Statistica Sinica **30** (2020), 1285-1311
doi:https://doi.org/10.5705/ss.202017.0550

SUFFICIENT DIMENSION REDUCTION WITH SIMULTANEOUS ESTIMATION OF EFFECTIVE DIMENSIONS FOR TIME-TO-EVENT DATA

Ming-Yueh Huang and Kwun Chuen Gary Chan

Academia Sinica and University of Washington

Abstract: When researchers do not have enough scientific knowledge to assume a particular regression model, sufficient dimension reduction is a flexible yet parsimonious nonparametric framework to study how covariates are associated with an outcome. We propose a novel estimator of low-dimensional composite scores that summarizes the contribution of covariates on a right-censored survival outcome. The proposed estimator determines the degree of dimension reduction adaptively from the data; it estimates the structural dimension, the central subspace, and a rate-optimal smoothing bandwidth parameter simultaneously from a single criterion. The methodology is formulated in a counting process framework. Furthermore, the estimation is free of the inverse probability weighting employed in existing methods, which often leads to instability in small samples. We derive the large-sample properties for the estimated central subspace with a data-adaptive structural dimension and bandwidth. The estimation can be implemented easily using a forward selection algorithm; this implementation is justified by the asymptotic convexity of the criterion in working dimensions. Numerical simulations and two real examples are given to illustrate the proposed method.

Key words and phrases: Central subspace, counting process, data-adaptive bandwidth, higher-order kernel, structural dimension.

1. Introduction

A primary goal of survival analyses is to predict or explain the association between survival times and interesting covariates when the survival time is subjected to censorship caused by the termination of a follow-up study or patients dropping out. In the literature, semiparametric models for right-censored survival data include Cox's proportional hazards model (Cox (1972)), the proportional odds model (Bennett (1983)), and the accelerated failure time model (Cox and Oakes (1984)), among many others. Although semiparametric models do not impose full distributional assumptions, certain parametric structures are still specified for the relation between the response and the covariates. In practice, there is often not enough scientific knowledge to assume a particular transformation or link function. A possible solution is provided by a fully nonparametric regression, such as Beran's estimator (Beran (1981)), for the conditional survival function. When the number of covariates increases, the nonparametric estimator usually suffers from the curse of dimensionality. To consider a more flexible, yet parsimonious model formulation between parametric and nonparametric frameworks, sufficient dimension reduction (Li (1991)) arises as an appealing middle ground, in which the model complexity is controlled by the structural dimension. To obtain the best results, we need to estimate the structural dimension jointly with the central subspace, for which we provide a vigorous solution in this paper for censored survival outcomes collected in biomedical studies.

For uncensored data, various methods have been proposed to estimate the central subspace of the sufficient dimension reduction model with a fixed dimension. These methods include the inverse regression (Li (1991); Li and Wang (2007); Zhu, Zhu and Feng (2010)), minimum average variance estimation coupled with average derivatives (Zhu and Zeng (2006); Xia (2007); Wang and Xia (2008); Yin and Li (2011)), semiparametric framework (Ma and Zhu (2012, 2013)), and reproducing kernel approaches (Fukumizu, Bach and Jordan (2009); Fukumizu and Leng (2014)). To determine the structural dimension, commonly used methods are sequential testing (Li (1991)), a BIC-type criterion (Zhu, Miao and Peng (2006); Ma and Zhang (2015)), cross-validation (Wang and Xia (2008)), and bootstrap (Dong and Li (2010)). Under a right-censoring mechanism, the data structure may not permit direct extensions of these approaches, and only a limited number of methods have been studied. Using an imputation technique, Li, Wang and Chen (1999) proposed a consistent estimator for the central subspace by calculating the conditional expectation of the unobserved part of the response in the sliced inverse regression. Another method, proposed by Lu and Li (2011), used inverse censoring probability weighting (ICPW) to remove the bias caused by censoring; the structural dimension is determined by a BIC-type criterion. Similarly, Nadkarni, Zhao and Kosorok (2011) proposed a minimum discrepancy approach, coupled with inverse censoring weighting, to build a more efficient inverse regression estimator. In addition, they used bootstrapping to estimate the structural dimension. To relax strong assumptions, such as linearity and constant variance conditions, on the design matrices from the conventional sliced inverse regression, Xia, Zhang and Xu (2010) proposed using inverse survival weighting and double kernel smoothing techniques, and used the minimum average variance estimation based on hazard functions (hMAVE). To obtain the structural dimension, the authors applied a cross-validation criterion for the conditional hazard function.

Among these methods, an inverse weighting technique is required to adjust the censored response. However, in practice, the inverse weights often lead to unstable estimators, especially when the values of the weights are close to zero. In this work, we propose a new criterion that focuses directly on the mean function of the counting process for the observed failure event, rather than treating the partially observed failure time as a missing data problem. Hence, no inverse weights are required, and the resulting estimator is more stable than existing estimators. In addition, existing methodologies consider the basis estimation and dimension determination as separate problems, and require different criteria to estimate the parameters of interest. Instead, we use a single criterion to simultaneously estimate the effective dimension, central subspace, and a rate-optimal bandwidth for the estimation of conditional cumulative hazard and survival functions, which eases the burden of computation, in practice. The data-adaptive bandwidth is another important contribution, because existing nonparametric methods often involve a subjective bandwidth which could compromise performance. In addition, no subjective tuning parameters are required.

The rest of this article is organized as follows. Section 2 introduces the model structure. The proposed estimator is introduced in Section 3, and its asymptotic properties are established. In Section 4, a series of simulation studies are conducted and two empirical examples are given in Section 5 to illustrate the proposed methods. Concluding remarks are provided in Section 6. All technical proofs are given in the Appendix.

2. Sufficient Dimension Reduction Model for Censored Survival Data

Let T denote the failure time of interest and $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a covariate vector of interest. The sufficient dimension reduction model is of the form:

$$T \perp\!\!\!\perp \mathbf{X} \,|\, B^{\mathrm{T}}\mathbf{X} \tag{2.1}$$

for some full-rank $p \times d$ parameter matrix B with $d \leq p$, where $\perp\!\!\!\perp$ denotes independence. The column space of B is called a sufficient dimension reduction subspace and is denoted by $\mathrm{span}(B)$. Obviously, (2.1) holds trivially when $d = p$ and B is equal to the $p \times p$ identity matrix, because $T \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{X}$. Moreover, when span$(B_1)$ is a sufficient dimension reduction subspace and span$(B_2) \supseteq$ span(B_1), it is easy to see that span(B_2) is also a sufficient dimension reduction subspace. Thus, the model with fixed d_1 is a submodel of that with fixed $d_2 > d_1$. Owing to this nested structure, the primary parameter of interest is the sufficient dimension subspace with the smallest dimension, which is called the central subspace, and is denoted by $\mathcal{S}_{T|\mathbf{X}}$. The corresponding basis matrix is denoted by B_0, and its dimension d_0 is called the structural dimension. For further discussion on the existence and uniqueness of the central subspace, see Cook (1998).

Another equivalent form of (2.1) is

$$F_T(t \,|\, \mathbf{x}) = F(t, B^{\mathrm{T}}\mathbf{x}) \tag{2.2}$$

for some unknown function $F(\cdot,\cdot)$, where $F_T(t \,|\, \mathbf{x})$ is the conditional distribution function of T given $\mathbf{X} = \mathbf{x}$. Expression (2.2) shows that the sufficient dimension reduction is indeed a distribution regression problem, and that the central subspace can capture all information between T and $\mathbf{X}$. Let $\lambda_T(t \,|\, \mathbf{x})$ be the conditional hazard function of T given $\mathbf{X} = \mathbf{x}$. By the one-to-one relationship between the distribution and the hazard function, (2.2) is equivalent to

$$\lambda_T(t \,|\, \mathbf{x}) = \lambda(t, B^{\mathrm{T}}\mathbf{x}) \tag{2.3}$$

for some unspecified function $\lambda(\cdot,\cdot)$ (Xia, Zhang and Xu (2010)). Under (2.2) and (2.3), $F_T(t \,|\, \mathbf{x})$ and $\lambda_T(t \,|\, \mathbf{x})$ remain the same for any basis matrix B with the same column space. In fact, there are infinitely many basis matrices spanning the same space, which are isomorphic up to a linear transformation. The parameter space of B is a subspace of $\mathbb{R}^{p\times d}$, called the Grassmann manifold $Gr(d, \mathbb{R}^p)$ (Ma and Zhu (2013)).

In survival analyses, the failure time is often censored by a censoring time C. One can only observe $Y = T \wedge C = \min(T, C)$ and the noncensoring indicator $\delta = 1(T \le C)$, where $1(\cdot)$ represents the indicator function. For identifiability, conditional independence between T and C is assumed; that is,

$$T \perp\!\!\!\perp C \,|\, \mathbf{X}. \tag{2.4}$$

The condition in (2.4) is a common assumption in regression analyses of survival data. Let $S_Y(t \,|\, \mathbf{x})$, $S_T(t \,|\, \mathbf{x})$, and $S_C(t \,|\, \mathbf{x})$ be the conditional survival functions of Y, T, and C, respectively, given $\mathbf{X} = \mathbf{x}$. From (2.4), it is easy to see that $S_Y(t \,|\, \mathbf{x}) = S_T(t \,|\, \mathbf{x}) S_C(t \,|\, \mathbf{x})$ and $\mathrm{pr}(\delta = 1 \,|\, \mathbf{X} = \mathbf{x}) = \int_0^\infty S_C(t- \,|\, \mathbf{x}) \mathrm{d}F_T(t \,|\, \mathbf{x})$.

These properties further ensure that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}} \subseteq \mathcal{S}_{T|\mathbf{X}} + \mathcal{S}_{C|\mathbf{X}}$, where the sum $L_1 + L_2$ of two linear subspaces L_1 and L_2 is defined as $\{\mathbf{v}_1 + \mathbf{v}_2 : \mathbf{v}_1 \in L_1, \mathbf{v}_2 \in L_2\}$. Because only Y and (Y, δ) are observable, existing methods for uncensored data can be applied directly in order to obtain $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{(Y,\delta)|\mathbf{X}}$. However, these subspaces cannot recover $\mathcal{S}_{T|\mathbf{X}}$ directly. Thus, we have to investigate the relationship between $\mathcal{S}_{(Y,\delta)|\mathbf{X}}$ and $\mathcal{S}_{T|\mathbf{X}}$ to target the primary parameter of interest.

Because the hazard function can only be identified up to the maximal support of the survival function of Y, denoted by τ, one can only estimate the central subspace of T up to τ, such that B_0 satisfies (2.2) and (2.3) for $t \in [0, \tau]$. For example, when $\lambda_T(t \,|\, \mathbf{x}) = \lambda(t, B_0^{\mathrm{T}}\mathbf{x})$ for $t \in [0, \tau]$, and $\lambda_T(t \,|\, \mathbf{x}) = \lambda(t, \tilde{B}^{\mathrm{T}}\mathbf{x})$ for $t > \tau$ with $\tilde{B} \notin \mathrm{span}(B_0)$, the overall central subspace is $\mathcal{S}_{T\,|\,\mathbf{X}} = \mathrm{span}(B_0) + \mathrm{span}(\tilde{B})$. In such cases, $\tilde{B}$ can never be identified from the right-censored data observable up to τ. However, the proposed method is still able to estimate B_0. Because our method can be applied to finite or infinite τ, for simplicity, we set τ to be $+\infty$ in the following discussion so that the parameter of interest $\mathcal{S}_{T|\mathbf{X}}$ is as the same as $\mathrm{span}(B_0)$.

3. The Proposed Estimator

We propose an estimation criterion based on the counting process $N_t = 1(Y \le t, \delta = 1)$ for the observed failure event. Let $R_t = 1(Y \ge t)$. From (2.3), we have the following.

Proposition 1. $E(\mathrm{d}N_t \,|\, R_t, \mathbf{X} = \mathbf{x}) = R_t \lambda(t, B_0^{\mathrm{T}}\mathbf{x})\mathrm{d}t.$

Proposition 1 transforms the original sufficient dimension reduction problem into a mean-regression problem, using the counting process for the observed failure event as the outcome. Although Proposition 1 seems standard, as do many common methods in survival analyses, our objective of estimating d_0 and B_0 simultaneously poses a unique challenge. Here, we must consider a prediction criterion, shown in (3.3), based on a least squares criterion

$$E\left\{\int_0^\infty \left(N_t - \int_0^t R_s \lambda(s, B^{\mathrm{T}}\mathbf{x})\mathrm{d}s\right)^2 \mathrm{d}F_Y(t)\right\}, \tag{3.1}$$

for the estimation of B_0, where $F_Y(t)$ is the marginal distribution of Y. Note that the expectation is taken with respect to the joint distribution of $(Y, \delta, \mathbf{X})$. Instead of using $E\{\mathrm{d}N_t / E(R_t \,|\, \mathbf{X}) \,|\, \mathbf{X} = \mathbf{x}\} = \lambda(t, B_0^{\mathrm{T}}\mathbf{x})\mathrm{d}t$, as in the existing methods, our approach puts R_t in the conditional mean and, hence, no inverse weight $E(R_t \,|\, \mathbf{x})$ is required. A simple calculation shows that this criterion can be decomposed into

$$\begin{aligned} &E\left\{\int_0^\infty \left(N_t - \int_0^t R_s\lambda(s, B_0^{\mathrm{T}}\mathbf{x})\mathrm{d}s\right)^2 \mathrm{d}F_Y(t)\right\} \\ &\quad + E\left[\int_0^\infty S_Y(t \,|\, \mathbf{X})\{\Lambda(t, B_0^{\mathrm{T}}\mathbf{X}) - \Lambda(t, B^{\mathrm{T}}\mathbf{X})\}^2 \mathrm{d}F_Y(t)\right] \\ &\quad + E\left[\int_0^\infty \int_0^t \{\Lambda(s, B_0^{\mathrm{T}}\mathbf{X}) - \Lambda(s, B^{\mathrm{T}}\mathbf{X})\}^2 \mathrm{d}F_Y(s \,|\, \mathbf{X})\mathrm{d}F_Y(t)\right], \end{aligned} \tag{3.2}$$

where $\Lambda(t, B^{\mathrm{T}}\mathbf{x}) = \int_0^t \lambda(s, B^{\mathrm{T}}\mathbf{x})\mathrm{d}s$ and $F_Y(t \,|\, \mathbf{x}) = 1 - S_Y(t \,|\, \mathbf{x})$. Note that $\{\Lambda(s, B_0^{\mathrm{T}}\mathbf{X}) - \Lambda(s, B^{\mathrm{T}}\mathbf{X})\}^2$ is nonnegative. Thus, when both $S_Y(t \,|\, \mathbf{x})$ and $\lambda(t, B^{\mathrm{T}}\mathbf{x})$ are continuous in $t \in (0, \infty)$ and $S_Y(t \,|\, \mathbf{X}) > 0$ for $t \in [0, \tau]$, it can be shown that the last two terms in (3.2) are equal to zero if and only if $\mathrm{span}(B) \supseteq \mathrm{span}(B_0)$. Thus, the criterion in (3.1) attains its minimum if and only if the column space of B is a sufficient dimension reduction subspace. To further distinguish the overfitted models with $d > d_0$, we follow Huang and Chiang (2017), and propose the following leave-one-out cross-validation criterion for $\Lambda(t, B_0^{\mathrm{T}}\mathbf{x})$.

From Proposition 1, we have

$$\Lambda(t, B_0^{\mathrm{T}}\mathbf{x}) = \int_0^t \frac{E(\mathrm{d}N_s \,|\, B_0^{\mathrm{T}}\mathbf{X} = B_0^{\mathrm{T}}\mathbf{x})}{E(R_s \,|\, B_0^{\mathrm{T}}\mathbf{X} = B_0^{\mathrm{T}}\mathbf{x})}.$$

Thus, a nonparametric estimator for $\Lambda(t, B_0^{\mathrm{T}}\mathbf{x})$ can be

$$\widehat{\Lambda}(t, B_0^{\mathrm{T}}\mathbf{x}) = \int_0^t \frac{\mathrm{d}\widehat{H}(s, B_0^{\mathrm{T}}\mathbf{x})}{\widehat{R}(s, B_0^{\mathrm{T}}\mathbf{x})},$$

where

$$\begin{aligned} \widehat{H}(t, B^{\mathrm{T}}\mathbf{x}) &= \frac{\sum_{i=1}^n N_{it}\mathcal{K}_h(B^{\mathrm{T}}\mathbf{X}_i - B^{\mathrm{T}}\mathbf{x})}{\sum_{i=1}^n \mathcal{K}_h(B^{\mathrm{T}}\mathbf{X}_i - B^{\mathrm{T}}\mathbf{x})}, \\ \widehat{R}(t, B^{\mathrm{T}}\mathbf{x}) &= \frac{\sum_{i=1}^n R_{it}\mathcal{K}_h(B^{\mathrm{T}}\mathbf{X}_i - B^{\mathrm{T}}\mathbf{x})}{\sum_{i=1}^n \mathcal{K}_h(B^{\mathrm{T}}\mathbf{X}_i - B^{\mathrm{T}}\mathbf{x})}, \end{aligned}$$

$N_{it} = 1(Y_i \leq t, \delta_i = 1)$, $R_{it} = 1(Y_i \geq t)$, $\mathcal{K}_h(\mathbf{u}) = \prod_{k=1}^d K(u_k/h)/h$ with $\mathbf{u} = (u_1, \ldots, u_d)^{\mathrm{T}}$, h is a positive bandwidth, and K is a qth-order kernel function. Note that $\widehat{H}(t, B^{\mathrm{T}}\mathbf{x})$ and $\widehat{R}(t, B^{\mathrm{T}}\mathbf{x})$ are kernel smoothing estimators for $H(t, B^{\mathrm{T}}\mathbf{x}) = E[N_t \,|\, B^{\mathrm{T}}\mathbf{X} = B^{\mathrm{T}}\mathbf{x}]$ and $R(t, B^{\mathrm{T}}\mathbf{x}) = E[R_t \,|\, B^{\mathrm{T}}\mathbf{X} = B^{\mathrm{T}}\mathbf{x}]$, respectively. Here, we suggest taking $q = \max\{4, 2\lfloor (d+6)/4 \rfloor\}$ (see Remark 3). Now, let $(Y^0, \delta^0, \mathbf{X}^0)$ be a future run, independent of current data $\{(Y_i, \delta_i, \mathbf{X}_i)\}_{i=1}^n$, $N_t^0 = 1(Y^0 \le t, \delta^0 = 1)$, and $R_t^0 = 1(Y^0 \ge t)$. To perform the cross-validation, we consider a prediction risk

$$E\left[\int_0^\infty \left\{N_t^0 - \int_0^t R_s^0 \mathrm{d}\widehat{\Lambda}(s, B^{\mathrm{T}}\mathbf{X}^0)\right\}^2 \mathrm{d}F_Y(t)\right], \tag{3.3}$$

which can be decomposed into $\sigma_0^2 + b_0^2(B) + \mathrm{MISE}_B(h) + C(B,h)$, where

$$\begin{aligned}
\sigma_0^2 =& \ E\left[\int_0^\infty \left\{N_t^0 - \int_0^t R_s^0 \mathrm{d}\Lambda(s, B_0^{\mathrm{T}}\mathbf{X}^0)\right\}^2 \mathrm{d}F_Y(t)\right], \\
b_0^2(B) =& \ E\left[\int_0^\infty S_Y(t \,|\, \mathbf{X}^0)\{\Lambda(t, B_0^{\mathrm{T}}\mathbf{X}^0) - \Lambda(t, B^{\mathrm{T}}\mathbf{X}^0)\}^2 \mathrm{d}F_Y(t)\right] \\
&+E\left[\int_0^\infty \int_0^t \{\Lambda(s, B_0^{\mathrm{T}}\mathbf{X}^0) - \Lambda(s, B^{\mathrm{T}}\mathbf{X}^0)\}^2 \mathrm{d}F_Y(s \,|\, \mathbf{X}^0) \mathrm{d}F_Y(t)\right], \qquad (3.4) \\
\mathrm{MISE}_B(h) =& \ E\left(\int_0^\infty \left[\int_0^t R_s^0 \mathrm{d}\{\Lambda(s, B^{\mathrm{T}}\mathbf{X}^0) - \widehat{\Lambda}(s, B^{\mathrm{T}}\mathbf{X}^0)\}\right]^2 \mathrm{d}F_Y(t)\right), \text{ and} \\
C(B,h) =& \ E\left(\int_0^\infty \left[\int_0^t R_s^0 \mathrm{d}\{\Lambda(s, B_0^{\mathrm{T}}\mathbf{X}^0) - \Lambda(s, B^{\mathrm{T}}\mathbf{X}^0)\}\right]\right. \\
&\left.\cdot \left[\int_0^t R_s^0 \mathrm{d}\{\Lambda(s, B^{\mathrm{T}}\mathbf{X}^0) - \widehat{\Lambda}(s, B^{\mathrm{T}}\mathbf{X}^0)\}\right] \mathrm{d}F_Y(t)\right). \qquad (3.5)
\end{aligned}$$

Note that the expectation is taken with respect to the joint distribution of $\{(Y_i, \delta_i, \mathbf{X}_i)\}_{i=1}^n$ and $(Y^0, \delta^0, \mathbf{X}^0)$. When $h \to 0$ and $nh^d \to \infty$, we can show that both $\mathrm{MISE}_B(h)$ and $C(B,h)$ converge to zero and, thus, (3.3) is dominated by $\sigma_0^2 + b_0^2(B)$. Because model (2.3) has a nested structure, $b_0^2(B)$ decreases with an increase of the working dimension when the working dimension is less than the structural dimension d_0. Furthermore, as discussed in relation to (3.2), $b_0^2(B) \ge 0$, and the equality holds if and only if span(B) is a sufficient dimension reduction subspace. Thus, the minimum of the prediction risk occurs only when span$(B) \supseteq$ span(B_0). In this case, $C(B,h) = 0$, and (3.3) reduces to $\sigma_0^2 + \mathrm{MISE}_B(h)$. In addition, to minimize $\mathrm{MISE}_B(h) = O_p\{h^{2q} + 1/(nh^d)\}$, the optimal rate of h is $O\{n^{-1/(2q+d)}\}$. Thus, once the working dimension d is greater than or equal to the structural dimension when span$(B) \supseteq$ span(B_0), the prediction risk has an asymptotic order of $\sigma_0^2 + O_p\{n^{-2q/(2q+d)}\}$, which starts to increase in d. In summary, we have the following proposition.

Proposition 2. *Under model* (2.1)*, the basis matrix* B_0 *of the central subspace* $\mathcal{S}_{T|\mathbf{X}}$ *and the optimal bandwidth* $h_0 = c_{d_0} n^{-1/(2q+d_0)}$ *minimize the prediction in* (3.3) *as* $h \to 0$*,* $nh^{d_0} \to \infty$*, and* $n \to \infty$*, where the constant* c_{d_0} *is given in Appendix* A.1*.*

Based on Proposition 2, the proposed estimator for (B_0, h_0) is the minimizer of the sample analogue

$$\text{CV}(B,h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\left\{N_{it} - \int_0^t R_{is}\mathrm{d}\widehat{\Lambda}^{-i}(s, B^{\mathrm{T}}\mathbf{X}_i)\right\}^2 \mathrm{d}\widehat{F}_Y(t),$$

where $\widehat{F}_Y(t)$ is the empirical distribution of $\{Y_i\}_{i=1}^n$, and the superscript $-i$ indicates an estimator based on a sample with the ith subject excluded. Note that $\widehat{\Lambda}^{-i}(t, B^{\mathrm{T}}\mathbf{X}_i)$ and $\widehat{F}_Y(t)$ are both step functions in t, and that the integrals in $\text{CV}(B,h)$ indeed have closed forms for computation. More precisely,

$$\text{CV}(B,h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\{1(Y_i \le Y_k, \delta_i = 1) - \widehat{\Lambda}^{-i}(Y_i \wedge Y_k, B^{\mathrm{T}}\mathbf{X}_i)\}^2.$$

Because the prediction risk is asymptotically convex in d, we utilize the following procedure to obtain the estimator.

Step a. For $d = 0$, calculate

$$\widehat{\text{CV}}(0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty}\left\{N_{it} - \int_0^t R_{is}\mathrm{d}\widehat{\Lambda}(s)\right\}^2 \mathrm{d}\widehat{F}_Y(t),$$

where

$$\widehat{\Lambda}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{\mathrm{d}N_{is}}{1 - \widehat{F}_Y(s^-)}.$$

Step b. For $d \ge 1$, define $(\widehat{B}_d, \widehat{h}_d)$ as the minimizer of $\text{CV}(B,h)$ over all $B \in Gr(d, \mathbb{R}^p)$ and $h \in \mathbb{R}$; then, calculate $\widehat{\text{CV}}(d) = \text{CV}(\widehat{B}_d, \widehat{h}_d)$. Because B is identifiable only up to its column space, we use an iterative procedure for separated B and h to implement the optimization problem.

Step b1. Choose a proper initial value $(\widehat{B}_d^{(0)}, \widehat{h}_d^{(0)})$. A possible choice of $\widehat{h}_d^{(0)}$ is $n^{-1/(2q+d)}$. The choice of $\widehat{B}_d^{(0)}$ is discussed further in Remark 1.

Step b2. For $k = 1, 2, \ldots$, define $\widehat{h}_d^{(k)}$ as the minimizer of $\text{CV}(\widehat{B}_d^{(k-1)}, h)$.

This step is a univariate optimization problem, which can be solved using common methods such as the gradient descent and Newton-type algorithms.

Step b3. Define $\widehat{B}_d^{(k)}$ as the minimizer of $\mathrm{CV}(B, \widehat{h}_d^{(k)})$. The practical implementation is discussed in Remark 1.

Step b4. Repeat Steps b2–b3 until $|\mathrm{CV}(\widehat{B}_d^{(k)}, \widehat{h}_d^{(k)}) - \mathrm{CV}(\widehat{B}_d^{(k-1)}, \widehat{h}_d^{(k-1)})| < \epsilon$, for some pre-chosen $\epsilon > 0$.

Step c. Repeat Step b until $d = \widehat{d}$ with $\widehat{\mathrm{CV}}(\widehat{d}+1) > \widehat{\mathrm{CV}}(\widehat{d})$. The proposed estimator $(\widehat{B}, \widehat{h})$ is then defined as $(\widehat{B}_{\widehat{d}}, \widehat{h}_{\widehat{d}})$.

In Appendix A.4, we show that $\mathrm{CV}(B, h)$ converges to the prediction risk in (3.3) as $n \to \infty$. Thus, its minimizer provides a valid estimator for the central subspace. A distinguishing feature of our estimation procedure is that it estimates the basis matrix and the dimension of the central subspace simultaneously. Thus, it requires less computing time than existing methods do (Xia, Zhang and Xu (2010); Nadkarni, Zhao and Kosorok (2011)). Moreover, the bandwidth used in the estimation criterion is also selected at the same time, and can be used to estimate the conditional survival functions after obtaining the estimated central subspace. Although the cross-validation criterion may not be convex in d for small samples, it is asymptotically convex in d, such that the stopping rule of the forward-searching procedure ensures convergence of the proposed estimator to the global optimum in large samples.

Remark 1. Step b3 can be performed in two ways. First, the Newton-type optimization algorithms for Grassmann manifolds (Edelman, Arias and Smith (1999); Adragni, Cook and Wu (2012)) can be applied to solve the minimization problem. We suggest using the method of Xia, Zhang and Xu (2010) to determine the initial value, which can be computed quickly and does not rely on additional distributional assumptions, as is required in some other existing methods. An alternative way to implement Step b3 is to employ a local coordinate system of the Grassmann manifold (Ma and Zhu (2013)), which transforms the Grassmann manifold optimization to an unconstrained optimization of $(p-d) \times d$ free parameters. The transformation is possible through Gaussian elimination, given a consistent initial value, and a Newton-type algorithm (Fletcher and Reeves (1964)) can be employed directly in the resulting optimization problem. In limited simulations, we found that two methods exhibit similar performance. However, the latter requires slightly less computing time and, thus, is recommended.

Remark 2. Although a cross-validation criterion has also been considered (Xia, Zhang and Xu (2010)), the cross-validation values can be unbounded and are sensitive to bandwidth selection. On the other hand, our proposed method fits the observed failure process using its conditional mean. As a result, the proposed cross-validation function is bounded.

Based on the notation and assumptions in Appendix A.2, the large-sample properties of our proposed estimator are established in the following theorem.

Theorem 1. *Suppose that Assumptions* A1–A5 *are satisfied. Then,* $\mathrm{pr}(\widehat{d} = d_0) \to 1$*,* $\widehat{h}_{\widehat{d}} = O_p\{n^{-1/(2q+d_0)}\}$*, and*

$$\sqrt{n}\mathrm{vec}(\widehat{B} - B_0)1\{\widehat{d} = d_0\} \xrightarrow{d} \mathcal{N}_{pd_0}(0, V^{-1}(B_0)E\{S^{\otimes 2}(B_0)\}V^{-1}(B_0))$$

as $n \to \infty$*. The asymptotic variance is defined in the Appendix.*

Remark 3. We show that $\widehat{h}_d = O_p\{n^{-1/(2q+d)}\}$ for each fixed d in the proof of Theorem 1. Coupled with the restriction in Assumption A3, the order of the kernel function should satisfy $q > \max\{2, (d + 2)/2\}$. Because we always use a symmetric kernel function with an even order, and require that the order be as small as possible, a suggested choice is to take $q = \max\{4, 2\lfloor (d+6)/4 \rfloor\}$ for each working dimension d. Because $q \geq 4$, in the practical implementation, we use the bi-weight kernel $K(u) = (105/64)(1-3u^2)(1-u^2)^2 1(|u| \leq 1)$. More details about higher-order kernel functions can be found in the literature (Hansen (2005)).

4. Simulation Studies

In this section, we investigate the finite-sample performance of our proposed estimator, and compare it with that of the hMAVE (Xia, Zhang and Xu (2010)) and ICPW (Lu and Li (2011)) estimators. We also performed additional simulations using the IRE estimator (Nadkarni, Zhao and Kosorok (2011)); the results were qualitatively similar to the ICPW estimator, and are not presented here. We first consider two different settings, which are slight modifications to existing examples (Xia, Zhang and Xu (2010)). The first is a proportional hazard model

$$\text{M1. } T = \Lambda_0^{-1}\{\varepsilon \exp(6B_0^{\mathrm{T}}\mathbf{X} + 1)\},$$

where $\varepsilon \sim \mathrm{Exp}(1)$ and $\mathbf{X} = (X_1, \ldots, X_7) \sim N(0, I_7)$ are independent, $B_0 = (-0.5, 0, 0.5, 0, -0.5, 0, 0.5)^{\mathrm{T}}$, and $\Lambda_0^{-1}(v) = \Phi\{5(v - 2)\}$ with $\Phi(\cdot)$ being the cumulative distribution function of the standard normal distribution. The censoring time follows $C = \Phi(2X_2 + 2X_3) + c_1$, where c_1 is a constant used to control the proportions of the censoring. The second setting is a nonlinear model

$$\text{M2. } T = \exp\{5 - 10(1 - 2^{1/2}B_0^{\mathrm{T}}\mathbf{X})^2 + \varepsilon\},$$

where $\varepsilon \sim N(0, 0.2^2)$, $X_k \sim \text{Uniform}(0,1)$ independently, for $k = 1, \ldots, 7$, and $B_0 = 2^{-1/2}(1,0,0,0,1,0,\ldots,0)^{\mathrm{T}}$. Furthermore, the censoring time is set as $C = c_2 2^{1/2}\beta_c^{\mathrm{T}}\mathbf{X}$, where $\beta_c = 2^{-1/2}(0,1,0,0,1,0,\ldots,0)^{\mathrm{T}}$ and c_2 is used to control the censoring rate. A more complicated model setting is also considered:

$$\begin{aligned}\text{M3. } \lambda_T(t \,|\, \mathbf{X}) = \ & 10(\phi(t-4)\exp(-\mathbf{X}^{\mathrm{T}}\beta_1) + \phi(t-8)\exp(-0.5\mathbf{X}^T\beta_2) \\ & + \phi(t-14)\exp(2\mathbf{X}^{\mathrm{T}}\beta_3)),\end{aligned}$$

where $\mathbf{X} = (X_1, \ldots, X_{20})$, ϕ is the standard normal density function, $\beta_1 = (1,0,0,0.1,\ldots,0.1)$, $\beta_2 = (0,1,0,0.1,\ldots,0.1)$, $\beta_3 = (0,0,1,0.1,\ldots,0.1)$, and $X_k \sim \text{Uniform}(0,10)$ are independently generated, for $k = 1, \ldots, 20$. The true basis matrix is hence $B_0 = (\beta_1, \beta_2, \beta_3)$. The censoring time $C = c_2 2^{1/2}\beta_c^{\mathrm{T}}\mathbf{X}$, where $\beta_c = 2^{-1/2}(0,1,0,0,1,0,\ldots,0)^{\mathrm{T}}$, and c_3 is used to control the censoring rate. All settings are implemented through 1,000 simulations and the estimation errors for an estimator $\widehat{B}$ are measured by the Frobenius norm of $\widehat{B}(\widehat{B}^{\mathrm{T}}\widehat{B})^{-1}\widehat{B}^{\mathrm{T}} - B_0(B_0^{\mathrm{T}}B_0)^{-1}B_0^{\mathrm{T}}$.

The simulation results are displayed in Tables 1–2, and show that the proposed method selects the correct structure dimension very often. For all settings, the proportion of simulations that select the true dimension increases with the sample size. In addition, our proposed estimator has a smaller average estimation error than those of the hMAVE and ICPW estimators, whereas the variabilities of the estimation errors are fairly comparable. In an ICPW estimation, the conditional censoring distribution is estimated by a flexible kernel-weighted local Kaplan-Meier estimator, which suffers from the curse of dimensionality, and is highly variable when the censoring rate is low; a related conclusion can be found in Lu and Li (2011). Moreover, the poor performance of the ICPW estimator under M2 is probably caused by an additional violation of a linearity condition in covariate distributions.

Because the estimation of the conditional survival function of the observed time is not required in the proposed method, the final estimator is more robust to a misspecification of the censoring distribution, censoring rate, and/or dimension of the covariates. From the computational time shown in Table 2, the proposed method is comparable to hMAVE, and is often faster. Even though hMAVE adopts a local linear regression to estimate the gradient of the conditional hazard function and avoid the nonlinear minimization in the estimation, the method needs an iterative refinement procedure to update the estimator to deal with the curse of dimensionality. In our method, we adopt a forward-selection procedure from a lower dimension to avoid high-dimensional smoothing and estimate the cumulative hazard functions directly, conditioning on fixed subspaces. Because there is no additional refinement, the proposed estimation procedure often performs faster than hMAVE.

5. Applications

5.1. Worcester heart attack study data

The first example is based on the Worcester heart attack study data, which were collected from 1975 to 2001 on all acute myocardial infarction patients admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. The main goal of this study is to describe factors associated with trends over time in the incidence and survival rates following hospital admission for acute myocardial infarction. Because the data set is not fully released, we use a random subsample of 500 patients (Hosmer, Lemeshow and May (2008)) and consider all 13 variables, which are displayed in the first two columns of Table 3. In addition, all variables are standardized to have mean zero and unit variance. There are 215 observed deaths in the study; hence the censoring rate is 57%.

The cross-validation values for the working dimensions $d = 0, 1, 2$ are 0.302, 0.247, and 0.264, respectively, with corresponding standard errors 0.022, 0.024, and 0.023. The standard errors are obtained from Hájek projections, because $\textsc{cv}(B, h)$ has an asymptotic representation as a U-statistic. Thus, the estimated structural dimension is one, and the estimated coefficients of the linear index $\widehat{b}^{\mathrm{T}}X$, with corresponding standard errors, are shown in the third column of Table 3. The estimated bandwidth is 2.538. In general, we detect the same covariates as those detected by hMAVE $(\check{b}_1, \check{b}_2, \check{b}_3, \check{b}_4)$ (Xia, Zhang and Xu (2010)), except for age and complete heart block. Our estimated structural dimension is much smaller than that obtained by hMAVE, and a central subspace with a smaller dimension is preferred in practice. The sample correlation of $\widehat{b}^{\mathrm{T}}X$ and $(\check{b}_1^{\mathrm{T}}X, \check{b}_2^{\mathrm{T}}X, \check{b}_3^{\mathrm{T}}X, \check{b}_4^{\mathrm{T}}X)$ is $(-0.112, 0.416, -0.399, -0.747)$. Thus, the fourth direction is more significant when selected using our cross-validation criterion. To assess the model fitting for the observed failure process, we also calculate the cross-validation value based on hMAVE (0.335), which is 36% larger than our cross-validation value of 0.247, with a 95% confidence interval of (0.200, 0.295). Thus, our method arrives at a more parsimonious estimate and a better fit to the observed data.

5.2. AIDS clinical trials group study 175

The second example is taken from randomized clinical trial that compared the effects of different treatments on adults infected with the human immunodeficiency virus type I (HIV-1), whose CD4 T cell counts were between 200 and 500 per cubic millimeter, at baseline. The patients were randomly assigned to four treatment groups: zidovudine, zidovudine plus didanosine, zidovudine plus zalcitabine, and didanosine, where zidovudine (i.e., the first group) is considered as the baseline comparison group. There are 2,467 patients in this data set. Excluding subjects with missing values or unrecorded relevant information, we consider a subset of 2,139 subjects from the original data, which is found in the R package `speff2trial`. A detailed description of the data can be found in the literature (Hammer et al. (1996)). The events of interest are the diagnosed acquired immune deficiency syndrome (AIDS), which is defined as first occurrence of a decline in CD4 T cell count of at least 50, or death. In this work, we are interested in assessing the effects of baseline covariates in addition to the treatments $\mathbf{X} = (X_1, \ldots, X_{17})$ on the patients' time to event T. The events are observed for 521 subjects (24.4%). The covariates considered are listed in the first two columns of Table 4. The covariates X_1 and X_2 are calculated as log(CD4 counts + 1) and log(CD8 counts + 1), respectively, and are then centralized and standardized. The covariates X_6, X_7, and X_{11} are log-transformed, centralized, and standardized, and X_{14} is centralized and standardized from the original covariates. In the literature, some studies have found that log(CD4 counts + 1) and log((CD4 counts + 1)/(CD8 counts + 1)) may be better predictors than the original log(CD4 counts + 1) and log(CD8 counts + 1). However, under the proposed semiparametric model, the new designed covariates are simply linear combinations of the originals. Thus, they lead to the same conditional survival model and prediction values for the survival time. For the sake of convenience, we simply choose log(CD4 counts + 1) and log(CD8 counts + 1) as our design covariates.

The cross-validation values for the working dimensions $d = 0, 1, 2, 3$ are 0.193, 0.190, 0.188, and 0.189, respectively, with standard errors 0.010, 0.010, 0.010, and 0.010. Our proposed method reveals two linear indices to explain the relationship between T and $\mathbf{X}$. The coefficients of the indices $(\widehat{b}_1^{\mathrm{T}}\mathbf{X}, \widehat{b}_2^{\mathrm{T}}\mathbf{X})$ and the corresponding standard errors are shown in the third and fourth columns of Table 4. The standard errors are obtained by estimating the asymptotic covariance matrix in Theorem 1. The estimated bandwidth is 4.251. The 95% confidence intervals of the treatment arms do not include zero in both central subspace directions, but have opposite signs. To further understand the direction-of-treatment effects, we examine the survival probabilities $\mathrm{pr}\{Y > t \,|\, \widehat{b}_1^{\mathrm{T}}\mathbf{X} = \widehat{b}_1^{\mathrm{T}}\bar{\mathbf{X}} + r\mathbf{e_1}, \widehat{b}_2^{\mathrm{T}}\mathbf{X} = \widehat{b}_2^{\mathrm{T}}\bar{\mathbf{X}}\}$ and $\mathrm{pr}\{Y > t \,|\, \widehat{b}_1^{\mathrm{T}}\mathbf{X} = \widehat{b}_1^{\mathrm{T}}\bar{\mathbf{X}}, \widehat{b}_2^{\mathrm{T}}\mathbf{X} = \widehat{b}_2^{\mathrm{T}}\bar{\mathbf{X}} + r\mathbf{e_2}\}$, where $\bar{\mathbf{X}}$ is the sample mean of $\mathbf{X}$, $\mathbf{e_1} = (1, 0, 0, \ldots, 0)$, $\mathbf{e_2} = (0, 1, 0, \ldots, 0)$, and r is a perturbation parameter. We plot the estimates for $t = 1$ and 2 years in Figure 1. As shown in the solid lines, the survival probabilities increase with an increase in CD4 counts, but remain constant or decrease with an increase in CD8 counts, holding other factors constant. This also shows that survival increases with the first linear index, but decreases, in general, with the second. Therefore, the three treatment arms are associated with improved survival compared with that of the zidovudine only group. Moreover, the relationship between the second linear index and the conditional survival function is nonlinear, which may not be discovered by common regression models.

We also implement the hMAVE method for this data set, yielding an one-dimensional central subspace with basis matrix $\check{b} = (0.027, -0.029, -0.318, 0.178, -0.237, 0.024, -0.053, -0.490, -0.049, 0.327, 0.155, -0.388, -0.071, 0.343, -0.269, -0.085, -0.282)^{\mathrm{T}}$. The cross-validation criterion based on the linear index $\check{b}^{\mathrm{T}}\mathbf{X}$ gives a value of 0.193, which shows a slightly poorer prediction accuracy than that of our estimated linear indices.

6. Discussion

Sufficient dimension reduction is a flexible alternative to regression models to summarize the relationship between a response and a covariate vector when there is not enough prior knowledge to assume a particular regression model. This is well studied for completely observable response data, but far less so for survival data. In this work, we consider dimension reduction for survival data and propose a novel estimation method to estimate the central subspace. This method requires no inverse probability weighting and outperforms existing methods in numerical studies. An appealing feature of the sufficient dimension reduction model is that it does not assume any stringent structure for the conditional survival function. Furthermore, we estimate the effective dimensions simultaneously with the basis matrix.

To estimate the central subspace, the existing literature often suggests using separate criteria to estimate the basis matrix and the structural dimension. Thus, a longer computation is required to calculate the different criteria. The main advantage of the proposed method is that we can estimate the basis and dimension using a single criterion. Thus, the computation burden is eased, in practice. Moreover, the tuning bandwidth can be selected at the same time. We have shown that the estimated bandwidth is $O_p\{n^{-1/(2q+d_0)}\}$, which reaches the optimal rate for a nonparametric estimation of the conditional survival function (Dabrowska (1992)). Indeed, this bandwidth minimizes the integrated mean squared error $\textsc{mise}_{B_0}(h)$, asymptotically. Examining the weak convergence of the estimated conditional survival function is left to future research.

The investigation of the semiparametric efficiency bound for the central subspace under the survival regression setting remains an open problem. A profile likelihood approach may reach the semiparametric efficiency bound for a fixed dimension, but would be unable to select the structural dimension simultaneously, because the associated bandwidth estimator would be sub-optimal, for reasons given in the literature (Hall (1987)). Thus, it becomes a major challenge to find a simple criterion that can be used to estimate the structural dimension and the basis matrix simultaneously and efficiently.

Owing to its connection with the counting process framework, this study provides a novel way to extend the idea into different survival data structures, for example, to left-truncated response or recurrent events. These topics, too, are left to future research.

Acknowledgment

The authors thank the editor, associate editor, and two reviewers for their constructive comments. Both authors are partially supported by the United States National Institutes of Health grant R01HL122212, and the second author is partially supported by the United States National Science Foundation grant DMS1711952.

A. Appendix

A.1. Proof of Proposition 2

Proof. In Section 3 we have seen that the minimum of the prediction risk in (5) is attained if and only if $\text{span}(B) \supseteq \text{span}(B_0)$, which reduces the prediction risk into $\sigma_0^2 + \textsc{mise}_B(h)$. By paralleling the proof steps of Du and Akritas (2002), we

Table 1. The proportion of estimated structural dimension ($\widehat{d}$) and the means and standard deviations (s.d.) of estimated bandwidths. c.r. denotes censoring rate (%).

model	c.r.	n	$\widehat{d}$								bandwidth	
			0	1	2	3	4	5	6	7	mean	s.d.
M1	20	100	0.000	0.924	0.076	0.000	0.000	0.000	0.000	0.000	0.010	0.0351
		200	0.000	0.950	0.050	0.000	0.000	0.000	0.000	0.000	0.012	0.0385
		400	0.000	0.978	0.022	0.000	0.000	0.000	0.000	0.000	0.025	0.1082
	50	100	0.000	0.934	0.066	0.000	0.000	0.000	0.000	0.000	0.009	0.0330
		200	0.000	0.941	0.059	0.000	0.000	0.000	0.000	0.000	0.007	0.0292
		400	0.000	0.962	0.038	0.000	0.000	0.000	0.000	0.000	0.007	0.0274
M2	20	100	0.010	0.875	0.114	0.001	0.000	0.000	0.000	0.000	0.019	0.0720
		200	0.001	0.957	0.040	0.002	0.000	0.000	0.000	0.000	0.012	0.0986
		400	0.000	0.981	0.019	0.000	0.000	0.000	0.000	0.000	0.003	0.0191
	50	100	0.013	0.799	0.183	0.005	0.000	0.000	0.000	0.000	0.030	0.0773
		200	0.000	0.919	0.078	0.003	0.000	0.000	0.000	0.000	0.014	0.0727
		400	0.000	0.976	0.023	0.001	0.000	0.000	0.000	0.000	0.006	0.0645
M3	20	100	0.000	0.000	0.248	0.571	0.172	0.008	0.001	0.000	1.650	0.5278
		200	0.000	0.000	0.176	0.688	0.131	0.005	0.000	0.000	1.551	0.4705
		400	0.000	0.000	0.069	0.848	0.080	0.002	0.001	0.000	1.456	0.3111
	50	100	0.000	0.000	0.290	0.566	0.136	0.007	0.001	0.000	1.553	0.4978
		200	0.000	0.000	0.252	0.618	0.120	0.009	0.000	0.001	1.434	0.4258
		400	0.000	0.000	0.200	0.699	0.099	0.002	0.000	0.000	1.389	0.3921

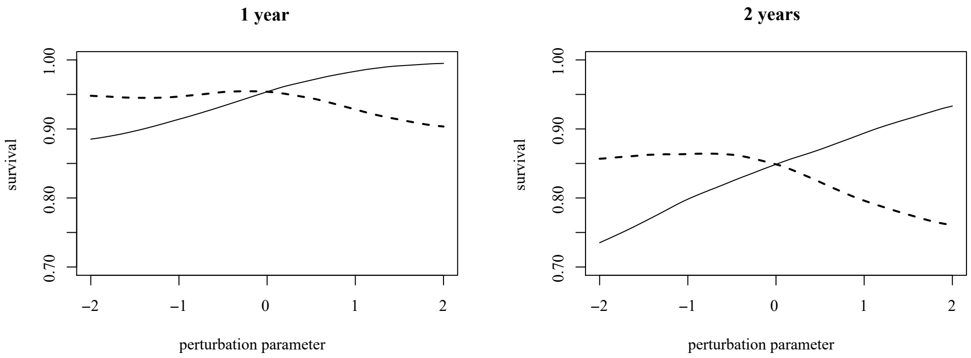

Figure 1. The estimated conditional survival probabilities as functions of covariates, perturbed along the leading coefficient of the first (solid line) and second (dashed line) linear indices around the mean covariates.

Table 2. The means and standard deviations (s.d.) of the basis estimation errors, and the averaged computing time (in seconds). c.r. denotes censoring rate (%).

model	c.r.	n	proposed			hMAVE			ICPW		
			mean	s.d.	time	mean	s.d.	time	mean	s.d.	time
M1	20	100	0.087	0.0860	3.05	0.127	0.1126	3.19	0.489	0.4957	0.03
		200	0.051	0.0522	13.05	0.073	0.0638	11.81	0.217	0.2854	0.14
		400	0.038	0.0461	57.66	0.049	0.0576	97.00	0.077	0.1003	0.57
	50	100	0.076	0.0892	3.11	0.109	0.1143	2.93	0.245	0.1495	0.04
		200	0.039	0.0540	14.56	0.055	0.0694	11.80	0.113	0.0675	0.16
		400	0.028	0.0414	54.59	0.036	0.0509	51.96	0.057	0.0341	0.61
M2	20	100	0.140	0.3303	4.06	0.147	0.3309	19.36	0.911	0.0934	0.04
		200	0.016	0.1184	15.92	0.019	0.1193	65.64	0.912	0.0749	0.18
		400	0.001	0.0004	63.28	0.002	0.0009	278.08	0.910	0.0558	0.68
	50	100	0.173	0.3413	3.69	0.198	0.3379	23.33	0.960	0.0514	0.04
		200	0.040	0.1727	15.33	0.055	0.1737	72.37	0.961	0.0415	0.18
		400	0.004	0.0435	62.22	0.010	0.0443	266.22	0.962	0.0299	0.63
M3	20	100	1.714	0.6722	10.45	3.824	0.2984	23.42	3.146	0.2642	0.06
		200	1.555	0.7308	45.33	3.677	0.2791	86.61	2.937	0.2217	0.22
		400	1.410	0.7672	197.58	3.670	0.3039	335.90	2.825	0.2053	0.87
	50	100	1.865	0.5943	9.26	4.432	0.4375	24.89	3.375	0.3008	0.06
		200	1.677	0.6275	44.05	4.003	0.3615	94.08	3.043	0.2497	0.23
		400	1.550	0.6666	178.37	3.792	0.3633	333.52	2.880	0.2086	0.89

can derive that

$$\begin{aligned}\widehat{\Lambda}(t, B^{\mathrm{T}}\mathbf{x}) - \Lambda(t, B^{\mathrm{T}}\mathbf{x}) =& \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{N_{it}}{R(t, B^{\mathrm{T}}\mathbf{x})} - \int_0^t \frac{R_{is}}{R(s, B^{\mathrm{T}}\mathbf{x})}\mathrm{d}\Lambda(s, B^{\mathrm{T}}\mathbf{x})\right\} \\ & \frac{K_h\{B^{\mathrm{T}}(\mathbf{X}_i - \mathbf{x})\}}{f_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x})} + O_p(h^{2q} + 1/(nh^d)^{3/4}), \qquad \text{(A.1)}\end{aligned}$$

where $f_{B^{\mathrm{T}}\mathbf{X}}(\mathbf{u})$ is the density of $B^{\mathrm{T}}\mathbf{X}$. By substituting (A.1) into $\textsc{mise}_B(h)$ and from the arguments of Härdle and Marron (1985); Härdle, Hall and Marron (1988), we can show that $\textsc{mise}_B(h) = \textsc{amise}_B(h)\{1 + o_p(1)\}$, where

$$\begin{aligned}\textsc{amise}_B(h) = & \; h^{2q}\int_0^\infty\int \mathcal{B}^2(t, \mathbf{x}; B)dF_{\mathbf{X}}(\mathbf{x})dF_Y(t) \\ & + (nh^d)^{-1}\int_0^\infty\int \mathcal{V}(t, \mathbf{x}; B)dF_{\mathbf{X}}(\mathbf{x})dF_Y(t),\end{aligned}$$

Table 3. The estimated coefficients and corresponding standard errors for Worcester heart attack study data.

collected variable	covariate	$\widehat{b}$
initial systolic blood	X_1	1
initial diastolic pressure	X_2	0.836(0.0954)
congestive heart complications	X_3	−0.486(0.0845)
age (in years)	X_4	0.125(0.0792)
myocardial infarction order	X_5	−0.173(0.0917)
body mass index	X_6	−0.528(0.1181)
gender	X_7	−0.510(0.0683)
initial heart rate	X_8	0.553(0.0888)
history of cardiovascular disease	X_9	−0.062(0.0727)
atrial fibrillation	X_{10}	−0.086(0.0816)
cardiogenic shock	X_{11}	0.621(0.0802)
complete heart bolck	X_{12}	0.054(0.0319)
myocardial infarction type	X_{13}	−0.506(0.1061)

$\mathcal{B}^2(t,\mathbf{x};B) = \int\int_0^t S_Y(s-\mid\mathbf{x})\mathrm{d}\mathcal{B}^2_\Lambda(s,\mathbf{x};B)$, and $\mathcal{V}(t,\mathbf{x};B) = \int_0^t S_Y(s-\mid\mathbf{x})\mathrm{d}\mathcal{V}(s,\mathbf{x};B)$ with $F_{\mathbf{X}}(\mathbf{x})$ being the marginal distribution function of $\mathbf{X}$. When $h = c_d n^{-1/(2q+d)}$ with

$$c_d = \left\{ \frac{d\int_0^\infty \int \mathcal{V}(t,\mathbf{x};B)dF_{\mathbf{X}}(\mathbf{x})dF_Y(t)}{2q\int_0^\infty \int \mathcal{B}^2(t,\mathbf{x};B)dF_{\mathbf{X}}(\mathbf{x})dF_Y(t)} \right\}^{1/(2q+d)},$$

$\mathrm{AMISE}_B(h)$ has minimum $\int_0^\infty \int \{c_d^{2q}\mathcal{B}^2(t,\mathbf{x};B) + c_d^{-d}\mathcal{V}(t,\mathbf{x};B)\}dF_{\mathbf{X}}(\mathbf{x})dF_Y(t)$, which is increasing in d. Thus, the prediction risk in (5) attains its minimum when $B = B_0$ and $h = c_{d_0} n^{-1/(2q+d_0)}$.

A.2. Notations and Assumptions

Let $(\cdot)^{\otimes}$ be the Kronecker power of a vector. Define

$$f^{[m]}(\mathbf{x};B) = \partial^m_{B^{\mathrm{T}}\mathbf{x}}[E\{(\mathbf{X}-\mathbf{x})^{\otimes m} \mid B^{\mathrm{T}}\mathbf{X} = B^{\mathrm{T}}\mathbf{x}\} f_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x})],$$

$$G_R^{[m]}(t,\mathbf{x};B) = \partial^m_{B^{\mathrm{T}}\mathbf{x}}[E(R_t \mid B^{\mathrm{T}}\mathbf{X} = B^{\mathrm{T}}\mathbf{x})E\{(\mathbf{X}-\mathbf{x})^{\otimes m} \mid B^{\mathrm{T}}\mathbf{X} = B^{\mathrm{T}}\mathbf{x}\} f_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x})],$$

$$G_H^{[m]}(t,\mathbf{x};B) = \partial^m_{B^{\mathrm{T}}\mathbf{x}}[E(N_t \mid B^{\mathrm{T}}\mathbf{X} = B^{\mathrm{T}}\mathbf{x})E\{(\mathbf{X}-\mathbf{x})^{\otimes m} \mid B^{\mathrm{T}}\mathbf{X} = B^{\mathrm{T}}\mathbf{x}\} f_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x})],$$

for $m = 0, 1, 2$. The estimators $\widehat{\Lambda}(t, B^{\mathrm{T}}\mathbf{x})$ and its derivatives $\partial^m_{\mathrm{vec}(B)}\widehat{\Lambda}(t, B^{\mathrm{T}}\mathbf{x})$

will be shown to converge uniformly to $\Lambda(t, B^{\mathrm{T}}\mathbf{x})$ and $\Lambda^{[m]}(t, \mathbf{x}; B) = \sum_{\ell=0}^{m} \binom{m}{\ell}$ $R^{[\ell-m]}(s, \mathbf{x}; B) dH^{[\ell]}(s, \mathbf{x}; B)$, where

$$
\begin{aligned}
R^{[m]}(t, \mathbf{x}; B) =& \sum_{\ell=0}^{m} \binom{m}{\ell} G_R^{[\ell]}(t, \mathbf{x}; B) f^{[\ell-m]}(\mathbf{x}; B),\\
H^{[m]}(t, \mathbf{x}; B) =& \sum_{\ell=0}^{m} \binom{m}{\ell} G_H^{[\ell]}(t, \mathbf{x}; B) f^{[\ell-m]}(\mathbf{x}; B),\\
f^{[-1]}(\mathbf{x}; B) =& -\frac{f^{[1]}(\mathbf{x}; B)}{f^2_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x})}, \ \ f^{[-2]}(\mathbf{x}; B) = \frac{2(f^{[1]}(\mathbf{x}; B))^2}{f^3_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x})} - \frac{f^{[2]}(\mathbf{x}; B)}{f^2_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x})},
\end{aligned}
$$

for $m = 1, 2$. Moreover, to derive the asymptotic normality of our proposed estimator, we also define the corresponding score vectors and information matrices of $\mathrm{CV}(B, h)$:

$$
\begin{aligned}
S(B) =& \int_0^\infty \left\{ \left(N_t - \int_0^t R_s \mathrm{d}\Lambda(s, B^{\mathrm{T}}\mathbf{X}) \right) \int_0^t R_s \mathrm{d}\Lambda^{[1]}(s, \mathbf{X}; B) \right\} \mathrm{d}F_Y(t),\\
V(B) =& \ E\left(\int_0^\infty \left[\left(\int_0^t R_s \mathrm{d}\Lambda^{[1]}(s, \mathbf{X}; B) \right)^{\otimes 2} \right.\right.\\
& \left.\left. - \left\{ N_t - \int_0^t R_s \mathrm{d}\Lambda(s, B^{\mathrm{T}}\mathbf{X}) \right\} \int_0^t R_s \mathrm{d}\Lambda^{[2]}(s, \mathbf{X}; B) \right] \mathrm{d}F_Y(t) \right).
\end{aligned}
$$

The following regularity conditions are imposed for our theorem:

A1 $\partial_{\mathbf{u}}^{q+2} E(R_t \,|\, B^{\mathrm{T}}\mathbf{X} = \mathbf{u})$, $\partial_{\mathbf{u}}^{q+2} E(N_t \,|\, B^{\mathrm{T}}\mathbf{X} = \mathbf{u})$, $\partial_{\mathbf{u}}^{q+m} E\{(\mathbf{X} - \mathbf{x})^{\otimes m} \,|\, B^{\mathrm{T}}\mathbf{X} = \mathbf{u}\}$, and $\partial_{\mathbf{u}}^{q+2} f_{B^{\mathrm{T}}\mathbf{X}}(\mathbf{u})$ are Lipschitz continuous in u with the Lipschitz constants being independent of $(t, \mathbf{x}, B)$.

A2 $\inf_{(x,B)} f_{B^{\mathrm{T}}\mathbf{X}}(B^{\mathrm{T}}\mathbf{x}) > 0$ and $\inf_{(t,\mathbf{x},B)} R(t, B^{\mathrm{T}}\mathbf{x}) > 0$.

A3 For $d \geq 1$, there exist $\delta \in (1/(4q), 1/\max\{2d+2, d+4\})$ and positive constants $h_{l,d}$ and $h_{u,d}$ such that both ς and h fall in the interval $H_{\delta,n} = [h_{l,d} n^{-\delta}, h_{u,d} n^{-\delta}]$.

A4 $\inf_{\{B: d < d_0\}} b_0^2(B) > 0$ and $b_0^2(B) = 0$ if and only if $B = B_0$ when $d = d_0$.

A5 $V(B_{d,0})$ is non-singular for $d \geq d_0$.

Assumptions A1–A2 are smoothness and boundedness conditions for the population functions to ensure the uniform convergence of kernel estimators used in $\mathrm{CV}(B, h)$. Moreover, assumption A3 is used to remove the remainder term in

Table 4. The estimated coefficients and corresponding standard errors for ACTG175 data.

collected variable	covariate	$\widehat{b}_1$	$\widehat{b}_2$
CD4 T cell count	X_1	1	0
CD8 T cell count	X_2	0	1
treatment arm			
zidovudine and didanosine	X_3	2.964(0.1327)	−1.727(0.2006)
zidovudine and zalcitabine	X_4	1.666(0.0833)	−1.584(0.1852)
didanosine	X_5	1.284(0.1188)	−2.204(0.1188)
v.s. zidovudine			
age (in years)	X_6	−0.172(0.0408)	0.281(0.0536)
weight (in kg)	X_7	−0.217(0.0256)	0.862(0.0475)
hemophilia	X_8	1.002(0.1592)	2.045(0.2823)
homosexual activity	X_9	0.007(0.0955)	0.232(0.1490)
history of intravenous drug use	X_{10}	0.340(0.0885)	−1.933(0.1589)
Karnofsky score	X_{11}	0.489(0.0205)	−1.086(0.0729)
prior treatment			
non-zidovudine antiretroviral	X_{12}	0.947(0.2025)	3.012(0.2981)
zidovudine use in the 30 days	X_{13}	1.686(0.1117)	7.248(0.2151)
number of days of antiretroviral	X_{14}	0.115(0.0628)	1.964(0.1229)
race	X_{15}	−0.035(0.0597)	−1.061(0.1269)
gender	X_{16}	1.192(0.1372)	4.020(0.1503)
symptomatic indicator	X_{17}	−1.033(0.0445)	0.891(0.0894)

the approximation of $\textsc{cv}_Y(B,h)$ and $\textsc{cv}(B,h)$ to their target functions and to establish the $n^{1/2}$-consistency of $\widehat{B}$. Assumptions A4–A5 are made to ensure the identifiability of B_0. One should also note that only $B^{\mathrm{T}}\mathbf{X}$ are required to have continuous density. Thus, some discrete covariates are allowed in our proposal as long as there exists at least one continuous covariate. Related conditions can also be found in assumption A1 in Ma and Zhu (2013).

A.3. Preliminary Lemmas

We first derive the large sample properties of $\partial^m_{\mathrm{vec}(B)}\widehat{\Lambda}(t,B^{\mathrm{T}}\mathbf{x})$ for $m=0,1,2$. To simplify our presentation, the following notations are further introduced:

$$\widehat{G}^{[m]}_{R,c}(t,\mathbf{x};B)=\partial^m_{\mathrm{vec}(B)}\{\widehat{R}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x})\}-G^{[m]}_{R}(t,\mathbf{x};B),$$

$$
\begin{aligned}
\widehat{G}_{H,c}^{[m]}(t,\mathbf{x};B) &= \partial^m_{\mathrm{vec}(B)}\{\widehat{H}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x})\} - G_H^{[m]}(t,\mathbf{x};B),\\
\widehat{H}_c^{[m]}(t,\mathbf{x};B) &= \partial^m_{\mathrm{vec}(B)}\widehat{H}(t,B^{\mathrm{T}}\mathbf{x}) - H^{[m]}(t,\mathbf{x};B),\\
\widehat{R}_c^{[m]}(t,\mathbf{x};B) &= \partial^m_{\mathrm{vec}(B)}\widehat{R}(t,B^{\mathrm{T}}\mathbf{x}) - R^{[m]}(t,\mathbf{x};B),\\
\widehat{\Lambda}_c^{[m]}(t,\mathbf{x};B) &= \partial^m_{\mathrm{vec}(B)}\widehat{\Lambda}(t,B^{\mathrm{T}}\mathbf{x}) - \Lambda^{[m]}(t,\mathbf{x};B)\\
\widehat{f}_c^{[m]}(\mathbf{x};B) &= \partial^m_{\mathrm{vec}(B)}\widehat{f}(B^{\mathrm{T}}\mathbf{x}) - f^{[m]}(\mathbf{x};B),
\end{aligned}
$$

where $\widehat{f}(B^{\mathrm{T}}\mathbf{x}) = n^{-1}\sum_{i=1}^n \mathcal{K}_h\{B^{\mathrm{T}}(\mathbf{X}_i-\mathbf{x})\}$. Moreover, we define the strong representations for $\partial^m_{\mathrm{vec}(B)}\widehat{\Lambda}(t,B^{\mathrm{T}}\mathbf{x})$, $m=0,1$, as follows:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \xi_{\Lambda,i}(t,\mathbf{x};B) &= \int_0^t \frac{\mathrm{d}\widehat{H}_c^{[0]}(s,\mathbf{x};B)}{R(s,B^{\mathrm{T}}\mathbf{x})} - \frac{\widehat{R}_c^{[0]}(s,\mathbf{x};B)\mathrm{d}\Lambda(s,B^{\mathrm{T}}\mathbf{x})}{R(s,B^{\mathrm{T}}\mathbf{x})},\\
\frac{1}{n}\sum_{i=1}^n \xi^{[1]}_{\Lambda,i}(t,\mathbf{x};B) &= \int_0^t \frac{\mathrm{d}\widehat{H}_c^{[1]}(s,\mathbf{x};B)}{R(s,B^{\mathrm{T}}\mathbf{x})} + \frac{\widehat{R}_c^{[1]}(s,\mathbf{x};B)\mathrm{d}\Lambda(s,B^{\mathrm{T}}\mathbf{x})}{R(s,B^{\mathrm{T}}\mathbf{x})}\\
&\quad - \frac{\widehat{R}^{[1]}(s,\mathbf{x};B)\mathrm{d}\widehat{H}_c^{[0]}(s,\mathbf{x};B)}{R^2(s,B^{\mathrm{T}}\mathbf{x})}\\
&\quad - \frac{\widehat{R}_c^{[0]}(s,\mathbf{x};B)\{\mathrm{d}H^{[1]}(s,\mathbf{x};B) + 2R^{[1]}(s,\mathbf{x};B)\mathrm{d}\Lambda(s,B^{\mathrm{T}}\mathbf{x})\}}{R^2(s,B^{\mathrm{T}}\mathbf{x})}.
\end{aligned}
$$

Since the VC-indices of $\{1(Y\le y): y\in\mathbb{R}\}$, $\{a_1K^{(k)}(\mathbf{X}^{\mathrm{T}}b+a_2): a_1,a_2\in\mathbb{R}, b\in\mathbb{R}^d\}$, and $\{(\mathbf{X}-\mathbf{x})^{\otimes k}:\mathbf{x}\in\mathbb{R}^d\}$ are 1, d, and 1, respectively for $k=0,1,2$, these classes is ensured to be Euclidean by Lemma 2.12 of Pakes and Pollard (1989). Coupled with Lemma 2.14 of Pakes and Pollard (1989) and Theorem II.37 of Pollard (1984), we can show that

$$
\begin{aligned}
&\sup_{\mathbf{x},B}\|\partial^m_{\mathrm{vec}(B)}\widehat{f}(B^{\mathrm{T}}\mathbf{x}) - E\{\partial^m_{\mathrm{vec}(B)}\widehat{f}(B^{\mathrm{T}}\mathbf{x})\}\| = o\left\{\frac{\log n}{n^{1/2}h^{(d+2m)/2}}\right\},\\
&\sup_{\mathbf{x},B}\|\partial^m_{\mathrm{vec}(B)}(\widehat{R}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x})) - E[\partial^m_{\mathrm{vec}(B)}(\widehat{R}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x}))]\|\\
&\qquad = o\left\{\frac{\log n}{n^{1/2}h^{(d+2m)/2}}\right\},\\
&\sup_{\mathbf{x},B}\|\partial^m_{\mathrm{vec}(B)}(\widehat{H}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x})) - E[\partial^m_{\mathrm{vec}(B)}(\widehat{H}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x}))]\|\\
&\qquad = o\left\{\frac{\log n}{n^{1/2}h^{(d+2m)/2}}\right\}
\end{aligned}
$$

almost surely. By assumption A1, one can further derive that

$$\begin{aligned}\sup_{\mathbf{x},B}\|E\{\partial^m_{\mathrm{vec}(B)}\widehat{f}(B^{\mathrm{T}}\mathbf{x})\}-f^{[m]}(\mathbf{x};B)\|&=O(h^q),\\ \sup_{\mathbf{x},B}-\|E[\partial^m_{\mathrm{vec}(B)}\{\widehat{R}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x})\}]-G^{[m]}_R(t,\mathbf{x};B)\|&=O(h^q),\\ \sup_{\mathbf{x},B}\|E[\partial^m_{\mathrm{vec}(B)}\{\widehat{H}(t,B^{\mathrm{T}}\mathbf{x})\widehat{f}(B^{\mathrm{T}}\mathbf{x})\}]-G^{[m]}_H(t,\mathbf{x};B)\|&=O(h^q).\end{aligned}$$

Coupled with the triangular inequality, we obtain the following lemma:

Lemma 1. *Suppose that assumption* A1 *is satisfied. Then,*

$$\begin{aligned}\sup_{\mathbf{x},B}\|\widehat{f}^{[m]}_c(\mathbf{x};B)\|&=O(h^q)+o\left\{\frac{\log n}{n^{1/2}h^{(d+2m)/2}}\right\},\\ \sup_{\mathbf{x},B}\|\widehat{G}^{[m]}_{R,c}(t,\mathbf{x};B)\|&=O(h^q)+o\left\{\frac{\log n}{n^{1/2}h^{(d+2m)/2}}\right\},\\ \sup_{\mathbf{x},B}\|\widehat{G}^{[m]}_{H,c}(t,\mathbf{x};B)\|&=O(h^q)+o\left\{\frac{\log n}{n^{1/2}h^{(d+2m)/2}}\right\}\end{aligned}$$

almost surely.

By applying the Taylor expansion thoerem and the results in Lemma 1, one can further ensure from assumptions A2–A3 that

Lemma 2. *Suppose that assumptions* A1–A3 *are satisfied. Then,*

$$\begin{aligned}\sup_{\mathbf{x},B}\left\|\widehat{\Lambda}^{[0]}_c(t,\mathbf{x};B)-\frac{1}{n}\sum_{i=1}^n\xi_{\Lambda,i}(t,\mathbf{x};B)\right\|&=o_p(n^{-1/2}),\\ \sup_{\mathbf{x},B}\left\|\widehat{\Lambda}^{[1]}_c(t,\mathbf{x};B)-\frac{1}{n}\sum_{i=1}^n\xi^{[1]}_{\Lambda,i}(t,\mathbf{x};B)\right\|&=o_p(n^{-1/2}).\end{aligned}$$

A.4. Proof of Theorem 1

Proof. The proof is very similar to that in Huang and Chiang (2017). Thus, we only outline the steps here. Let $\mathrm{ECV}(B,h)=\sigma_0^2+b_0^2(B)+\mathrm{AMISE}_B(h)$. The first step is to show the uniform convergence of $\mathrm{CV}(B,h)$ to $\mathrm{ECV}(B,h)$. By substituting $E_{it}=N_{it}-\int_0^t R_{is}\mathrm{d}\Lambda(s,B_0^{\mathrm{T}}\mathbf{x})$, $M_{it}(B)=\int_0^t R_{is}\mathrm{d}\{\Lambda(s,B_0^{\mathrm{T}}\mathbf{x})-\Lambda(s,B^{\mathrm{T}}\mathbf{x})\}$, and $P_{it}(B)=\int_0^t R_{is}\mathrm{d}\{\Lambda(s,B^{\mathrm{T}}\mathbf{x})-\widehat{\Lambda}(s,B^{\mathrm{T}}\mathbf{x})\}$ into the proof of Theorem 1 in

Huang and Chiang (2017), we have

$$\sup_{B,h}\frac{|\text{CV}(B,h)-\text{ECV}(B,h)|}{\text{AMISE}_B(h)}=o(1)\ \ a.s.\text{ for span}(B)\supseteq\text{span}(B_0),\tag{A.2}$$

$$\sup_{B,h}\frac{|\text{CV}(B,h)-\text{ECV}(B,h)|}{b_0(B)\text{AMISE}_B^{1/2}(h)}=O(1)\ \ a.s.\text{ for span}(B)\not\supseteq\text{span}(B_0).\tag{A.3}$$

The second step is to show that the underestimated dimension will be asymptotically excluded. Denote $\text{DCV}(B,h)=\text{CV}(B,h)-\text{ECV}(B,h)$. By virtue of the minimizer $(\widehat{B},\widehat{h})$ of $\text{CV}(B,h)$ and the Boole's inequality, we have the following inequalities:

$$\begin{aligned}1\leq&\text{pr}\{b_0^2(\widehat{B})<\varepsilon\}\\&+\text{pr}\left\{b_0^2(\widehat{B})\geq\varepsilon,\frac{\text{DCV}(\widehat{B},\widehat{h})}{b_0(\widehat{B})}+\frac{\text{DCV}(B_0,h_0)}{\varepsilon^{1/2}}\geq\varepsilon^{1/2}-\frac{\text{AMISE}_{B_0}(h_0)}{\varepsilon^{1/2}}\right\}\end{aligned}\tag{A.4}$$

for any $\varepsilon>0$. Since $\text{DCV}(\widehat{B},\widehat{h})/b_0(\widehat{B})=O_p\{\text{AMISE}_{\widehat{B}}^{1/2}(\widehat{h})\}\to 0$, $\text{DCV}(B_0,h_0)/\varepsilon^{1/2}=o_p\{\text{AMISE}_{B_0}(h_0)\}\to 0$, and $\text{AMISE}_{B_0}(h_0)\to 0$, one has $\text{pr}\{b_0^2(\widehat{B})<\varepsilon\}\to 1$ for any $\varepsilon>0$. Now by taking $\varepsilon=\inf_{\{B:d<d_0\}}b_0^2(B)/2$ and using the Boole's inequality again, we have $\text{pr}(\widehat{d}\geq d_0)\to 1$.

In the third step, we derive the asymptotic properties of $\widehat{B}_d$ for $d\geq d_0$. Similar to the derivation in the second step, we can also show that

$$\text{pr}\{b_0^2(\widehat{B}_d)<\varepsilon\}\to 1\text{ as }n\to\infty\text{ for any }\varepsilon>0.\tag{A.5}$$

Since $\text{span}(\widehat{B}_d)\supseteq\text{span}(B_0)$ implies that $b_0^2(\widehat{B}_d)=0$, we now consider the case when $\text{span}(\widehat{B}_d)\not\supseteq\text{span}(B_0)$ and, hence, $\widehat{B}_d\xrightarrow{p}B_{d,0}$. By the first-order Taylor expansion of $\partial_{\text{vec}(B)}\text{CV}(B,h)$ at $B=B_{d,0}$ and $\partial_{\text{vec}(B)}\text{CV}(\widehat{B}_d,\widehat{h}_d)=0$, it yields that

$$\begin{aligned}&[I_{pd}+V^{-1}(B_{d,0})\{\partial^2_{\text{vec}(B)}\text{CV}(\widehat{B}^*_d,\widehat{h}_d)-V(B_{d,0})\}]n^{1/2}\text{vec}(\widehat{B}_d-B_{d,0})\\&\qquad=n^{1/2}V^{-1}(B_{d,0})\partial_{\text{vec}(B)}\text{CV}(B_{d,0},\widehat{h}_d)\end{aligned}\tag{A.6}$$

where $\text{vec}(\widehat{B}^*_d)$ lies on the line segment between $\text{vec}(\widehat{B}_d)$ and $\text{vec}(B_{d,0})$. Similar to the approximation in the proof of Theorem 2 in Huang and Chiang (2017), we have

$$n^{1/2}\text{vec}(\widehat{B}_d-B_{d,0})\xrightarrow{d}N(0,V^{-1}(B_{d,0})E\{S^{\otimes 2}(B_{d,0})\}V^{-1}(B_{d,0})),\tag{A.7}$$

and, hence, $b_0^2(\widehat{B}) = O_p(n^{-1})$. Coupled with assumption A3, it further implies that

$$\frac{b_0^2(\widehat{B})}{\text{AMISE}_{\widehat{B}}(\widehat{h})} = o_p(1). \tag{A.8}$$

To show the consistency of $(\widehat{d}, \widehat{h})$ and asymptotic normality of $\widehat{B}$, we define the following sets first:

$$E_0 = \left\{ b_0^2(\widehat{B}) < \log\frac{n}{n}, \widehat{h} \in H_{1/(2q+\widehat{d}),n}, \widehat{d} = d_0 \right\}, \quad E_1 = \left\{ b_0^2(\widehat{B}) \geq \log\frac{n}{n} \right\},$$

$$E_2 = \{\widehat{d} < d_0\}, \quad E_3 = \left\{ b_0^2(\widehat{B}) < \log\frac{n}{n}, \widehat{d} \geq d_0, \widehat{h} \in H_{\widehat{\delta},n} \text{ with } \widehat{\delta} \neq \frac{1}{(2q+\widehat{d})} \right\},$$

$$E_4 = \left\{ b_0^2(\widehat{B}) < \log\frac{n}{n}, \widehat{h} \in H_{1/(2q+\widehat{d}),n}, \widehat{d} > d_0 \right\},$$

and $E_{con} = \{\text{DCV}(\widehat{B}, \widehat{h}) + \text{DCV}(B_0, h_0) \geq \text{ECV}(\widehat{B}, \widehat{h}) - \text{ECV}(B_0, h_0)\}$.

By the minimizer $(\widehat{B}, \widehat{h})$ of $\text{CV}(B, h)$ and the Boole's inequality, one has

$$1 = \text{pr}\{\text{CV}(\widehat{B}, \widehat{h}) \leq \text{CV}(B_0, h_0)\} \leq \text{pr}(E_0) + \sum_{m=1}^{4} \text{pr}(E_{con} \cap E_m). \tag{A.9}$$

From $b_0^2(\widehat{B}) = O_p(n^{-1})$, we have

$$\text{pr}(E_{con} \cap E_1) \leq \text{pr}(E_1) \to 0 \text{ as } n \to \infty. \tag{A.10}$$

Moreover, from $\text{pr}(\widehat{d} \geq d_0) \to 1$ we have

$$\text{pr}(E_{con} \cap E_2) \leq \text{pr}(\widehat{d} < d_0) \to 0 \text{ as } n \to \infty. \tag{A.11}$$

Since $\text{AMISE}_B(h) = O\{h^{2q} + 1/(nh^d)\}$, $\text{AMISE}_{B_0}(h_0) = O\{n^{-2q/(2q+d_0)}\} \leq Cn^{-2q/(2q+\widehat{d})}$ for some constant C when $\widehat{d} \geq d_0$ as $n \to \infty$. Thus,

$$\begin{aligned} &\text{pr}(E_{con} \cap E_3) \\ &\quad \leq \text{pr}\left\{ \frac{\text{DCV}(\widehat{B}, \widehat{h}) + \text{DCV}(B_0, h_0)}{-n^{2q/(2q+\widehat{d})}} \geq \overline{\mathcal{B}}^2 n^{-2q\left(\widehat{\delta} - 1/(2q+\widehat{d})\right)} + \overline{\mathcal{V}} n^{-\widehat{d}\left(1/(2q+\widehat{d}) - \widehat{\delta}\right)} - C, \right. \\ &\qquad \left. \widehat{d} \geq d_0 \right\} \to \ 0, \end{aligned} \tag{A.12}$$

where $\overline{\mathcal{B}}^2 = \int_0^\infty \int \mathcal{B}^2(t,\mathbf{x};B_0)dF_{\mathbf{X}}(\mathbf{x})dF_Y(t)$ and $\overline{\mathcal{V}} = \int_0^\infty \int \mathcal{V}(t,\mathbf{x};B_0)dF_{\mathbf{X}}(\mathbf{x})$ $dF_Y(t)$, since the left-hand side converges to zero by (A.2), (A.3), and (A.8) and the right-hand side tends to infinity when $\widehat{\delta} \neq 1/(2q+\widehat{d})$ as $n \to \infty$. Further, we also have

$$\begin{aligned}
&\mathrm{pr}(E_{con} \cap E_4) \\
&\quad \leq \mathrm{pr}\left\{ \frac{\mathrm{DCV}(\widehat{B},\widehat{h}) + \mathrm{DCV}(B_0,h_0)}{n^{-2q/(2q+d_0)}} \geq C_{\widehat{d}} n^{2q/(2q+d_0)-2q/(2q+\widehat{d})} - C_{d_0}, \widehat{d} > d_0 \right\} \\
&\quad \to 0, \qquad\qquad (\mathrm{A}.13)
\end{aligned}$$

since the left-hand side converges to zero by (A.2), (A.3), and (A.8) and the right-hand side tends to infinity when $\widehat{d} > d_0$ as $n \to \infty$. By substituting (A.10)–(A.13) into (A.9), we immediately have

$$\mathrm{pr}(E_0) \to 1 \text{ as } n \to \infty. \qquad (\mathrm{A}.14)$$

Finally, the asymptotic normality in Theorem 1 is ensured by (A.14) and (A.7).

References

Adragni, K. P., Cook, D. R. and Wu, S. (2012). Grassmannoptim: An R package for Grassmann manifold optimization. *Journal of Statistical Software* **50**, 1–18.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–277.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley.

Cook, D. R. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. John Wiley & Sons.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34**, 187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* CRC Press.

Dabrowska, D. M. (1992). Variable bandwidth conditional Kaplan-Meier estimate. *Scand. J. Statist.* **19**, 351–361.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97**, 279–294.

Du, Y. and Akritas, M. G. (2002). Uniform strong representation of the conditional Kaplan-Meier process. *Math. Methods Statist.* **11**, 152–182.

Edelman, A., Arias, T. A. and Smith, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303–353.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *Comput. J.* **7**, 149–154.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37**, 1871–1905.

Fukumizu, K. and Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *J. Amer. Statist. Assoc.* **109**, 359–370.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Ann. Statist.* **15**, 1491–1519.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M. ,Hirsch, M. S. and Merigan, T. C. (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.

Hansen, B. E. (2005). Exact mean integrated squared error of higher order kernel estimators. *Econometric Theory* **21**, 1031–1057.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83**, 86–101. With comments by David W. Scott and Iain Johnstone and a reply by the authors.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465–1481.

Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-event Data*. Wiley-Interscience.

Huang, M.-Y. and Chiang, C.-T. (2017). An effective semiparametric estimation approach for the sufficient dimension reduction model. *J. Amer. Statist. Assoc.* **112**, 1296–1310.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316–342. With discussion and a rejoinder by the author.

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

Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regression. *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.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107**, 168–179.

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41**, 250–268.

Nadkarni, N. V., Zhao, Y. and Kosorok, M. R. (2011). Inverse regression estimation for censored data. *J. Amer. Statist. Assoc.* **106**, 178–190.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027–1057.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103**, 811–821.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654–2690.

Xia, Y., Zhang, D. and Xu, J. (2010). Dimension reduction and semiparametric estimation of survival models. *J. Amer. Statist. Assoc.* **105**, 278–290.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39**, 3392–3416.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101**, 630–643.

Zhu, L.-P., Zhu, L.-X. and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.* **105**, 1455–1466.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Assoc.* **101**, 1638–1651.

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

E-mail: myh0728@stat.sinica.edu.tw

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.

E-mail: kcgchan@uw.edu

(Received February 2018 ; accepted August 2018)