $n^{-1} \sum_{i=1}^n Y_i(t)(r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0))$ can be approximated by*

$$n^{-1} \sum_{i=1}^n Y_i(t)\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{Z}_i(\boldsymbol{\theta}_0) + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T(-\boldsymbol{Z}_{2i} I(\boldsymbol{W}_i > \boldsymbol{c_0}))(\boldsymbol{c} - \boldsymbol{c_0})\}$$
$$+\tfrac{1}{2}(\boldsymbol{c} - \boldsymbol{c}_0)^T E\{Y(t)\boldsymbol{\gamma}_0^T \boldsymbol{Z}_2 f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{c}_0)\}(\boldsymbol{c} - \boldsymbol{c}_0) \tag{4}$$

with a difference of $o_p(||\boldsymbol{\theta}-\boldsymbol{\theta}_0||^2)$ *term and* $S^{(0)}(t,\boldsymbol{\theta})-S^{(0)}(t,\boldsymbol{\theta}_0)$ *can be approximated by*

$$(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T S^{(1)}(t,\boldsymbol{\theta}_0)+\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\{S^{(2)}(t,\boldsymbol{\theta}_0)+S^{(2)*}(t,\boldsymbol{\theta}_0)\}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \quad (5)$$

with a difference of $o_p(||\boldsymbol{\theta}-\boldsymbol{\theta}_0||^2)$ *term, where* $S^{(2)*}(t,\boldsymbol{\theta}_0)$ *comes from the second order terms in the approximation (4) and is defined in the appendix. Expression for* $S^{(2)*}(t,\boldsymbol{\theta}_0)$ *in a single-index model is also provided in the appendix.*

Lemma 3. *For* $n\to\infty$ *and* $\boldsymbol{V}_n=\boldsymbol{V}_{n,1}+\boldsymbol{V}_{n,2}$ *defined below, the Taylor-type expansion of* $X_n(\boldsymbol{\theta})$ *is*

$$X_n(\boldsymbol{\theta})=(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\dot{X}_n(\boldsymbol{\theta}_0)+\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{V}_n(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T+o_p(||\boldsymbol{\theta}-\boldsymbol{\theta}_0||^2),$$

where

$$\boldsymbol{V}_{n,1}=-\frac{1}{n}\sum_i\{S^{(2)}(T_i,\boldsymbol{\theta}_0)S^{(0)}(T_i,\boldsymbol{\theta}_0)-S^{(1)}(T_i,\boldsymbol{\theta}_0)S^{(1)}(T_i,\boldsymbol{\theta}_0)^T\}/S^{(0)}(T_i,\boldsymbol{\theta}_0)^2,$$

$$\boldsymbol{V}_{n,2}=\begin{bmatrix} 0_{p\times p} & 0_{p\times p^*} & 0_{p\times q} \\ 0_{p^*\times p} & 0_{p^*\times p^*} & \boldsymbol{V}_{n,2,\boldsymbol{\gamma},\boldsymbol{c}} \\ 0_{q\times p} & \boldsymbol{V}_{n,2,\boldsymbol{\gamma},\boldsymbol{c}}^T & \boldsymbol{V}_{n,2,\boldsymbol{c},\boldsymbol{c}} \end{bmatrix} \quad (6)$$

with

$$\boldsymbol{V}_{n,2,\boldsymbol{\gamma},\boldsymbol{c}}=n^{-1}\textstyle\sum_i\delta_i\left\{Y_i(t)(-\boldsymbol{Z}_{2i}I(\boldsymbol{W}_i>\boldsymbol{c_0}))-S_{\boldsymbol{\gamma},\boldsymbol{c}}^{(2)*}(T_i,\boldsymbol{\theta}_0)/S^{(0)}(T_i,\boldsymbol{\theta}_0)\right\},$$

$$\boldsymbol{V}_{n,2,\boldsymbol{c},\boldsymbol{c}}=n^{-1}\textstyle\sum_i\delta_i\left\{Y_i(t)\boldsymbol{\gamma}_0'\boldsymbol{Z}_{2i}\hat{f}_{\boldsymbol{W}|\boldsymbol{Z}_{2i}}(\boldsymbol{c_0})-S_{\boldsymbol{c},\boldsymbol{c}}^{(2)*}(T_i,\boldsymbol{\theta}_0)/S^{(0)}(T_i,\boldsymbol{\theta}_0)\right\},$$

where $\hat{f}_{\boldsymbol{W}|\boldsymbol{Z}_{2i}}(\boldsymbol{c}_0)$ *is a* $q \times q$ *diagonal matrix of the estimated conditional probability density of* $\boldsymbol{W}$ *given* $\boldsymbol{Z}_{2i}$ *and* $S^{(2)*}_{\boldsymbol{\gamma},\boldsymbol{c}}(t, \boldsymbol{\theta}_0)$, $S^{(2)*}_{\boldsymbol{c},\boldsymbol{c}}(t, \boldsymbol{\theta}_0)$ *are the corresponding block matrix of* $S^{(2)*}(t, \boldsymbol{\theta}_0)$ *in (5).*

Note here that the matrix $\boldsymbol{V}_{n,2}$ reflects the second term of $S^{(2)*}(t, \boldsymbol{\theta}_0)$ in Lemma 2. Similar results for $\boldsymbol{V}_{n,2,\boldsymbol{\gamma},\boldsymbol{\eta}}$ and $\boldsymbol{V}_{n,2,\boldsymbol{\eta},c}$ can also be derived for the single-index model.

Proof. Since $X_n(\boldsymbol{\theta}) = 1/n \sum_i \delta_i \left\{ r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0) - \log \frac{S^{(0)}(\boldsymbol{\theta})}{S^{(0)}(\boldsymbol{\theta}_0)} \right\}$, the result of Lemma 3 follows Lemma 2 and the Taylor expansion of

$$\log(\frac{x}{x_0}) = \log(1 + \frac{x - x_0}{x_0}) = \frac{x - x_0}{x_0} - \frac{(x - x_0)^2}{2x_0^2} + o((x - x_0)^2).$$

■

Since the linear predictor $\boldsymbol{\beta}^T \boldsymbol{Z}_{\boldsymbol{c}i}$ and function $S^{(0)}(t, \boldsymbol{\theta})$ are not continuously differentiable with respect to $\boldsymbol{c}$, we cannot use Taylor expansion directly to the log likelihood function. However, with Lemma 3, $X_n(\boldsymbol{\theta})$ admits a Taylor-type expansion when $\boldsymbol{\theta}$ is close to $\boldsymbol{\theta}_0$. The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ depends on the Hessian matrix of $X(\boldsymbol{\theta})$, which is twice differentiable. To obtain the Hessian matrix, we define the following matrix,

$$\boldsymbol{V}_1 = -E\left[\int_0^\tau Y(t) \left\{ \frac{s^{(2)}(t;\boldsymbol{\theta}) s^{(0)}(t;\boldsymbol{\theta}) - s^{(1)}(t;\boldsymbol{\theta}) s^{(1)}(t;\boldsymbol{\theta})^T}{(s^{(0)}(t;\boldsymbol{\theta}))^2} \right\} d\Lambda(t;\boldsymbol{\theta}_0)\right].$$

We can partition matrix $\boldsymbol{V}_1$ into blocks of $\boldsymbol{V}_{1,\boldsymbol{\beta},\boldsymbol{\beta}}$, $\boldsymbol{V}_{1,\boldsymbol{\beta},\boldsymbol{c}}$ and $\boldsymbol{V}_{1,\boldsymbol{c},\boldsymbol{c}}$ the same way as we partition $s^{(2)}(t;\boldsymbol{\theta})$. With the details of derivations of $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ in the appendix, finally we have matrix

$$\boldsymbol{V} = \ddot{X}(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{V}_1 + \boldsymbol{V}_2,$$

where $\ddot{X}(\boldsymbol{\theta})$ is the Hessian matrix of $X(\boldsymbol{\theta})$ and $\boldsymbol{V}_2$ has the same structure as $\boldsymbol{V}_{n,2}$ as in equation (6) with the sample mean replaced by expected value of $\boldsymbol{V}_{2,\gamma,\boldsymbol{c}}$ and $\boldsymbol{V}_{2,\boldsymbol{c},\boldsymbol{c}}$. By the law of large numbers and regularity condition C5, $\boldsymbol{V}_{n,1}$ and $\boldsymbol{V}_{n,2}$ converge to $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ in probability, respectively. When the model is correctly specified, by Theorem 1, $X(\boldsymbol{\theta})$ is a strictly concave function within a neighbourhood of $\boldsymbol{\theta}_0$. Therefore, $\boldsymbol{V}$ is a non-singular matrix.

Theorem 2. *Under regularity conditions C1 – C5 and the same assumptions in Theorem 1, let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$, then we have*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N(\mathbf{0}, \boldsymbol{V}^{-1}\boldsymbol{\Gamma}\boldsymbol{V}^{-1})$$

in distribution, where $\boldsymbol{\Gamma} = \left\{\dot{X}(\boldsymbol{\theta})\dot{X}(\boldsymbol{\theta})^T\right\} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$.

The proof of Theorem 2 is similar to that in Theorem 5.23 of van der Vaart (2000) and we provide the details in the appendix.

When the Cox proportional hazards model $\Lambda(t;\boldsymbol{\theta}) = e^{r(\boldsymbol{\theta})}\Lambda_0(t;\boldsymbol{\theta})$ is correctly specified, for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, it is easy to show that $\boldsymbol{V}_{2,\gamma,\boldsymbol{c}} = \boldsymbol{V}_{2,\boldsymbol{c},\boldsymbol{c}} = 0$ and $\boldsymbol{\Gamma} = -\boldsymbol{V}$. Therefore, the variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1}$. If the model is misspecified, we have $\boldsymbol{\Sigma} = \boldsymbol{V}^{-1}\boldsymbol{\Gamma}\boldsymbol{V}^{-1}$. The estimated information matrix $\hat{\boldsymbol{\Gamma}}$ is obtained by

$$\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{\tau} Y_i(t)\left\{\boldsymbol{Z}_i(\hat{\boldsymbol{\theta}}) - \frac{S^{(1)}(T_i,\hat{\boldsymbol{\theta}})}{S^{(0)}(T_i,\hat{\boldsymbol{\theta}})}\right\}\left\{\boldsymbol{Z}_i(\hat{\boldsymbol{\theta}}) - \frac{S^{(1)}(T_i,\hat{\boldsymbol{\theta}})}{S^{(0)}(T_i,\hat{\boldsymbol{\theta}})}\right\}^{T} e^{r_i(\hat{\boldsymbol{\theta}})} d\hat{\Lambda}_0(t),$$

where $\hat{\Lambda}_0(t) = \sum_{i=1}^{n}\int_0^{\tau} \frac{dN_i(s)}{\sum_{j=1}^{n} Y_j(s)\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{\hat{c}j})}$ is the Breslow's estimator of the baseline cumulative hazard function. The matrix $\boldsymbol{V}$ is estimated by $\hat{\boldsymbol{V}}_{n,1} + \hat{\boldsymbol{V}}_{n,2}$ with the unknown parameter $\boldsymbol{\theta}$ replaced by the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Following the same argument of Andersen and Gill (1982), it is straightforward to show that $\sqrt{n}\{\hat{\Lambda}_0(t) - \Lambda_0(t)\}$ converges weakly to a zero-mean Gaussian process and is asymptotically independent with $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

4. SIMULATION STUDIES

In this section, we evaluate performances of the proposed model (1) using simulated datasets. Different sample sizes n from 200 to 1000 are considered. We simulate the biomarker variable w_i with a normal distribution with mean = 0.2 and standard deviation = 2.0. Subjects are randomly assigned to either the treatment group ($z_{1i} = 1$) or the control group ($z_{1i} = 0$) with a Bernoulli distribution and $P(z_{1i} = 1)$ = 0.5. The survival time T_i has an exponential distribution with hazard function

$$h_i(t) = h_0(t)e^{\beta_1 z_{1i} + \beta_2 (w_i - c)^+ + \beta_3 z_{1i}(w_i - c)^+}$$

where the baseline hazard $h_0(t) = 0.5$. For any given cut-point c, we have the covariate vector $\boldsymbol{Z}_i(c) = \{z_{1i}, (w_i - c)^+, z_{1i}(w_i - c)^+\}$. Let C_i be the random censoring time generated from a uniform distribution $C_i \sim U(0,5)$, which gives a censoring proportion of 30% to 40% in the simulated data. The simulated dataset contains the censored survival time (T_i, δ_i), treatment covariate z_{1i} and biomarker covariate w_i.

When the model is correctly specified, we consider the following settings of true values for $\boldsymbol{\theta}$. We let c take values -1.0, -0.5, 0.5, 1.0, and 1.5. For $\boldsymbol{\beta}$, $\beta_1 = 0$, β_2 takes values 0.6 and 0.8, and β_3 takes values -0.4, -0.2, and 0.3. To study the robustness of the proposed model, we further fit model (1) with the simulated data from $h_i(t) = h_0(t)e^{\boldsymbol{Z}_i^* \boldsymbol{\beta}}$, which contains a misspecified biomarker covariate effect of w_i. That is, for the covariate vector, $\boldsymbol{Z}_i^* = \{z_{1i}, f(w_i), z_{1i}f(w_i)\}$. One of the misspecified biomarker effects is discussed by a shifted softplus function, which is a continuous differentiable version of the ReLU function:

$$f(w) = 0.2 \log\left[1 + \exp\left\{5(w + 0.5)\right\}\right]. \tag{7}$$

In each combination of these parameter settings, we replicate the simulation 600 times to evaluate the finite sample properties of $\boldsymbol{\theta}$ in the proposed model.

We first check the absolute empirical bias of $\hat{c}$ versus sample size n when c=-0.5, $\beta_1 = 0$, $\beta_2 = 0.6$, and $\beta_3 = -0.4$ with 600 replications. The absolute empirical biases are 0.082, 0.023, 0.022, 0.015, and 0.011 for sample sizes 200, 400, 600, 800, and 1000, respectively. It shows that when the sample size increases, there is an overall decreasing trend, which is within the theoretical expectation that using a larger sample size reduces the empirical bias of estimation. When the sample size reaches around 1000, the empirical bias becomes quite small and it provides justification to use sample size 1000 in the simulation.

Table 1 gives the empirical bias and standard errors (se_1 and se_2) of the estimation by the proposed gradient descent method. Here, se_1 is the estimated standard error based on the proposed model and se_2 is the empirical standard error based on the simulation. From the table, we observe these two standard errors of $\hat{\beta}_3$ for each combination of parameters $\boldsymbol{\theta}$ are close to each other. For example, when $c = -0.5$, $\beta_2 = 0.6$, and $\beta_3 = -0.4$, se_1 and se_2 are 0.056 and 0.058 respectively. Similar performances are observed for the other threshold parameters β_1, β_2 and c.

In terms of empirical bias, for the threshold parameter c, the biases are very small in all settings we consider, which confirms the conclusion that $\hat{c}$ is unbiased and consistent. We also find the biases of $\boldsymbol{\beta}$ are small by the proposed method. For example, even when the bias of $\hat{c}$ is 0.021, which is the highest shown in Table 1, the biases of $\hat{\beta}_1$ (0.003), $\hat{\beta}_2$ (0.016) and $\hat{\beta}_3$ (-0.008) remain very small. This indicates that our proposed gradient descent method provides unbiased estimators of $\boldsymbol{\beta}$. Even when the model is misspecified as a model with the softplus function, we still have low empirical biases of these estimated $\boldsymbol{\beta}$. The average of $\hat{c}$ in the misspecified model is around -0.5, which agrees with $f(z_2)$, with a 0.5 unit shifted to the left in the softplus function. This demonstrates that the proposed method is robust to model misspecification.

TABLE 1: Empirical biases, estimated standard error (se_1) and empirical standard error (se_2) of $\hat{\boldsymbol{\theta}}$. These results are based on sample size n = 1000, $\beta_1 = 0$ with 600 replications.

							$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{c}$	
β_2	β_3	c	$\hat{c}^\dagger$	$\hat{\beta}_1{}^\dagger$	$\hat{\beta}_2{}^\dagger$	$\hat{\beta}_3{}^\dagger$	se_1	se_2	se_1	se_2	se_1	se_2	se_1	se_2
0.6	-0.4	-1	-3	9	5	-2	121	121	42	45	50	52	244	269
0.6	-0.4	-0.5	7	2	8	-4	110	109	48	50	56	58	231	262
0.6	-0.4	0.5	9	0	9	-1	96	101	65	68	74	76	236	273
0.6	-0.4	1	-10	3	12	-9	92	97	79	91	90	96	253	319
0.6	-0.4	1.5	-2	-2	16	-8	88	87	99	115	112	122	280	324
0.6	-0.2	-1	-12	4	7	-2	111	116	41	44	47	48	213	241
0.6	-0.2	-0.5	1	-2	4	1	103	105	46	48	52	51	203	231
0.6	-0.2	0.5	21	3	16	-8	93	90	62	65	68	68	205	219
0.6	-0.2	1	-1	-2	8	0	89	91	75	79	81	82	222	254
0.6	-0.2	1.5	5	-1	18	-5	87	88	93	103	100	110	244	289
0.6	0.3	-1	-6	-8	2	3	110	115	38	41	47	48	143	156
0.6	0.3	-0.5	-1	-4	3	2	103	107	42	45	52	54	136	155
0.6	0.3	0.5	-8	-1	1	3	93	94	54	57	67	68	138	156
0.6	0.3	1	2	-5	4	3	89	89	65	66	80	78	147	174
0.6	0.3	1.5	11	-6	10	6	86	85	80	83	99	98	161	174
0.8	-0.4	-1	-3	-4	3	-1	114	117	44	45	49	51	172	195
0.8	-0.4	-0.5	2	-2	7	-2	105	104	49	53	54	54	163	182
0.8	-0.4	0.5	-4	1	6	1	94	97	65	65	70	67	166	188
0.8	-0.4	1	10	3	13	-6	90	93	79	88	84	87	177	207
0.8	-0.4	1.5	1	0	13	-2	87	86	97	108	103	111	196	223
0.8	-0.2	-1	-6	8	3	-2	109	110	42	48	46	46	155	166
0.8	-0.2	-0.5	4	-7	3	2	102	103	47	49	51	50	147	152
0.8	-0.2	0.5	-3	4	6	-2	92	93	62	68	66	67	149	159
0.8	-0.2	1	2	0	9	1	89	90	74	81	79	86	158	173
0.8	-0.2	1.5	9	-1	16	-4	86	90	92	100	97	103	175	198
0.8	0.3	-1	0	-2	4	1	108	111	40	40	47	48	115	122
0.8	0.3	-0.5	-6	-5	3	0	102	98	44	44	52	49	109	116
0.8	0.3	0.5	0	1	5	3	92	92	57	58	67	66	110	121
0.8	0.3	1	3	1	11	-3	89	90	68	72	79	86	116	130
0.8	0.3	1.5	4	-6	10	10	86	87	82	87	97	102	128	143
Mis			$\hat{c}$											
0.6	-1.0		-514	0	2	-9	134	149	47	51	89	97	217	269
0.6	-0.8		-501	-11	5	-7	129	140	48	54	74	82	239	297
0.6	-0.6		-509	-12	-1	1	121	128	49	52	64	66	248	288
0.8	-1.0		-516	-10	2	-2	128	136	52	55	75	79	182	214
0.8	-0.8		-511	-3	0	0	120	127	51	57	66	71	186	211
0.8	-0.6		-524	-2	-2	0	113	120	51	59	59	68	180	236

simulation results are multiplied by 10^3.

† represents the empirical bias.

Table 2 summarizes the results of the coverage probabilities of the 95% confidence intervals of $\boldsymbol{\theta}$ and the probability of rejecting the null hypothesis $H_0 : \beta_3 = 0$. Here, hypothesis testing is possible because we set β_2 to be 0.6 or 0.8 in our simulation. Therefore, the regularity condition C3 ($\boldsymbol{\gamma} = (\beta_2, \beta_3)^T \neq 0$) holds. For the coverage probability of c, most of the settings have coverage probabilities around the nominal level of 0.95 despite a few settings having the coverage probabilities lower than 0.90. For example, when $\beta_2 = 0.6$, $\beta_3 = -0.4$, and $c = 1.5$, the coverage probability of c is 0.88. This may be because at this value of c, most subjects do not have threshold-defined subset effect because $(w - c) < 0$.

The overall coverage probabilities of $\boldsymbol{\beta}$ are close to the nominal level of 0.95, which indicates that the estimated standard deviations are appropriate. For the testing of null hypothesis $H_0 : \beta_3 = 0$, most combinations of the true values under an alternative hypothesis indicate that we have a high probability to reject the null hypothesis. The proposed method shows that when the absolute value of β_3 is small (e.g. $\beta_3 = -0.2$) and c increases, the power to reject $H_0 : \beta_3 = 0$ drops quickly. The reason is that when we increase the true value of c, the proportion of 0 in the biomarker covariate $(w - c)^+$ increases.

When the threshold model is misspecified as a softplus function (7), the coverage probabilities for $\boldsymbol{\beta}$ are still around the 95% nominal level. The shifted value of the softplus function -0.5 is included in all the confidence intervals of c. This implies the robust estimation of $\boldsymbol{\theta}$. The power of the test of the null hypothesis testing $\beta_3 = 0$ is 1.00 in all the settings.

We further evaluate the empirical bias and standard errors when an indicator threshold model (Chen, Jiang & Tu, 2014),

$$h^*(t) = h_0(t)e^{\beta_1 z_{1i} + \beta_2 I(w_i > c) + \beta_3 z_{1i} I(w_i > c)}$$

is misspecified as a continuous threshold model (Model 1) and a continuous threshold model is misspecified as an indicator threshold model. The upper half of Table 3 presents the simulation results in a misspecified model which simulates the biomarker by indicator function but fits the model with ReLU function. As the results show, even though the standard errors are relatively small, the biases of $\hat{\beta}_2$ and $\hat{\beta}_3$ are quite large in all the settings. Since β_1 is the coefficient of the treatment, which is not restricted by indicator or ReLU functions, the biases of $\hat{\beta}_1$ are small as expected. We utilize the R package *bhm* (Chen, 2020), which is the realization of the indicator threshold model defined above. As the lower half of Table 3 shows, we simulate the biomarker with ReLU function but fit a naive model in which an indicator function of the biomarker is exploited. We also find the same phenomenon of very large biases and relatively low standard errors of $\boldsymbol{\beta}$. Based on the lower half of Table 3, we can conclude that if the biomarker of the true data has a ReLU shape, then the model with indicator function fails. Therefore, for a given dataset, whether one shall fit a biomarker threshold model with an indicator function or a continuous threshold function depends on the

TABLE 2: Empirical coverage probabilities for parameter $\boldsymbol{\theta}$, 95% confidence interval of c and power of test of $\beta_3 = 0$. These results are based on sample size n = 1000, $\beta_1 = 0$ with 600 replications.

β_2	β_3	c	$\hat{c}$	$95\% CI(c)$	$CP(c)$	$CP(\beta_1)$	$CP(\beta_2)$	$CP(\beta_3)$	Power(β_3)
0.6	-0.4	-1.0	-1.02	(-1.48, -0.53)	0.92	0.95	0.93	0.94	1.00
0.6	-0.4	-0.5	-0.53	(-0.95, -0.04)	0.92	0.95	0.95	0.95	1.00
0.6	-0.4	0.5	0.49	(0.05, 0.97)	0.91	0.94	0.96	0.94	1.00
0.6	-0.4	1.0	0.99	(0.50, 1.49)	0.88	0.95	0.91	0.93	1.00
0.6	-0.4	1.5	1.47	(0.95, 2.05)	0.88	0.96	0.91	0.94	0.98
0.6	-0.2	-1.0	-1.00	(-1.43, -0.60)	0.91	0.96	0.94	0.95	0.98
0.6	-0.2	-0.5	-0.48	(-0.90, -0.10)	0.92	0.96	0.95	0.95	0.97
0.6	-0.2	0.5	0.49	(0.12, 0.92)	0.90	0.95	0.93	0.95	0.85
0.6	-0.2	1.0	0.98	(0.56, 1.43)	0.90	0.95	0.94	0.94	0.72
0.6	-0.2	1.5	1.48	(1.03, 1.98)	0.88	0.95	0.93	0.95	0.54
0.6	0.3	-1.0	-1.00	(-1.29, -0.73)	0.92	0.93	0.95	0.94	1.00
0.6	0.3	-0.5	-0.50	(-0.77, -0.24)	0.92	0.96	0.94	0.95	1.00
0.6	0.3	0.5	0.50	(0.22, 0.76)	0.92	0.95	0.96	0.94	1.00
0.6	0.3	1.0	1.00	(0.71, 1.29)	0.94	0.95	0.95	0.95	0.97
0.6	0.3	1.5	1.50	(1.20, 1.83)	0.91	0.94	0.93	0.94	0.89
0.8	-0.4	-1.0	-1.00	(-1.34, -0.67)	0.93	0.97	0.95	0.95	1.00
0.8	-0.4	-0.5	-0.49	(-0.82, -0.18)	0.90	0.94	0.95	0.95	1.00
0.8	-0.4	0.5	0.50	(0.17, 0.82)	0.92	0.94	0.94	0.94	1.00
0.8	-0.4	1.0	1.00	(0.66, 1.36)	0.92	0.96	0.94	0.96	1.00
0.8	-0.4	1.5	1.49	(1.12, 1.89)	0.91	0.94	0.92	0.95	0.99
0.8	-0.2	-1.0	-1.00	(-1.31, -0.70)	0.92	0.94	0.94	0.94	0.99
0.8	-0.2	-0.5	-0.50	(-0.78, -0.21)	0.94	0.96	0.95	0.94	0.98
0.8	-0.2	0.5	0.49	(0.21, 0.79)	0.93	0.96	0.93	0.95	0.88
0.8	-0.2	1.0	0.99	(0.69, 1.31)	0.91	0.95	0.92	0.95	0.73
0.8	-0.2	1.5	1.50	(1.17, 1.85)	0.90	0.95	0.94	0.95	0.57
0.8	0.3	-1.0	-1.00	(-1.22, -0.78)	0.94	0.94	0.96	0.94	1.00
0.8	0.3	-0.5	-0.50	(-0.72, -0.29)	0.93	0.96	0.95	0.96	1.00
0.8	0.3	0.5	0.50	(0.28, 0.71)	0.93	0.95	0.95	0.96	1.00
0.8	0.3	1.0	1.00	(0.78, 1.23)	0.94	0.95	0.94	0.95	0.97
0.8	0.3	1.5	1.49	(1.25, 1.75)	0.91	0.93	0.93	0.95	0.89
Misspecified									
0.6	-1.0	-	-0.51	(-0.94, -0.07)	-	0.93	0.93	0.93	1.00
0.6	-0.8	-	-0.50	(-0.99, -0.03)	-	0.95	0.93	0.94	1.00
0.6	-0.6	-	-0.51	(-1.03, -0.03)	-	0.93	0.91	0.93	1.00
0.8	-1.0	-	-0.52	(-0.87, -0.15)	-	0.94	0.93	0.93	1.00
0.8	-0.8	-	-0.51	(-0.88, -0.15)	-	0.92	0.95	0.94	1.00
0.8	-0.6	-	-0.52	(-0.89, -0.18)	-	0.94	0.94	0.95	1.00

TABLE 3: Biases of $\hat{\boldsymbol{\theta}}$, standard error of $\hat{\boldsymbol{\beta}}$ and 95% confidence interval of c in the misspecified and naive models. These results are based on sample size n = 1000, $\beta_1 = 0$ with 600 replications.

β_2	β_3	c	$\hat{c}^\dagger$	$\hat{\beta}_1^\dagger$	$\hat{\beta}_2^\dagger$	$\hat{\beta}_3^\dagger$	$\hat{c}$	$95\%CI(c)$	$SE(\hat{\beta}_1)$	$SE(\hat{\beta}_2)$	$SE(\hat{\beta}_3)$
Indicator biomarker, ReLU model fitting											
0.6	-0.2	-1.0	-329	-0.5	-50.4	17.2	-429.2	(-1016.0, 157.6)	25.1	3.0	4.1
0.6	-0.2	-0.5	-310	2.0	-48.6	16.4	-359.8	(-825.8, 106.3)	22.7	3.1	4.2
0.6	0.3	-1.0	-354	-1.5	-50.4	-25.3	-454.1	(-938.5, 30.2)	26.0	2.8	3.8
0.6	0.3	-0.5	-303	-3.7	-49.0	-24.3	-352.9	(-697.6, -8.2)	21.6	2.9	4.0
0.8	-0.2	-1.0	-355	0.7	-67.2	16.9	-455.0	(-900.4, -9.4)	23.2	2.8	3.9
0.8	-0.2	-0.5	-307	2.2	-64.7	16.2	-357.0	(-697.0, -16.9)	20.1	3.0	4.0
0.8	0.3	-1.0	-355	-1.1	-67.4	-25.5	-454.6	(-857.6, -51.6)	24.3	2.8	3.8
0.8	0.3	-0.5	-308	-3.6	-65.3	-24.4	-358.4	(-608.8, -108.1)	19.4	2.9	4.0
ReLU biomarker, indicator model fitting											
0.6	-0.2	-1.0	189.9	-69.3	150.9	-27.6	89.9	(83.1, 97.1)	8.9	10.3	13.9
0.6	-0.2	-0.5	166.3	-66.6	145.5	-26.4	116.3	(108.5, 124.6)	8.7	10.8	14.5
0.6	0.3	-1.0	182.2	-49.6	141.2	31.0	82.2	(77.5, 87.7)	8.9	10.1	13.5
0.6	0.3	-0.5	161.3	-52.1	136.9	28.2	111.3	(105.6, 117.2)	8.5	10.5	14.1
0.8	-0.2	-1.0	186.0	-88.4	185.5	-22.7	86.0	(80.7, 91.9)	8.8	10.4	13.6
0.8	-0.2	-0.5	164.1	-85.9	179.4	-21.8	114.1	(108.1, 120.5)	8.5	10.7	14.1
0.8	0.3	-1.0	177.2	-71.0	173.8	25.0	77.2	(73.0, 81.9)	8.8	10.3	13.4
0.8	0.3	-0.5	156.8	-73.1	169.8	22.1	106.8	(102.4, 111.6)	8.5	10.6	13.9

simulation results are multiplied by 10^2.

† represents the empirical bias.

functional form of the biomarker effect, which shall be explored carefully.

5. APPLICATION

We use a prostate cancer clinical trial conducted by the second Veterans Adminstration Cooperative Urologic Research Group (Byar & Corle, 1977; Andrews & Herzberg, 1985) to demonstrate the proposed continuous threshold model. In this trial, there is a total of $n = 505$ patients with prostate cancer, in which 128 patients were randomly assigned to receive a placebo $(z_{1i} = 0)$ and 377 patients to receive a diethylstilbestrol treatment $(z_{1i} = 1)$. The primary outcome for this trial is overall survival, which measures the time from randomization to the time of death from any causes. At the end of the study, 355 out of the 505 patients died and the other 150 patients (30%) were censored. Patients' statuses represent that patient i has an event $(\delta_i = 1)$ or that patient i is censored $(\delta_i = 0)$. Serum prostatic acid phosphatase (AP) is a predictive biomarker (Byar & Corle, 1977; Jiang et al., 2007) with values range from 1 to 5960 units. We observe that the distribution of AP is skewed and apply a log-transformation to scale down the large tails. The transformed logAP ranges from 0 to 8.69 units with median 1.95 units and interquartile range from 1.61 to 3.37 units. Gutman, Sproul & Gutman (1936) observe that serum prostatic acid phosphatase activity is significantly higher in prostate cancer patients. The primary interest is to identify the threshold and investigate how the AP biomarker is related to the prostate cancer patients' mortality.

For Table 4, β_1, β_2 and β_3 are the coefficients of rectified biomarker, treatment, and the interaction of treatment and rectified biomarker, respectively. The results of Table 4 indicate that there is a significant subset treatment effect with the *p*-value for interaction = 0.005. The parameter β_3 is estimated to be -0.256, with 95% CI [-0.434, -0.078] for the transformed AP biomarker. The estimated threshold parameter $\hat{c}$ is 1.857 (95% CI = [0.513, 3.201]) based on the log-transformation on the biomarker AP. In its original scale, we have AP = $\exp(1.857)$ = 6.405 (95% CI = [1.670, 24.557]) as the cut-off (threshold effect). The hazard ratio of treatment is estimated to be $\mathrm{HR} = \exp(\hat{\beta}_1)$ = 1.06 (95% CI: 0.78 to 1.44) for patients with AP less than 6.405. For patients with AP value greater than 6.405, the hazard ratio of death (diethylstilbestrol versus placebo) was decreased by $(1 - \exp(\hat{\beta}_3)) \times 100\% = 22.6\%$ (95% CI: 7.5% to 35.2%), with each unit increase in the log AP value.

Figure 1 plots the estimated survival curves for four different subgroups by the Kaplan-Meier method. For the AP biomarker values below 6.405, which includes about 43% of the patients, the survival function of patients with diethylstilbestrol is similar to those with placebo (Hazard ratio (HR) = 1.09, 95% CI = [0.76, 1.55], p-value = 0.65). While for biomarker values above 6.405, the treatment benefit depends on the AP value and the hazard ratio is a function of AP: HR = $\exp\{0.056 - 0.256 \times (\log(AP) - 1.857)\}$. On average, for $\mathrm{AP} > 6.405$, which includes about 57% of the patients, the survival function of patients with diethylstilbestrol is significantly higher than those with a placebo (HR = 0.73, 95% CI = [0.53, 0.99], p-value = 0.045), which indicates patients benefit more from the treatment of diethylstilbestrol than a placebo for the large AP biomarker values.

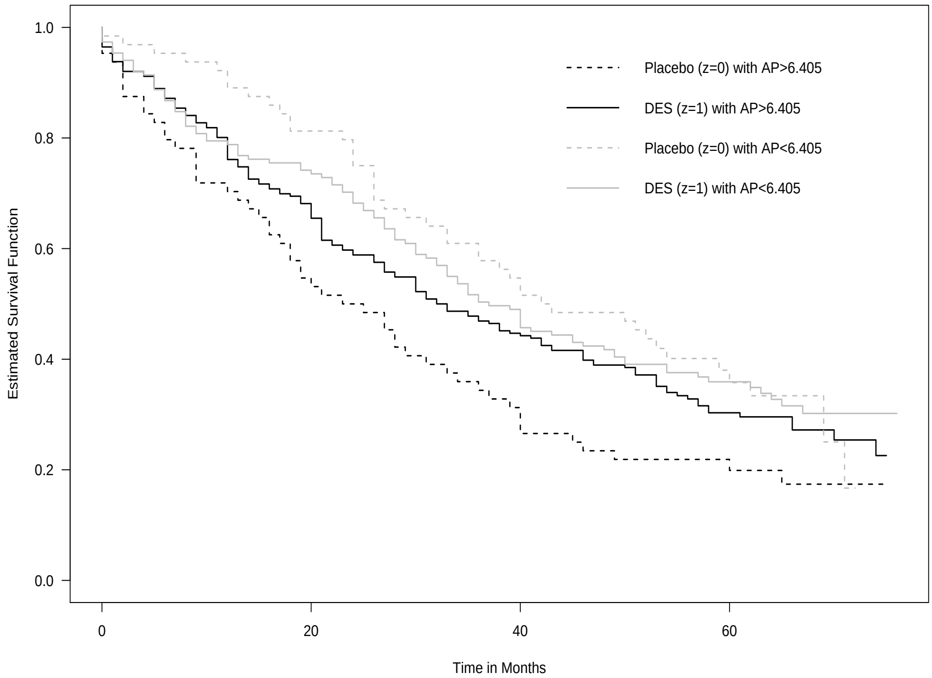

FIGURE 1: Estimated survival curves for subgroups. The dark dash and dark solid lines show the survival functions for patients with treatment of placebo and diethylstilbestrol (DES) when AP is greater than 6.405, respectively. The light dash and light solid lines show the survival functions for patients with treatment of placebo and diethylstilbestrol when AP is less than 6.405, respectively.

To assess the validity of the fitted model, we consider the Cox-Snell residuals and the Martingale residuals. For the Cox-Snell residual plot, the residuals are expected to equal the estimated cumulative hazard for each observation. As Figure 2 shows, there is an overall linear relation between the estimated cumulative hazards and the residuals, which indicates an excellent overall fitness of the pro-

TABLE 4: Summary table for model parameter $\boldsymbol{\theta}$ by gradient descent.

Parameter	Estimate	SE	Z value	95% C.I.		p-value
β_1 : $(\log AP - c)^+$	0.351	0.088	4.003	0.179	0.523	< 0.0001
β_2 : treatment	0.056	0.157	0.359	-0.251	0.363	0.720
β_3 : treatment$\times(\log AP - c)^+$	-0.256	0.091	-2.817	-0.434	-0.078	0.005
c : threshold	1.857	0.686	2.708	0.513	3.201	-

posed model. For the martingale residuals, the residuals are expected to have zero mean and range between $-\infty$ and 1. In large samples, the martingale residuals should be approximately uncorrelated. Figure 2 illustrates that the residuals have mean 0 with all values below 1 and that there are no obvious patterns. Therefore, the martingale residuals plot also indicates a good model fit.

Moreover, we compare different models by calculating the probability of concordance between the linear predictors and the censored survival time. For the model with the ReLU threshold effect, we have the concordance probability = 0.57, while for models with indicator function as threshold effect and linear function without threshold effect, the concordance probabilities are 0.54 and 0.55, respectively. It is reasonable to choose the proposed continuous threshold model, which gives the highest concordance probability among these three models.

Finally, the indicator threshold model suggests that for $\mathrm{AP} \leq 54.05$, which includes 81.8% of the patients, there is no treatment benefit (the hazard ratio = 0.96, 95% CI: [0.74, 1.25]). For patients with AP $>$ 54.05, which includes 18.2% of the patients, the hazard ratio for the treatment is 0.39 with 95% CI: [0.21, 0.71]. It may be unethical to treat one patient with AP $= 54.06$ (HR = 0.39) but not to treat the other patient with AP $= 54.05$ (HR = 0.96) because the difference of the AP value between these two patients is just 0.01. However, in a continuous threshold model, we do not have to face such a problem. For a patient with AP = 54.05, the HR = 0.61261, and for a patient with AP = 54.06, the HR = 0.61258, therefore, both patients benefit from the drug. For patients with biomarker value around the cut-point of a continuous threshold model, we understand that the benefit is small. In this case, the patients will be recommended to receive the treatment for the potential benefit if the drug is tolerable with little side effect.

6. DISCUSSION

In this paper, we derive the exact mathematical form of the asymptotic distribution under certain regularity conditions for both correctly specified and misspecified multivariable biomarker threshold models when the likelihood function is not continuously differentiable. Our proposed continuous threshold model gives guidance on personalized treatment when there is a subset effect in the biomarker. Our model can be expressed as the Cox hinge and segmented models as Fong et al. (2016) do in the logistic model. The proof of asymptotic results provides theoretical justification for the use of deep learning method in Cox proportional hazards model for survival data (Katzman et al., 2018). We have made the source code of the proposed method available as part of the biomarker threshold models R package `bhm` (https://CRAN.R-project.org/package=bhm

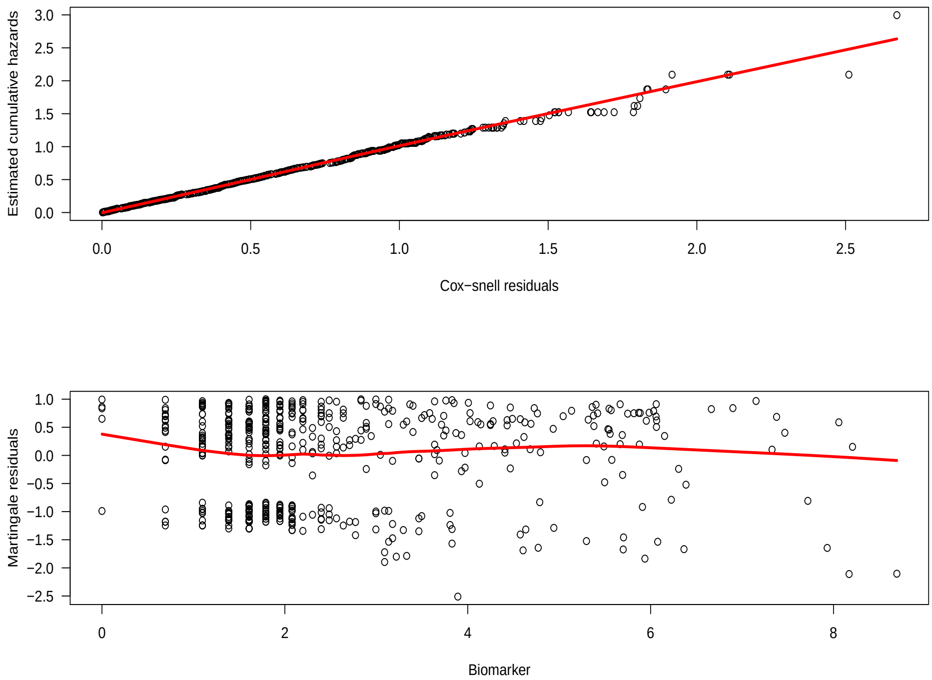

FIGURE 2: Cox-Snell residuals plot and Martingale residuals plot of the fitted model.

and https://github.com/statapps/bhm).

We use the shifted softplus function in the misspecified model in numerical simulations. The softplus function, defined as $f(x) = \log(1 + \exp(x))$, is a similar function to the ReLU function, and is popular in deep learning as an activation function. Moreover, the derivative of the softplus function is $f'(x) = \frac{e^x}{e^x+1} = \frac{1}{1+e^{-x}}$, i.e., the logistic function. Thus, the properties of this continuously differentiable function in the proportional hazards model may be investigated for further study. However, the softplus function has no advantage of sparsity in high-dimensional model and requires exponential evaluations, which is computationally expensive. In the future, we may study the trade-off between these functions to achieve continuous differentiation and computational efficiency.

The regularity conditions require the existence of a change point c and a non-zero regression coefficient $\boldsymbol{\gamma}$ for the ReLU function. It is of interest to test the hypothesis that either c does not exist or $\boldsymbol{\gamma} = \mathbf{0}$. Since the model is not identifiable under this null hypothesis, this will be an interesting future research topic.

ACKNOWLEDGEMENTS

The authors wish to thank both the associate editor and the two reviewers for their insightful comments and suggestions. This work was supported by the grant from the Natural Sciences and Engineering Research Council of Canada. The computation was made possible by the facilities of Compute/Calcul Canada.

BIBLIOGRAPHY

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012). Statistical models based on counting processes. Springer Science & Business Media.

Andersen, P. K., & Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4), 1100-1120.

Andrews, D., & Herzberg, A. (1985). Data. New York (NY) : Springer.

Byar, D. P., & Corle, D. K. (1977). Selecting optimal treatment in clinical trials using covariate information. *Journal of Chronic Diseases*, 30(7), 445459. http://doi.org/10.1016/0021-9681(77)90037-6

Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438), 477-489.

Chen B. E. (2020). A Package for Biomarker Threshold Models. R package version 1.15, URL: https://CRAN.R-project.org/package=bhm.

Chen, B. E., Jiang, W., & Tu, D. (2014). A hierarchical Bayes model for biomarker subset effects in clinical trials. *Computational Statistics & Data Analysis*. 71. 324334. 10.1016/j.csda.2013.05.015.

Chen, D., Li, J., & Chong, J. K. (2017). Hazards regression for freemium products and services: a competing risks approach. *Journal of Statistical Computation and Simulation*, 87(9), 1863-1876.

Chiu, G., Lockhart, R., and Routledge, R. (2006). Bent-cable regression theory and applications. *Journal of American Statistics Association*, 101:542-553.

Fang, T., Mackillop, W., Jiang, W., Hildesheim, A., Wacholder, S., & Chen, B. E. (2017). A Bayesian method for risk window estimation with application to HPV vaccine trial. *Computational Statistics & Data Analysis*, 112, 5362.

Fong, Y., Di, C., Huang, Y., & Gilbert, P. B. (2016). Model-robust inference for continuous threshold regression models. *Biometrics*, 73(2), 452462. doi:10.1111/biom.12623

Gutman, E. B., Sproul, E. E., & Gutman, A. B. (1936). Significance of increased phosphatase activity of bone at the site of osteoplastic metastases secondary to carcinoma of the prostate gland. *The American Journal of Cancer*, 28(3), 485-495.

Hahnloser, R.H., Sarpeshkar, R., Mahowald, M.A., Douglas, R.J., & Seung, H.S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405, 947-951.

Hansen, B. E. (2017). Regression kink with an unknown threshold. *Journal of Business & Economic Statistics* 35(2), 228240.

He, Y., Lin, H., & Tu, D. (2018). A singleindex threshold Cox proportional hazard model for identifying a treatmentsensitive subset based on multiple biomarkers. *Statistics in medicine*, 37(23), 3267-3279.

Hinkley, D. (1971). Inference in Two-Phase Regression. *Journal of the American Statistical Association*, 66(336), 736-743. doi:10.2307/2284220

Jensen, U., & Lütkebohmert, C. (2008). A Cox-type regression model with change-points in the covariates. *Lifetime data analysis*, 14(3), 267-285.

Jiang, W., Freidlin, B., Simon, R. (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute* 99, 10361043.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*. 18. 10.1186/s12874-018-0482-1.

Li, J., & Jin, B. (2018). Multi-threshold accelerated failure time model. *The Annals of Statistics*, 46(6A), 2657-2682.

Lin, H., Zhou, L., Peng, H., & Zhou, X. H. (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *Canadian Journal of Statistics*, 39(2), 324-343.

Liu, Y., Jiang, W., & Chen, B.E. (2015). Testing for treatment-biomarker interaction based on local partial-likelihood. *Statistics in medicine*, 34 27, 3516-30 .

Luo, X. & Boyett, J. (1997). Estimation of a threshold parameter in cox regression. *Communications in Statistics - Theory and Methods*, 26, 23292346.

Pons, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Annals of Statistics*. 31. 10.1214/aos/1051027876.

Prentice, R. L. & Self, S. G. (1983). Asymptotic distribution theory for cox-type regression models with general relative risk form. *The Annals of Statistics* 11(3), 804813.

Seo, M. H., & Linton, O. (2007). A smoothed least squares estimator for threshold regression models. *Journal of Econometrics*, 141(2), 704-735.

Spencer, A. V., Harbron, C., Mander, A., Wason, J., & Peers, I. (2016). An adaptive design for updating the threshold value of a continuous biomarker. *Statistics in Medicine*, 35(27), 49094923.

Subtil, F., & Rabilloud, M. (2014). Estimating the optimal threshold for a diagnostic biomarker in case of complex biomarker distributions. *BMC Medical Informatics and Decision Making*, 14(1).

Team, R. D. C. (2020). *R: A language and environment for statistical computing R Foundation for Statistical Computing.*

Tucker, H. G. (1959). A Generalization of the Glivenko-Cantelli Theorem. *The Annals of Mathematical Statistics*, 30(3), 828830.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

APPENDIX

Derivation of the first order derivative. The partial derivative of $X(\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial X(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = E\left[\int_0^\tau Y(t)\left\{\boldsymbol{Z_c} - \frac{s_{\boldsymbol{\beta}}^{(1)}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta})}\right\} d\Lambda(t;\boldsymbol{\theta}_0)\right],$$

where $s^{(1)}_{\boldsymbol{\beta}}(t;\boldsymbol{\theta})$ are the first $(p+p^*)$ elements of $s^{(1)}(t;\boldsymbol{\theta})$. To obtain the partial derivative of $X(\boldsymbol{\theta})$ with respect to $\boldsymbol{c}$, we rewrite

$$E\{Y(t)\boldsymbol{\gamma}^T\boldsymbol{Z}_2(\boldsymbol{W}-\boldsymbol{c})^+\} = E\{Y(t)\boldsymbol{\gamma}^T\boldsymbol{Z}_2\int_{\boldsymbol{c}}^{\infty}(\boldsymbol{w}-\boldsymbol{c})f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{w})d\boldsymbol{w}\}.$$

We then have the following partial derivative

$$\frac{\partial E\{Y(t)(\boldsymbol{\gamma}^T\boldsymbol{Z}_2)(\boldsymbol{W}-\boldsymbol{c})^+\}}{\partial \boldsymbol{c}} = -E\left\{Y(t)(\boldsymbol{\gamma}^T\boldsymbol{Z}_2)I(\boldsymbol{W}>\boldsymbol{c})\right\},$$

which is a continuous function of $\boldsymbol{c}$. Furthermore, because

$$E\left\{e^{\boldsymbol{\gamma}^T\boldsymbol{Z}_2(\boldsymbol{W}-\boldsymbol{c})^+}\right\} = E\int_{-\infty}^{\boldsymbol{c}} f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{w})d\boldsymbol{w} + E\left\{\int_{\boldsymbol{c}}^{\infty} e^{\boldsymbol{\gamma}^T\boldsymbol{Z}_2(\boldsymbol{w}-\boldsymbol{c})} f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{w})d\boldsymbol{w}\right\},$$

we have

$$\frac{\partial s^{(0)}(t;\boldsymbol{\theta})}{\partial \boldsymbol{c}} = -E\left\{Y(t)(\boldsymbol{\gamma}^T\boldsymbol{Z}_2)e^{\boldsymbol{\beta}^T\boldsymbol{Z}_{\boldsymbol{c}}}I(\boldsymbol{W}>\boldsymbol{c})\right\} = s^{(1)}_{\boldsymbol{c}}(t;\boldsymbol{\theta}),$$

where $s^{(1)}_{\boldsymbol{c}}(t;\boldsymbol{\theta})$ are the last q elements of $s^{(1)}(t;\boldsymbol{\theta})$. Therefore, $X(\boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{c}$,

$$\frac{\partial X(\boldsymbol{\theta})}{\partial \boldsymbol{c}} = E\left[\int_0^{\tau} Y(t)\left\{-\boldsymbol{\gamma}^T\boldsymbol{Z}_2 I(\boldsymbol{W}>\boldsymbol{c}) - \frac{s^{(1)}_{\boldsymbol{c}}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta})}\right\}d\Lambda(t;\boldsymbol{\theta}_0)\right].$$

Overall, the first order partial derivatives of $X(\boldsymbol{\theta})$ and $X_n(\boldsymbol{\theta})$ with respective to $\boldsymbol{\theta}$ are derived as Equations (3). ■

Derivation of the Hessian matrix. For the partition of $\boldsymbol{V}_1$ in the main text, it is obvious that $\frac{\partial^2 X(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \boldsymbol{V}_{1,\boldsymbol{\beta},\boldsymbol{\beta}}$ is the first $(p+p^*)\times(p+p^*)$ matrix of $\boldsymbol{V}_1$.

Denote the partial derivative of $s_{\boldsymbol{c}}^{(1)}(t;\boldsymbol{\theta})$ with respect to $\boldsymbol{\gamma}$ as $s_{\boldsymbol{\gamma},\boldsymbol{c}}^{*}(t;\boldsymbol{\theta})$, then

$$\begin{aligned} s_{\boldsymbol{\gamma},\boldsymbol{c}}^{*}(t;\boldsymbol{\theta}) &= -E\left\{Y(t)\boldsymbol{Z}_2 I(\boldsymbol{W}>\boldsymbol{c})(\boldsymbol{\gamma}^T\boldsymbol{Z}_2(\boldsymbol{W}-\boldsymbol{c})^+ e^{r(\boldsymbol{\theta})}+1)\right\} \\ &= s_{\boldsymbol{\gamma},\boldsymbol{c}}^{(2)}(t;\boldsymbol{\theta}) - E\left\{Y(t)\boldsymbol{Z}_2 I(\boldsymbol{W}>\boldsymbol{c})e^{r(\boldsymbol{\theta})}\right\}. \end{aligned}$$

The second order partial derivative of $X(\boldsymbol{\theta})$ with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{c}$ is $\frac{\partial^2 X(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{c}} =$

$$E\left[\int_0^\tau Y(t)\left\{-\boldsymbol{Z}_2 I(\boldsymbol{W}>\boldsymbol{c}) - \frac{s_{\boldsymbol{\gamma},\boldsymbol{c}}^{*}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta})} + \frac{s_{\boldsymbol{c}}^{(1)}(t;\boldsymbol{\theta})s_{\boldsymbol{\gamma}}^{(1)}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta})^2}\right\} d\Lambda(t;\boldsymbol{\theta}_0)\right],$$

which can be written as $\boldsymbol{V}_{1,\boldsymbol{\gamma},\boldsymbol{c}} + \boldsymbol{V}_{2,\boldsymbol{\gamma},\boldsymbol{c}}$, where $\boldsymbol{V}_{1,\boldsymbol{\gamma},\boldsymbol{c}}$ is defined as a block of $\boldsymbol{V}_1$ and $\boldsymbol{V}_{2,\boldsymbol{\gamma},\boldsymbol{c}} =$

$$-E\left[\int_0^\tau Y(t)\left\{\boldsymbol{Z}_2 I(\boldsymbol{W}>\boldsymbol{c}) - \frac{E(Y(t)\boldsymbol{Z}_2 I(\boldsymbol{W}>\boldsymbol{c})e^{r(\boldsymbol{\theta})}}{s^{(0)}(t;\boldsymbol{\theta})}\right\} d\Lambda(t;\boldsymbol{\theta}_0)\right].$$

Denote the partial derivative of $s_{\boldsymbol{c}}^{(1)}(t;\boldsymbol{\theta})$ with respect to $\boldsymbol{c}$ as $s_{\boldsymbol{c},\boldsymbol{c}}^{*}(t;\boldsymbol{\theta})$, then

$$\begin{aligned} s_{\boldsymbol{c},\boldsymbol{c}}^{*}(t;\boldsymbol{\theta}) &= E\left[Y(t)\boldsymbol{\gamma}^T\boldsymbol{Z}_2\left\{f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{c}) + \boldsymbol{\gamma}^T\boldsymbol{Z}_2 I(\boldsymbol{W}>\boldsymbol{c})\right\}e^{r(\boldsymbol{\theta})}\right] \\ &= s_{\boldsymbol{c}\boldsymbol{c}}^{(2)}(t;\boldsymbol{\theta}) + E\left\{Y(t)\boldsymbol{\gamma}^T\boldsymbol{Z}_2 f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{c})e^{r(\boldsymbol{\theta})}\right\}. \end{aligned}$$

We have that the second order partial derivative with respect to $\boldsymbol{c}$ is

$$\frac{\partial^2 X(\boldsymbol{\theta})}{\partial\boldsymbol{c}^2} = E\left[\int_0^\tau Y(t)\left\{\boldsymbol{\gamma}^T\boldsymbol{Z}_2 f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{c}) - \frac{s_{\boldsymbol{c},\boldsymbol{c}}^{*}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta})} + \frac{s_{\boldsymbol{c}}^{(1)}(t;\boldsymbol{\theta})^2}{s^{(0)}(t;\boldsymbol{\theta})^2}\right\} d\Lambda(t;\boldsymbol{\theta}_0)\right],$$

which can be written as $\boldsymbol{V}_{1,\boldsymbol{c},\boldsymbol{c}} + \boldsymbol{V}_{2,\boldsymbol{c},\boldsymbol{c}}$, where $\boldsymbol{V}_{1,\boldsymbol{c},\boldsymbol{c}}$ is defined as a block of $\boldsymbol{V}_1$ and $\boldsymbol{V}_{2,\boldsymbol{c},\boldsymbol{c}} =$

$$E\left[\int_0^\tau Y(t)\left\{(\boldsymbol{\gamma}^T\boldsymbol{Z}_2)f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{c}) - \frac{E(Y(t)\boldsymbol{\gamma}^T\boldsymbol{Z}_2 f_{\boldsymbol{W}|\boldsymbol{Z}_2}(\boldsymbol{c})e^{r(\boldsymbol{\theta})})}{s^{(0)}(t;\boldsymbol{\theta})}\right\} d\Lambda(t;\boldsymbol{\theta}_0)\right].$$

Finally, we have $\boldsymbol{V} = \ddot{X}(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{V}_1 + \boldsymbol{V}_2$. ■

Proof of Theorem 1. The strict concaveness of $X(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ within a local neighbourhood of $\boldsymbol{\theta}_0$ can be proved by showing that the negative Hessian matrix is strictly positive definite. When the model is correctly specified, it is obvious that $\boldsymbol{V}_2 = \mathbf{0}$, therefore it is sufficient to show that $-\boldsymbol{V}_1$ is strictly positive definite.

Denote $\bar{\boldsymbol{Z}}(\boldsymbol{\theta}) = \frac{s^{(1)}(t;\boldsymbol{\theta})}{s^{(0)}(t;\boldsymbol{\theta})} = \frac{E\{Y(t)\boldsymbol{Z}(\boldsymbol{\theta})e^{r(\boldsymbol{\theta})}\}}{E\{Y(t)e^{r(\boldsymbol{\theta})}\}}$, then we have

$$\begin{aligned}\frac{s^{(2)}(t;\boldsymbol{\theta})s^{(0)}(t;\boldsymbol{\theta})-s^{(1)}(t;\boldsymbol{\theta})s^{(1)}(t;\boldsymbol{\theta})^T}{(s^{(0)}(t;\boldsymbol{\theta}))^2} &= \frac{E\{Y(t)\boldsymbol{Z}(\boldsymbol{\theta})\boldsymbol{Z}(\boldsymbol{\theta})^T e^{r(\boldsymbol{\theta})}\}}{E\{Y(t)e^{r(\boldsymbol{\theta})}\}} - \bar{\boldsymbol{Z}}(\boldsymbol{\theta})\bar{\boldsymbol{Z}}(\boldsymbol{\theta})^T \\ &= \frac{E\{Y(t)(\boldsymbol{Z}(\boldsymbol{\theta})-\bar{\boldsymbol{Z}}(\boldsymbol{\theta}))(\boldsymbol{Z}(\boldsymbol{\theta})-\bar{\boldsymbol{Z}}(\boldsymbol{\theta}))^T e^{r(\boldsymbol{\theta})}\}}{E\{Y(t)e^{r(\boldsymbol{\theta})}\}} \\ &= E\{w(\boldsymbol{Z}(\boldsymbol{\theta}) - \bar{\boldsymbol{Z}}(\boldsymbol{\theta}))(\boldsymbol{Z}(\boldsymbol{\theta}) - \bar{\boldsymbol{Z}}(\boldsymbol{\theta}))^T\} \\ &= \bar{V}(\boldsymbol{\theta}),\end{aligned}$$

where $\bar{V}(\boldsymbol{\theta})$ is the weighted covariance and therefore positive definite with weight $w = \frac{Y(t)e^{r(\boldsymbol{\theta})}}{E\{Y(t)e^{r(\boldsymbol{\theta})}\}}$. By condition C4, for any vector $\boldsymbol{a} \neq \mathbf{0}$, $P(\boldsymbol{a}^T\boldsymbol{Z}(\boldsymbol{\theta}) \neq 0) = 1$, so $\bar{V}(\boldsymbol{\theta})$ must be strictly positive definite, which implies $-\boldsymbol{V}_1$ is strictly positive definite.

■

Proof of Lemma 2. For each $1 \leq i \leq n$, we have $r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0) =$

$$\begin{aligned}&(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{Z}_{1i} + \boldsymbol{\gamma}^T\boldsymbol{Z}_{2i}(\boldsymbol{W}_i - \boldsymbol{c})^+ - \boldsymbol{\gamma}_0^T\boldsymbol{Z}_{2i}(\boldsymbol{W}_i - \boldsymbol{c}_{\mathbf{0}})^+ \\ &= (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{Z}_{1i} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T\boldsymbol{Z}_{2i}(\boldsymbol{W}_i - \boldsymbol{c}_{\mathbf{0}})^+ - \boldsymbol{\gamma}_0^T\boldsymbol{Z}_{2i}\boldsymbol{I}(\boldsymbol{W}_i > \boldsymbol{c}_0)(\boldsymbol{c} - \boldsymbol{c}_0) \\ &\quad + \boldsymbol{\gamma}_0^T\boldsymbol{Z}_{2i}\{\boldsymbol{I}(\boldsymbol{W}_i > \boldsymbol{c}_0)(\boldsymbol{c} - \boldsymbol{c}_0) + \boldsymbol{\gamma}^T\boldsymbol{Z}_{2i}\{(\boldsymbol{W}_i - \boldsymbol{c})^+ - (\boldsymbol{W}_i - \boldsymbol{c}_{\mathbf{0}})^+\} \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\boldsymbol{Z}_i(\boldsymbol{\theta}_0) + \boldsymbol{\gamma}_0^T\boldsymbol{Z}_{2i}\{\boldsymbol{I}(\boldsymbol{W}_i > \boldsymbol{c}_0)(\boldsymbol{c} - \boldsymbol{c}_0) + ((\boldsymbol{W}_i - \boldsymbol{c})^+ - (\boldsymbol{W}_i - \boldsymbol{c}_{\mathbf{0}})^+)\} \\ &\quad + (\boldsymbol{\gamma}^T - \boldsymbol{\gamma}_0^T)\boldsymbol{Z}_{2i}\{(\boldsymbol{W}_i - \boldsymbol{c})^+ - (\boldsymbol{W}_i - \boldsymbol{c}_{\mathbf{0}})^+\}.\end{aligned}$$

For each of $j = 1, \ldots, q$, by simple algebra, we can rewrite $(w_{ij} - c_j)^+ - (w_{ij} - c_{0j})^+$ as

$$-I(w_{ij} > c_{0j})(c_j - c_{0j}) + (c_j - w_{ij})\{I(w_{ij} > c_{0j}) - I(w_{ij} > c_j)\}.$$

Therefore, the difference between $(r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0))$ and $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{Z}_i(\boldsymbol{\theta}_0)$ is

$$\begin{aligned}
&-(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T \boldsymbol{Z}_{2i} \boldsymbol{I}(\boldsymbol{W}_i > \boldsymbol{c_0})(\boldsymbol{c} - \boldsymbol{c_0}) \\
&+\boldsymbol{\gamma}_0^T \boldsymbol{Z}_{2i}(\boldsymbol{c} - \boldsymbol{w}_i)\{I(\boldsymbol{W}_i > \boldsymbol{c}_0) - I(\boldsymbol{W}_i > \boldsymbol{c})\} \\
&+(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T \boldsymbol{Z}_{2i}(\boldsymbol{c} - \boldsymbol{w}_i)\{I(\boldsymbol{W}_i > \boldsymbol{c}_0) - I(\boldsymbol{W}_i > \boldsymbol{c})\}.
\end{aligned}$$

Since $I(w_{ij} > c_{0j}) - I(w_{ij} > c_j)$ is non-zero only when $\min(c_j, c_{0j}) < w_{ij} < \max(c_j, c_{0j})$, the third term above is of the order $O_p(||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \times ||\boldsymbol{c} - \boldsymbol{c}_0|| \times ||\boldsymbol{c} - \boldsymbol{c}_0||) = o_p(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2)$. By the law of large numbers, we have that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Y_i(t)\boldsymbol{Z}_{2ij}(c_j - w_{ij})\{I(w_{ij} > c_{0j}) - I(w_{ij} > c_j)\} \\
=&\frac{1}{n}\sum_{i=1}^{n} Y_i(t)\boldsymbol{Z}_{2ij}(c_j - c_{0j})^2 \frac{c_j - w_{ij}}{c_j - c_{0j}} \frac{I(w_{ij} > c_{0j}) - I(w_{ij} > c_j)}{c_j - c_{0j}}
\end{aligned}$$

converges in probability to

$$\frac{1}{2}(c_j - c_{0j})^2 E\{Y(t)\boldsymbol{Z}_{2j} f_{w_j|\boldsymbol{Z}_{2j}}(c_{0j})\} + o_p(|c_j - c_{0j}|^2),$$

where $\boldsymbol{Z}_{2j}$ is the random variable that generates the independent identical distributed observations $\boldsymbol{Z}_{2ij}$. Taking the average summation, the difference of approximation (4) is of the same order as a $o_p(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2)$ term. The approximation of $S^{(0)}(t,\boldsymbol{\theta}) - S^{(0)}(t,\boldsymbol{\theta}_0)$ can be obtained by applying the Taylor expansion of

$$e^{r_i(\boldsymbol{\theta})} - e^{r_i(\boldsymbol{\theta}_0)} = e^{r_i(\boldsymbol{\theta}_0)} \left\{ (r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0)) + \frac{1}{2}(r_i(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}_0))^2 \right\} + o_p(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2).$$

The additional matrix for the second order term in (5) is defined by $S^{(2)*}(t,\boldsymbol{\theta}_0) =$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(t) \begin{bmatrix} 0_{p\times p} & 0_{p\times p^*} & 0_{p\times q} \\ 0_{p^*\times p} & 0_{p^*\times p^*} & -\boldsymbol{Z}_{2i} I(\boldsymbol{W}_i > \boldsymbol{c_0}) \\ 0_{q\times p} & -(\boldsymbol{Z}_{2i} I(\boldsymbol{W}_i > \boldsymbol{c_0}))^T & \boldsymbol{\gamma}_0^T \boldsymbol{Z}_{2i} \hat{f}_{\boldsymbol{W}|\boldsymbol{Z}_{2i}}(\boldsymbol{c}_0) \end{bmatrix} \times e^{r_i(\boldsymbol{\theta}_0)}$$

For a single-index model, using similar method, we can derive the approximation (5) with the matrix in the above formula replaced by

$$\begin{bmatrix} 0_{p\times p} & 0_{p\times p_1} & 0_{p\times q} & 0_{p\times 1} \\ 0_{p_1\times p} & 0_{p_1\times p_1} & \boldsymbol{Z}_{2i1}\boldsymbol{W}_i^T I(\boldsymbol{\eta}_0^T\boldsymbol{W}_i > c_0) & -\boldsymbol{Z}_{2i1} I(\boldsymbol{\eta}_0^T\boldsymbol{W}_i > c_0) \\ 0_{q\times p} & \boldsymbol{W}_i\boldsymbol{Z}_{2i1}^T I(\boldsymbol{\eta}_0^T\boldsymbol{W}_i > c_0) & \boldsymbol{\gamma}_0^T\boldsymbol{Z}_{2i1}\boldsymbol{W}_i\boldsymbol{W}_i^T \hat{f}_{\boldsymbol{\eta}_0^T\boldsymbol{W}|\boldsymbol{Z}_{2i1}}(c_0) & -\boldsymbol{W}_i\boldsymbol{Z}_{2i1}^T\gamma_0 \hat{f}_{\boldsymbol{\eta}_0^T\boldsymbol{W}|\boldsymbol{Z}_{2i1}}(c_0) \\ 0_{1\times p} & -\boldsymbol{Z}_{2i1}^T I(\boldsymbol{\eta}_0^T\boldsymbol{W}_i > c_0) & -\boldsymbol{\gamma}_0^T\boldsymbol{Z}_{2i1}\boldsymbol{W}_i^T \hat{f}_{\boldsymbol{\eta}_0^T\boldsymbol{W}|\boldsymbol{Z}_{2i1}}(c_0) & \boldsymbol{\gamma}_0^T\boldsymbol{Z}_{2i1} \hat{f}_{\boldsymbol{\eta}_0^T\boldsymbol{W}|\boldsymbol{Z}_{2i1}}(c_0) \end{bmatrix}$$

■

Proof of Theorem 2. For any random sequence $\tilde{\boldsymbol{\theta}}$ that is $n^{1/2}$ consistent, let $\tilde{\boldsymbol{u}} = n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Since $\boldsymbol{V}_n$ converges to $\boldsymbol{V}$ in probability, together with Lemma 3, we have

$$nX_n(\tilde{\boldsymbol{\theta}}) = \tilde{\boldsymbol{u}}^T n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0) + \frac{1}{2}\tilde{\boldsymbol{u}}^T\boldsymbol{V}\tilde{\boldsymbol{u}} + o_p(1).$$

Both the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \boldsymbol{V}^{-1}\dot{X}_n(\boldsymbol{\theta}_0)$ are $n^{1/2}$ consistent, hence

$$nX_n(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{u}}^T n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0) + \tfrac{1}{2}\hat{\boldsymbol{u}}^T\boldsymbol{V}\hat{\boldsymbol{u}} + o_p(1),$$
$$nX_n(\tilde{\boldsymbol{\theta}}) = -\tfrac{1}{2}\{n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0)\}^T\boldsymbol{V}^{-1}\{n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0)\} + o_p(1).$$

Because $\hat{\boldsymbol{\theta}}$ maximizes $X_n(\boldsymbol{\theta})$, with a difference of an $o_p(1)$ term, we have

$$\hat{\boldsymbol{u}}^T n^{1/2} \dot{X}_n(\boldsymbol{\theta}_0) + \frac{1}{2}\hat{\boldsymbol{u}}^T \boldsymbol{V} \hat{\boldsymbol{u}} + \frac{1}{2}\{n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0)\}^T \boldsymbol{V}^{-1}\{n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0)\} + o_p(1) \geq 0.$$

Completing the square gives

$$\frac{1}{2}\{\hat{\boldsymbol{u}} + \boldsymbol{V}^{-1} n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0)\}^T \boldsymbol{V}\{\hat{\boldsymbol{u}} + \boldsymbol{V}^{-1} n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0)\} + o_p(1) \geq 0.$$

By regularity conditions C3-C4 and Theorem 1, $-\boldsymbol{V}$ is strictly positive-definite, the quadratic form must converge to 0 in probability. Hence we have

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\boldsymbol{V}^{-1} n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0) + o_p(1).$$

Theorem 2 follows the Martingale central limit theorem of

$$n^{1/2}\dot{X}_n(\boldsymbol{\theta}_0) = n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\left\{\boldsymbol{Z}_i - \frac{S^{(1)}(T_i, \boldsymbol{\theta}_0)}{S^{(0)}(T_i, \boldsymbol{\theta}_0)}\right\} dM_i(t),$$

which follows a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{V}^{-1}\boldsymbol{\Gamma}\boldsymbol{V}^{-1}$. ■

Received 7 May 2019

Accepted 29 March 2020