Published in final edited form as:

Stat Med. 2013 December 10; 32(28): 4980–4994. doi:10.1002/sim.5885.

A General Semiparametric Hazards Regression Model: Efficient Estimation and Structure Selection

Xingwei Tong, Liang Zhu, Chenlei Leng*, Wendy Leisenring, and Leslie L. Robison

Abstract

We consider a general semiparametric hazards regression model that encompasses Cox's proportional hazards model and the accelerated failure time model for survival analysis. To overcome the nonexistence of the maximum likelihood, we derive a kernel-smoothed profile likelihood function, and prove that the resulting estimates of the regression parameters are consistent and achieve semiparametric efficiency. In addition, we develop penalized structure selection techniques to determine which covariates constitute the accelerated failure time model and which covariates constitute the proportional hazards model. The proposed method is able to estimate the model structure consistently and model parameters efficiently. Furthermore, variance estimation is straightforward. The proposed estimation performs well in simulation studies and is applied to the analysis of a real data set. Copyright

Keywords

Accelerated failure time model; Cox's proportional hazards model; Efficiency; Kernel-smoothed profile likelihood function; Model selection; Penalized likelihood

1. Introduction

The proportional hazards model proposed by Cox ([1], [2]) and the accelerated failure time model (AFT) are the two major approaches to analyze censored data in survival studies [3]. Cox's model specifies that the covariates act multiplicatively on the hazard function, while the AFT model specifies that the effect of the covariates is multiplicative on time. Clearly, the covariates in these two models provide different measures of the effect on survival.

Efficient estimation and inference for the proportional hazard model are based on the partial likelihood that is easy to implement and gives attractive statistical properties [3]. On the other hand, efficient inference for the AFT model with unspecified error distribution is much more difficult ([4], [5], [6]). Jin et al. [7] proposed to use linear programming for the rank-based estimation method that can be suboptimal in terms of efficiency. Zeng and Lin [8] proposed an approximate nonparametric maximum likelihood method via maximizing a kernel-smoothed profile likelihood function, and showed that this estimator is efficient. Lu [9] further studied this model with a cure fraction.

In practice, when given a dataset, we need to choose whether to use Cox's model or the AFT model. A usual technique would be to fit one of the two models and then to conduct some model testing to assess the lack-of-fit. However, due to the finite sample size and other data characteristics, the ability to check the model assumptions may be limited. In addition, it is

not unusual that both Cox's model and the AFT model may fit the data reasonably well, if appropriate time-dependent covariates are included [8].

Instead of deciding which model to fit, we consider a general semiparametric hazard model in this paper first studied by Chen and Jewell [10]. This model encompasses Cox's and the AFT model as special cases, and is more flexible in modeling survival data. Chen and Jewell used the estimating equations method for inference that is not efficient in general. Numerical challenges also arise due to the non-smoothness of Chen and Jewell's estimating equations. Chen and Jewell confined application of their model to low dimensional cases maybe due to the computational complexity. We, on the other hand, propose a kernel-smoothed profile likelihood function aiming at maximum likelihood estimation, motivated by Zeng and Lin [8]. The maximization can be easily implemented using any optimization algorithm developed for smooth objective functions, such as the Newton-Raphson algorithm. We show that the resulting estimators are asymptotically normal and attain the semiparametric efficiency bound. The limiting covariance matrix of the estimates can be readily computed as an output of the algorithm. More importantly, we develop penalized model selection techniques to determine which covariates constitute the accelerate failure time model and which covariates contribute to the proportional hazards model. The resulting estimators are shown to have the oracle property ([11], [12]). Namely, the final chosen model correctly distinguishes Cox covariates and AFT covariates with probability tending to one, and the resulting estimates are as efficient as if we were to use the true model structure for estimation and inference. Zhang and Lu [12] studied variable selection for the Cox model while Cai et al. [13] studied it for the AFT model. See also Lu and Zhang [14] and Johnson [15]. Our model selection has an added implication in terms of distinguishing these two kinds of covariates.

The rest of the paper is organized as follows. We discuss the general semiparametric hazards model in Section 2 and discuss using a smoothed profile likelihood function for estimation. Section 3 gives the large sample properties of the proposed estimators. We study penalized model selection techniques in Section 4. Simulation studies are presented in Section 5, followed by an analysis of Childhood Cancer Survivor study data in Section 6. We make some concluding remarks in Section 7. All the proofs are relegated to the appendix.

2. A general hazards model

Let T denote the failure time of interest and X be the p -dimensional covariate that is time invariant. If T follows the proportional hazards model, then the conditional hazards function of T given X has the form

$$\lambda_{T|X}(t) = \lambda_0(t) \exp(X' \beta), \quad (1)$$

where β is the regression parameter and $\lambda_0(t)$ is an unspecified baseline hazard function. If T follows the accelerated failure time model, then T satisfies that

$$\log T = -X' \beta + \varepsilon,$$

where ε is a measurement error independent of X . If we denote the cumulative hazards function of ε as Λ_ε and set $\Lambda_0(\cdot) = \Lambda_\varepsilon(\log(\cdot))$, the cumulative hazards function of T given X in the AFT model takes the form

$$\Lambda_{T|X}(t) = \Lambda_0(\text{texp}(X' \beta)).$$

Differentiating both sides of the above equality, one obtains the conditional hazards function of T given X

$$\lambda_{T|X}(t) = \lambda_0(\text{texp}(X' \beta)) \exp(X' \beta). \quad (2)$$

In this paper, we explore a general hazards regression model which includes Cox's model and the AFT model as special cases. We assume that the conditional hazards function of T given X takes the form

$$\lambda_{T|X}(t) = \lambda_0(\text{texp}(X' \beta_1)) \exp(X' \beta_1) \exp(X' \beta_2), \quad (3)$$

where β_1 and β_2 are two regression parameters. Alternatively, this general hazards model implies that the failure time T follows the model

$$\log T = -X' \beta_1 + \varepsilon, \quad (4)$$

where ε follows the proportional hazards model with its cumulative hazards function given by $\Lambda_0(t) \exp(X' \beta_2)$.

The two parameters, β_1 and β_2 , measure two different effects of the covariate on the survival time. The first parameter β_1 can be interpreted as the rate that the covariate either accelerates or degrades the rate at which an individual proceeds along the time line. The second parameter β_2 identifies multiplicative effects of the covariate on the hazard function after adjusting for β_1 . This model reduces to the accelerated failure time model when $\beta_2 = 0$, and it becomes the proportional hazards model when $\beta_1 = 0$. For a covariate X_j , if $\beta_{1j} = \beta_{2j} = 0$, and this variable has no effects on the survival time; if $\beta_{1j} = 0$ and $\beta_{2j} \neq 0$, this variable is a Cox covariate; if $\beta_{1j} \neq 0$ and $\beta_{2j} = 0$, this variable is an AFT variable; and if $\beta_{1j} \neq 0$ and $\beta_{2j} \neq 0$, this variable is a mixed variable that contributes to both models. Clearly, this general hazards model is broader than the two models and is more flexible. The structure of the model raises two important questions: one is whether efficient estimation of β_1 and β_2 is possible, and the other is which covariates play the role in Cox's model and which covariates play the role in the accelerated failure time model.

Let C be the censoring time and τ be the largest study time. Suppose that we have a data set with n subjects. For the i th subject, we observe the censored event time $Y_i = \min(C_i, T_i)$, the censoring indicator $\delta_i = I(T_i \leq C_i)$ and covariate X_i , where $I(\cdot)$ is the indicator function. We assume that C_i is independent of T_i conditional on X_i and that its distribution does not depend on $\beta = (\beta_1', \beta_2')$. Then given i.i.d observations $O_i = (Y_i, \delta_i, X_i)$, $i = 1, \dots, n$, from $O = (Y, \delta, X)$, the log-likelihood function of β and λ is given by

$$\frac{1}{n} \sum_{i=1}^n \left[\delta_i X_i' (\beta_1 + \beta_2) + \delta_i \log \lambda(Y_i \exp(X_i' \beta_1)) - \exp(X_i' \beta_2) \Lambda(Y_i \exp(X_i' \beta_1)) \right]. \quad (5)$$

Generally, the maximum likelihood estimation can be obtained by maximizing the above log-likelihood function. An argument similar to that in Zeng and Lin [8] shows, however,

that the maximum of (5) does not exist, because the estimator of Λ is very non-smooth. To overcome this difficulty, we use a piecewise constant function to approximate λ . That is, we divide the interval that contains all the possible values $Y_i \exp(X_i' \beta_1)$ into K_n equally space intervals and then assumes that the function λ is a constant on each interval. More specifically, we write

$$\lambda(t) = \sum_{k=1}^{K_n} \alpha_k I(t \in [t_{k-1}, t_k]),$$

where $0 = t_0 < t_1 < \dots < t_{K_n} = \eta$, $\eta = \sup_{i, \beta_1 \in \mathcal{B}_1} |Y_i \exp(X_i' \beta_1)|$. Here $\mathcal{B}_1$ is a compact subset of R^p , $\alpha_1, \dots, \alpha_{K_n}$ are the corresponding unknown constants on the intervals. Then the cumulative hazards function Λ can be written as

$$\Lambda(t) = \sum_{k=1}^{K_n} \alpha_k (t - t_k) I(t \in [t_{k-1}, t_k]) + \frac{\eta}{K_n} \sum_{k=1}^{K_n} \alpha_k I(t \geq t_{k-1}).$$

Therefore the log-likelihood function (5) can be re-expressed as

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \delta_i X_i' (\beta_1 + \beta_2) + \frac{1}{n} \sum_{k=1}^{K_n} \log \alpha_k \left[\sum_{i=1}^n \delta_i I(R_i(\beta_1) \in [t_{k-1}, t_k]) \right] - \frac{1}{n} \sum_{k=1}^{K_n} \alpha_k \\ & \times \left[\sum_{i=1}^n \exp(X_i' \beta_2) (R_i(\beta_1) - t_k) I(R_i(\beta_1) \in [t_{k-1}, t_k]) + \frac{\eta}{K_n} \sum_{i=1}^n \exp(X_i' \beta_2) I(R_i(\beta_1) \geq t_{k-1}) \right], \end{aligned} \quad (6)$$

where $R_i(\beta_1) = Y_i \exp(X_i' \beta_1)$.

Differentiating with respect to α_k , one obtain that

$$\begin{aligned} \alpha_k = & \sum_{i=1}^n \delta_i I(R_i(\beta_1) \in [t_{k-1}, t_k]) \left[\sum_{i=1}^n \exp(X_i' \beta_2) (R_i(\beta_1) \right. \\ & \left. - t_k) I(R_i(\beta_1) \in [t_{k-1}, t_k]) + \frac{\eta}{K_n} \sum_{i=1}^n \exp(X_i' \beta_2) I(R_i(\beta_1) \geq t_{k-1}) \right]^{-1} \end{aligned}$$

Therefore, the profile log-likelihood function of β is

$$\begin{aligned} l_n^p(\beta) = & \frac{1}{n} \sum_{i=1}^n \delta_i X_i' (\beta_1 + \beta_2) + \sum_{k=1}^{K_n} \left[\frac{1}{n} \sum_{i=1}^n \delta_i I(R_i(\beta_1) \in [t_{k-1}, t_k]) \right] \\ & \times \log \left[\frac{K_n}{n\eta} \sum_{j=1}^n \delta_j I(R_j(\beta) \in [t_{k-1}, t_k]) \right] \\ & - \sum_{k=1}^{K_n} \left[\frac{1}{n} \sum_{i=1}^n \delta_i I(R_i(\beta_1) \in [t_{k-1}, t_k]) \right] \times \log \left[\frac{K_n}{n\eta} \sum_{i=1}^n \exp(X_i' \beta_2) (R_i(\beta_1) - t_k) \right. \\ & \left. \times I(R_i(\beta_1) \in [t_{k-1}, t_k]) + \frac{1}{n} \sum_{i=1}^n \exp(X_i' \beta_2) I(R_i(\beta_1) \geq t_{k-1}) \right]. \end{aligned} \quad (7)$$

Since the profile log-likelihood function is not smooth due to the many indicator functions in (7). As a result, direct maximization is difficult or impossible. As $n \rightarrow \infty$, $K_n \rightarrow \infty$, $K_n/n \rightarrow 0$, we have

$$l_n^p(\beta) \rightarrow l(\beta), \quad \text{uniformly for } \beta \in \mathcal{B},$$

where

$$l(\beta) = E[\delta X'(\beta_1 + \beta_2)] + \delta \log \left[\frac{dP(\delta=1, R(\beta_1) \leq t)/dt}{E \exp(X' \beta_2) I(R(\beta_1) \geq t)} \Big|_{t=R(\beta_1)} \right].$$

This motivates to use the kernel-smoothed approximation to $l(\beta)$, given by

$$l_n^s(\beta) = \frac{1}{n} \sum_{i=1}^n \delta_i X_i'(\beta_1 + \beta_2) + \frac{1}{n} \sum_{i=1}^n \delta_i \log \left[\frac{\frac{1}{nh} \sum_{j=1}^n \delta_j K((R_j(\beta_1) - R_i(\beta_1))/h)}{\frac{1}{n} \sum_{j=1}^n \exp(X_j' \beta_2) \int_{-\infty}^{(R_j(\beta_1) - R_i(\beta_1))/h} K(s) ds} \right], \quad (8)$$

where $h = h(n)$ is a bandwidth parameter converging to zero for $n \rightarrow \infty$.

The proposed estimator $\hat{\beta}_n = (\hat{\beta}_{1n}', \hat{\beta}_{2n}')'$ was obtained by maximizing the kernel-smoothed likelihood function $l_n^s(\beta)$ over $\beta \in \mathcal{B}$ for a compact set $\mathcal{B} \in \mathbb{R}^{2p}$. Given $\hat{\beta}_n$, the cumulative hazard function can be estimated as

$$\hat{\Lambda}_n(t) = \int_0^t \frac{\frac{1}{nh} \sum_{i=1}^n \delta_i K((R_i(\hat{\beta}_{1n}) - s)/h)}{\frac{1}{n} \sum_{i=1}^n \exp(X_i' \hat{\beta}_{2n}) \int_{-\infty}^{(R_i(\hat{\beta}_{1n}) - s)/h} K(u) du} ds, \quad (9)$$

and the corresponding baseline function can be estimated by

$$\hat{\lambda}_n(t) = \frac{\frac{1}{nh} \sum_{i=1}^n \delta_i K((R_i(\hat{\beta}_{1n}) - t)/h)}{\frac{1}{n} \sum_{i=1}^n \exp(X_i' \hat{\beta}_{2n}) \int_{-\infty}^{(R_i(\hat{\beta}_{1n}) - t)/h} K(u) du}.$$

These are kernel-smoothed estimates of the hazard functions motivated by the need to smooth the likelihood.

For bandwidth selection, we can use cross-validation such as the K-fold cross-validation approach. An application of this method in the current setting is to divide the data into K equal-sized groups. Let D_k denote the k th subgroup of data, then the k th prediction error is given by

$$PE_k(h) = \sum_{i \in D_k} [N_i(\tau) - \int_0^\tau I(Y_i \geq t) \exp(X_i' \beta_2^{(-k)}) d\hat{\Lambda}_n(t \exp(X_i' \hat{\beta}_1^{(-k)}))]^2$$

for $k = 1, \dots, K$, where $\hat{\beta}_1^{(-k)}$ and $\hat{\beta}_2^{(-k)}$ are the estimators of β_1 and β_2 based on the data without the subgroup D_k . The optimal bandwidth is the minimizer of the total prediction

$$\text{error } PE(h) = \sum_{k=1}^K PE_k(h).$$

3. Large sample property

In this section, we study the asymptotic behavior of the proposed estimators. The regularity conditions and proofs are given in the Appendix. First, we present the consistent result.

Denote the true value of β as $\beta_0 = (\beta_{10}', \beta_{20}')'$

Theorem 1

Suppose that Conditions A1–A6 in the Appendix hold. If $h = O(n^\nu)$ with $-1/2 < \nu < 0$, then $\hat{\beta}_n \rightarrow \beta_0$, a.s..

Theorem 1 states the strong consistency of the proposed estimator $\hat{\beta}_n$. Intuitively, this can be seen from the uniform convergence of $l_n^s(\beta)$ over $\beta \in \mathcal{B}$ and the fact that β_0 uniquely maximizes the limiting function of $l_n^s(\beta)$. For the asymptotic normality of the estimates, we have the following result.

Theorem 2

Suppose that Conditions A1–A6 in the Appendix hold. If the bandwidth satisfies $nh^{2m} \rightarrow 0$ and $nh^6 \rightarrow \infty$, then $\sqrt{n}(\hat{\beta}_n - \beta_0)$ converges weakly to a normal random variable with mean zero and covariance achieving the semi-parametric efficient bounds of β_0 . Here m is defined in condition A6 in the Appendix.

The inverse negative second derivative of $l_n^s(\hat{\beta}_n)$ can be used for estimating the asymptotic covariance of $\sqrt{n}(\hat{\beta}_n - \beta_0)$, because $\hat{\beta}_n$ is efficient. We note that the estimating equations based estimator in Chen and Jewell [10] is not efficient in general. From this theorem, we see that in order to obtain asymptotic normality, oversmoothing is needed. This result is similar to that in Zeng and Lin [8]. The cross validation choice of the bandwidth, aiming at minimizing the expected prediction error, tends to choose a bandwidth that optimizes the overall performance with respect to the kernel smoothed hazard function and the parameters, and thus does not satisfy the rate specified in Theorem 2. However, in our simulation studies and the data analysis, we observe that the difference between the a prespecified bandwidth and a cross validation based bandwidth is often small. Zeng and Lin [8] also reported that their result was not sensitive to the bandwidth. We now make a connection of our estimates to those in Cox's model and the AFT model. In the proof of Theorem 2 in the Appendix, we show that the score function $U_n(\beta)$ of β takes the form

$$U_n(\beta) = \frac{1}{n} \sum_{i=1}^n l_{\beta}^s(O_i) + o_p(n^{-1/2}),$$

where

$$\begin{aligned}
\dot{l}_\beta^s(O) = & \int \left(X_2 - \frac{EX_2 \exp(X' \beta_{20}) I(Y \geq t)}{E \exp(X' \beta_{20}) I(Y \geq t)} \right) dM(t) \\
& + \int \left(X_2 - \frac{EX_2 \exp(X' \beta_{20}) I(Y \geq t)}{E \exp(X' \beta_{20}) I(Y \geq t)} \right) \frac{t \exp(X' \beta_{10}) \dot{\lambda}(t \exp(X' \beta_{10}))}{\lambda(t \exp(X' \beta_{10}))} dM(t) \\
& + \int \left(X_3 - \frac{EX_3 \exp(X' \beta_{20}) I(Y \geq t)}{E \exp(X' \beta_{20}) I(Y \geq t)} \right) dM(t),
\end{aligned} \quad (10)$$

for $X_2 = (X', 0_p)'$, $X_3 = (0_p, X')'$, with 0_p being the p -dimensional zero vector, and $dM(t) = d\delta/(Y - t) - \exp(X' \beta_{20})/(Y - t) d\Lambda(t \exp(X' \beta_{10}))$. Here and in the sequel, $\dot{f} = df(t)/dt$ denotes the first derivative function of f provided it exists.

From (10), if the AFT model holds, that is $\beta_2 = 0$, then the third term in (10) disappears. This gives the same score function as the AFT model derived in Zeng and Lin [8]. When the proportional hazards model holds, that is $\beta_1 = 0$, then the first two terms in (10) disappear. The resulting score function of β_2 reduces to

$$\frac{1}{n} \sum_{i=1}^n \int \left(X_i - \frac{EX_i \exp(X' \beta_{20}) I(Y \geq t)}{E \exp(X' \beta_{20}) I(Y \geq t)} \right) dM_i(t),$$

which coincides with the score function derived from the partial likelihood function in the proportional hazards model. These arguments illustrate clearly that the efficient score functions used in Cox's and the AFT models are special cases of those of the proposed general model.

To analyze the asymptotic properties of $\hat{\Lambda}_n(\cdot)$, we write $R_0 = Y \exp(X' \beta_{10})$, $G_0(s) = E[\exp(X' \beta_{20})/(R_0 > t)]$, $G_1(s) = E[\exp(X' \beta_{20})/(R_0 > t)X_2]$, $G_2(s) = G_1(s)/G_0(s)$, and $G_3(s) = E[\exp(X' \beta_{20})/(R_0 > t)X_3]$. Denote $f_{X/Y}(x)$ as the conditional density function of X given Y .

Theorem 3

If the conditions in Theorem 2 hold, then $\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$ follows a zero-mean Gaussian processes asymptotically with covariance function $EH(O; t)H(O, s)$ at (t, s) when n tends to infinity, where

$$\begin{aligned}
H(O; t) = & \int_0^t \frac{\delta f_{R_0|s, X}(s) - \lambda(s) \exp(X' \beta_{20}) I(R_0 > s)}{G_0(s)} ds \\
& - \int_0^t \frac{[G_1'(s) + G_3'(s)] \lambda(s) + s G_1'(s) \dot{\lambda}(s)}{G_0(s)} ds \times [El_\beta^s(O) l_\beta^s(O)']^{-1} l_\beta^s(O).
\end{aligned}$$

4. Model selection

Model selection is important in regression analysis and many procedures have been developed. In this section, we apply the model selection idea to analyze the model structure. For our model, the goal is to determine which covariates play role in Cox's model and the accelerate failure time model. Thus, developing model selection strategies is very important and meaningful in elucidating the roles of the covariates. To this end, we propose to maximize the penalized smoothed profile likelihood function

$$Q(\beta) = l_n^s(\beta) - \sum_{m=1}^2 \sum_{j=1}^p p_{w_{mj}}(|\beta_{mj}|),$$

where $p(\cdot)$ is a penalty function and w_{mj} , $m = 1, 2$; $j = 1, \dots, p$ are the tuning parameters.

Denote $\tilde{\beta}_n$ as the maximizer of $Q(\beta)$ over $\beta \in \mathcal{B}$. Let $V(\beta_0) = E \dot{l}_\beta^s(O) \dot{l}_\beta^s(O)' = (V_{ij})$ evaluated at $\beta = \beta_0$ be the information matrix. To derive the asymptotic properties of $\tilde{\beta}_n$, denote the nonzero index set as $B_0 = \{mj: \beta_{mj}^{(0)} = 0\}$ and $B_1 = \{mj: \beta_{mj}^{(0)} \neq 0\}$, where $\beta_{mj}^{(0)}$ is the j th component of β_{m0} . For a subset B of $\{11, \dots, 1p, 21, \dots, 2p\}$, let $\beta_B = \{\beta_{mj}, mj \in B\}$ be a sub-vector of β . We have the following theorem that states the structure selection consistency, also known as the oracle properties [11]. It is clear that the conditions in Theorem 1 in [16] all hold, one obtains the following Oracle properties,

Theorem 4

Suppose that the conditions of Theorem 2 and Condition A7 hold and that for $m = 1, 2, j = 1, \dots, p$, $w_{mj} \rightarrow 0$ and $\sqrt{n}w_{mj} \rightarrow \infty$ when $n \rightarrow \infty$. Then we have

- a. Selection consistency: $\tilde{\beta}_{nB_0} = 0$ with probability tending to one;
- b. Asymptotical normality: $\sqrt{n}(\tilde{\beta}_{nB_1} - \beta_{0B_1}) \rightarrow N(0, \{V(\beta_{B_1})\}^{-1})$, where $V(\beta_{B_1})$ is the Fisher information matrix for the nonzero set B_1 .

Condition A7 put some requirements on the penalty functions that can be used in $Q(\beta)$. Here we use the smoothly clipped absolute deviation (SCAD) penalty function whose first derivative is given by

$$\dot{p}_w(|\theta|) = w \{I(|\theta| \leq w) + \frac{(aw - |\theta|)_+}{(a-1)w} I(|\theta| > w)\}, \quad (11)$$

where $a = 3.7$ is a constant, and $c_+ = \max(c, 0)$ for any real scalar c . The SCAD penalty satisfies condition A7 in the Appendix and enjoys favorable properties as outlined in Fan and Li [11]. Alternatively we may use the adaptive Lasso penalty [17], an extension of the Lasso [18]. By setting $w_{mj} = w$, we choose w by the Bayesian information criterion (BIC), which minimizes with respect to w

$$BIC(w) = -l_n^s(\tilde{\beta}_w) + \#\tilde{\beta}_w \cdot \log(n)/n,$$

where $\#\tilde{\beta}_w$ denotes the number of nonzero coefficients in $\tilde{\beta}_w$. Following Wang, Li and Tsai [19], we can show that the resulting penalized estimators are model selection consistent. Thus, the asymptotic covariance matrix of $\tilde{\beta}_{nB_1}$ can be consistently estimated by

$\frac{1}{n} \{V(\tilde{\beta}_{nB_1})\}^{-1}$. Since the proof of this theorem is now standard [12], we omit the proof to save space.

5. Simulation

We conduct simulation studies in this section to evaluate the finite sample performance of the proposed method by comparing to the estimating equations method in Chen and Jewell

[10], as well as the Cox's and the AFT model. We also assess the accuracy of the inference procedure and the performance in model structure determination.

Example 1

In the first study, we compare the performance of our proposed estimator with those proposed by Chen and Jewell [10]. For that purpose, we simulate data motivated by the simulation setup in Chen and Jewell [10]. In particular, we consider the one-covariate case where X_i follows the Bernoulli distribution with success probability 0.5. The censoring time C_i follows the uniform distribution $U(0, \tau)$ with $\tau = 9$. Given X_i , we generate the failure time T_i under model (3) with $\lambda_0(t) = 1/(bt + 1)$ and different values of β_1 and β_2 . For this simulation, we use four possible configurations for $(\beta_1, \beta_2)'$ from $(1, -1)'$, $(1, 0)'$, $(0, 1)'$, $(0, 0)'$. When $\beta_{10}=0$, Cox's model should fit the data very well, and when $\beta_{20}=0$, the AFT model should fit the data very well. When both β_{10} and β_{20} are 0, both the AFT model and Cox's model would fit the data well. Two sample sizes $n = 100$ and $n = 200$ are used. For each simulation setup, the simulation is repeated 1000 times.

The results for the proposed method, Cox's model and the AFT model, using the efficient estimation approach proposed by Zeng and Lin [8], are reported in Table 1. In our simulation studies, we take $h = 0.85n^{-1/7}$. For our proposed method, we report the average bias (Bias), the sample standard deviations of the estimates across 1000 replications (SSD), the averages of the estimated standard errors (ESE), and the 95% empirical coverage probabilities (CP). For the other two methods, we only provide the Bias and SSD's. These results indicate that the estimation procedure of the proposed method gives almost unbiased estimates. The ESEs are all reasonably close to SSDs, and the coverage probability is close to the nominal probability 95%. These indicate that the proposed estimation and inference procedures work well for finite samples. The Cox model works well when $\beta_{10} = 0$, but has significant bias when β_{10} is not 0. Similarly, the AFT model works well only when $\beta_{20} = 0$, but has significant bias when β_{20} is not 0. One reviewer asked why the results of the proposed estimator for bias and SSD are different from the Cox (AFT) model when $\beta_{10} = 0$ ($\beta_{20} = 0$). The reason is that in that situation, Cox model only estimates β_2 while the proposed method still estimates two parameters. If one use variable selection in the next step, then β_1 will be set to be 0 with high probability and the estimates for β_2 should be the same as that in the Cox model. From the comparison with the estimators proposed by Chen and Jewell [10], all the Bias' of both estimators are very small, but our estimators have smaller SSDs than the corresponding estimators by Chen and Jewell [10] in general. This implies that our estimators is more efficient, which agrees with the statement in Theorem 2. We also try the simulation when $b = 0.5$ and obtain the similar results (not shown). To evaluate the performance for estimating Λ_0 , we plot the estimated curves in Figure 1 for $b = 0.5, 1$ and $n = 100, 200$. Figure 1 shows that the estimated curves are very close to the true curve of $\Lambda_0(t)$. In terms of bandwidth selection, we have also tried K-fold cross-validation to choose the optimal bandwidth and the results are similar.

Example 2

In the second study, we consider a setup similar to the first study except that there are six covariates. These covariates are generated from a multivariate normal distribution with mean zero and the covariance between X_i and X_j being $\rho^{|i-j|}$. For this example, we take $\rho = 0, 0.2, 0.5, 0.8$. The true coefficients are taken as $\beta_{10} = (1, 1, 1, 0, 0, 0)'$, and $\beta_{20} = (0, 0, 0, 1, 1, 1)'$. The largest follow up time is taken to be $\tau = 6$ and the censoring time is exponentially distributed with rate 4, which gives a censoring rate around 45%. We apply the variable selection method with the SCAD penalty and compare the result with the estimate without penalization and the estimates assuming the true structure is known. We also try other penalties, such as LASSO, hard thresholding penalty function (HARD), and obtain the

similar results. Table 2 reports the relative error, defined as the median ratio of the MSE versus the MSE from the method without variable selection. The average numbers of 0 coefficients are also reported in Table 2, in which the column labeled “Correct” presents the average number of six true coefficients correctly estimated as 0 and the column labeled “Incorrect” depicts the average of six nonzero coefficients erroneously estimated as 0.

Based on 1000 replications with sample sizes $n = 100$ and 200 , we can see from Table 2 that the proposed methods outperform the method without penalization. This improvement is more significant when the sample size is large. On the other hand, the proposed method performs well in structure selection, as the number of correctly estimated 0 is close to 6, and the number of incorrectly estimated 0 is close to 0. In addition, we can see that the proposed method gives a relative error very close to the oracle relative error, especially when the sample size is large. This confirms the theoretical results in Theorem 4 empirically. We also provide the simulation results for the case when there are covariates in both AFT and Cox part. Specifically we set $\beta_{10} = (1, 1, 1, 1, 0, 0)'$, and $\beta_{20} = (0, 0, 1, 1, 1, 1)'$. From Table 2, we can see that the performance in terms of relative errors and variable selection is still satisfactory, although the relative error part is worse than that when there are no overlapped covariates in both AFT and Cox models.

6. Application

We apply the proposed methods to a subset of data from the Childhood Cancer Survivor Study (CCSS). This study followed up around 14000 survivors who were diagnosed with childhood cancers and survived at least 5 years since the primary diagnosis to investigate late effects of cancer and cancer treatments ([20], [21]). The diagnoses include leukemia, central nervous system (CNS) malignancy, Hodgkin lymphoma, non-Hodgkin lymphoma, Wilm's tumor, neuroblastoma, soft-tissue sarcoma and bone tumors. During the follow-up, detailed demographic and treatment information was collected via questionnaires and the date of death was obtained from the death certificate if the patient died during the study.

Among all diseases, patients with Hodgkin lymphoma (HL) has the highest death rate during the follow-up (16.8%). So here we focus the subgroup of survivors who have been diagnosed HL during their childhood and focus on the study of risk factors on death. In total there were 1559 HL survivors. For the survivors who were alive at the end of study, the dates of last contact were obtained as censoring time. The covariates we are interested in are age at diagnosis (X_1 , years), gender (X_2 : 0=male, 1=female), radiation dose in the cancer treatment (X_3 , 0: <2500 cGy, 1: 2500 cGy) and cumulative anthracycline exposure (X_4 , g/ m^2). Now we use model (3) to fit the data. Table 3 shows the estimation results for the estimates without penalty and with the SCAD penalty. The SCAD penalty here is

$$\hat{p}_{w_j}(|\beta_j|) = w_j \{I(|\beta_j| \leq w_j) + \frac{(aw_j - |\beta_j|)_+}{(a-1)w_j} I(|\beta_j| > w_j)\},$$
 with $w_j = w/|\beta_j|$, $a = 3.7$, $w = 0.3$ and β_j being j th component of the estimator without penalty. Using K-fold cross validation, the bandwidth is chosen to be 0.2. From Table 3, the covariate information of anthracycline exposure plays an important role in constructing the AFT model since only anthracycline exposure is significant. On the other hand, radiation exposure is important to building Cox's model since other covariates in β_2 are insignificant. More specifically, higher cumulative anthracycline dose and higher radiation exposure are likely to increase the risk of death in HL survivors, but age at diagnosis and gender do not have significant influence. We also note that the standard errors in the model with penalization are markedly smaller than those in the unpenalized model.

Conditional on the covariate information X , one can estimate the conditional survival function by $\exp(-\Lambda_n(t \exp(X'\hat{\beta}_{1n})) \exp(X'\hat{\beta}_{2n}))$. Then we can estimate the marginal survival function for a subgroup by averaging the conditional survival function estimates. In Figure

2, we plot the estimated marginal survival function for two groups: the first group with $X_4 = 0$ corresponding to those without exposure to anthracycline and the other group with $X_4 > 0$ corresponding to those with exposure to anthracycline. For comparison, we plot the nonparametric Kaplan-Meier estimates, our proposed estimates and the estimates by Chen and Jewell [10] for the marginal survival functions for these two subgroups. From Figure 2, one can see that our proposed estimates agree better with the Kaplan-Meier estimates. The large discrepancy of the Chen and Jewell's estimates to the Kaplan-Meier estimates is probably due to the numerical instability in implementing their algorithm.

7. Conclusion

We have proposed an efficient procedure for estimation and inference for a general semiparametric hazards model and studied penalized likelihood methods for model structure selection. Comparing to the estimation approach in Chen and Jewell [10], our proposal is semiparametric efficient and enjoys better numerical properties. In addition, the inference procedure is much simpler than the approach taken by Chen and Jewell [10]. Our method is potentially useful, especially at the initial model building stage, to elucidate the roles of the covariates in time to event analysis.

In this paper, we have focused on univariate failure time. An extension to multivariate failure time model can be useful. Another interesting topic worth pursuing is to include cure fraction [9] which is more natural for many problems.

References

1. Cox DR. Regression Models and Life Tables. *Journal of the Royal Statistical Society Series B*. 1972; 34:187–220.
2. Cox DR. Partial likelihood. *Biometrika*. 1975; 62:269–276.
3. Kalbfleisch, J.; Prentice, R. *The Statistical Analysis of Failure Time Data*. Wiley; New York: 2002.
4. Ritov Y. Estimation in a linear regression model with censored data. *Annals of Statistics*. 1990; 18:303–328.
5. Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank tests. *Biometrika*. 1990; 77:845–851.
6. Lai TL, Ying Z. Rank regression methods for left-truncated and right-censored data. *Annals of Statistics*. 1991; 19:531–556.
7. Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003; 90:341–353.
8. Zeng D, Lin D. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*. 2007; 102:1387–1396.
9. Lu W. Efficient estimation for accelerated failure time model with a cure fraction. *Statistica Sinica*. 2010; 20:661–674. [PubMed: 20414460]
10. Chen YQ, Jewell NP. On a general class of semiparametric hazards regression models. *Biometrika*. 2001; 88:687–702.
11. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
12. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*. 2007; 94:691–703.
13. Cai T, Huang J, Tian L. Regularized estimation for the accelerated failure time model. *Biometrics*. 2009; 65:394–404. [PubMed: 18573133]
14. Lu W, Zhang HH. Variable selection for proportional odds model. *Statistics in Medicine*. 2007; 26:3771–3781. [PubMed: 17266170]
15. Johnson BA. Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society, Series B*. 2008; 70:351–370.

16. Johnson B, Lin DY, Zeng D. Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association*. 2008; 103(482):672–680.
17. Zou H. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
18. Tibshirani RJ. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
19. Wang H, Li R, Tsai C. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. 2007; 94:553–568. [PubMed: 19343105]
20. Robison LL, Mertens AC, Boice JD, et al. Study design and cohort characteristics of the childhood cancer survivor study: A multiinstitutional collaborative project. *Medical and Pediatric Oncology*. 2002; 38(4):229–239. [PubMed: 11920786]
21. Robison LL, Armstrong G, Boice JD, et al. The Childhood Cancer Survivor Study: a National Cancer Institute-supported resource for outcome and intervention research. *Journal of Clinical Oncology*. 2009; 27(14):2308–2318. [PubMed: 19364948]
22. Van der Vaart, AW.; Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer; New York: 1996.
23. Bickel, PJ.; Klaassen, C.; Ritov, R.; Wellner, J. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer; New York: 1993.

Appendix

To derive the asymptotic results of the proposed estimator, we need the following regular conditions. For a function h , let $\dot{h}$ be its first derivative function and $\ddot{h}$ be its second derivative function. Throughout the paper, we assume that the identifiability condition outlined in Chen and Jewell ([10], Proposition 1) holds.

- A1** The true value $\beta_0 = (\beta'_{10}, \beta'_{20})'$ belongs to the interior of a known compact set $\mathcal{B}$ in R^{2p} .
- A2** Conditional on X , C is independent of T and ε , and X is bounded.
- A3** The censoring time C has a positive and twice continuously differentiable density in $[0, \tau]$ and $\inf P(C_i = \tau | X_i) > 0$.
- A4** For $t = 0$, $\lambda_0(t)$ is positive and thrice continuously differentiable and $\lambda_0(0) > 0$.
- A5** If there exists a vector β and one deterministic function G , such that $X'\beta = G(\varepsilon)$ with probability one, then $\beta = 0$ and $G = 0$.
- A6** The kernel function $K(\cdot)$ is thrice-continuously differentiable and its r th ($r = 0, 1, 2, 3$) derivative functions have bounded variations. In addition, the first $(m - 1)$ moments of m th derivative function of $K(\cdot)$ are 0 for some $m > 3$.
- A7** $\lim_{n \rightarrow \infty} \inf_{\theta \rightarrow 0^+} \liminf_{w_{jn}} \dot{p}_{w_{jn}}(\theta) / w_{jn} > 0$, $\max_{j \in B_1} \dot{p}_{w_{jn}}(|\beta_j^{(0)}|) = O(n^{-1/2})$, $\max_{j \in B_1} \ddot{p}_{w_{jn}}(|\beta_j^{(0)}|) = o(1)$.

For $i = 1, \dots, n$, define the counting processes $N_i(t) = \delta_i(Y_i \leq t)$ and their compensation martingales

$$M_i(t) = N_i(t) - \int_0^t I(Y_i \geq s) \exp(X_i' \beta_{20}) d\Lambda_0(\exp(X_i' \beta_{10}))$$

for $t \in [0, \tau]$.

Proof of Theorem 1

The uniform convergence of $l_n^s(\beta)$ to $l(\beta)$ over $\beta \in \mathcal{B}$ follows from the following two results

$$\sup_{\beta_1 \in \mathcal{B}, s \in [0, \tau]} \left| \frac{1}{nh} \sum_{i=1}^n \delta_i K((R_i(\beta_1) - s)/h) - dP(\delta=1, R_i(\beta_1) \leq s)/ds \right| \rightarrow 0, \quad \text{a.s.}$$

and

$$\sup_{\beta \in \mathcal{B}, s \in [0, \tau]} \left| \frac{1}{n} \sum_{j=1}^n \exp(X_j' \beta_2) \int_{-\infty}^{(R_j(\beta_1) - s)/h} K(t) dt - E \exp(X' \beta_2) I(R(\beta_1) \geq s) \right| \rightarrow 0, \quad \text{a.s.}$$

Now we show that β_0 is the unique maximizer of $l(\beta)$. To this end, define

$$\Lambda(t; \beta) = \int_0^t \frac{dP(\delta=1, R(\beta_1) \leq s)/ds}{E \exp(X' \beta_2) I(R(\beta_1) > s)} ds.$$

Note that

$$\Lambda(t; \beta_0) = \int_0^t \frac{EdN(s \exp(-X' \beta_{10}))}{E \exp(X' \beta_{20}) I(R(\beta_{10}) > s)} = \Lambda_0(t)$$

and

$$E \exp(X' \beta_2) \Lambda(R(\beta_1); \beta) = P(\delta=1).$$

Suppose that β maximizes $l(\beta)$, that is

$$\delta X'(\beta_1 + \beta_2) + \delta \log \lambda(R(\beta_1); \beta) - \exp(X' \beta_2) \Lambda(R(\beta_1); \beta) \geq \delta X'(\beta_{10} + \beta_{20}) + \delta \log \lambda_0(R(\beta_{10})) - \exp(X' \beta_{20}) \Lambda_0(R(\beta_{10})),$$

where $\lambda(t; \beta) = d\Lambda(t; \beta)/dt$.

Taking expectation on both sides and then using the nonnegativity of Kullback-Leibler information, one obtains that

$$\begin{aligned} & \exp \left[\delta X'(\beta_1 + \beta_2) + \delta \log \lambda(R(\beta_1); \beta) - \exp(X' \beta_2) \Lambda(R(\beta_1); \beta) \right] \\ &= \exp \left[\delta X'(\beta_{10} + \beta_{20}) + \delta \log \lambda_0(R(\beta_{10})) - \exp(X' \beta_{20}) \Lambda_0(R(\beta_{10})) \right]. \end{aligned}$$

First, we let $\delta = 0$ and $Y = \tau$, then we obtain that $\exp(-\exp(X'\beta_2)\Lambda(\tau \exp(X'\beta_1); \beta)) = \exp(-\exp(X'\beta_2)\Lambda_0(\tau \exp(X'\beta_1)))$. Letting $\delta = 1$ and integrating $Y = y$ to τ , then one obtains that

$$\begin{aligned} & \exp(-\exp(X'\beta_2)\Lambda(y \exp(X'\beta_1); \beta)) - \exp(-\exp(X'\beta_2)\Lambda(\tau \exp(X'\beta_1); \beta)) \\ &= \exp(-\exp(X'\beta_2)\Lambda_0(y \exp(X'\beta_1))) - \exp(-\exp(X'\beta_2)\Lambda_0(\tau \exp(X'\beta_1))). \end{aligned}$$

These two equalities yield that

$$\exp(X'\beta_2)\Lambda(y \exp(X'\beta_1); \beta) = \exp(X'\beta_2)\Lambda_0(y \exp(X'\beta_1)).$$

Then there exists an increasing and differentiable function H_0 , such that

$$T \exp(X'\beta_1) = H_0(\exp(X'(\beta_{20} - \beta_2))\Lambda_0(\exp(\varepsilon))). \quad (\text{A.1})$$

From the definition of T and ε , one obtains that $T = \exp(-X'\beta_{10}) \exp(\varepsilon)$ and thus $X'(\beta_1 - \beta_{10}) + \varepsilon = H_1(X'(\beta_{20} - \beta_2) + H_2(\varepsilon))$, where $H_1(\cdot) = \log H_0(\exp(\cdot))$ and $H_2(\cdot) = \log(\Lambda_0(\exp(\cdot)))$. Taking derivatives on both sides with respect to ε , one obtains that $X'(\beta_{20} - \beta_2) = H_3(\varepsilon)$ for some function H_3 , which yields that $\beta_2 = \beta_{20}$ from assumption A5. Again using assumptions A5, (A.1) and this result gives that $X'(\beta_1 - \beta_{10}) = H_1(H_2(\varepsilon)) - \varepsilon$ and $\beta_1 = \beta_{10}$. This completes the proof.

Proof of Theorem 2

Let $X_1 = (X', X')'$, $X_2 = (X', 0_p)'$, $X_3 = (0_p', X')'$, where 0_p denotes the p -dimensional zero vector. Let P_n and P represent the empirical measure and probability measure. Then the score function of β is

$$U_n(\beta) = P_n \left[\delta X_1 + \delta \frac{A_{3n}(Y, X; \beta)}{A_{1n}(Y, X; \beta)} - \delta \frac{A_{4n}(Y, X; \beta)}{A_{2n}(Y, X; \beta)} \right]$$

where $A_{kn}(y, x; \beta) = P_n A_{ki}(y, x; \beta)$, $k = 1, 2, 3, 4$ and

$$\begin{aligned} A_{1i}(y, x; \beta) &= \frac{1}{h} \delta_i K((R_i(\beta_1) - R(y, x; \beta_1))/h), \quad R(y, x; \beta_1) = y \exp(x' \beta_1), \\ A_{2i}(y, x; \beta) &= \exp(X'_i \beta_2) \int_{-\infty}^{(R_i(\beta_1) - R(y, x; \beta_1))/h} K(s) ds, \\ A_{3i}(y, x; \beta) &= \delta_i \dot{K} \left(\frac{R_i(\beta_1) - R(y, x; \beta_1)}{h} \right) \frac{R_i(\beta_1) X_{2i} - R(y, x; \beta_1) x_2}{h^2}, \\ A_{4i}(y, x; \beta) &= \exp(X'_i \beta_2) X_{3i} \int_{-\infty}^{(R_i(\beta_1) - R(y, x; \beta_1))/h} K(s) ds \\ &\quad + \exp(X'_i \beta_2) K \left(\frac{R_i(\beta_1) - R(y, x; \beta_1)}{h} \right) \frac{R_i(\beta_1) X_{2i} - R(y, x; \beta_1) x_2}{h}. \end{aligned}$$

Let $R_0 = R(Y, X; \beta_0)$, $R_0(y, x) = R(y, x; \beta_0)$. After some calculation, one obtains that

$$A_{jn}(y, x; \theta_0) \rightarrow EB_j(O; y, x)$$

for $j = 1, 2, 3, 4$, where

$$\begin{aligned} B_1(O; y, x) &= I(R_0 \geq R_0(y, x)) \lambda_0(R_0(y, x)) \exp(X' \beta_{20}), \\ B_2(O; y, x) &= I(R_0 \geq R_0(y, x)) \exp(X' \beta_{20}), \\ B_3(O; y, x) &= f_{R_0|y, X}(R_0(y, x)) \lambda_0(R_0(y, x)) R_0(y, x) (X_2 - x_2) - I(R_0 \geq R_0(y, x)) \lambda_0(R_0(y, x)) \\ &\quad \times \exp(X' \beta_{20}) R_0(y, x) (X_2 - x_2) - I(R_0 \geq R_0(y, x)) \lambda_0(R_0(y, x)) \exp(X' \beta_{20}) X_2, \\ B_4(O; y, x) &= I(R_0 \geq R_0(y, x)) \exp(X' \beta_{20}) X_3 + f_{R_0|y, X}(R_0(y, x)) \exp(X' \beta_{20}) R_0(y, x) (X_2 - x_2). \end{aligned}$$

Let $\tilde{O} = (\tilde{Y}, \tilde{\delta}, \tilde{X})$ be i.i.d copy of O , and the notation $E_O h(O)$ is the expectation of some function $h(\cdot)$ of O with respect to the joint density of O . Using the consistency of the estimator $\hat{\beta}_n$ and conditions A3–A6, it follows from the proof of Theorem 2 in Zeng et al. ([8]) and Theorem 2.11.23 in van der Vaart and Wellner [22] that

$$\sqrt{n} P_n \delta \left\{ \frac{A_{3n}(Y, X; \hat{\beta}_n)}{A_{1n}(Y, X; \hat{\beta}_n)} - \frac{A_{4n}(Y, X; \hat{\beta}_n)}{A_{2n}(Y, X; \hat{\beta}_n)} - \frac{\mathcal{P}_n B_3(\tilde{O}; Y, X)}{\mathcal{P}_n B_1(\tilde{O}; Y, X)} + \frac{\mathcal{P}_n B_4(\tilde{O}; Y, X)}{\mathcal{P}_n B_2(\tilde{O}; Y, X)} \right\} = o_p(1).$$

Since $U_n(\hat{\beta}_n) = 0$, one obtains that

$$\begin{aligned} 0 &= \sqrt{n} U_n(\hat{\beta}_n) = \sqrt{n} P_n \left\{ \delta X_1 + \delta \frac{A_{3n}(Y, X; \hat{\beta}_n)}{A_{1n}(Y, X; \hat{\beta}_n)} - \delta \frac{A_{4n}(Y, X; \hat{\beta}_n)}{A_{2n}(Y, X; \hat{\beta}_n)} \right\} \\ &= \sqrt{n} P_n \left\{ \delta X_1 + \delta \frac{P_n B_3(\tilde{O}; Y, X)}{P_n B_1(\tilde{O}; Y, X)} - \delta \frac{P_n B_4(\tilde{O}; Y, X)}{P_n B_2(\tilde{O}; Y, X)} \right\} + o_p(1) \\ &= \sqrt{n} P_n \int \left\{ X_1 + \frac{P_n B_3(\tilde{O}; t, X)}{P_n B_1(\tilde{O}; t, X)} - \frac{P_n B_4(\tilde{O}; t, X)}{P_n B_2(\tilde{O}; t, X)} \right\} dM(t) \\ &\quad + \sqrt{n} P_n \int \left\{ X_1 + \frac{P_n B_3(\tilde{O}; t, X)}{P_n B_1(\tilde{O}; t, X)} - \frac{P_n B_4(\tilde{O}; t, X)}{P_n B_2(\tilde{O}; t, X)} \right\} \\ &\quad \times I(R_0 \geq R_0(t, x)) \lambda_0(R_0(t, X)) \exp(X' \beta_{20}) dR_0(t, X) + o_p(1) \\ &= \sqrt{n} P_n \int \left\{ X_1 + \frac{E_{\tilde{O}} B_3(\tilde{O}; t, X)}{E_{\tilde{O}} B_1(\tilde{O}; t, X)} - \frac{E_{\tilde{O}} B_4(\tilde{O}; t, X)}{E_{\tilde{O}} B_2(\tilde{O}; t, X)} \right\} dM(t) + o_p(1). \end{aligned}$$

By calculating the expectation of B_j , we can simplify the expressions as

$$0 = \sqrt{n} \mathcal{P}_n \dot{l}_{\beta}^s(O) + o_p(1),$$

where

$$\begin{aligned} \dot{l}_{\beta}^s(O) &= \int \left\{ X_1 - \frac{E_{\tilde{O}} \{ \tilde{X}_1 \exp(\tilde{X}' \beta_{20}) I(\tilde{R}_0 \geq R_0(t, X)) \}}{E_{\tilde{O}} \{ I(\tilde{R}_0 \geq R_0(t, X)) \exp(\tilde{X}' \beta_{20}) \}} \right\} dM(t) \\ &\quad + \int \left\{ X_2 - \frac{E_{\tilde{O}} \{ \tilde{X}_2 I(\tilde{R}_0 \geq R_0(t, X)) \exp(\tilde{X}' \beta_{20}) \}}{E_{\tilde{O}} \{ I(\tilde{R}_0 \geq R_0(t, X)) \exp(\tilde{X}' \beta_{20}) \}} \right\} \times \frac{\dot{\lambda}_0(R_0(t, X)) R_0(t, X)}{\lambda_0(R_0(t, X))} dM(t). \end{aligned}$$

Here $\dot{l}_\beta^s$ is easily shown to be the efficient score function for β_0 , which is given by Bickel, et al. [23]. Write $H(t) = -\frac{E_{\tilde{O}}\{\tilde{X}_1 \exp(\tilde{X}'\beta_{20})I(\tilde{R}_0 \geq t)\}}{E_{\tilde{O}}\{\exp(\tilde{X}'\beta_{20})I(\tilde{R}_0 \geq t)\}} - \frac{E_{\tilde{O}}\{\tilde{X}_2 \exp(\tilde{X}'\beta_{20})I(\tilde{R}_0 \geq t)\}}{E_{\tilde{O}}\{\exp(\tilde{X}'\beta_{20})I(\tilde{R}_0 \geq t)\}} \frac{\dot{\lambda}_0(t)t}{\lambda_0(t)}$. Applying the definition of the $M(t)$, the score function can be rewritten as

$$\begin{aligned}\dot{l}_\beta^s(O) &= X_1 \delta - \int X_1 I(R_0 \geq t) \lambda_0(t) \exp(X' \beta_{20}) dt + \delta X_2 \frac{\dot{\lambda}_0(R_0) R_0}{\lambda_0(R_0)} \\ &\quad - \int X_2 \frac{\dot{\lambda}_0(t)t}{\lambda_0(t)} I(R_0 \geq t) \lambda_0(t) \exp(X' \beta_{20}) dt + \int_0^{R_0} H(t) \lambda_0(t) \exp(X' \beta_{20}) dt + \delta H(R_0) \\ &= X_2 \delta + X_2 R_0 \left\{ \delta \frac{\dot{\lambda}_0(R_0)}{\lambda_0(R_0)} - \lambda_0(R_0) \right\} + \delta H(R_0) + X_3 \delta - X_3 \int_0^{R_0} \lambda_0(t) \exp(X' \beta_{20}) dt \\ &\quad - \int_0^{R_0} H(t) \lambda_0(t) \exp(X' \beta_{20}) dt.\end{aligned}$$

Now we prove that $E[\dot{l}_\beta^s(\dot{l}_\beta^s)']$ is nonsingular. Otherwise, there exists a vector α , such that $\alpha E[\dot{l}_\beta^s(\dot{l}_\beta^s)']\alpha' = 0$, which yields that $\alpha' \dot{l}_\beta^s = 0$. After multiplying both sides by $d \exp(-\Lambda_0(R_0(t), X))$ and then integrating t from 0 to t , we obtain that $X_2' \alpha = A_1(\varepsilon)$ and $X_3' \alpha = A_2(\varepsilon)$ for two differentiable functions A_1 and A_2 by solving the two equations obtained by letting $\delta = 1$ and $\delta = 0$. Let $\alpha = (\alpha_1', \alpha_2')'$, where α_1 is the first p -dimensional vector of α , α_2 is the last p -dimensional vector of α . By the definitions of X_2 and X_3 , one obtains that $X' \alpha_1 = A_1(\varepsilon)$ and $X' \alpha_2 = A_2(\varepsilon)$ with probability one. Finally, Condition A5 entails that $\alpha_1 = \alpha_2 = 0$, and thus $\alpha = 0$. This completes the proof.

Proof of Theorem 3

Denote $\mathcal{G}_n = \sqrt{n}(P_n - P)$, and $R = Y \exp(X' \hat{\beta}_{1n})$. Note that

$$\begin{aligned}\sqrt{n} \hat{\Lambda}_n(t) &= \sqrt{n} \int_0^t \frac{\frac{1}{nh} \sum_{i=1}^n \delta_i K((Y_i \exp(X_i' \hat{\beta}_{1n}) - s)/h)}{\frac{1}{n} \sum_{i=1}^n \exp(X_i' \hat{\beta}_{2n}) \int_{-\infty}^{(Y_i \exp(X_i' \hat{\beta}_{1n}) - s)/h} K(u) du} ds \\ &= \mathcal{G}_n \int_0^t \frac{\delta K((R_0 - s)/h)}{G_0(s)} ds + \mathcal{G}_n \int_0^t \frac{\frac{\delta}{h} [K(\frac{R_0 - s}{h}) - K(\frac{R_0 - s}{h})]}{G_0(s)} ds \\ &\quad - \mathcal{G}_n \int_0^t \frac{E_{\tilde{O}}[\tilde{\delta} f_{\tilde{R}_0}(\tilde{\delta}(s))]}{G_0(s)^2} \exp(X' \beta_{20}) \int_{-\infty}^{\frac{R_0 - s}{h}} K(u) du ds \times X_3' (\hat{\beta}_n - \beta_0) \\ &\quad - \mathcal{G}_n \int_0^t \frac{E_{\tilde{O}}[\tilde{\delta} f_{\tilde{R}_0}(\tilde{\delta}(s))]}{G_0(s)^2} \exp(X' \beta_{20}) \int_{-\infty}^{\frac{R_0 - s}{h}} K(u) du ds \\ &\quad - \mathcal{G}_n \int_0^t \frac{E_{\tilde{O}}[\tilde{\delta} f_{\tilde{R}_0}(\tilde{\delta}(s))]}{G_0(s)^2} \frac{1}{h} \exp(X' \beta_{20}) K(\frac{R_0 - s}{h}) R_0 ds \times X_2' (\hat{\beta}_n - \beta_0) \\ &\quad + \sqrt{n} \int_0^t \frac{E_{\tilde{O}}[\tilde{\delta} f_{\tilde{R}_0}(\tilde{\delta}(s))]}{G_0(s)} ds + o_p(1).\end{aligned}$$

By the asymptotical normality of $\hat{\beta}_n$, one obtains that

$$\begin{aligned}
& \mathcal{G}_n \int_0^t \frac{\frac{\delta}{h} [K(\frac{\tilde{R}-s}{h}) - K(\frac{R_0-s}{h})]}{G_0(s)} ds = \mathcal{G}_n \int_0^t \frac{\frac{\delta}{h^2} K(\frac{R_0-s}{h}) R_0}{G_0(s)} X_2' (\hat{\beta}_n - \beta_0) ds + o_p(1) \\
& = E_{\tilde{O}} \int_0^t \frac{\frac{\delta}{h^2} K(\frac{\tilde{R}_0-s}{h}) \tilde{R}_0}{G_0(s)} \tilde{X}_2' ds \times [E_{\tilde{O}} \dot{l}_{\beta}^s(\tilde{O}) \dot{l}_{\beta}^s(\tilde{O})']^{-1} \mathcal{G}_n \dot{l}_{\beta}^s(O) + o_p(1) \\
& = \int_0^t ds E_{\tilde{O}} \int \frac{I(\tilde{R}_0 \geq y) K(\frac{y-s}{h}) y \exp(\tilde{X}' \beta_{20}) \lambda_0(y) \tilde{X}_2'}{h^2 G_0(s)} dy \times [E_{\tilde{O}} \dot{l}_{\beta}^s(\tilde{O}) \dot{l}_{\beta}^s(\tilde{O})']^{-1} \mathcal{G}_n \dot{l}_{\beta}^s(O) + o_p(1) \\
& = \int_0^t ds \int \frac{\dot{K}(\frac{y-s}{h}) y G_1'(y) \lambda_0(y)}{h^2 G_0(s)} dy + [E_{\tilde{O}} \dot{l}_{\beta}^s(\tilde{O}) \dot{l}_{\beta}^s(\tilde{O})']^{-1} \mathcal{G}_n \dot{l}_{\beta}^s(O) + o_p(1) \\
& = - \int_0^t \frac{[G_1'(s) + s G_1'(s)] \lambda_0(s) + s G_1'(s) \dot{\lambda}_0(s)}{G_0(s)} ds \times [E_{\tilde{O}} \dot{l}_{\beta}^s(\tilde{O}) \dot{l}_{\beta}^s(\tilde{O})']^{-1} \mathcal{G}_n \dot{l}_{\beta}^s(O) + o_p(1)
\end{aligned}$$

In addition, by observing that

$$\begin{aligned}
E_{\tilde{O}} [\tilde{\delta} f_{\tilde{R}_0|\tilde{\delta}}(s)] &= E \int \frac{1}{h} K(\frac{y-s}{h}) \lambda_0(y) I(\tilde{R}_0 \geq y) \exp(\tilde{X}' \beta_{20}) dy \\
&= \int \frac{1}{h} K(\frac{y-s}{h}) \lambda_0(y) G_0(y) dy \rightarrow \lambda_0(s) G_0(s),
\end{aligned}$$

$$\int_0^t \frac{E_{\tilde{O}} [\tilde{\delta} f_{\tilde{R}_0|\tilde{\delta}}(s)]}{G_0(s)} ds = \Lambda_0(t) \text{ and}$$

$$E \frac{1}{h} \exp(\tilde{X}' \beta_{20}) K(\frac{\tilde{R}_0-s}{h}) \tilde{R}_0 \tilde{X}_2' \rightarrow -s \dot{G}_1(s),$$

one obtains that

$$\begin{aligned}
\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t)) &= \mathcal{G}_n \int_0^t \frac{\delta f_{R_0|\delta, X}(s) - \lambda_0(s) \exp(X' \beta_{20}) I(R_0 > s)}{G_0(s)} ds \\
&\quad - \int_0^t \frac{[G_1'(s) + G_3'(s)] \lambda_0(s) + s G_1'(s) \dot{\lambda}_0(s)}{G_0(s)} ds \\
&\quad \times [E_{\tilde{O}} \dot{l}_{\beta}^s(\tilde{O}) \dot{l}_{\beta}^s(\tilde{O})']^{-1} \mathcal{G}_n \dot{l}_{\beta}^s(O) + o_p(1).
\end{aligned}$$

The proof is completed.

Figure 1.

Estimated curve of $\Lambda_0(t) = b^{-1} \log(1 + bt)$ with $b = 0.5$ and $b = 1$ for $n=100$ (left panel) and $n=200$ (right panel). Solid line: True curve $\Lambda_0(t)$ with $b = 0.5$ (Upper) and $b = 1$ (Lower); Dashed line: the proposed estimated curve $\Lambda_n(t)$ with $b = 0.5$ (Upper) and $b = 1$ (Lower).

Figure 2. Estimated marginal survival functions for $X_4 > 0$ (left panel) and $X_4 = 0$ (right panel). Solid line: Kaplan-Meier estimates; Dashed line: Chen and Jewell's estimates; Dotted lines: Our proposed estimates.

Table 1

Summary of Simulation studies in Example 1 with $h_n = 0.85n^{-1/7}$.

β		Proposed method				CJ's method				Cox's model				AFT model			
		True	Bias	SSD	ESE	CP	Bias	SSD	Bias	SSD	Bias	SSD	Bias	SSD			
$n = 100$																	
β_1	1	0.051	0.536	0.534	0.965	-0.016	0.572	-	-	-0.714	0.930	-	-	-	-	-	-
β_2	-1	0.007	0.359	0.391	0.974	0.011	0.506	0.517	0.255	-	-	-	-	-	-	-	-
β_1	-1	-0.018	0.547	0.538	0.97	-0.104	0.606	-	-	-0.058	0.411	-	-	-	-	-	-
β_2	0	0.007	0.345	0.347	0.968	-0.025	0.508	-0.652	0.254	-	-	-	-	-	-	-	-
β_1	0	-0.069	0.550	0.538	0.953	-0.102	0.488	-	-	0.992	0.843	-	-	-	-	-	-
β_2	1	-0.006	0.398	0.433	0.970	0.043	0.465	0.017	0.244	-	-	-	-	-	-	-	-
β_1	0	0.039	0.535	0.501	0.943	-0.093	0.512	-	-	0.024	0.396	-	-	-	-	-	-
β_2	0	-0.016	0.352	0.378	0.968	0.024	0.440	0.060	0.236	-	-	-	-	-	-	-	-
$n = 200$																	
β_1	1	-0.001	0.456	0.402	0.939	-0.009	0.509	-	-	-1.114	0.828	-	-	-	-	-	-
β_2	-1	0.030	0.289	0.284	0.941	-0.003	0.412	0.516	0.178	-	-	-	-	-	-	-	-
β_1	-1	0.029	0.444	0.408	0.950	-0.058	0.552	-	-	-0.045	0.301	-	-	-	-	-	-
β_2	0	-0.050	0.311	0.314	0.954	0.076	0.439	-0.653	0.179	-	-	-	-	-	-	-	-
β_1	0	0.032	0.438	0.405	0.942	-0.097	0.417	-	-	1.271	0.389	-	-	-	-	-	-
β_2	1	-0.061	0.309	0.316	0.946	0.046	0.39	0.004	0.171	-	-	-	-	-	-	-	-
β_1	0	0.014	0.425	0.376	0.915	-0.067	0.393	-	-	0.006	0.295	-	-	-	-	-	-
β_2	0	-0.012	0.279	0.276	0.956	0.031	0.343	-0.005	0.169	-	-	-	-	-	-	-	-

CJ's method refers to the estimator proposed by Chen and Jewell (2001) with the function $G(\cdot; Z, \beta)$, required for defining the estimating equations, specified as tZ .

Table 2

Summary of Simulation studies in Example 2

n	ρ	Method	$\beta_0 = (1, 1, 1, 0, 0, 0, 0, 1, 1, 1)$				$\beta_0 = (1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1)$			
			Relative Error		Aver. no. of 0 coeff		Relative Error		Aver. no. of 0 coeff	
			Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
100	0	PROPOSED	0.422	0.68	5.76	0.54	0.861	7.56	0.54	
		ORACLE	0.403	0	6	0	0.404	8	0	
	0.2	PROPOSED	0.507	0.90	5.72	0.57	0.883	7.48	0.57	
		ORACLE	0.403	0	6	0	0.431	8	0	
	0.5	PROPOSED	0.549	0.87	5.71	0.54	0.861	7.56	0.54	
		ORACLE	0.355	0	6	0	0.404	8	0	
	0.8	PROPOSED	0.460	0.76	5.80	0.57	0.883	7.48	0.57	
		ORACLE	0.442	0	6	0	0.431	8	0	
	0	PROPOSED	0.351	0.77	5.82	0.55	0.737	7.68	0.55	
		ORACLE	0.321	0	6	0	0.327	8	0	
200	0.2	PROPOSED	0.365	0.69	5.84	0.72	0.819	7.63	0.72	
		ORACLE	0.354	0	6	0	0.297	8	0	
	0.5	PROPOSED	0.484	0.80	5.87	0.58	0.761	7.69	0.58	
		ORACLE	0.385	0	6	0	0.290	8	0	
	0.8	PROPOSED	0.322	0.88	5.86	0.72	0.841	7.52	0.72	
		ORACLE	0.295	0	6	0	0.305	8	0	

Table 3
Estimation results for the Childhood Cancer Survivor Study with standard errors in parentheses.

	Without penalty	With penalty
β_1	Age	-0.041 (0.020)
	Gender	-0.529 (0.139)
	radiation dose	-0.888 (0.256)
	anthracycline exposure	3.244 (0.627)
β_2		4.016 (0.342)
	Age	0.125 (0.034)
	Gender	0.591 (0.249)
	radiation dose	1.824 (0.354)
	anthracycline exposure	-0.344 (1.053)
		1.028 (0.192)
		-