Statistica Sinica **32** (2022), 2359-2380
doi:https://doi.org/10.5705/ss.202020.0404

EFFICIENT ESTIMATION FOR DIMENSION REDUCTION WITH CENSORED SURVIVAL DATA

Ge Zhao, Yanyuan Ma and Wenbin Lu

Portland State University, Penn State University, and North Carolina State University

Abstract: We propose a general index model for survival data, that generalizes many commonly used semiparametric survival models and belongs to the framework of dimension reduction. Using a combination of a geometric approach in semiparametrics and a martingale treatment in survival data analysis, we devise estimation procedures that are feasible and do not require covariate-independent censoring, as assumed in many dimension-reduction methods for censored survival data. We establish the root-n consistency and asymptotic normality of the proposed estimators and derive the most efficient estimator in this class for the general index model. Numerical experiments demonstrate the empirical performance of the proposed estimators, and an application to an AIDS data set further illustrates the usefulness of the work.

Key words and phrases: Dimension reduction, general index model, kernel estimation, semiparametric theory, survival analysis.

1. Introduction

The Cox proportional hazards model (Cox (1972)) is probably the most widely used semiparametric model for analyzing survival data. In the Cox model, the covariate effect is described by a single linear combination of the covariates in an exponential function, and is multiplicative in modeling the hazard function. Although this special way of modeling the hazard function permits a convenient estimation procedure, such as the maximum partial likelihood estimation (Cox (1975)), it has limitations. As widely studied in the literature, there are many situations where the Cox model may not be proper. Owing to the limitations of the Cox model, many other semiparametric survival models have been proposed, such as the accelerated failure time model (Buckley and James (1979)), proportional odds model (McCullagh (1980)), and linear transformation model (Dabrowska and Doksum (1988)), among others. Despite of all these efforts, the link between the summarized covariate effect, typically in the form of a linear

Corresponding author: Ge Zhao, Department of Mathematics and Statistics, Portland State University, Portland, OR 97201, USA. E-mail: gzhao@pdx.edu.

combination of the covariates, and the possibly transformed event time remains to have a predetermined form and, hence, can be restrictive.

The single-index feature of the above-mentioned semiparametric survival models is appealing because the covariate effect has a nice interpretation. It also naturally achieves dimension reduction when there is a large number of covariates. However, the specific model form that links the covariate index to the event time may be restrictive, and it is often difficult to check the goodness-of-fit of the specific link function form. To achieve a model that is flexible and feasible in practice, we borrow and extend the idea of a linear summary of the covariate effects, while freeing up the specific functional relation between the event time and the linear summaries. Thus, we propose the following general index model:

$$\mathrm{pr}(T \le t \mid \mathbf{X}) = \mathrm{pr}(T \le t \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}), \quad t > 0, \tag{1.1}$$

where T is the survival time of interest, $\mathbf{X}$ is the p-dimensional covariates, and $\boldsymbol{\beta}_0 \in \mathcal{R}^{p\times d}$ is the regression coefficient matrix, with $p > d$. Several properties of model (1.1) are worth mentioning. First, instead of a single linear summary, we allow d linear summaries, described by the d columns of $\boldsymbol{\beta}_0$. This increases the flexibility of how the covariate effects are combined. We can view this as a generalization from a single-index to a multi-index covariate summary. Imagine an extreme case when $d = p$. This model degenerates to the restriction-free case, where the dependence of T on $\mathbf{X}$ is arbitrary. Of course, in practice, when d is large, the estimation will encounter difficulties, and it is not feasible to carry out the analysis. However, conceptually, this provides a way of appreciating the flexibility of the model. In addition, in practice, when d is often smaller than p, this model framework allows us to find and incorporate the suitable number of indices d. Second, we do not specify a functional form of the conditional probability. Thus, the conditional probability in (1.1) is simply a function of t and $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}$. This relaxes both the exponential form of the covariate relation and the multiplicative form of the hazard function in the Cox model. It is also more flexible than other popular semiparametric survival models, such as the accelerated failure time and linear transformation models. Despite the flexibility of the model in (1.1), we show that by properly incorporating a semiparametric treatment and martingale techniques, estimation and inference are still possible. Third, the analysis can be carried out under the usual conditional independent censoring assumption, where the censoring time is allowed to depend on the covariates. It is common to have competing events that share partial risk factors as the event of interest. Hence relaxing the restrictive covariate-independent

censoring assumption allows us to work on the original censoring distribution assumption (Tsiatis (1975); Li, Wang and Chen (1999); Lu and Li (2011); Lopez, Patilea and Van Keilegom (2013)) and is valuable in practice.

The proposed general index model and associated semiparametric estimation method naturally provide a dimension-reduction tool for survival data. It has two main advantages over existing dimension-reduction methods for survival data. First, many existing methods require a stronger assumption on the censoring time, such as the covariate-independent censoring assumption (Li, Wang and Chen (1999); Lu and Li (2011)), or require a nonparametric estimation of the conditional survival function of censored survival times (Xia, Zhang and Xu (2010)) or censoring times (Li, Wang and Chen (1999)), given all the covariates, which may suffer from the curse of dimensionality. All these drawbacks are avoided here. Second, most existing methods (Xia, Zhang and Xu (2010); Li, Wang and Chen (1999)) are constructed based on general inverse probability weighted estimation techniques, and are thus not efficient. In contrast, our proposed method is built on semiparametric theory (Tsiatis (2006)) and achieves optimal semiparametric efficiency.

The rest of the paper is organized as follows. In Section 2, we develop estimation procedures for the index parameters in $\boldsymbol{\beta}$ and the functional relation between the event time and multiple indices. In Section 3, we establish the large-sample properties to enable inference. We perform extensive numerical experiments in Section 4, including a simulation and an analysis of an AIDS data set. We conclude the paper with a discussion in Section 5. All technical details are provided in the Supplementary Material.

2. Methodology Development

2.1. Semiparametric analysis

We first define some notation. Define $Z = \min(T, C)$ and $\Delta = I(T \leq C)$, where C is the censoring time. Assume $C \perp\!\!\!\perp T \mid \mathbf{X}$ and the relation between T and $\mathbf{X}$ follows the model in (1.1), where $\perp\!\!\!\perp$ stands for independence. The observed data consist of $(\mathbf{X}_i, Z_i, \Delta_i)$, for $i = 1, \ldots, n$, which are independent copies of $(\mathbf{X}, Z, \Delta)$. Note that even without censoring, $\boldsymbol{\beta}_0$ in (1.1) is not identifiable, because for any $d \times d$ full-rank matrix $\mathbf{A}$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0\mathbf{A}$ suit model (1.1) equally well. Thus, we fix a parameterization of $\boldsymbol{\beta}_0$ by assuming the upper $d \times d$ block of $\boldsymbol{\beta}_0$ to be the identity matrix $\mathbf{I}_d$, and the first d components of $\mathbf{X}$ to be continuous. This ensures the unique identification of $\boldsymbol{\beta}_0$, except for some pathological cases (Ichimura (1993)). Here, we consider a fixed d, and focus on estimating the

lower block of $\boldsymbol{\beta}_0$, which has dimension $(p-d)\times d$. In the event that the first d components of $\mathbf{X}$ happen to contain covariates that are irrelevant, numerical issues will arise, and one should rearrange the covariates in $\mathbf{X}$. We then proceed to estimate the conditional distribution function in (1.1). For convenience, write $\mathbf{X}=(\mathbf{X}_u^{\mathrm{T}},\mathbf{X}_l^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{X}_u\in\mathcal{R}^d$ and $\mathbf{X}_l\in\mathcal{R}^{p-d}$. Note that under the assumption of $C \perp\!\!\!\perp T \mid \mathbf{X}$ and (1.1), we can easily obtain

$$E\{f_1(C)f_2(T)\mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\}=E\{f_1(C)\mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\}E\{f_2(T)\mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\},$$

for any functions f_1, f_2; hence, $C \perp\!\!\!\perp T \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}$. This turns out to be an important property in the subsequent technical derivations.

Next, we derive the probability density function (pdf) of the model in (1.1). Write $S_c(z,\mathbf{x})=\mathrm{pr}(C\ge z\mid \mathbf{X}=\mathbf{x})$, $\Lambda_c(z,\mathbf{x})=-\log S_c(z,\mathbf{x})$, $\lambda_c(z,\mathbf{x})=\partial\Lambda_c(z,\mathbf{x})/\partial z$, and $f_c(z,\mathbf{x})=-\partial S_c(z,\mathbf{x})/\partial z$. Let $\tau<\infty$ be the maximum follow-up time. Here, $\lambda_c(z,\mathbf{x})$ and $f_c(z,\mathbf{x})$ are absolutely continuous on both $(0,\tau)$ and (τ,∞), have a discontinuity point at τ. Specifically, let $p(\mathbf{x})\equiv\mathrm{pr}(C=\tau\mid\mathbf{x})$; then, $\lambda_c(\tau,\mathbf{x})=p(\mathbf{x})S_c(\tau-,\mathbf{x})$ and $f_c(\tau,\mathbf{x})=p(\mathbf{x})$. The maximum follow-up time τ indicates that all surviving subjects are censored at the end of the study τ. This naturally leads to a point mass at τ. Our analysis below is adapted to the discontinuity of the censoring process, using the fact that the discontinuity at τ does not destroy the martingale structure (Fleming and Harrington (1991); Prentice and Kalbfleisch (2003)). Similarly, to describe the event process, for any parameter matrix $\boldsymbol{\beta}$, define $S(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})=\mathrm{pr}(T\ge z\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, $f(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})=-\partial S(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})/\partial z$, $\Lambda(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})=-\log S(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, and $\lambda(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})=\partial\Lambda(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})/\partial z$. Using this notation, the pdf of the model in (1.1) is

$$\begin{aligned} f_{\mathbf{X},Z,\Delta}(\mathbf{x},z,\delta,\boldsymbol{\beta},\lambda,\lambda_c,f_X)=f_{\mathbf{X}}(\mathbf{x})\lambda(z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})^{\delta}e^{-\int_0^z\lambda(s,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})ds}\\ \times\lambda_c(z,\mathbf{x})^{1-\delta}e^{-\int_0^z\lambda_c(s,\mathbf{x})ds}, \end{aligned} \tag{2.1}$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the pdf of $\mathbf{X}$. Here, for convenience, we assume the existence of the conditional pdfs of T, C given $\mathbf{X}$ and the marginal pdf $f_{\mathbf{X}}(\mathbf{x})$. However, the existence of $f_{\mathbf{X}}(\mathbf{x})$ is not essential, and our subsequent derivations still hold with suitable modifications. We assume the true data-generation process is based on $f_{\mathbf{X},Z,\Delta}(\mathbf{x},z,\delta,\boldsymbol{\beta}_0,\lambda_0,\lambda_{c0},f_{X0})$.

We now view (2.1) as a semiparametric model, where $\boldsymbol{\beta}$ is a finite-dimen -sional parameter of interest, and all remaining unknown components of the model are treated as infinite-dimensional nuisance parameters. We use a geometric approach to derive the efficient score based on (2.1). In survival analysis, the most popular approaches to estimation are martingale-based estimators

(Fleming and Harrington (1991)) and nonparametric maximum likelihood estimators (NPMLEs) (Zeng and Lin (2007)). Here, we find that an NPMLE is not well suited without adaption, owing to the inseparable relation between the hazard function and the covariates. The martingale approach may enable us to obtain a specific estimator for $\boldsymbol{\beta}$, while we aim at obtaining a more comprehensive understanding of the estimation of $\boldsymbol{\beta}$. The geometrical treatment in semiparametrics allows us to take advantage of the efficient score, the variance of which attains the semiparametric efficiency bound. The efficient score is the projection of the score vector with respect to $\boldsymbol{\beta}$ onto the orthogonal complement of the nuisance tangent space. In order to obtain the efficient score, we project the score vector onto the nuisance tangent space and calculate its residual. Here, the nuisance tangent space is the mean squared closure of all nuisance score functions of any parametric submodel of the semiparametric model that we are studying.

Following the geometric approach, we first characterize the nuisance tangent space as described in Proposition 1. The proof uses properties of martingale integration; see the Supplementary Material for details. Define the filtration $\mathcal{F}_n(t) \equiv \sigma\{\mathbf{X}_i, I(Z_i \le u, \Delta_i = 1), I(Z_i \le u, \Delta_i = 0), 0 \le u \le t, i = 1, \ldots, n\}$. Define $M_i(t, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i) \equiv N_i(t) - \int_0^t Y_i(s)\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)ds$ and $M_{ic}(t, \mathbf{X}_i) \equiv N_{ic}(t) - \int_0^t Y_i(s)\lambda_c(s, \mathbf{X}_i)ds$, where $N_i(t) = \Delta_i I(Z_i \le t)$, $N_{ic}(t) = (1 - \Delta_i)I(Z_i \le t)$ and $Y_i(t) = I(Z_i \ge t)$. Then, $M_i(t, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)$ and $M_{ic}(t, \mathbf{X}_i)$ are mean-zero martingale processes with respect to the filtration $\mathcal{F}_n(t)$. In the following, we eliminate the subindex $_i$ whenever it does not cause confusion.

Proposition 1. *The nuisance tangent space* $\Gamma = \Gamma_1 \oplus \Gamma_2 \oplus \Gamma_3$, *where*

$$\Gamma_1 = \left[\mathbf{a}(\mathbf{X}) : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}, \mathbf{a}(\mathbf{X}) \in \mathcal{R}^{(p-d)d}\right],$$
$$\Gamma_2 = \left\{\int_0^\infty \mathbf{h}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})dM(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) : \forall \mathbf{h}(Z, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)d}\right\},$$
$$\Gamma_3 = \left\{\int_0^\infty \mathbf{h}(s, \mathbf{X})dM_c(s, \mathbf{X}) : \forall \mathbf{h}(Z, \mathbf{X}) \in \mathcal{R}^{(p-d)d}\right\},$$

and "$\oplus$" denotes the direct sum. Here, $M(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})$ *and* $M_c(s, \mathbf{X})$ *are* $M_i(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)$ *and* $M_{ic}(s, \mathbf{X}_i)$, *respectively with the subindex* $_i$ *omitted.*

Having found the nuisance tangent space, we can now identify the efficient score function by projecting the score function onto Γ and calculating the residual. The score function is defined as $\mathbf{S}_{\boldsymbol{\beta}}(\Delta, Z, \mathbf{X}) \equiv \partial \log f_{\mathbf{X},Z,\Delta}(\mathbf{x}, z, \delta, \boldsymbol{\beta}, \lambda, \lambda_c, f_X)/\partial\boldsymbol{\beta}$. Let $\boldsymbol{\lambda}_1(s, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \equiv \partial\lambda(s, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})/\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ be the partial derivative of $\lambda(s, \mathbf{v})$ with respect to the vector $\mathbf{v}$ evaluated at $\mathbf{v} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, and $\boldsymbol{\lambda}_{10}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) \equiv \partial\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})/\partial(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})$ be the partial derivative of $\lambda_0(s, \mathbf{v})$ with respect to the vector $\mathbf{v}$ evaluated

at $\mathbf{v} = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}$. Straightforward calculation yields

$$\mathbf{S}_{\boldsymbol{\beta}}(\Delta, Z, \mathbf{X}) = \int_0^\infty \frac{\boldsymbol{\lambda}_{10}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})}{\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})} \otimes \mathbf{X}_l dM(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}), \tag{2.2}$$

where "$\otimes$" denotes the matrix Kronecker product. Based on the score function, the efficient score is derived in Proposition 2. The proof is given in the Supplementary Material.

Proposition 2. *Let the score function at the observation* $(\mathbf{X}, Z, \Delta)$ *be given as in* (2.2)*, and the nuisance tangent space be given as in Proposition* 1*. Then, the efficient score is*

$$\mathbf{S}_{\text{eff}}(\Delta, Z, \mathbf{X}) = \int_0^\infty \frac{\boldsymbol{\lambda}_{10}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})}{\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})} \otimes \left[\mathbf{X}_l - \frac{E\left\{\mathbf{X}_l S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}\right] \times dM(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}). \tag{2.3}$$

We further simplify the efficient score before constructing the corresponding efficient estimating equation. We can verify that

$$E\int_0^\infty \frac{\boldsymbol{\lambda}_{10}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})}{\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})} \otimes \left[\mathbf{X}_l - \frac{E\left\{\mathbf{X}_l S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}\right] Y(s)\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) ds = \mathbf{0}.$$

As a consequence, writing $dM(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) = dN(s) - Y(s)\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})ds$ in (2.3), we get

$$E\int_0^\infty \frac{\boldsymbol{\lambda}_{10}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})}{\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})} \otimes \left[\mathbf{X}_l - \frac{E\left\{\mathbf{X}_l S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}\right] dN(s) = \mathbf{0}.$$

2.2. Estimation procedure

Based on the above analysis, we propose obtaining the efficient estimator by solving

$$\sum_{i=1}^n \Delta_i \frac{\widehat{\boldsymbol{\lambda}}_1(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta})}{\widehat{\lambda}(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta})} \otimes \left[\mathbf{X}_{li} - \frac{\widehat{E}\left\{\mathbf{X}_{li} Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\}}{\widehat{E}\left\{Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\}}\right] = \mathbf{0}, \tag{2.4}$$

which is simpler than directly using the efficient score. To emphasize that the function estimation of λ, $\boldsymbol{\lambda}_1$, and $E(\cdot)$ relies on the parameter $\boldsymbol{\beta}$ through the data $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j$, we include the last parameter $\boldsymbol{\beta}$. We use this more precise notation below whenever it helps to avoid ambiguity.

In forming (2.4), several nonparametric estimators are used. Specifically, the

hazard function and its derivative are estimated using the local Nelson–Aalen estimator, that is,

$$\begin{aligned}\widehat{\lambda}(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}) &= \int_0^\infty K_b(t - Z_i) d\widehat{\Lambda}(t|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}) \\ &= \sum_{j=1}^{n} K_b(Z_j - Z_i) \frac{\Delta_j K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{\sum_{k=1}^{n} I(Z_k \geq Z_j) K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}, \quad (2.5)\end{aligned}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\lambda}}_1(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}) &= \frac{\partial \widehat{\lambda}(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta})}{\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)} \\ &= -\sum_{j=1}^{n} K_b(Z_j - Z_i) \frac{\Delta_j \mathbf{K}_h'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{\sum_{k=1}^{n} I(Z_k \geq Z_j) K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)} \\ &\quad + \sum_{j=1}^{n} K_b(Z_j - Z_i) \Delta_j K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \\ &\quad \times \frac{\sum_{k=1}^{n} I(Z_k \geq Z_j) \mathbf{K}_h'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{\{\sum_{k=1}^{n} I(Z_k \geq Z_j) K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}^2}. \quad (2.6)\end{aligned}$$

In (2.5) and (2.6), $K(\cdot)$ is a kernel function and $K_h(\cdot) = K(\cdot/h)/h$, $\mathbf{K}_h'(\mathbf{v}) = \partial K_h(\mathbf{v})/\partial \mathbf{v}$ is the first derivative of K_h with respect to its variables (a vector), and h and b are bandwidths. The estimated expectation terms are

$$\widehat{E}\left\{Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\} = \frac{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) I(Z_j \geq Z_i)}{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}, \quad (2.7)$$

$$\widehat{E}\left\{\mathbf{X}_{li} Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\} = \frac{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \mathbf{X}_{lj} I(Z_j \geq Z_i)}{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}. \quad (2.8)$$

We use the Gaussian kernel function throughout the implementation, and obtain the solution of (2.4) using Powell's hybrid method, which is designed for solving nonlinear equations (Powell (1965, 1970)). The last parameter in (2.5), (2.6), (2.7), and (2.8) reflects the occurrence of $\boldsymbol{\beta}$ in $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j$ and $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k$.

The estimator obtained from (2.4) is shown to achieve the smallest possible variability; hence, this estimator is efficient and recommended. The efficient estimator is the focus of our study. We provide a detailed algorithm for the efficient estimation procedure below.

1. Obtain an initial estimator of $\boldsymbol{\beta}$ using, for example, hmave (Xia, Zhang and Xu (2010)). Denote the result $\widetilde{\boldsymbol{\beta}}$.

2. Replace $E\{Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$, $E\{\mathbf{X}_l Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$, $\lambda(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$, and $\boldsymbol{\lambda}_1(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$ with their nonparametric estimated versions given in (2.5), (2.6), (2.7), and (2.8) respectively. Write the resulting estimators as $\widehat{E}\{Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}\}$, $\widehat{E}\{\mathbf{X}_l Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}\}$, $\widehat{\lambda}(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$, and $\widehat{\boldsymbol{\lambda}}_1(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$, respectively.

3. Plug $\widehat{E}\{\mathbf{X}_l Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}\}$, $\widehat{E}\{Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}\}$, $\widehat{\lambda}(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$, and $\widehat{\boldsymbol{\lambda}}_1(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$ into (2.4), and solve the estimating equation to obtain the efficient estimator $\widehat{\boldsymbol{\beta}}$, using $\widetilde{\boldsymbol{\beta}}$ as the starting value.

Here, $E\{Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} \equiv E\{Y(t) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}|_{t=Z}$. Other terms are defined similarly.

Remark 1. According to the derivation, $E\{\mathbf{S}_{\mathrm{eff}}(\Delta, Z, \mathbf{X}) \mid \mathbf{X}\} = \mathbf{0}$ is ensured by $E\{dM(t, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) \mid \mathbf{X}\} = 0$. Hence to preserve the mean zero property, we can replace $\boldsymbol{\lambda}_{10}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})/\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})$ with any function of s and $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}$, say $\mathbf{g}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})$, and still obtain

$$E\int_0^{\infty} \mathbf{g}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) \otimes \left[\mathbf{X}_l - \frac{E\left\{\mathbf{X}_l S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{S_c(s, \mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}\right] dM(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) = \mathbf{0}.$$

This implies that if we are aiming only at a consistent estimator, we can use an arbitrary function $\mathbf{g}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})$ to replace $\boldsymbol{\lambda}_{10}(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})/\lambda_0(s, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})$ in the efficient score to get a more general martingale integration. Hence, a generic estimating equation is given by

$$\sum_{i=1}^{n} \Delta_i \mathbf{g}(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \left[\mathbf{X}_{li} - \frac{\widehat{E}\left\{\mathbf{X}_{li} Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\}}{\widehat{E}\left\{Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\}}\right] = \mathbf{0},$$

for any $\mathbf{g}$.

Remark 2. We can further generalize the estimating equation form to

$$\sum_{i=1}^{n} \Delta_i \mathbf{g}(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \left[\mathbf{a}(\mathbf{X}_{li}) - \frac{\widehat{E}\left\{\mathbf{a}(\mathbf{X}_{li}) Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\}}{\widehat{E}\left\{Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\}}\right] = \mathbf{0}$$

by taking advantage of the fact that

$$E\Delta \mathbf{g}(Z, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) \otimes \left[\mathbf{a}(\mathbf{X}_l) - \frac{E\left\{\mathbf{a}(\mathbf{X}_l) Y(Z) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{Y(Z) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}\right] = \mathbf{0},$$

for any $\mathbf{a}(\mathbf{X}_l)$.

Remark 3. In the algorithm, we used the hmave estimator $\widetilde{\boldsymbol{\beta}}$ as a starting value when solving our efficient estimating equation. This is a choice out of convenience. One can use any other estimator as a starting value, such as the Cox model estimator when $d = 1$, or use any of the estimators described in Remarks 1 and 2.

Remark 4. When solving the estimating equation (2.4) based on data with finite sample size, we may be unable to find a solution. In this, we use the minimizer of

$$\left\| \sum_{i=1}^{n} \Delta_i \frac{\widehat{\boldsymbol{\lambda}}_1(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta})}{\widehat{\lambda}(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta})} \otimes \left[\mathbf{X}_{li} - \frac{\widehat{E}\left\{ \mathbf{X}_{li} Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta} \right\}}{\widehat{E}\left\{ Y_i(Z_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta} \right\}} \right] \right\|_2$$

with respect to $\boldsymbol{\beta}$, where $\| \cdot \|_2$ is the l_2 norm. The proposed numerical procedure does not change because the hybrid method actually solves the estimating equation by minimizing its l_2 norm (Powell (1965, 1970)).

Remark 5. In performing the nonparametric estimation, bandwidths need to be selected. Note that the final estimator is insensitive to the bandwidths, as indicated in Condition C2, Lemma 1, and Theorems 1 and 2, where a range of different bandwidths all lead to the same asymptotic property. Therefore we suggest selecting the corresponding bandwidths by taking the sample size n to its suitable power to satisfy C2, multiplying the standard deviation of the covariate to adjust the range, and then multiplying a constant to scale it. For example, when $d = 1$, we can let h be $n^{-1/3}$ multiplied by the standard deviation of $\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i$ and a constant, and let b be $n^{-1/3}$ multiplied by the standard deviation of Z_i and a constant. Here, the constant can simply be one or any other values, typically in the range of $[0.1, 10]$. The selection of the bandwidths in each problem is discussed in Section 4. In general, using the above construction, there is not much effect when changing the constant in estimating (2.5), (2.6), (2.7), and (2.8). Finally, when the sample size is small, a nonparametric estimator may generate a null value in the denominator. We can either increase the bandwidth or replace it with a small value (Delecroix, Hristache and Patilea (2006)) to facilitate the computation.

3. Asymptotics

We show that the efficient estimator described in Section 2 is root-n consistent, asymptotically normally distributed, and achieves optimal efficiency. Let the parameter space of $\boldsymbol{\beta}$ be $\mathcal{B}$. We first list some regularity conditions.

C1 (*The kernel function.*) The univariate kernel function $K(x)$ is symmetric,

differentiable, bounded, and with bounded derivative. In addition, $K(x)$ is an order ν kernel (i.e., $\int x^j K(x)dx = 0$, for $1 \le j < \nu$, $0 < \int x^\nu K(x)dx < \infty$), and it satisfies $\int K^2(x)dx < \infty$, $\int x^2K^2(x)dx < \infty$, $\int K'^2(x)dx < \infty$, $\int x^2K'^2(x)dx < \infty$, $\int K''^2(x)dx < \infty$, $\int x^2K''^2(x)dx < \infty$. The d-dimension kernel function is a product of d univariate kernel functions; that is, $K(\mathbf{u}) = \prod_{j=1}^d K(u_j)$, for $\mathbf{u} = (u_1, \ldots, u_d)^{\mathrm{T}}$. For simplicity, we use the same K for both univariate and multivariate kernel functions.

C2 (*The bandwidths.*) The bandwidths satisfy $h = n^{-\alpha_h}$, $b = n^{-\alpha_b}$, $\alpha_h > 0$, $\alpha_b > 0$, $1 - \alpha_h(d+2) - \alpha_b > 0$, and $1 - 2\alpha_h\nu < 0$, where $2\nu > d + 1$.

C3 (*The boundedness.*) The parameter space $\mathcal{B}$ is bounded and $\boldsymbol{\beta}_0$ is an interior point of $\mathcal{B}$.

C4 (*The density of index.*) Uniformly for any $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$, the density function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, that is, $f_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}(\cdot)$, has compact support, is bounded away from zero and infinity on its support, and its first four derivatives are bounded.

C5 (*The smoothness.*) For all $\mathbf{X}$ and Z, the absolute values of $E\{\mathbf{X}_j I(Z_j \ge Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Z\}$ and $E\{I(Z_j \ge Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Z\}$, and their first four derivatives, are bounded uniformly component wise. The absolute value of $E\{\mathbf{X}_j\mathbf{X}_j^{\mathrm{T}} I(Z_j \ge Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Z\}$ and its first two derivatives are bounded uniformly component wise.

C6 (*The survival function.*) The survival function $S_c(\tau, \mathbf{X})$ is bounded away from zero. In addition, $S(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$, $S_c(t, \mathbf{X})$, and $f(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ satisfy that $\partial^{i+j}S(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})/\partial t^i\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^j$, $\partial^{i+j}f(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})/\partial t^i\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^j$, and $\partial^{i+j}E\{S_c(t, \mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}/\partial t^i\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^j$ exist and are bounded and bounded away from zero on $[0, \tau]$, for all $i \ge 0, j \ge 0, i + j \le 4$. Here, $\partial^{i+j}E\{S_c(\tau, \mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}/\partial\tau^i\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^j$ is defined as $\lim_{t\to\tau-} \partial^{i+j}E\{S_c(t, \mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}/\partial t^i\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^j$.

C7 (*The uniqueness.*) The equation

$$E\left(\Delta\frac{\boldsymbol{\lambda}_1(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})}{\lambda(Z, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})} \otimes \left[\mathbf{X}_l - \frac{E\left\{\mathbf{X}_l Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{Y(Z) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right\}}\right]\right) = \mathbf{0}$$

has a unique solution on $\mathcal{B}$. Because the true parameter $\boldsymbol{\beta}_0$ satisfies the equation, the unique solution is $\boldsymbol{\beta}_0$.

Here, we included $\boldsymbol{\beta}$ in $\lambda(\cdot)$ and $\boldsymbol{\lambda}(\cdot)$ in Condition C7 to emphasize that the functional forms differ as $\boldsymbol{\beta}$ changes. These conditions are quite commonly imposed in nonparametrics, survival analysis, and estimating equations, and are generally mild. Conditions C1 and C2 contain some basic requirements on the kernel function and the bandwidths, that are common in kernel-related works and can be guaranteed to be satisfied. The boundedness of the parameter space $\mathcal{B}$ in C3 is also satisfied in general. Conditions C4–C6 impose certain boundedness conditions on the event time, censoring time, covariates, their expectations, and their corresponding derivatives that are very mild and usually satisfied (Silverman (1978); Claeskens and Van Keilegom (2003)). Indeed, Condition C6 requires that both the event and the censoring process survival functions be bounded away from zero. This is widely required in the literature to control the tail behavior of survival functions, and it implies that at least some subjects are censored at the end of the study. Note that $S_c(t;\mathbf{X})$ is continuous on $t \in (0,\tau)$, but has a jump at $t=\tau$. To take into account this discontinuity, we define the derivative $\partial^{i+j}E\{S_c(\tau,\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}/\partial\tau^i\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^j$ as $\lim_{t\to\tau-}\partial^{i+j}E\{S_c(t,\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}/\partial t^i\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^j$. In the proofs, such definitions for the derivatives based on the left limits do not alter the derivations, because all the related integration terms have integration limits on $(0,\tau)$. Moreover, Condition C4 can be modified as follows.

C4′ (*The density of index, relaxed.*) Uniformly for any $\boldsymbol{\beta}$ in a local neighborhood of $\boldsymbol{\beta}_0$, the density function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, that is, $f_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}(\mathbf{v})$, is bounded and satisfies the following requirement: there exists a constant $\epsilon>0$, such that $\int_{\{\mathbf{v}:f_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}(\mathbf{v})\le d_n\}} f_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}(\mathbf{v})d\mathbf{v} < n^{-\epsilon}$ for sufficiently large n. Here, $d_n\to 0$ as $n\to\infty$, and $n^{-\epsilon}=O(h^2+n^{-1/2}h^{-1/2})$, where h satisfies Condition C2. In addition, the first four derivatives of $f_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}(\cdot)$ are bounded.

Condition C4′ is a weaker version of Condition C4. It requires that the tail of $f_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}$ be sufficiently thin that the near-zero values of $f_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}(\cdot)$ do not affect the overall performance of our estimator. Under Condition C4′, a trimmed version of the nonparametric estimator is applied to avoid the zero-denominator issue, and it retains the same asymptotic properties. The trimmed estimators of (2.5), (2.6) and (2.7), (2.8) are

$$\widehat{\lambda}(Z_i,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i,\boldsymbol{\beta}) = \sum_{j=1}^{n} \frac{K_b(Z_j-Z_i)\Delta_j K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{\sum_{k=1}^{n} I(Z_k\ge Z_j)K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)} \times I\left\{\frac{1}{n}\sum_{k=1}^{n} I(Z_k\ge Z_j)K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)>d_n\right\}, \quad (3.1)$$

$$\widehat{\boldsymbol{\lambda}}_1(Z_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}) = -\sum_{j=1}^{n} \frac{K_b(Z_j - Z_i)\Delta_j \mathbf{K}_h'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{\sum_{k=1}^{n} I(Z_k \geq Z_j) K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}$$
$$\times I\left\{\frac{1}{n}\sum_{k=1}^{n} I(Z_k \geq Z_j) K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) > d_n\right\}$$
$$+ \sum_{j=1}^{n} K_b(Z_j - Z_i)\Delta_j K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$$
$$\times \frac{\sum_{k=1}^{n} I(Z_k \geq Z_j)\mathbf{K}_h'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{\{\sum_{k=1}^{n} I(Z_k \geq Z_j) K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}^2}$$
$$\times I\left\{\frac{1}{n}\sum_{k=1}^{n} I(Z_k \geq Z_j) K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) > d_n\right\}, \quad (3.2)$$

$$\widehat{E}\left\{Y_i(Z_i) \,|\, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\} = \frac{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) I(Z_j \geq Z_i)}{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}$$
$$\times I\left\{\frac{1}{n}\sum_{k=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) > d_n\right\}, \quad (3.3)$$

$$\widehat{E}\left\{\mathbf{X}_{li} Y_i(Z_i) \,|\, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}\right\} = \frac{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\mathbf{X}_{lj} I(Z_j \geq Z_i)}{\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}$$
$$\times I\left\{\frac{1}{n}\sum_{k=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_k - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) > d_n\right\}. \quad (3.4)$$

Similar estimators are used in Mack and Silverman (1982), Collomb and Härdle (1986), Härdle and Stoker (1989), and Ichimura and Todd (2007). The unique solution requirement in Condition C7 is needed to ensure the convergence of the estimator, and can be relaxed to local uniqueness if needed.

Before presenting the main results, we first summarize several preliminary results. These results highlight the theoretical properties of the kernel-based estimators of several conditional expectations, as well as the estimation properties of the hazard function and its derivative, and hence are of independent interest. These properties also play an important role in the proofs of Theorems 1 and 2.

Lemma 1. *Assume the regularity conditions* C1–C7 *hold. For any* Z, $\mathbf{X}$, $Y(Z)$, *and* $\boldsymbol{\beta}$ *in the parameter space, the estimators defined in* (2.5), (2.6), (2.7), *and* (2.8) *satisfy the following results uniformly for all* $\boldsymbol{\beta}$ *in a local neighborhood of* $\boldsymbol{\beta}_0$*:*

$$\widehat{E}\left\{Y(Z) \,|\, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}\right\} = E\{Y(Z) \,|\, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} + O_p\{(nh)^{-1/2}(log\, n)^{1/2} + h^2\}, \quad (3.5)$$

$$\widehat{E}\left\{\mathbf{X}Y(Z)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\beta}\right\} = E\{\mathbf{X}Y(Z)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} + O_p\{(nh)^{-1/2}(\log n)^{1/2} + h^2\}, \tag{3.6}$$

$$\frac{\partial\widehat{E}\left\{Y(Z)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\beta}\right\}}{\partial\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}} = \frac{\partial E\{Y(Z)\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}}{\partial\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}} + O_p\{(nh^3)^{-1/2}(\log n)^{1/2} + h^2\}, \tag{3.7}$$

$$\frac{\partial\widehat{E}\left\{\mathbf{X}Y(Z)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\beta}\right\}}{\partial\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}} = \frac{\partial E\{\mathbf{X}Y(Z)\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}}{\partial\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}} + O_p\{(nh^3)^{-1/2}(\log n)^{1/2} + h^2\}, \tag{3.8}$$

$$\widehat{\lambda}(Z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\beta}) = \lambda(Z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\beta}) + O_p\{(nhb)^{-1/2}(\log n)^{1/2} + h^2 + b^2\}, \tag{3.9}$$

$$\widehat{\boldsymbol{\lambda}}_1(Z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\beta}) = \boldsymbol{\lambda}_1(Z,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\beta}) + O_p\{(nbh^3)^{-1/2}(\log n)^{1/2} + h^2 + b^2\}. \tag{3.10}$$

If Condition C4 *is replaced by Condition* C4′*, the trimmed estimators* (3.1), (3.2), (3.3)*, and* (3.4) *retain the same results.*

The proof of Lemma 1 is given in the Supplementary Material. We note that the convergence in Lemma 1 holds uniformly with respect to $\boldsymbol{\beta}$ in a local neighborhood of $\boldsymbol{\beta}_0$ and for any bandwidth that satisfies Condition C2.

Theorem 1. *Assume the regularity conditions* C1–C7 *hold, or with Condition* C4 *replaced by Condition* C4′*. The estimator obtained from solving* (2.4) *is consistent, that is,* $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \to \mathbf{0}$ *in probability when* $n \to \infty$.

Theorem 2. *Assume the regularity conditions* C1–C7 *hold, or with Condition* C4 *replaced by Condition* C4′*. The estimator obtained from solving* (2.4) *satisfies*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, [E\{\mathbf{S}_{\mathrm{eff}}^{\otimes 2}(\Delta, Z, \mathbf{X})\}]^{-1})$$

in distribution when $n \to \infty$*. Here,* $\mathbf{S}_{\mathrm{eff}}(\Delta, Z, \mathbf{X})$ *is the efficient score function given in* (2.3)*. Thus, the estimator is efficient.*

Note that because $\mathbf{S}_{\mathrm{eff}}$ is a martingale, we have

$$\begin{aligned}
&E\{\mathbf{S}_{\mathrm{eff}}^{\otimes 2}(\Delta, Z, \mathbf{X})\} \\
&= E\int_0^\infty \left(\frac{\boldsymbol{\lambda}_{10}(s,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})}{\lambda_0(s,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})} \otimes \left[\mathbf{X}_l - \frac{E\left\{\mathbf{X}_l S_c(s,\mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{S_c(s,\mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}\right]\right)^{\otimes 2} \lambda(s,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})Y(s)ds \\
&= E\int_0^\infty \left(\frac{\boldsymbol{\lambda}_{10}(s,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})}{\lambda_0(s,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})} \otimes \left[\mathbf{X}_l - \frac{E\left\{\mathbf{X}_l S_c(s,\mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}{E\left\{S_c(s,\mathbf{X}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}\right\}}\right]\right)^{\otimes 2} dN(s).
\end{aligned}$$

Therefore, a natural estimator of the estimation variance is the inverse of

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\frac{\widehat{\boldsymbol{\lambda}}_1(z_i,\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})}{\widehat{\lambda}(z_i,\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})}\otimes\left[\mathbf{x}_{il}-\frac{\widehat{E}\left\{\mathbf{X}_l S_c(z_i,\mathbf{X})\mid\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}}\right\}}{\widehat{E}\left\{S_c(z_i,\mathbf{X})\mid\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}}\right\}}\right]\right)^{\otimes 2}.$$

4. Numerical Experiments

4.1. Simulation

To evaluate the finite-sample performance of our method, we perform four simulation studies. In the first study, we generate event times from

$$T=\Phi\left[\epsilon\left\{\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right)+1\right\}-3\right],$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution, ϵ has an exponential distribution with parameter 1, and $\mathbf{X}$ follows a standard normal distribution independent of ϵ. We consider $d=1$ and $p=7$, and the true parameter values are taken to be $\boldsymbol{\beta}=(1,0,-1,0,1,0,-1)^{\mathrm{T}}$. We further generate the covariate dependent censoring times using $C=\Phi(2X_2+2X_3)+U$, where U denotes a random variable uniformly distributed on $(0,c_1)$, where c_1 is a constant controlling the censoring proportion.

In the second study, we generate the event times from

$$T=\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}+\epsilon),$$

where ϵ follows a Gumbel distribution with location 0 and rate 5 and each component in $\mathbf{X}$ follows an independent uniform distribution on $(-0.2,\ 0.2)$. We consider $d=1$ and $p=7$, and set the true parameter value to be $\boldsymbol{\beta}=(1,1.3,-1.3,1,-0.5,0.5,-0.5)^{\mathrm{T}}$. We generate the censoring time from a uniform distribution on $(0,c_2)$, where different values of c_2 are used to achieve various censoring rates.

In the third study, we generate the event times from

$$T=\exp\left\{1-(1-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^2+\epsilon\right\},$$

where $\epsilon\sim\mathrm{Normal}(0,1)$, and each component of $\mathbf{X}$ is independently distributed with a uniform distribution on (0, 1). We consider $d=1$ and $p=10$, and set the true parameter value to be $\boldsymbol{\beta}=(1,-0.6,0,-0.3,-0.1,0,0.1,0.3,0,0.6)^{\mathrm{T}}$. The censoring time is generated from $C=U\boldsymbol{\beta}_c^{\mathrm{T}}\mathbf{X}$, where $\boldsymbol{\beta}_c=(0,0,0,1,1,0,0,0,0,0)^{\mathrm{T}}$ and U is uniformly distributed on $(0,c_3)$, and c_3 is a constant controlling the cen-

soring proportion.

In the last simulation study, we increase d to 2 to further evaluate the performance of the proposed method. We set the event times

$$T = \exp\left\{5 - 10\sum_{j=1}^{2}(1 - \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{X})^2 + \epsilon\right\},$$

where $\epsilon \sim \text{Normal}(0, 1)$, and each component of $\mathbf{X}$ is independently distributed with a uniform distribution on (0, 1), and $\boldsymbol{\beta}_j$, for $j = 1, 2$, denotes the jth column of $\boldsymbol{\beta}$ with $p = 6$. We set the true parameter value to be $\boldsymbol{\beta} = \{(1, 0, 2.75, -0.75, -1, 2)^{\mathrm{T}}; (0, 1, -3.125, -1.125, 1, -2)^{\mathrm{T}}\}^{\mathrm{T}}$. The censoring time is generated from a uniform distribution on $(0, c_4)$, where c_4 controls the censoring rate.

These studies are designed to resemble and extend the simulation studies considered in Xia, Zhang and Xu (2010), which proposed hmave, the best method so far in the literature to achieve dimension reduction for censored data. The method hmave minimizes

$$n^{-3}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{\widehat{\lambda}_i(Z_k) - a_{jk} - \mathbf{d}_{jk}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j)\right\}^2 w_{ij}$$

with respect to a_{jk}, $\mathbf{d}_{jk}$, and $\boldsymbol{\beta}$, and extracts $\widehat{\boldsymbol{\beta}}$. Here, $\widehat{\lambda}_i(Z_k)$ is a nonparametric estimator of the conditional hazard function given $\mathbf{X}_i$ evaluated at Z_k, $a_{jk} \in \mathbf{R}$, $\mathbf{d}_{jk} \in \mathbf{R}^d$, and $w_{ij} \equiv K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j)$ is a kernel-based weight. We can understand it as a local linear estimator of $\lambda(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ based on data $\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \widehat{\lambda}_i(t)\}$, for $i = 1, \ldots, n$. The local linear estimator minimizes

$$\sum_{i=1}^{n}\left\{\widehat{\lambda}(t, \mathbf{X}_i) - a_{t,\mathbf{X}} - \mathbf{d}_{t,\mathbf{X}}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}^2 K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}).$$

Now, selecting the set of $t, \mathbf{X}$ values as $t = Z_k$, for $k = 1, \ldots, n$, and $\mathbf{X} = \mathbf{X}_j$, for $j = 1, \ldots, n$ and summing them leads to hmave. Because hmave is parameterized differently, we reparameterize it using $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}\mathbf{A}^{-1}$, where $\widetilde{\boldsymbol{\beta}}$ is the raw hmave estimator, and $\mathbf{A}$ is the upper $d \times d$ submatrix of $\widetilde{\boldsymbol{\beta}}$.

We compare our estimation of both parameters and survival functions with those from hmave, the Cox proportional model (Cox), and the accelerated failure time model (AFT). In terms of estimating the survival function, the semiparametric method calculates $\widehat{S}(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}) = \exp\{-\widehat{\Lambda}(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})\}$ using a local Nelson–Aalen estimator of $\Lambda(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$. In contrast, hmave estimates $S(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$ differently by using a local polynomial regression (Masry (1996)). Cox and AFT

estimate the survival function based on the corresponding fitted models.

In all of the aforementioned studies, we generate 1,000 data sets. In the first study, the sample size $n = 100$ is considered. We set the sample sizes to $n = 200$ for all the remaining studies. In all the nonparametric regression estimators, we set the bandwidths to be $n^{-1/3}$ times the standard deviation of the regressors, multiplied by a constant c. We find that for constants in the range 0.1 to 10, the final results are similar. The results of the first simulation study are given in Table S1 and Figures S1 and S2, where we consider three different censoring rates, 0%, 20%, and 40% respectively. From these results, we can see that the proposed semiparametric method performs better, in general, and sometimes much better, in that it has smaller absolute biases and sample standard errors when estimating $\boldsymbol{\beta}$. To compare our method with hmave, we further compute the estimated projection matrix $\widehat{\mathbf{P}} \equiv \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}$ and the true projection matrix $\mathbf{P} \equiv \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$. The largest singular value of $\widehat{\mathbf{P}} - \mathbf{P}$ serves as another criterion to measure the closeness of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. We find that the mean and the variance of the largest singular value of $\widehat{\mathbf{P}} - \mathbf{P}$ are much smaller based on the semiparametric method than they are based on hmave. The same results are also presented in Figure S1 to provide a quick visual inspection. In Figure S2, for each method, we further plot the average of the 1,000 estimated survival functions $\widehat{S}(t, \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}, \widehat{\boldsymbol{\beta}})$ as a function of t, where we fix $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$ at the empirical mean of the covariate index $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\overline{\mathbf{X}}$. We can see that among all methods, the semiparametric estimator performs best in terms of estimating the survival function as well. We also report the Harrell's concordance index (Harrell, Lee and Mark (1996)) in Table S2, showing that, except for AFT, all methods yield very large values.

The results of the second study are presented in Tables S3 and S4 and Figures S3 and S4. In this study, AFT performs very well in estimating both $\boldsymbol{\beta}$ and $S(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$, and in terms of Harrell's index. This is expected, because the data are generated from an AFT model. The semiparametric method performs better than hmave and Cox in estimating $\boldsymbol{\beta}$, and has competitive performance in estimating $S(t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta})$. It also yields a better Harrell's concordance index than that of Cox. The superiority of the semiparametric method over hmave, Cox, and AFT is more prominent in the third study, as reflected in Tables S5 and S6 and Figures S5 and S6. Here, the semiparametric method is substantially more accurate in estimating each component in $\boldsymbol{\beta}$, yielding smaller biases and variances. The largest singular value of the difference between the estimated and the true projection matrices is also much smaller for the semiparametric method in comparison with the others. Harrell's concordance index is also better or competitive.

When we increase d to 2 in the last simulation, the semiparametric method continues to generate satisfactory results; see Tables S7 and S8 and Figures S7 and S8. In this case, the performance of hmave is rather concerning, possibly because of the difficulties associated with multiple indices. In order to illustrate the performance of the semiparametric method when the number of indices is misspecified, we perform additional estimations by fixing $d = 1$ and $d = 3$, although the true number of indices is 2. The results in Table S9 and Figure S9, respectively, show that the estimation of the survival function at $d = 3$ is similar to that of $d = 2$; thus including a redundant index is wasteful, but does not cause bias. In contrast, the survival function is estimated with large bias when $d = 1$, owing to the model misspecification. In practice, we suggest using the validated information criterion (VIC) (Ma and Zhang (2015)) to determine the suitable number of indices and, thus, protect against model misspecification.

We also perform an additional experiment to further assess the finite-sample performance of the asymptotic results established in Section 3. To this end, we generate covariates $\mathbf{X}$ from a standard normal distribution and event times T from a distribution with hazard function

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \left\{ \sum_{j=1}^{2} \exp\left(\boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{X}\right) \right\},$$

where the baseline hazard $\lambda_0(t) = t$ and the dimension of $\boldsymbol{\beta}$ is $d = 2, p = 6$. We use the parameter values $\boldsymbol{\beta} = \{(1, 0, 2.75, -0.75, -1, 2)^{\mathrm{T}}; (0, 1, -3.125, -1.125, 1, -2)^{\mathrm{T}}\}^{\mathrm{T}}$, and adopt the same censoring process as in the second study to yield a 40% censoring rate. We carry out 1,000 simulations and consider sample sizes $n = 100$, 500, and 1000. The estimation results, together with the sample standard errors, average of the estimated standard deviations, and coverage probabilities of the 95% confidence intervals, are given in Table S10. These results indicate that the large-sample properties of the estimator require a sample size greater than 1,000. However, the general trend is that when the sample size increases, the results approach what we expect based on the asymptotic results, in that the sample standard errors and their estimated versions become closer to each other, and the 95% coverage probabilities get closer to the nominal level. That the asymptotic result requires a very large sample size to illustrate itself is quite common in survival data analysis, and is not unique to our semiparametric method. In the event of a limited sample size in practice, we recommend using the bootstrap method to assess the estimation variability.

4.2. AIDS application

We apply the proposed method to analyze HIV data from AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al. (1996)). In this study, 2,137 HIV-infected subjects are randomized to receive one of four treatments: zidovudine (ZDV) monotherapy, ZDV plus didanosine, ZDV plus zalcitabine, and ddI monotherapy. As in Geng, Zhang and Lu (2015) and Jiang et al. (2017), the survival time of interest is chosen as the time to having a larger than 50% decline in the CD4 count, or progressing to AIDS or death, whichever comes first. In addition to the treatments, there are 12 covariates included in our study, specifically, patient age in years at baseline (X_1), patient weight in kilograms at baseline (X_2), hemophilia indicator (X_3), homosexual activity (X_4), history of IV drug use (X_5), Karnofsky score on a scale of 0–100 (X_6), race (X_7), gender (X_8), antiretroviral history (X_9), symptomatic indicator (X_{10}), number of CD4 at baseline (X_{11}), number of CD8 at baseline (X_{12}), treatment indicator (X_{13}), where $X_{13} = 0$ for treatment ZDV+ddl, and $X_{13} = 1$ for treatment ZDV+Zal. As in Jiang et al. (2017), we only analyze data from the two composite treatments, ZDV plus didanosine and ZDV plus zalcitabine, which have been shown to have significantly better survival than the other two treatments (Geng, Zhang and Lu (2015)). This subset of data contains 1,046 subjects with a censoring rate around 75%. In addition, each covariate is standardized, with no obvious outliers and no missing values.

To determine the proper reduced space dimension d, we employ the VIC (Ma and Zhang (2015)). The VIC is a procedure for determining d that is consistent and applies to general dimension-reduction procedures as long as an estimating equation-based estimator for the parameter is available under any candidate dimension, where d corresponding to the smallest VIC value is selected. In the example, the VIC value at $d = 1$ is 90.38. Further, when $d \geq 2$, the VIC values are all greater than 180.7, which results from the penalty term alone. Hence, we choose $d = 1$ as the final model. Table S11 contains the estimated coefficient $\widehat{\boldsymbol{\beta}}$ under the selected model, with the corresponding estimation standard errors and p-values. Here, we implement the semiparametric estimator to obtain these results owing to its superior theoretical and numerical performance.

The results in Table S11 indicate that in forming the index described by $\widehat{\boldsymbol{\beta}}_{\cdot,1}$, all covariates are significant except the hemophilia indicator (X_3), gender (X_8) and the number of CD4 at baseline (X_{11}). The estimated cumulative hazard functions are reported in Figure S10, plotted as a function of time (upper left panel), a function of the covariate index $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ (upper right panel),

and as a function of both (bottom panel). Specifically, in plotting the cumulative hazard as a function of time t, we fix the covariate index at three different sets of covariate values, $X_{1:12} = (40, 60, 1, 0, 0, 80, 0, 0, 0, 1, 200, 800)^{\mathrm{T}}$, $X_{1:12} = (20, 70, 1, 0, 0, 80, 0, 1, 0, 1, 200, 800)^{\mathrm{T}}$, and $X_{1:12} = (60, 70, 1, 0, 0, 20, 0, 0, 0, 1, 200, 200)^{\mathrm{T}}$, together with the treatment indicator of $X_{13} = 0$ and $X_{13} = 1$. Based on the plots, the estimated cumulative hazard of the treatment ZDV+ddl is slightly larger than that of the treatment ZDV+Zal in all scenarios. In plotting the estimated cumulative hazard $\widehat{\Lambda}$ as a function of the index $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$, we fix the time at $t = 100, 500$, and 1000. Finally, we also plot the cumulative hazard as a function of the two variables t and $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ using a contour plot, where the hazard values are explicitly written out on each contour. We also implemented hmave, Cox, and AFT on the data set for comparison. Specifically, using each method, we performed the analysis on 80% of the individuals, and then calculated the predicted survival times for the remaining 20% of the individuals. The mean residual square (MSE) of the semiparametric method is 63,359.4, which is smaller than Cox (109,394.0), AFT (132,821.8), and hmave (87,713.9). In Figure S11, we provide a box plot of the residuals. We repeated the analysis 20 times with different 80%–20% splits of the data, finding that the MSE of the semiparametric method is always the smallest.

5. Discussion

We have considered a very general model for analyzing time-to-event data subject to censoring. The model allows the event times to link to the covariate indices in an unspecified fashion. Because both the number of indices and the functional form of the linkage to the indices are determined by data, conceptually, the model is maximally flexible. In practice, a relatively low number of indices is expected to avoid the curse of dimensionality. This work was conducted without requiring covariate independent censoring. Instead, we required only event-independent censoring conditional on covariates, which is the minimum requirement for identification. We derived a class of estimators that are consistent and asymptotically normal. We also proposed a procedure for constructing a semiparametric efficient estimator that achieves the optimal estimation variability among all possible consistent estimators.

There are also several limitations, fundamentally due to the dimension-reduction modeling. First, to circumvent the general identifiability issue, we have proposed fixing the upper block of the parameter matrix as the identity. This is a valid choice if the first d components of the covariates are indeed active

in the model. In contrast, if any one of the first d components happens to be inactive, convergence issue will occur during the estimation process. This can be used as a way of selecting the first d components. In other words, one can permutate the covariates, and then use any covariate ordering that does not lead to numerical issues. Second, in practice, when the sample size does not exceed hundreds, the method may yield poor performance when the dimensions p and d are large, say, $p > 30$ and $d > 3$. This is the price paid for model flexibility. If this situation arises, one may consider obtaining more observations or imposing additional model assumptions to enrich the model structure.

Supplementary Material

The Supplementary Material includes detailed proofs of Proposition 1 and 2, Lemma 1, Theorem 1 and 2, and tables and figures of the numerical experiments.

Acknowledgments

This work was supported by the Faculty Development Program, Portland State University (FEAGXZ). The authors thank the editor, associate editor, editorial assistant, and the anonymous reviewers for many helpful comments.

References

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics* **31**, 1852–1884.

Collomb, G. and Härdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Processes and their Applications* **23**, 77–89.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B (Methodological)* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15**, 1–23.

Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference* **136**, 730–769.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Geng, Y., Zhang, H. H. and Lu, W. (2015). On optimal treatment regimes selection for mean survival time. *Statistics in Medicine* **34**, 1169–1184.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H. et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New*

England Journal of Medicine **335**, 1081–1090.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* **84**, 986–995.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Ichimura, H. and Todd, P. E. (2007). Implementing nonparametric and semiparametric estimators. *Handbook of Econometrics* **6**, 5369–5468.

Jiang, R., Lu, W., Song, R. and Davidian, M. (2017). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **79**, 1165–1185.

Li, K.-C., Wang, J.-L. and Chen, C.-H. (1999). Dimension reduction for censored regression data. *The Annals of Statistics* **27**, 1–23.

Lopez, O., Patilea, V. and Van Keilegom, I. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli* **19**, 721–747.

Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regressions. *Biometrics* **67**, 513–523.

Ma, Y. and Zhang, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika* **102**, 409–420.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Probability Theory and Related Fields* **61**, 405–415.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* **17**, 571–599.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodological)* **42**, 109–142.

Powell, M. J. D. (1965). A method for minimizing a sum of squares of non-linear functions without calculating derivatives. *The Computer Journal* **7**, 303–307.

Powell, M. J. D. (1970). A hybrid method for nonlinear equations. In *Numerical Methods for Nonlinear Algebraic Equations* (Edited by Rabinowitz, P.). Gordon and Breach, London.

Prentice, R. L. and Kalbfleisch, J. D. (2003). Mixed discrete and continuous Cox regression model. *Lifetime Data Analysis* **9**, 195–210.

Silverman, B. (1978). Choosing the window width when estimating a density. *Biometrika* **65**, 1–11.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *PNAS* **72**, 20–22.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

Xia, Y., Zhang, D. and Xu, J. (2010). Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association* **105**, 278–290.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **69**, 507–564.

Ge Zhao
Department of Mathematics and Statistics, Portland State University, Portland, OR 97201, USA.
E-mail: gzhao@pdx.edu
Yanyuan Ma
Department of Statistics, Penn State University, University Park, PA 16802, USA.
E-mail: yzm63@psu.edu
Wenbin Lu
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
E-mail: wlu4@ncsu.edu

(Received September 2020; accepted February 2021)