Semiparametric Additive Transformation Model under Current Status Data

Guang Cheng* and Xiao Wang†

Purdue University

Abstract

We consider the efficient estimation of the semiparametric additive transformation model with current status data. A wide range of survival models and econometric models can be incorporated into this general transformation framework. We apply the B-spline approach to simultaneously estimate the linear regression vector, the nondecreasing transformation function, and a set of nonparametric regression functions. We show that the parametric estimate is semiparametric efficient in the presence of multiple nonparametric nuisance functions. An explicit consistent B-spline estimate of the asymptotic variance is also provided. All nonparametric estimates are smooth, and shown to be uniformly consistent and have faster than cubic rate of convergence. Interestingly, we observe the convergence rate interfere phenomenon, i.e., the convergence rates of B-spline estimators are all slowed down to equal the slowest one. The constrained optimization is not required in our implementation. Numerical results are used to illustrate the finite sample performance of the proposed estimators.

Key Words: B-spline; Consistent variance estimation; Current status data; Efficient estimation; Semiparametric transformation models

1 Introduction

We consider the efficient estimation of the following semiparametric additive transformation model:

$$H(U) = Z'\beta + \sum_{j=1}^{d} h_j(W_j) + \epsilon, \tag{1}$$

*Corresponding Author, Department of Statistics, Purdue University, West Lafayette, IN 47907, Email: chengg@purdue.edu.

†Department of Statistics, Purdue University, West Lafayette, IN 47907, Email: wangxiao@purdue.edu.

where $H(\cdot)$ is a monotone transformation function, $h_j(\cdot)$'s are smooth regression functions (with possibly different degrees of smoothness), and ϵ has a known distribution $F(\cdot)$ with support $\mathbb{R}$. A wide range of survival models and econometric models can be incorporated into the above general transformation framework, e.g., (Huang & Rossini, 1997; Shen, 1998; Huang, 1999; Banerjee et al., 2006, 2009). In particular, the model (1) can be readily applied to a failure time T by letting $U = \log T$. We can obtain the partly linear additive Cox model, i.e., Huang (1999), by assuming $F(s) = 1-\exp(-e^s)$ and $H(u) = \log A(e^u)$, where A is an unspecified cumulative hazard function. Specifically, the hazard function of T, given the covariates (z, w), has the form

$$\lambda(t|z, w) = a(t) \exp(\tilde{\beta}' z + \sum_{j=1}^{d} \tilde{h}_j(w_j)), \tag{2}$$

where $a(t)$ is the baseline hazard function, $\tilde{\beta} = -\beta$ and $\tilde{h}_j = -h_j$. However, if we change the form of $F(s)$ to $e^s/(1 + e^s)$, the model (1) just becomes the partly linear additive proportional odds model.

Motivated by the close connection with survival models, we focus on the current status data in this paper which arises not only in survival analysis but also in demography, epidemiology, econometrics and bioassay. More specifically, we observe $X = (V, \Delta, Z, W)$, where $V \in \mathbb{R}$ is a random examination time and $\Delta = 1\{U \leq V\}$. We assume that U and V are independent given (Z, W). Under current status data, the model (1) is also related to the semiparametric binary model studied in econometrics. Using the link function $F(\cdot)$, we assume that the probability of $\Delta = 1$, given the covariates (Z, W, V), is of the expression:

$$P(\Delta = 1|Z, W, V) = F\left(\tilde{\beta}' Z + \sum_{j=1}^{d} \tilde{h}_j(W_j) + H(V)\right). \tag{3}$$

Note that Banerjee et al. (2006) and Banerjee et al. (2009) have done a great deal of statistical estimation and hypothesis testing on the model (3) (without $\widetilde{h}_j$ terms) by assuming $F(\cdot)$ to be log-log function and logistic function, respectively. An extensive discussions on the relation between (3) and survival models can be found in Doksum & Gasko (1990). Recently a similar transformation model has been considered by Chen & Tong (2010) but for the *right censored data*. They showed that the monotone transformation function is root-n estimable which will never be achieved in the case of current status data. This is the key theoretical difference between the two types of survival data.

In this paper, we employ the B-spline approach to simultaneously estimate the vector β, monotone H and smooth h_j's. The corresponding estimates are denoted as $\widehat{\beta}$, $\widehat{H}$ and $\widehat{h}_j$. In contrast, Ma & Kosorok (2005a) apply the penalized NPMLE approach to (1) (with $d = 1$) which yields a non-smooth step function $\check{H}$ and the penalized estimate $\check{h}$. Our B-spline framework

has the following theoretical and computational advantages over the existing penalized NPMLE approach:

1. Our B-spline estimate $\widehat{H}$ is smooth and uniformly consistent. However, $\check{H}$ is always discontinues (regardless of the smoothness of its true function H_0) and has a bias which does not vanish asymptotically; see Page 2258 of Ma & Kosorok (2005a). We can also obtain the faster rate of convergence for $\widehat{H}$ than that for $\check{H}$, i.e., $O_P(n^{-1/3})$, by using the B-spline estimation approach. Therefore, we expect more accurate inferences drawn from $\widehat{H}$.

2. We are able to give an explicit B-spline estimate for the asymptotic variance of $\widehat{\beta}$ based on which the asymptotic confidence interval of β can be easily constructed. Under very weak conditions, its consistency is proven. However, the block jackknife approach in Ma & Kosorok (2005a) requires more computation, and is even not theoretically justified.

3. Our spline estimation algorithm requires much less computation than the isotonic type algorithm used in Ma & Kosorok (2005a) since the order of jumps in the step function is supposed to be much larger than the order of knots we choose for estimating H and h_j's.

In contrast with Huang (1999), we deal with the current status data rather than the right censored data, and thus we also need to estimate the monotone transformation function which has been profiled out in their partial likelihood framework. Despite the non-root-n convergence rates of $\widehat{H}$ and $\widehat{h}_j$'s, we are able to show that $\widehat{\beta}$ is root-n consistent, asymptotically normal and semiparametric efficient. We derive the efficient information bound by taking the general two-stage projection approach from Sasieni (1992) which is needed due to the involvement of multiple nonparametric functions in semiparametric models. Interestingly, we observe the convergence rate interfere phenomenon for the B-spline estimators, i.e., the convergence rates of nonparametric estimators are all slowed down to equal the slowest one. Moreover, by approximating $\log \dot{H}$ with the B-spline, we can avoid the monotonicity constraint in the implementation, which is usually required in the literature, e.g., Zhang et al. (2010).

The remainder of the paper is organized as follows. Section 2 describes the B-spline estimation procedure. The asymptotic properties such as consistency and convergence rates of the estimates are obtained in Section 3. The asymptotic distribution of the parametric component is studied in Section 4, and its efficient information and the corresponding explicit B-spline estimate are given in Section 5. Simulation studies are presented in Section 6.1. We close with an appendix containing technical details.

2 Semiparametric B-spline Estimation

2.1 Assumptions

We first define some notations. For any vector v, $v^{\otimes 2} = vv'$. The notations $\gtrsim$ and $\lesssim$ mean greater than, or smaller than, up to a universal constant. We denote $A_n \asymp B_n$ if $A_n \lesssim B_n$ and $A_n \gtrsim B_n$. The notations $\mathbb{P}_n$ and $\mathbb{G}_n$ are used for the empirical distribution and the empirical process of the observations, respectively. Furthermore, we use the operator notation for evaluating expectation. Thus, for every measurable function f and true probability P,

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \quad Pf = \int f dP \quad \text{and} \quad \mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - Pf).$$

We next present some model assumptions.

M1. U and V are independent given (Z, W).

M2. (a) The covariates (Z, W) are assumed to belong to a bounded subset in $\mathbb{R}^{l+d}$, say $[0,1]^l \times [0,1]^d$. The support for V is $[l_v, u_v]$, where $-\infty < l_v < u_v < +\infty$; (b) The joint density for (Z, V, W) w.r.t. Lebesgue measure stays away from zero, and the joint density for (V, W) stays away from infinity.

M3. $E(Z - E(Z|V, W))^{\otimes 2}$ is strictly positive definite.

M4. The residual error distribution $F(\cdot)$ is assumed to be known and has support $\mathbb{R}$. Denote the first, second and third derivative of F as f, $\dot{f}$ and $\ddot{f}$, respectively. We assume that (a) $(f(u) \vee |\dot{f}(u)| \vee |\ddot{f}(u)|) \leq M < \infty$ over the whole $\mathbb{R}$ and $f(u)$ stays away from zero in any compact set of $\mathbb{R}$; (b) $[f^2(v) - \dot{f}(v)F(v)] \wedge [f^2(v) + \dot{f}(v)(1 - F(v))] > 0$, for all $v \in \mathbb{R}$.

Since we employ the smooth B-spline estimation rather than the penalized NPML estimation, our residue error Condition M4 is much less restrictive than that in Ma & Kosorok (2005a). Note that Condition M4(b) ensures the concavity of the function $s \mapsto \delta \log F(s) + (1-\delta)\log(1-F(s))$ for $\delta = 0, 1$.

It is easy to verify that the above Condition M4 is satisfied in the following two general classes of residue error distribution functions after some algebra.

F1. $F(s) = \gamma[2\Gamma(\gamma^{-1})]^{-1}\int_{-\infty}^{s}\exp(-|t|^{\gamma})dt$ for $\gamma > 1$ is a family of distributions, which includes the standard normal distribution after appropriate rescaling ($\gamma = 2$). This corresponds to the probit model Kalbfleisch & Prentice (1980).

F2. $F(s) = 1 - [1 + \gamma e^s]^{-1/\gamma}$ is a Pareto distribution with parameter $\gamma \in (0, \infty)$ and corresponds to the odds-rate transformation family, see Dabrowska & Doksum (1988a,b). It includes the following two well-known special cases:

(a). Given $\gamma \to 0$, it yields the extreme value distribution, i.e. $F(s) = 1 - \exp(-e^s)$, which corresponds to the complementary log-log transformation, see Banerjee et al. (2006);

(b). Given $\gamma = 1$, it gives the logistic distribution, i.e. $F(s) = e^s/(1 + e^s)$, which corresponds to the logit transformation, see Banerjee et al. (2009).

2.2 B-spline Estimation Framework

From now on, we change the signs of β and h_j for simplicity of exposition. In addition, we re-center $H(v)$ to $H(v) - H(l_v)$ so that $H(l_v) = 0$ for the purpose of identifiability. The additional parameter $H(l_v)$ will be absorbed into the vector β, i.e., the first coordinate of z is set as one. Given a single observation at $x = (v, \delta, z, w)$, the log-likelihood of model (1) is written as

$$\begin{aligned} \ell(\beta, h_1, \ldots, h_d, H) &= \delta \log \left\{ F\left[H(v) + \beta' z + \sum_{j=1}^{d} h_j(w_j) \right] \right\} \\ &\quad + (1-\delta) \log \left\{ 1 - F\left[H(v) + \beta' z + \sum_{j=1}^{d} h_j(w_j) \right] \right\}. \end{aligned} \tag{4}$$

We assume that $\beta \in \mathcal{B}$, which is a bounded open subset in $\mathbb{R}^l$, and that its true value β_0 is an interior point of $\mathcal{B}$. Before specifying the parameter spaces for H and h_j's, we first introduce the Hölder ball $\mathbf{H}_c^r(\mathcal{Y})$, which is a class of smooth functions widely used in the nonparametric estimation, e.g., Stone (1982, 1985). For any $f \in \mathbf{H}_c^r(\mathcal{Y})$, it is $J < r$ times continuously differentiable on $\mathcal{Y}$ and its J-th derivative is uniformly Hölder continuous with exponent $\kappa \equiv r - J \in (0, 1]$, i.e.,

$$\sup_{y_1, y_2 \in \mathcal{Y}, y_1 \neq y_2} \frac{|f^{(J)}(y_1) - f^{(J)}(y_2)|}{|y_1 - y_2|^\kappa} \leq c.$$

The functions in the Hölder ball can always be approximated by a basis expansion, i.e.,

$$f(t) \approx \sum_{k=1}^{K} \gamma_k B_k(t) = \gamma' \mathbf{B}(t), \tag{5}$$

where $\gamma = (\gamma_1, \ldots, \gamma_K)'$ and $\mathbf{B}(t) = (B_1(t), \ldots, B_K(t))'$. Actually, if the degree d of the B-spline satisfies $d \geq (r-1)$, we have

$$\|f - \gamma' \mathbf{B}\|_\infty \asymp K^{-r} \quad \text{as } K \to \infty, \tag{6}$$

where $\|\cdot\|_\infty$ denotes the supremum norm..

Assume the following parameter space Condition P1 for the smooth h_j.

P1. For $j = 1, \ldots, d$ and some known c_j, we assume that the parameter space for h_j is $\mathcal{H}_j$, where

$$\mathcal{H}_j = \left\{ h_j : h_j \in \mathbf{H}_{c_j}^{r_j}[0,1] \text{ with } r_j > 1/2 \text{ and } \int_0^1 h_j(w_j)dw_j = 0 \right\},$$

and that the corresponding spline space is

$$\mathcal{H}_{jn} = \left\{ h_j : h_j(w) = \gamma_j' \mathbf{B}_j(w) \text{ with } \|h_j\|_\infty \leq c_j \text{ and } \int_0^1 h_j(w_j)dw_j = 0 \right\},$$

based on a system of basis functions $\mathbf{B}_j = (B_{j1}, \ldots, B_{jK_j})'$ of degree $d_j \geq (r_j - 1)$.

As seen from the previous examples, it is reasonable to assume that $H(\cdot)$ is differentiable and strictly increasing over $[l_v, u_v]$, i.e., $\dot{H}(v) \geq C_0 > 0$. Considering that $H(l_v) = 0$, we can thus write $H(v) = \int_{l_v}^v \exp(g(s))ds$, where $g(v) \equiv \log \dot{H}(v)$ is well defined. Such reparametrization can get around the strict monotonicity and positivity constraints of H, and thus avoids the constrained optimization in the computation. The parameter space Condition P2 for g is specified below.

P2. For some known c_0, we assume that the parameter space for g is $\mathcal{G}$, where

$$\mathcal{G} = \left\{ g : g \in \mathbf{H}_{c_0}^{r_0}[l_v, u_v] \text{ with } r_0 > 1/2 \right\},$$

and that the corresponding spline space is

$$\mathcal{G}_n = \{ g : g(v) = \gamma_0' \mathbf{B}_0(v) \text{ and } \|g\|_\infty \leq c_0 \}$$

based on a system of basis functions $\mathbf{B}_0 = (B_{01}, \ldots, B_{0K_0})$ of degree $d_0 \geq (r_0 - 1)$.

Similarly, we define $\mathcal{G}_n' = \{ H(v) = \int_{l_v}^v \exp(g(s))ds : g \in \mathcal{G}_n \}$. By some algebra, we can show that $H \in \mathbf{H}_{c_0'}^{r_0+1}[l_v, u_v]$ for some $c_0' < \infty$.

REMARK 1. *Note that in the theoretical proofs and numerical calculations the exact values of c_j are not necessary. Instead, only the boundedness condition, equivalently the compactness of parameter spaces and spline spaces, is needed. Here we assume this boundedness condition, which can be relaxed by invoking the chaining arguments, only for simplifying our theoretical derivations.*

In this paper, we propose the B-spline approach to estimate H and h_j's as follows. Let $\mathcal{A} = \mathcal{B} \times \mathcal{G} \times \Pi_{j=1}^d \mathcal{H}_j$ and $\mathcal{A}_n = \mathcal{B} \times \mathcal{G}_n \times \Pi_{j=1}^d \mathcal{H}_{jn}$. Denote α as $(\beta', g, h_1, \ldots, h_d)'$ and its true value α_0 as $(\beta_0', g_0, h_{10}, \ldots, h_{d0})'$, where $g_0(\cdot) = \log \dot{H}_0(\cdot)$. The log-likelihood (4) for the observation i can thus be reparametrized as

$$\begin{aligned} \ell_i(\alpha) &= \delta_i \log \left\{ F\left[\beta' z_i + \int_{l_v}^{v_i} \exp(g(s))ds + \sum_{j=1}^d h_j(w_{ij}) \right] \right\} \\ &\quad + (1-\delta_i) \log \left\{ 1 - F\left[\beta' z_i + \int_{l_v}^{v_i} \exp(g(s))ds + \sum_{j=1}^d h_j(w_{ij}) \right] \right\}. \end{aligned} \tag{7}$$

The corresponding B-spline estimate $\widehat{\alpha}$ is defined as

$$\widehat{\alpha} = \arg\max_{\alpha \in \mathcal{A}_n} \sum_{i=1}^n \ell_i(\alpha). \tag{8}$$

We can also write $\widehat{\alpha} = (\widehat{\beta}', \widehat{g}, \widehat{h}_1, \ldots, \widehat{h}_d)' = (\widehat{\beta}', \widehat{\gamma}_0' \mathbf{B}_0, \widehat{\gamma}_1' \mathbf{B}_1, \ldots, \widehat{\gamma}_d' \mathbf{B}_d)'$. Then, the estimate $\widehat{H}(v) = \int_{l_v}^v \exp(\widehat{\gamma}_0' \mathbf{B}_0(s))ds$. Some tedious algebra reveals that the Hessian matrix of $\ell_i(\alpha)$ w.r.t. $(\beta', \gamma_0', \gamma_1', \ldots, \gamma_d')'$ is indeed negative semidefinite under Condition M4(b) which guarantees the existence of $\widehat{\alpha}$. See more discussions on the computation feasibility in the simulation section. The above estimation procedure also applies to other linear sieves approximating the Hölder ball (or more generally Hölder space), e.g., wavelets.

3 Consistency and Rates of Convergence

In this section, we show that our B-spline estimate is consistent and the convergence rate of each nonparametric estimate appears to interfere with each other. Define

$$d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_2 + \sum_{j=1}^d \|h_j - h_{j0}\|_2,$$

where $\|\cdot\|_2$ is the L_2 norm. Now we give the main Theorem of this section.

THEOREM 1. *Suppose that Conditions M1-M4 and P1-P2 hold. If $K_j/n \to 0$ for $j = 0, 1, \ldots, d$, then we have*

$$d(\widehat{\alpha}, \alpha_0) = o_P(1). \tag{9}$$

More specifically, we further prove that

$$d(\widehat{\alpha}, \alpha_0) = O_P\left(\max_{0 \le j \le d} \left\{ K_j^{-r_j} \vee \sqrt{K_j/n} \right\} \right). \tag{10}$$

If we further require that $K_j \asymp n^{1/(2r_j+1)}$ *for* $j = 0, \ldots, d$*, then we have*

$$d(\widehat{\alpha}, \alpha_0) = O_P(n^{-r/(2r+1)}), \tag{11}$$

where $r = \min_{0 \le j \le d}\{r_j\}$.

Under the right censored data, Huang (1999) derived similar convergence rate result (10) in the partly linear additive Cox model by assuming equal r_j's. According to Theorem 1, the smooth $\widehat{H}$ can achieve the rate of convergence, i.e., $O_P(n^{-r/(2r+1)})$, no slower than $n^{1/3}$-rate derived in the penalized estimation context, see Ma & Kosorok (2005a), when we assume that g_0 and h_{j0}'s are all at least continuously differentiable, i.e., $r \ge 1$. More importantly, we can further show that $\widehat{H}$ is uniformly consistent, i.e., $\|\widehat{H} - H_0\|_\infty = o_P(1)$, by applying Lemma 2 in Chen & Shen (1998) that $\|f\|_\infty \lesssim \|f\|_{L_2(Leb)}^{2r/(2r+d)}$ for any $f \in \mathbf{H}_c^r[a,b]^d$ and noting that $\widehat{H}, H_0 \in \mathbf{H}_{c_0'}^{r_0+1}[l_v, u_v]$ for some $c_0' > 0$.

The above theorem also holds when we employ the constrained monotone B-spline to approximate H_0, i.e., $\gamma_0' \mathbf{B}_0(v) \approx \log H(v)$ with $\gamma_{01} \le \gamma_{02} \le \ldots \le \gamma_{0K_0}$. However, such constrained optimization usually requires additional computational effort, see Zhang et al. (2010).

REMARK 2. *From the above Theorem 1, we observe the interesting convergence rate interfere phenomenon, i.e., the convergence rate for each B-spline estimate is forced to equal the slowest one. In Ma & Kosorok (2005a), they also show that the convergence rate of the penalized estimate* $\widetilde{h}$ *is unfortunately slowed down to* $O_P(n^{-1/3})$ *by the NPMLE* $\widetilde{H}$ *regardless of the smoothness degree of* h_0*. One possible solution in achieving the optimal rate for each nonparametric estimate is to extend the most recent mixed rate asymptotic results Radchenko (2008) to the semiparametric setup.*

Since we assume that $r > 1/2$, the convergence rate given in (11) is always $o_P(n^{-1/4})$. Such a rate is usually fast enough to guarantee the regular asymptotic behavior of $\widehat{\beta}$, i.e., $\sqrt{n}$-consistency and asymptotic normality. Indeed, we will improve the current suboptimal rate of $\widehat{\beta}$ in (11) to the optimal $\sqrt{n}$ rate, and further show that $\widehat{\beta}$ is semiparametric efficient in next section.

4 Weak Convergence of the Parametric Estimate

In this section, we study the weak convergence of the spline estimate $\widehat{\beta}$ in the presence of multiple nonparametric nuisance functions. We first calculate the semiparametric efficient information based on the projection onto the nonorthogonal sumspace.

Let

$$Q_\theta(x) = f(\theta)\left(\frac{\delta}{F(\theta)} - \frac{1-\delta}{1-F(\theta)}\right),$$

where $\theta(z,v,w) = \beta' z + H(v) + \sum_{j=1}^{d} h_j(w_j)$. Denote θ_0 as the true value of θ. The score functions (operators) for β, g and h_j are separately calculated as

$$\dot{\ell}_\beta(X;\alpha) = ZQ_\theta(X), \tag{12}$$

$$\dot{\ell}_g[a](X;\alpha) = \left[\int_{l_v}^{V} \exp(g(s))a(s)ds\right] Q_\theta(X), \tag{13}$$

$$\dot{\ell}_{h_j}[b_j](X;\alpha) = b_j(W_j)Q_\theta(X). \tag{14}$$

We assume that $a \in L_2(H) \equiv \{a : \int_{l_v}^{u_v} a^2(s)dH(s) < \infty\}$ and $b_j \in L_2^0(w_j) \equiv \{b_j : \int_0^1 b_j(w_j)dw_j = 0$ and $\int_0^1 b_j^2(w_j)dw_j < \infty\}$ so that all the score functions defined above are square integrable.

To calculate the efficient score function $\widetilde{\ell}_\beta$, we need to find the projection of $\dot{\ell}_\beta$ onto the sumspace $\mathbf{A} = A_g + A_{h_1} + \cdots + A_{h_d}$, where $A_g = \{\dot{\ell}_g[a] : a \in L_2(H)\}$ and $A_{h_j} = \{\dot{\ell}_{h_j}[b_j] : b_j \in L_2^0(w_j)\}$. For simplicity, we define $\dot{\ell}_\beta(X;\alpha_0)$ and $\dot{\ell}_\beta(X;\widehat{\alpha})$ as $\dot{\ell}_{\beta_0}$ and $\dot{\ell}_{\widehat{\beta}}$, respectively. The same notation rule applies to $\dot{\ell}_g[a](X;\alpha)$ and $\dot{\ell}_{h_j}[b_j](X;\alpha)$. We define

$$\widetilde{\ell}_\beta(X;\alpha) = \dot{\ell}_\beta(X;\alpha) - \dot{\ell}_g[\bar{a}^\dagger](X;\alpha) - \sum_{j=1}^{d} \dot{\ell}_{h_j}[\bar{b}_j^\dagger](X;\alpha),$$

where $\bar{a}^\dagger = (a_1^\dagger, \ldots, a_l^\dagger)'$ and $\bar{b}_j^\dagger = (b_{j1}^\dagger, \ldots, b_{jl}^\dagger)'$. And $(a_k^\dagger, b_{1k}^\dagger, \ldots, b_{dk}^\dagger)$ is the minimizer of

$$(a_k, b_{1k}, \ldots, b_{dk}) \mapsto E\left\{[\dot{\ell}_{\beta_0}]_k - \dot{\ell}_{g_0}[a_k] - \sum_{j=1}^{d} \dot{\ell}_{h_{j0}}[b_{jk}]\right\}^2$$

for $k = 1, \ldots, l$. Similarly, denote $\widetilde{\ell}_\beta(X;\alpha_0)$ and $\widetilde{\ell}_\beta(X;\widehat{\alpha})$ as $\widetilde{\ell}_{\beta_0}$ and $\widetilde{\ell}_{\widehat{\beta}}$, respectively. By taking the two-stage projection approach from Sasieni (1992), we have

$$\tilde{\ell}_{\beta_0}(X) = \left(Z - \bar{b}^\dagger(W) - \frac{E((Z - \bar{b}^\dagger(W))Q_{\theta_0}^2(X)|V)}{E(Q_{\theta_0}^2(X)|V)}\right) Q_{\theta_0}(X) \tag{15}$$

where $\bar{b}^\dagger(W) = \sum_{j=1}^{d} \bar{b}_j^\dagger(W_j)$ satisfies

$$E\left\{\left[Z - \bar{b}^\dagger(W) - \frac{E((Z - \bar{b}^\dagger(W))Q_{\theta_0}^2|V)}{E(Q_{\theta_0}^2|V)}\right]_k Q_{\theta_0}^2 b_{jk}(W_j)\right\} = 0 \tag{16}$$

for every $b_{jk} \in L_2^0(w_j)$, $j = 1, \ldots, d$ and $k = 1, \ldots, l$. By slightly modifying the proof of Lemma 4 in Ma & Kosorok (2005a), we can show that the above nonorthogonal projection is well defined and $\bar{b}^\dagger(\cdot)$ exists by the alternating projection Theorem A.4.2 in Bickel et al. (1993).

Define Π_j and Π_a as the projection operators

$$\Pi_j g \mapsto \frac{E[g(V,W)Q_{\theta_0}^2|W_j = w_j]}{E[Q_{\theta_0}^2|W_j = w_j]}, \quad \Pi_a g \mapsto \frac{E[g(V,W)Q_{\theta_0}^2|V = v]}{E[Q_{\theta_0}^2|V = v]},$$

respectively. Define

$$\begin{aligned} D(v,w) &= \frac{E[ZQ_{\theta_0}^2|V=v,W=w]}{E[Q_{\theta_0}^2|V=v,W=w]}, \quad S(v,w_j) = \frac{E[Q_{\theta_0}^2|V=v,W_j=w_j]}{E[Q_{\theta_0}^2|W_j=w_j]}, \\ T(w_i,w_j) &= \frac{E[Q_{\theta_0}^2|W_i=w_i,W_j=w_j]}{E[Q_{\theta_0}^2|W_j=w_j]}, \quad U(w_j,v) = \frac{E[Q_{\theta_0}^2|W_j=w_j,V=v]}{E[Q_{\theta_0}^2|V=v]}. \end{aligned}$$

We say a function $f(s,t)$ belongs to a uniform Hölder ball $\mathbf{H}_c^r(\mathcal{S}\times\mathcal{T})$ in t relative to s if it is $J<r$ continuously differentiable w.r.t. t and its J-th partial derivative satisfies, with $\kappa \equiv r-J$,

$$\sup_{s\in\mathcal{S}}\sup_{t_1\neq t_2}\frac{|f_t^{(J)}(s,t_1)-f_t^{(J)}(s,t_2)|}{|t_1-t_2|^\kappa} \leq c.$$

Define $Sf(v,w_j) = S(v,w_j)f_{V|W_j}(v,w_j)$, $Tf(w_i,w_j) = T(w_i,w_j)f_{W_i|W_j}(w_i,w_j)$ and $Uf(w_j,v) = U(w_j,v)f_{W_j|V}(w_j,v)$, where $f_{V|W_j}$, $f_{W_i|W_j}$ and $f_{W_j|V}$ are the conditional densities of V given W_j, W_i given W_j and W_j given V w.r.t. Lebesgue measure, respectively.

Here, we assume some model assumptions implying that both $b_{jk}^\dagger$ and $a_k^\dagger$ belong to some Hölder balls for any $j=1,\ldots,d$ and $k=1,\ldots,l$.

M5. We assume that $[\Pi_j D(v,w)]_k \in \mathbf{H}_{\bar{c}_j}^{r_j}[0,1]$, $Sf(v,w_j) \in \mathbf{H}_{\bar{c}_j}^{r_j}([l_v,u_v]\times[0,1])$ in w_j relative to v and $Tf(w_i,w_j) \in \mathbf{H}_{\bar{c}_j}^{r_j}[0,1]^2$ in w_j relative to w_i for some $0<\bar{c}_j<\infty$ and $j=1,\ldots,d$.

M6. We assume that $[\Pi_a D(v,w)]_k \in \mathbf{H}_{\bar{c}_0}^{r_0+1}[l_v,u_v]$ and $Uf(w_j,v) \in \mathbf{H}_{\bar{c}_0}^{r_0+1}([0,1]\times[l_v,u_v])$ in v relative to w_j for some $0<\bar{c}_0<\infty$.

Note that we can simplify $Sf(v,w_j)$ $(Tf(w_i,w_j))$ to $S(v,w_j)$ $(T(w_i,w_j))$ in Condition M5 and simplify $Uf(w_j,v)$ to $U(w_j,v)$ in Condition M6 when we assume that V and W are independent and that W is pairwise independent.

THEOREM 2. *Suppose that Conditions M1-M6 and P1-P2 hold. If $K_j \asymp n^{1/(2r_j+1)}$ and $\widetilde{I}_0$ is invertible, then we have*

$$\sqrt{n}(\widehat{\beta}-\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \widetilde{I}_0^{-1}\widetilde{\ell}_{\beta_0}(X_i) + o_P(1) \xrightarrow{d} N(0,\widetilde{I}_0^{-1}), \tag{17}$$

where $\widetilde{I}_0$ is the efficient information matrix defined as $E\widetilde{\ell}_{\beta_0}\widetilde{\ell}_{\beta_0}'$.

5 B-spline Estimate of the Efficient Information

In this section, we give an explicit B-spline estimate for the efficient information as a by-product of the establishment of asymptotic normality of $\widehat{\beta}$. Indeed, it is simply the observed information matrix if we treat the semiparametric model as a parametric one after the B-spline approximation, i.e., $\mathcal{H}_j = \mathcal{H}_{jn}$ and $\mathcal{G} = \mathcal{G}_n$. Specifically, we treat $\ell_i(\alpha)$ defined in (7) as if it were a parametric likelihood $\ell_i(\beta, \gamma_0, \gamma_1, \ldots, \gamma_d)$.

We construct the corresponding information estimator for $(\beta', \gamma_0, \gamma_1, \ldots, \gamma_2)'$:

$$\widehat{J} = \begin{pmatrix} \widehat{I}_{11} & \widehat{I}_{12} \\ \widehat{I}_{21} & \widehat{I}_{22} \end{pmatrix}_{(l+\sum_{j=0}^{d} K_j)\times(l+\sum_{j=0}^{d} K_j)},$$

where $\widehat{I}_{j,k} = \sum_{i=1}^{n} A_j(X_i;\widehat{\alpha})A_k'(X_i;\widehat{\alpha})/n$, for $j,k = 1,2$, and

$$\begin{aligned} A_1(X;\alpha) &= \dot{\ell}_\beta(X;\alpha), \\ A_2(X;\alpha) &= \left(\dot{\ell}_g[B_{01}], \ldots, \dot{\ell}_g[B_{0K_0}], \dot{\ell}_{h_1}[B_{11}], \ldots, \dot{\ell}_{h_d}[B_{dK_d}]\right)'. \end{aligned}$$

The parametric inferences imply that the information estimator for β is of the form

$$\widehat{I} = \widehat{I}_{11} - \widehat{I}_{12}\widehat{I}_{22}^{-1}\widehat{I}_{21}. \tag{18}$$

Some calculations further reveal that

$$\widehat{I} = \mathbb{P}_n \left[\dot{\ell}_{\widehat{\beta}} - \dot{\ell}_{\widehat{g}}[(\bar{\gamma}_0^\dagger)'\mathbf{B}_0] - \sum_{j=1}^{d} \dot{\ell}_{\widehat{h}_j}[(\bar{\gamma}_j^\dagger)'\mathbf{B}_j]\right]^{\otimes 2}, \tag{19}$$

where $[\bar{\gamma}_j^\dagger]_{K_j\times l} = (\gamma_{j1}^\dagger, \ldots, \gamma_{jl}^\dagger)$ for $j = 0, 1, \ldots, d$ and $(\gamma_{0k}^\dagger, \ldots, \gamma_{dk}^\dagger)^T = \widehat{I}_{22}^{-1}\widehat{I}_{21}1_k$ where 1_k represents the l-vector with its k-th element as one and others as zeros. We will use (18) as our estimator for $\widetilde{I}_0$.

We need the following additional assumption for Theorem 3.

M7. We assume that

$$E \sup_{a_k\in\mathcal{G}_n} \left[\int_{l_V}^{V} [\exp(g(s)) - \exp(g_0(s))]a_k(s)ds\right]^2 \lesssim \|H - H_0\|_2^2.$$

THEOREM 3. *Under Conditions M1-M7 and P1-P2, we have* $\widehat{I} \xrightarrow{P} \widetilde{I}_0$.

Note that the consistency of the similar random-sieve efficient information estimate was also proven in the linear regression models with current data; see Theorem 3 of Shen (2000).

6 Numerical Results

6.1 Simulations

We perform a Monte-Carlo study to assess the finite-sample performance of our proposed method. To compare with the penalized NPMLE in Ma & Kosorok (2005a), we adopt the same setting used in their paper. We simulate the current status data from the partly linear additive Cox model which is a special case of general transformation model. We choose $H(u) = \log A(e^u)$ where $A(u) = e^{k_0}(\exp(u/3) - 1)$ with $k_0 = 0.06516$. The errors ϵ follow an extreme value distribution with $F(s) = 1 - \exp(-e^s)$. The regression coefficients $\beta_1 = 0.3$ and $\beta_2 = 0.25$. The covariate Z_1 is Uniform$[0.5, 1.5]$ and Z_2 is Bernoulli with success probability 0.5. We choose W as Uniform$[1, 10]$ and $h(w) = \sin(w/1.2 - 1) - k_0$. Censoring times are standard exponential distribution conditional on being in the interval $[0.2, 1.8]$. The sample sizes are $n = 400$ and $n = 1600$. We simulate 400 realizations for both sample sizes.

In practice, the location and the numbers of knots for H and h_j need to be determined. For simplicity, we will use the equal-spaced knots for all functions. Common model selection methods such as the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) can be employed for selecting the number of knots. In this paper, we determine $K_0, K_1, \ldots, K_d$ by the AIC, given by

$$\text{AIC} = -2\sum_{i=1}^{n} \ell_i(\hat{\alpha}) + 2(\ell + \sum_{j=0}^{d} K_j)$$

In our simulation, we use a quadratic spline to approximate both function h and function g in H. Then, $\text{AIC} = -2\sum_{i=1}^{n} \ell_i(\hat{\alpha}) + 2(K_0 + K_1 + 2)$. Based on our experiences, it is generally adequate to choose less than ten knots to achieve reasonable approximation, provided that h and H are not overly erratic. Figure 1 shows the AIC scores under different combinations of K_0 and K_1 for one realization of the simulation with the sample size $n = 1600$. It shows that the optimal choices for K_0 and K_1 are 5 and 5, respectively. The estimated h and H with various values of K_0 and K_1 are plotted in Figure 2. In the left panel of Figure 2, we fix $K_0 = 5$ and plot the estimated h with $K_1 = 3, 5, 10$. When K_1 is small (e.g., $K_1 = 3$), there seems be to a big bias in our estimator. On the other hand, when K_1 is large (e.g., $K_1 = 10$), the estimator displays a wiggly behavior. In the right panel of Figure 2, we fix $K_1 = 5$ and plot the estimated H with $K_0 = 5, 7, 10$. As the number of knots is increasing, the estimated H shows a similar wiggly shape. Hence, the numbers of knots should be chosen with caution. It is worth noting that the selected values $K_0, K_1, \ldots, K_d$ based the AIC criterion can be regarded only as the minimum numbers of knots required. They may not be the optimal choices since the concept of optimality is not well defined here. See Xue et al. (2004) for similar discussions.

Simulation results show that our B-spline estimation procedure performs quite well in the

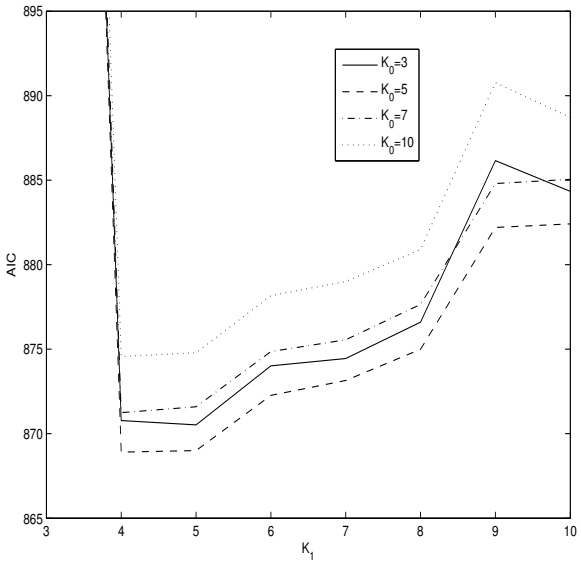

Figure 1: AIC scores under different combinations of K_0 and K_1

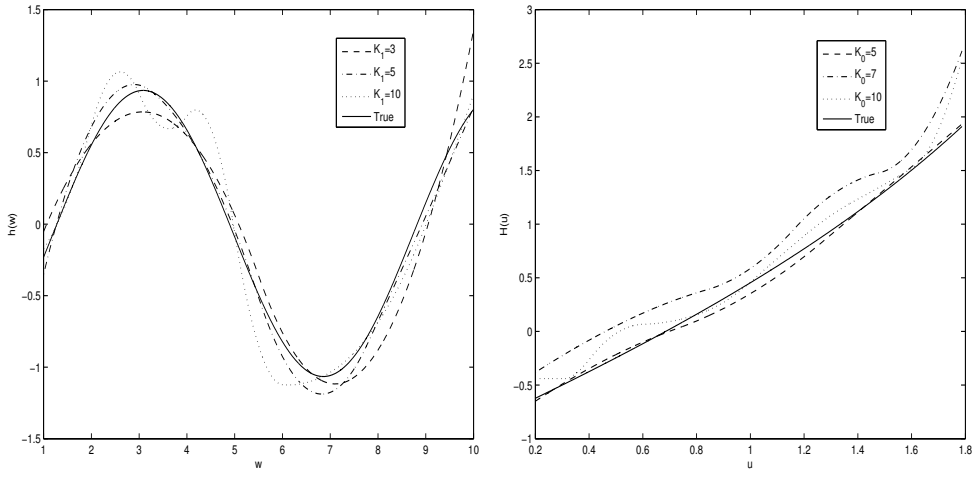

Figure 2: Plot of the estimated h and H with various values of K_0 and K_1.

Table 1: Monte Carlo results for the partly linear Cox model with current status data based on 400 replicates

		Sample size 400	Sample size 1600
$\widehat{\beta}_1$	Bias	0.0318	0.0100
	SD	0.2919	0.1246
	ESD	0.3102	0.1325
	Coverage	0.9620	0.9690
	ESD-WB	0.3547	0.1575
$\widehat{\beta}_2$	Bias	0.0168	0.0074
	SD	0.1533	0.0797
	ESD	0.1612	0.0803
	Coverage	0.9710	0.9680
	ESD-WB	0.1836	0.0936
Joint	Coverage	0.9620	0.9550

SD: Standard error; ESD: Estimated standard error; ESD-WB: Estimated standard error from the weighted bootstrap method

semiparametric transformation model. The bias and standard errors of the spline estimates of β_1 and β_2 are given in Table 1. The table shows that the sample biases of both $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are small. The ratio of the standard errors for the two sample sizes is close to 2, a result consistent with a $\sqrt{n}$-convergence rate for $\widehat{\beta}_1$ and $\widehat{\beta}_2$. The estimated standard errors from (18) (denoted as ESD) are also displayed in Table 1, which are very close to the simulation results. Although our proposed method tends to overestimate the standard error slightly but the overestimation lessens as sample size increases. We also compare our results with the weighted bootstrap method in Ma & Kosorok (2005b). The weights are from the exponential distribution with mean one. The estimated standard errors are also similar to the results obtained using our explicit B-spline estimate. The 95% confidence interval constructed from (18) generally have coverage close to the nominal value. Histograms of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are shown in Figure 3. It is clear that the marginal distributions of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are Gaussian. The left panel of Figure 4 displays the spline estimate of $h(w)$ and the monotone estimate $\widehat{H}$ is given in the right panel of Figure 4. The dashed line is the true function, the solid line is the average estimate over 400 realizations, and the dash-dotted line is the 95% pointwise confidence band for $h(w)$ or $H(v)$ when we know the true model, which is obtained by taking 2.5 percentile and 97.5 percentile of these 400 estimates at each w or v.

As suggested by one of the referees, we also perform a Monte-Carlo study by including two nonparametric functions in the model. Under the same setting as in the last study, the

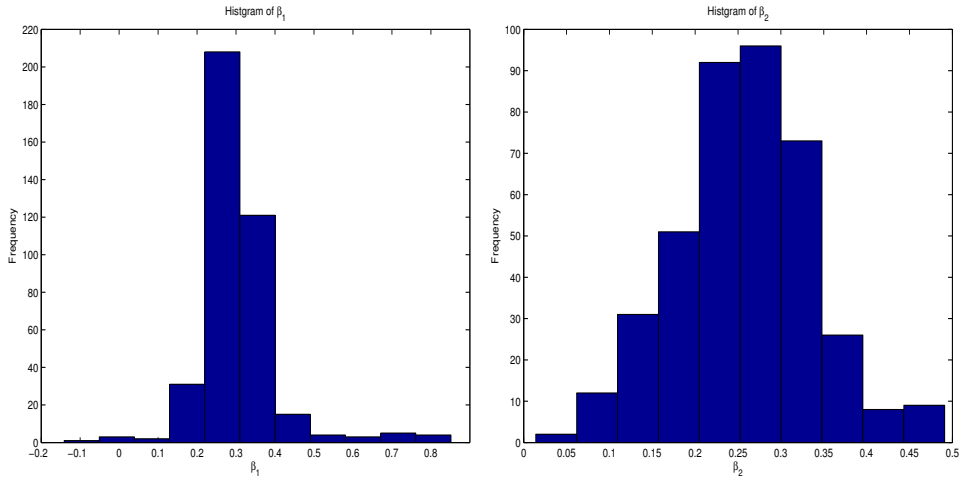

Figure 3: Histogram of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ based on 1600 samples and 400 replicates.

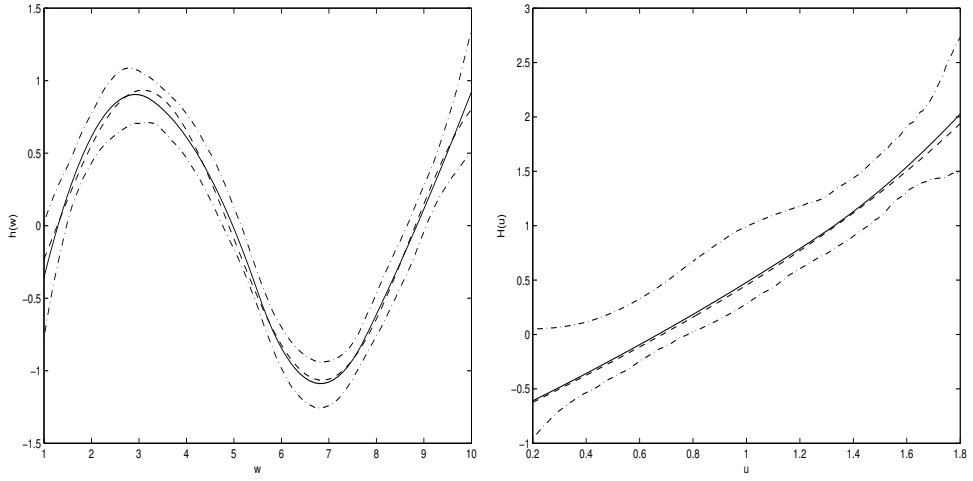

Figure 4: Left: Estimate and pointwise confidence interval for h. Right: Estimate and pointwise confidence interval for H. The solid line is the average estimate over 400 realizations from sample size $n = 1600$, and the dashed line is the true function. The dash-dotted lines are the 95% pointwise confidence interval.

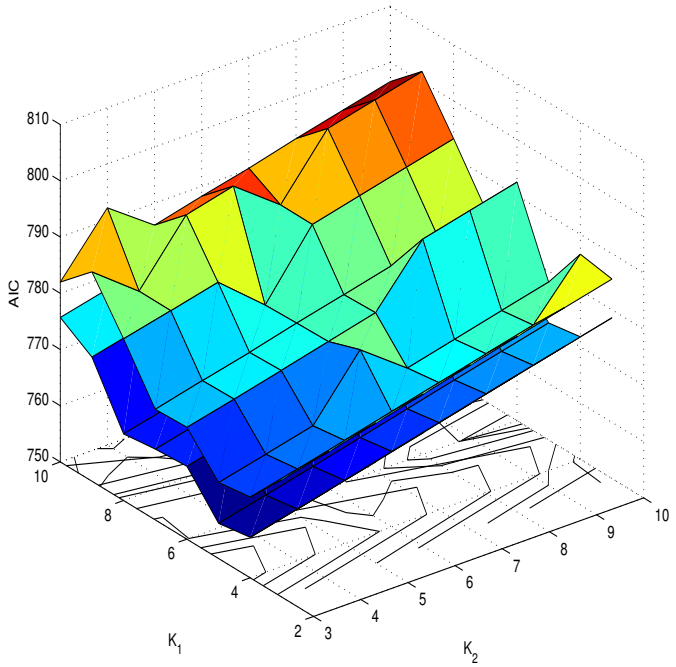

Figure 5: AIC scores under different combinations of K_0, K_1, and K_2

two nonparametric functions are $h_1(w_1) = \sin(w_1/1.2 - 1) - k_0$ with w_1 following a uniform distribution on $[1, 10]$ and $h_2(w_2) = 3w_2^2 - 1$ with w_2 following a uniform distribution on $[-1, 1]$. Figure 6 shows the AIC scores under different combinations of K_0, K_1 and K_2 for one realization of the simulation with the sample size $n = 1600$. For illustration, we only plot two choices of K_0 where the top surface is for $K_0 = 10$ and the bottom surface is for $K_0 = 4$. The optimal choice by the AIC criterion is $(K_0, K_1, K_2) = (4, 5, 3)$. The spline estimates of h_1, h_2 and H under the optimal number of knots are displayed in Figure 6, and the dotted lines are the true functions.

To compare our spline based method with the penalized method in Ma & Kosorok (2005a), there are four obvious advantages of our method. First, the computational cost of our spline estimate $\widehat{H}$ is much less expensive than that used in Ma & Kosorok (2005a), i.e. the cumulative sum diagram approach. This is because the number of basis B-splines, i.e., K_0, is often taken much smaller than the sample size n, thus the dimension of the estimation problem is greatly reduced. Secondly, our estimate of the transformation function H is smooth with a higher convergence rate. We obtain a narrower confidence interval for H shown in the right panel of Figure 4. Thirdly, we can obtain an explicit consistent estimate $\widehat{I}$. However, the block jackknife approach proposed in Ma & Kosorok (2005a) is not theoretically justified. At last, we do not require the constrained optimization in our implementations.

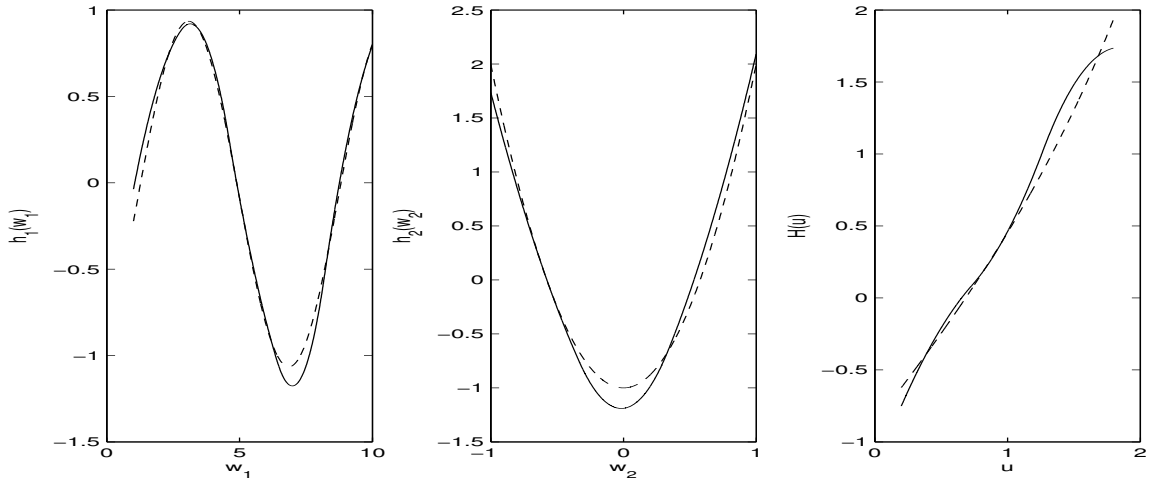

Figure 6: Estimates of h_1, h_2 and H. The dotted lines are true functions.

Table 2: The estimates and their corresponding estimated standard errors for the parametric part for the calcification data

	extreme value distribution	logistic distribution
$\hat{\beta}_1$	-0.1870	-0.2562
ESD($\hat{\beta}_1$)	0.2322	0.2119
$\hat{\beta}_2$	0.3502	0.3573
ESD($\hat{\beta}_2$)	0.3481	0.3280

ESD: Estimated standard error

6.2 Application: Calcification data

We illustrate the proposed method in a dataset from the calcification study. Yu et al. (2001) investigated the calcification of intraocular lenses, which is an infrequently reported complication of cataract treatment. Understanding the effect of some clinical variables on the time to calcification of the lenses after implantation is the objective of the study. The patients were examined by an ophthalmologist to determine the status of calcification at a random time ranging from zero to thirty six months after implantation of the intraocular lenses. The severity of calcification was graded into five categories ranging from zero to four. In our analysis, we simply treat those with severity > 1 as calcified and those with severity ≤ 1 as not calcified. This dataset can be treated as the current status dataset because only the examination time and the calcification status at examination are available. The covariates of interest include Z_1 incision length, Z_2 gender (0

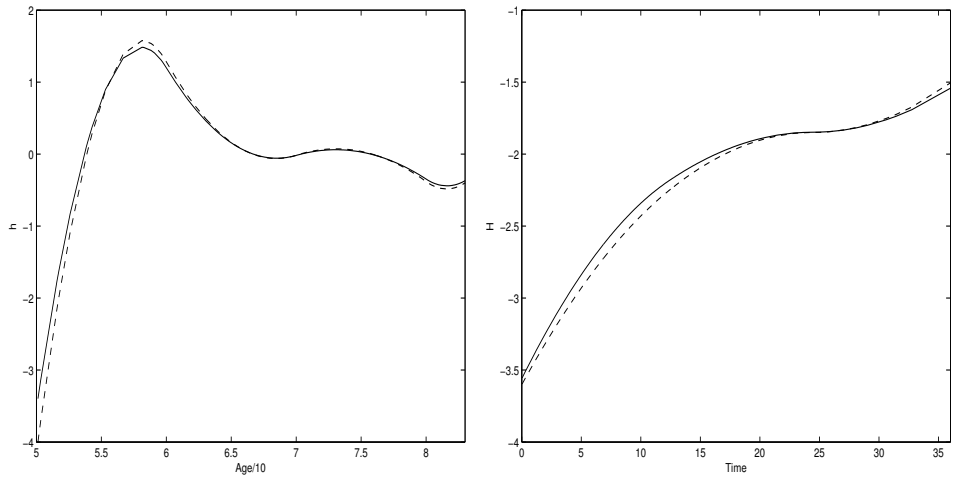

Figure 7: The spline estimates of $h(w)$ and $H(v)$ under two different assumptions of the error distribution: extreme value distribution (solid) and logistic distribution (small dashes).

for female and 1 for male), and W age at implantation/10. The original dataset has 379 records. We remove the one record with missing measurement, resulting the sample size $n = 378$. This dataset has been studied by Xue et al. (2004), Lam & Xue (2005), and Ma (2009). Xue et al. (2004) and Lam & Xue (2005) modeled the event time by the log-transformation. A straightforward estimation of the hazard function is not available. Ma (2009) used the cure model to fit the data, and assumed a generalized linear model for the cure probability. For subjects not cured, the linear and partly linear Cox proportional hazards models are used to model the survival risk.

We fit this dateset using the semiparametric additive transformation model. We assume the error distribution F to be one of the two distributions: extreme value distribution and logistic distribution. We approximate h and $\log \dot{H}$ by quadratic splines. The optimal choices of knots for h and $\log \dot{H}$ are 6 and 5, respectively. The estimates and their corresponding estimated standard errors for the parametric part are summarized in Table 2. The estimates for $h(w)$ based on different error distributions are displayed in the left panel of Figure 7, and the estimates of $H(v)$ are plotted in the right panel of Figure 7. The analysis shows very similar results for these two error distributions. From Table 2, both incision length and gender are insignificant at the 5% level of significance. From the left panel of Figure 7, $h(w)$ increases steadily from age 50, achieving a peak at age 60, decreasing gradually thereafter, which means that patients ages around 60 tend to enjoy a longer time to calcification. The estimated transformation function $\widehat{H}$ in the right panel of Figure 7 displays a nonlinear behavior and it shows that the transformation is necessary.

We can incorporate an unknown scale parameter into to the residual error distribution $F(\cdot)$ to further improve the above analysis. Our general B-spline estimation framework can also handle

this type of transformation models easily.

Acknowledgement

The first author's research is supported by the National Science Foundation under grant DMS-0906497. The second author's research is supported by the National Science Foundation under grant CMMI-1030246 and DMS-1042967. The authors would like to thank Professor Alexis K. F. Yu for providing the Calcification data and Professor Donglin Zeng for helpful discussions. The authors also would like to thank the editor, the associate editor, and two reviewers for their helpful comments and suggestions which led to a much improved presentation.

Appendix

Some useful Lemmas

We define ϵ-covering number (ϵ-bracketing number) as $N(\epsilon, \mathcal{A}, d)$ ($N_B(\epsilon, \mathcal{A}, d)$). The corresponding ϵ-entropy (ϵ-bracketing entropy) is defined as $H(\epsilon, \mathcal{A}, d) = \log N(\epsilon, \mathcal{A}, d)$ ($H_B(\epsilon, \mathcal{A}, d) = \log N_B(\epsilon, \mathcal{A}, d)$). Define $\mathcal{G}_n(\delta_0; \|\cdot\|) = \{g : g(v) = \gamma_0' \mathbf{B}_0(v)$ satisfying $\|g\| \leq \delta_0\}$ and $\mathcal{H}_{jn}(\delta_j; \|\cdot\|) = \{h_j : h_j(w_j) = \gamma_j \mathbf{B}_j(w_j)$ satisfying $\|h_j\| \leq \delta_j$ and $\int_0^1 h_j(w_j)dw_j = 0\}$. Obviously, $\mathcal{G}_n(c_0; \|\cdot\|_\infty) = \mathcal{G}_n$ and $\mathcal{H}_{jn}(c_j; \|\cdot\|_\infty) = \mathcal{H}_{jn}$. Lemma 1 follows from the B-spline approximation property (6). Lemma 2 is directly implied by Lemma 2.5 in (Van de Geer, 2000). Lemma 4 is adapted from Proposition 1 in (Cheng & Huang, 2010).

LEMMA 1. *There exist $g_n \in \mathcal{G}_n$ and $h_{jn} \in \mathcal{H}_{jn}$ such that*

$$\begin{aligned} \|g_n - g_0\|_\infty &\asymp K_0^{-r_0}, && \text{(A.1)} \\ \|H_n - H_0\|_\infty &= O(K_0^{-r_0}), && \text{(A.2)} \\ \|h_{jn} - h_{j0}\|_\infty &\asymp K_j^{-r_j}, && \text{(A.3)} \\ \left\| \sum_{j=1}^d h_{jn} - \sum_{j=1}^d h_{j0} \right\|_\infty &= O\left(\max_{j=1,\ldots,d} \{K_j^{-r_j}\} \right), && \text{(A.4)} \end{aligned}$$

where $H_n(v) = \int_{l_v}^v \exp(g_n(s))ds$.

LEMMA 2.

$$\begin{aligned} H(\epsilon, \mathcal{G}_n(\delta_0; \|\cdot\|), \|\cdot\|) &\lesssim K_0 \log(1 + 4\delta_0/\epsilon), && \text{(A.5)} \\ H(\epsilon, \mathcal{H}_{jn}(\delta_j; \|\cdot\|), \|\cdot\|) &\lesssim K_j \log(1 + 4\delta_j/\epsilon) && \text{(A.6)} \end{aligned}$$

for $1 \leq j \leq d$.

LEMMA 3. *Let* $\mathbf{h} = (h_1, \ldots, h_d)$. *Define* $\mathcal{K} = \{\zeta(\beta, \mathbf{h}, H) : \beta \in \mathcal{B}, \mathbf{h} \in \prod_{j=1}^d \mathcal{H}_{jn}, g \in \mathcal{G}_n\}$, *where the form of* ζ *is defined in (A.12). We have*

$$\sup_{\zeta \in \mathcal{K}} |\mathbb{G}_n \zeta| = O_P(\max_{j=0,1,\ldots,d} \{K_j^{1/2}\}). \tag{A.7}$$

Proof: Define $l^*(\beta, \mathbf{h}, H) = \delta F(\beta' z + \sum_{j=1}^d h_j(w_j) + H(v)) + (1-\delta)[1 - F(\beta' z + \sum_{j=1}^d h_j(w_j) + H(v))]$. The construction of $l^*(\cdot)$ implies that

$$\|l^*(\beta_0, \mathbf{h}_n, H_n) - l^*(\beta_0, \mathbf{h}_0, H_0)\|_\infty = O(\max_{j=0,1,\ldots,d} \{K_j^{-r_j}\}) \tag{A.8}$$

based on (A.2), (A.4) and M4. Thus, $l^*(\beta_0, \mathbf{h}_n, H_n)$ is bounded away from zero for sufficiently large n.

For any $\beta_1, \beta_2 \in \mathcal{B}$, $\mathbf{h}_1, \mathbf{h}_2 \in \prod_{j=1}^d \mathcal{H}_{jn}$ and $g_1, g_2 \in \mathcal{G}_n$, we have

$$\begin{aligned} & |\zeta(\beta_1, \mathbf{h}_1, H_1) - \zeta(\beta_2, \mathbf{h}_2, H_2)| \\ \lesssim \; & |l^*(\beta_1, \mathbf{h}_1, H_1) - l^*(\beta_2, \mathbf{h}_2, H_2)| \\ \lesssim \; & \|\beta_1 - \beta_2\| + \sum_{j=1}^d \|h_{1j} - h_{2j}\|_\infty + \|g_1 - g_2\|_\infty. \end{aligned} \tag{A.9}$$

The first and second inequalities in the above follow from the fact that $l^*(\beta_0, \mathbf{h}_n, H_n)$ is strictly positive for sufficiently large n by (A.8), and Condition M4(a), respectively. As shown in (A.9), the functions in the class $\mathcal{K}$ are Lipschitz continuous in $(\beta, \mathbf{h}, g)$. Therefore, by combining Lemma 2 and Theorem 2.7.11 in (Van de Geer & Wellner, 1996), we obtain that

$$H_B(\epsilon, \mathcal{K}, L_2(P)) \lesssim \max_{0 \le j \le d} \{K_j\} \log(1 + M/\epsilon),$$

where $M = \max_{0 \le j \le d} \{4c_j\}$. In the end, we apply Lemma 3.4.2 in (Van de Geer & Wellner, 1996) to this uniformly bounded class of functions $\mathcal{K}$ to obtain (A.7). □

LEMMA 4. *Suppose the following Conditions (B1)-(B3) hold.*

B1. $\mathbb{P}_n \dot{\ell}_{\widehat{\beta}} = o_P(n^{-1/2})$, $\mathbb{P}_n \dot{\ell}_{\widehat{g}}[\bar{a}^\dagger] = o_P(n^{-1/2})$ *and* $\mathbb{P}_n \dot{\ell}_{\widehat{h}_j}[\bar{b}_j^\dagger] = o_P(n^{-1/2})$*;*

B2. $\sup_{\{\alpha : d(\alpha, \alpha_0) \le C_1 n^{-r/(2r+1)}\}} \mathbb{G}_n(\widetilde{\ell}_\beta(X; \alpha) - \widetilde{\ell}_\beta(X; \alpha_0)) = o_P(1)$*;*

B3. $P(\widetilde{\ell}_\beta(X; \alpha) - \widetilde{\ell}_\beta(X; \alpha_0)) = -\widetilde{I}_0(\beta - \beta_0) + o(\|\beta - \beta_0\|) + o(n^{-1/2})$ *for* α *satisfying* $d(\alpha, \alpha_0) \le C_1 n^{-r/(2r+1)}$.

If $\widehat{\alpha}$ is consistent and $\widetilde{I}_0$ is invertible, then we have

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{I}_0^{-1} \widetilde{\ell}_{\beta_0}(X_i) + o_P(1) \xrightarrow{d} N(0, \widetilde{I}_0^{-1}).$$

LEMMA 5. *(i) If $a(\mathbf{s}, t) = a(\mathbf{s}_1, \mathbf{s}_2, t) \in \mathbf{H}_c^r(\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{T})$ in t relative to $\mathbf{s}_1$ and $\mathbf{s}_2$, then $\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1 \in \mathbf{H}_{c'}^r(\mathcal{S}_2 \times \mathcal{T})$ in t relative to $\mathbf{s}_2$.*

(ii) If $a(\mathbf{s}, t), b(\mathbf{s}, t) \in \mathbf{H}_c^r(\mathcal{S} \times \mathcal{T})$ in t relative to $\mathbf{s}$, then $c(\mathbf{s}, t) \equiv a(\mathbf{s}, t)b(\mathbf{s}, t) \in \mathbf{H}_{c'}^r(\mathcal{S} \times \mathcal{T})$ in t relative to $\mathbf{s}$.

(iii) If $a(\mathbf{s}, t) \in \mathbf{H}_c^r(\mathcal{S} \times \mathcal{T})$ in t relative to $\mathbf{s}$ and $f(\cdot) \in C^{\lceil \beta \rceil}$, then $f(a(\mathbf{s}, t)) \in \mathbf{H}_{c'}^r(\mathcal{S} \times \mathcal{T})$ in t relative to $\mathbf{s}$.

Proof: Let $\lfloor r \rfloor$ be the largest integer smaller than r. Denote the m-th derivative of $a(\mathbf{s}, t)$ w.r.t. t as $D_t^m a(\mathbf{s}, t)$ for $m = 0, 1, \ldots, \lfloor r \rfloor$.

(i) Note that $D_t^m a(\mathbf{s}_1, \mathbf{s}_2, t)$ is bounded for $0 \le m \le \lfloor r \rfloor$, by the dominated convergence theorem, we can take derivative inside the integral to obtain

$$D_t^m \Big(\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1 \Big) = \int_{\mathcal{S}_1} D_t^m a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1,$$

which implies that $D_t^m(\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1)$ is bounded for $0 \le m \le \lfloor r \rfloor$. Using this and the fact that

$$\frac{\big| D_t^{\lfloor r \rfloor} \big(\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t_2)\, d\mathbf{s}_1 \big) - D_t^{\lfloor r \rfloor} \big(\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t_1)\, d\mathbf{s}_1 \big) \big|}{|t_2 - t_1|^{r - \lfloor r \rfloor}}$$
$$\le \int_{\mathcal{S}_1} \sup_{\mathbf{s}_1, \mathbf{s}_2} \sup_{t_1 \ne t_2} \frac{|D_t^{\lfloor r \rfloor} a(\mathbf{s}_1, \mathbf{s}_2, t_2) - D_t^{m_\alpha} a(\mathbf{s}_1, \mathbf{s}_2, t_1)|}{|t_2 - t_1|^{r - \lfloor r \rfloor}}\, d\mathbf{s}_1 \le c' < \infty,$$

for all $\mathbf{s}_2$ and $t_1 \ne t_2$, we conclude that $\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1 \in \mathbf{H}_{c'}^r(\mathcal{S}_2 \times \mathcal{T})$ in t relative to $\mathbf{s}_2$ for some $c' < \infty$.

(ii) The result is true because

$$D_t^m c = \sum_{i+j=m} D_t^i a D_t^j b$$

is bounded for $0 \le m \le \lfloor r \rfloor$. Also we note that for $i < \lfloor r \rfloor$,

$$\frac{|D_t^i a(\mathbf{s}, t_2) - D_t^i a(\mathbf{s}, t_1)|}{|t_2 - t_1|^{r - \lfloor r \rfloor}} = \frac{|\int_{t_1}^{t_2} D_t^{i+1} a(\mathbf{s}, t)\, dt|}{|t_2 - t_1|^{r - \lfloor r \rfloor}}.$$

It can then be easily verified that

$$\sup_{\mathbf{s}} \sup_{t_1 \ne t_2} \frac{|D_t^{\lfloor r \rfloor} c(\mathbf{s}, t_2) - D_t^{\lfloor r \rfloor} c(\mathbf{s}, t_1)|}{|t_2 - t_1|^{r - \lfloor r \rfloor}} < \infty.$$

(iii) When $0 < \alpha \leq 1$, the result follows from the observation that

$$\frac{f(a(\mathbf{s},t_2)) - f(a(\mathbf{s},t_1))}{|t_2 - t_1|^\beta} = \frac{f(a(\mathbf{s},t_2)) - f(a(\mathbf{s},t_1))}{|a(\mathbf{s},t_2) - a(\mathbf{s},t_1)|} \cdot \frac{|a(\mathbf{s},t_2) - a(\mathbf{s},t_1)|}{|t_2 - t_1|^\beta}.$$

Using the chain rule, the above observation and part (ii) of the lemma, the desired result can be obtained by induction for general β. □

Denote

$$S_k(X;\alpha,w_k) = [\dot{\ell}_\beta(X;\alpha)]_k - \dot{\ell}_g[a_k](X;\alpha) - \sum_{j=1}^{d} \dot{\ell}_{h_j}[b_{jk}](X;\alpha),$$

where $w_k = (a_k, b_{1k}, \ldots, b_{dk})$. Let $\mathcal{W}_n = \mathcal{G}_n \times \prod_{j=1}^d \mathcal{H}_{jn}$ and $\mathcal{N}_0 = \{\alpha \in \mathcal{A} : d(\alpha,\alpha_0) = o(1)\}$.

LEMMA 6. *Under Conditions M1-M7 & P1-P2, we have*

$$E \sup_{w_k \in \mathcal{W}_n} |S_k(X;\alpha,w_k) - S_k(X;\alpha_0,w_k)|^2 \lesssim d^2(\alpha,\alpha_0) \tag{A.10}$$

for all $\alpha \in \mathcal{N}_0$ and $k = 1,\ldots,l$.

Proof: In view of (12)-(14) , we can bound the left hand side of (A.10) by

$$\begin{aligned}
\lesssim \quad & \|Q_\theta - Q_{\theta_0}\|_2^2 + E\left\{ \sup_{a_k \in \mathcal{G}_n} \left[\int_{l_v}^{V} (\exp(g(s)) - \exp(g_0(s)))a_k(s)ds \right]^2 (Q_\theta - Q_{\theta_0})^2 \right\} \\
& + E \sup_{a_k \in \mathcal{G}_n} \left[\int_{l_v}^{V} \exp(g_0(s))a_k(s)ds(Q_\theta - Q_{\theta_0}) \right]^2 \\
& + E \sup_{a_k \in \mathcal{G}_n} \left[\int_{l_v}^{V} (\exp(g(s)) - \exp(g_0(s)))a_k(s)ds Q_{\theta_0} \right]^2 \\
& + \sum_{j=1}^{d} E \sup_{b_{jk} \in \mathcal{H}_{jn}} \left[b_{jk}^2 (Q_\theta - Q_{\theta_0})^2 \right]
\end{aligned}$$

after some algebra. The compactness of $\mathcal{G}_n$ and $\mathcal{H}_{jn}$ imply that the third and fifth term in the above are both of the order $\|Q_\theta - Q_{\theta_0}\|_2^2$. For the second term, we can further bound it by

$$E\left[\sup_{a_k \in \mathcal{G}_n} \int_{l_V}^{V} a_k^2(s)ds \int_{l_V}^{V} [\exp(g(s)) - \exp(g_0(s))]^2 ds (Q_\theta - Q_{\theta_0})^2 \right].$$

Considering the compactness of $\mathcal{G}$ and $\mathcal{G}_n$, we know the second term is also of the order $\|Q_\theta - Q_{\theta_0}\|_2^2$. Assumption M4(a) together with Cauchy-Schwartz inequality implies that $\|Q_\theta - Q_{\theta_0}\|_2^2 \lesssim \|\beta - \beta_0\|^2 + \|H - H_0\|_2^2 + \|\sum_{j=1}^d (h_j - h_{j0})\|_2^2$. Since we assume that the density for W is bounded away from zero and infinity, we have that $\|\sum_{j=1}^d (h_j - h_{j0})\|_2^2 \lesssim \sum_{j=1}^d \|h_j - h_{j0}\|_2^2$ considering the identifiability condition $\int_0^1 h_j(w_j)dw_j = 0$. Assumption M7 implies that the fourth term is of the order $\|H - H_0\|_2^2$. Considering the form of $d(\alpha,\alpha_0)$, we conclude the whole proof. □

Proof of Theorem 1

We show the estimation consistency (9) by first establishing

$$P\left\{(\widehat{\beta}-\beta_0)'Z+\sum_{j=1}^{d}(\widehat{h}_j-h_{jn})(W_j)+\widehat{H}(V)-H_n(V)\right\}^2=o_P(1). \tag{A.11}$$

Combining (A.11) with the identifiability Condition M3, we directly obtain $(\widehat{\beta}-\beta_0)=o_P(1)$ which, in turn, implies that $P\left\{\sum_{j=1}^{d}(\widehat{h}_j-h_{jn})(W_j)+\widehat{H}(V)-H_n(V)\right\}^2=o_P(1)$. Considering the assumption M2(b) and that $\int_0^1 h_j(w_j)dw_j=0$ for $h_j\in\mathcal{H}_j\cup\mathcal{H}_{jn}$, we can further show $\sum_{j=1}^{d}\|\widehat{h}_j-h_{jn}\|_2+\|\widehat{H}-H_n\|_2=o_P(1)$. The spline approximation result (A.2) and (A.3) conclude the proof of (9).

In the below, we will show (A.11) to complete the proof of (9). Recall that $\mathbf{h}=(h_1,\ldots,h_d)$. Denote $\mathbf{h}_0$, $\mathbf{h}_n$ and $\widehat{\mathbf{h}}$ as the corresponding true value, B-spline approximation and sieve estimate, respectively. Recall that $l^*(\beta_0,\mathbf{h}_n,H_n)$ is bounded away from zero for sufficiently large n as implied by (A.8). Then, by the definition of $\widehat{\alpha}$, we have

$$\mathbb{P}_n\log\{l^*(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H})/l^*(\beta_0,\mathbf{h}_n,H_n)\}\geq 0,$$

which implies that, by the inequality that $\alpha\log(x)\leq\log(1+\alpha(x-1))$ for any $x>0$ and $\alpha\in(0,1)$,

$$0\leq\mathbb{P}_n\log\left[1+\alpha\left\{\frac{l^*(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H})}{l^*(\beta_0,\mathbf{h}_n,H_n)}-1\right\}\right]\equiv\mathbb{P}_n\zeta(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H}). \tag{A.12}$$

Lemma 3 implies that $(\mathbb{P}_n-P)\zeta(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H})=o_P(1)$ since $K_j/n=o(1)$ for any $j=0,1,\ldots,d$. Thus, $P\zeta(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H})\geq o_P(1)$ based on (A.12). Let $U_n(X)=l^*(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H})/l^*(\beta_0,\mathbf{h}_n,H_n)$. Based on (A.8) we know $PU_n(X)=1+o_P(1)$, which further implies $P\zeta(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H})\leq o_P(1)$ by the concavity of $s\mapsto\log(s)$. This in turn implies that $P\zeta(\widehat{\beta},\widehat{\mathbf{h}},\widehat{H})=o_P(1)$. This forces $P|(\beta_0'Z+\sum_{j=1}^{d}h_{jn}(W_j)+H_n(V))-(\widehat{\beta}'Z+\sum_{j=1}^{d}\widehat{h}_j(W_j)+\widehat{H}(V))|=o_P(1)$ by the strict concavity of $s\mapsto\log s$, Conditions M4(a), P1 and P2. It is easy to verify that $ER_n^2=o_P(1)$ if $E|R_n|=o_P(1)$. Thus, we have shown (A.11) in the end.

As for the convergence rate results (10) & (11), we first apply Theorem 3.2.5 in Van de Geer & Wellner (1996) to establish

$$\|\widehat{\theta}-\theta_0\|_2=O_P(\delta_{1n}\vee\delta_{2n}), \tag{A.13}$$

where $\widehat{\theta}$ is the plug-in sieve estimate of θ and

$$\delta_{1n}=\max_{0\leq j\leq d}\{\sqrt{K_j}\}/\sqrt{n}\quad\text{and}\quad\delta_{2n}=\max_{0\leq j\leq d}\{K_j^{-r_j}\}. \tag{A.14}$$

Following similar arguments in proving the consistency, we know that (A.13) implies (10) and (11) by choosing $K_j \asymp n^{1/(2r_j+1)}$.

In the below, we show (A.13) by verifying the conditions of Theorem 3.2.5 in Van de Geer & Wellner (1996). We first need to show that

$$P[\ell(\alpha_0) - \ell(\alpha)] \gtrsim \|\theta - \theta_0\|_2^2 \tag{A.15}$$

for every α in the neighborhood of α_0. Define $q(\delta, t) = \delta \log(F(t)) + (1-\delta)\log(1-F(t))$ and $\ddot{q}(\delta, t)$ as its second derivative w.r.t. t. Since α_0 maximizes $\alpha \mapsto P\ell(\alpha)$, we have

$$P[\ell(\alpha_0) - \ell(\alpha)] = P\left[\frac{-\ddot{q}(\delta, \widetilde{\theta})}{2}(\theta - \theta_0)^2\right],$$

where $\widetilde{\theta}$ is on the line segment between θ and θ_0. The compactness of the parameter spaces imply that $P[\ell(\alpha_0) - \ell(\alpha)] \asymp \|\theta - \theta_0\|_2^2$. This completes the proof of (A.15). We next calculate the order of $E\sup_{\|\theta-\theta_0\|_2 \le \delta} |\mathbb{G}_n(\ell(\alpha) - \ell(\alpha_0))|$ as a function of δ, denoted as $\phi_n(\delta)$, by the use of Lemma 3.4.2 of Van de Geer & Wellner (1996). Let $\mathcal{F}_{1n}(\delta) = \{\ell(\alpha) - \ell(\alpha_0) : g \in \mathcal{G}_n, h_j \in \mathcal{H}_{jn}, \|\theta - \theta_0\|_2 \le \delta\}$. Using the same argument as that in the proof of Lemma 3, we obtain that $H_B(\epsilon, \mathcal{F}_{1n}(\delta), L_2(P))$ is bounded by $C \max_{0\le j\le d}\{K_j\}\log(1+\delta/\epsilon)$. This leads to

$$J_B(\delta, \mathcal{F}_{1n}(\delta), L_2(P)) = \int_0^\delta \sqrt{1 + H_B(\epsilon, \mathcal{F}_{1n}(\delta), L_2(P))}d\epsilon \le C \max_{0\le j\le d}\{\sqrt{K_j}\}\delta.$$

The compactness of $\mathcal{G}_n$ and $\mathcal{H}_{jn}$ implies the uniform boundedness of any $f \in \mathcal{F}_{1n}(\delta)$. Thus, Lemma 3.4.2 of Van de Geer & Wellner (1996) gives

$$\phi_n(\delta) = \max_{0\le j\le d}\{\sqrt{K_j}\}\delta + \max_{0\le j\le d}\{K_j\}/\sqrt{n}.$$

By solving $\delta_{1n}^{-2}\phi_n(\delta_{1n}) \le \sqrt{n}$, we get the form of δ_{1n} in (A.14).

We next show that $\mathbb{P}_n\ell(\widehat{\alpha}) - \mathbb{P}_n\ell(\alpha_0) \ge -O_P(\delta_{2n}^2)$. The definition of $\widehat{\alpha}$ implies that $\mathbb{P}_n[\ell(\widehat{\alpha}) - \ell(\alpha_0)] \ge A_n + B_n$, where $A_n = (\mathbb{P}_n - P)\{\ell(\beta_0, H_n, \mathbf{h}_n) - \ell(\alpha_0)\}$ and $B_n = P\{\ell(\beta_0, H_n, \mathbf{h}_n) - \ell(\alpha_0)\}$. A straightforward Taylor expansion gives

$$A_n = (\mathbb{P}_n - P)\left\{\dot{\ell}_2(\beta_0, \widetilde{H}_n, \widetilde{\mathbf{h}}_n)(H_n - H_0) + \sum_{j=1}^d \dot{\ell}_{2+j}(\beta_0, \widetilde{H}_n, \widetilde{\mathbf{h}}_n)(h_{jn} - h_{j0})\right\},$$

where $\dot{\ell}_t$ is the Fréchet derivative of $\ell(\beta_0, H_n, \mathbf{h}_n)$ w.r.t. the t-th argument. Considering (A.2), (A.3) and the fact that $0 < \epsilon_1 \le |\dot{q}(\delta, t)| \le \epsilon_2 < \infty$ for t in some compacta of $\mathbb{R}^1$, we have

$$P\left\{\frac{\dot{\ell}_2(\beta_0, \widetilde{H}_n, \widetilde{\mathbf{h}}_n)(H_n - H_0) + \sum_{j=1}^d \dot{\ell}_{2+j}(\beta_0, \widetilde{H}_n, \widetilde{\mathbf{h}}_n)(h_{jn} - h_{j0})}{\max_{0\le j\le d}\{K_j^{-r_j}\}n^\epsilon}\right\}^2 \to 0 \tag{A.16}$$

for any $\epsilon > 0$. Let $\mathcal{F}_{2n} = \{\ell(\beta_0, H, \mathbf{h}) - \ell(\alpha_0) : g \in \mathcal{G}_n, h_j \in \mathcal{H}_{jn}, \|g - g_0\|_\infty \leq C_0 K_0^{-r_0}, \|h_j - h_{j0}\|_\infty \leq C_j K_j^{-r_j}\}$. Similar analysis in Lemma 3 show that the bracketing entropy integral (in terms of $L_2(P)$) for $\mathcal{F}_{2n}$ is finite, thus yields that $\mathcal{F}_{2n}$ is P-Donsker. Combining this P-Donsker result and (A.16), we use Corollary 2.3.12 of Van de Geer & Wellner (1996) to conclude that $\sqrt{n}A_n/(\max_{0\leq j\leq d}\{K_j^{-r_j}\}n^\epsilon) = o_P(1)$. By choosing some proper $0 < \epsilon < 1/2$ satisfying $n^{\epsilon-1/2} = \max_{0\leq j\leq d}\{K_j^{-r_j}\}$, we have $A_n = o_P(\max_{0\leq j\leq d}\{K_j^{-2r_j}\})$. We can also show $B_n \geq -O(\max_{0\leq j\leq d}\{K_j^{-2r_j}\})$ by similar analysis of (A.15). This gives the form of δ_{2n} in (A.14), and thus concludes the whole proof. □

Proof of Theorem 2

We apply Lemma 4 to prove this theorem by checking their Conditions B1 – B3. To facilitate the understanding, we first sketch the verification of Condition B1 and then provide the details. To verify B1, we first know that $\mathbb{P}_n\dot{\ell}_{\widehat{\beta}} = 0$ since $\widehat{\beta}$ maximizes $l(\beta, \widehat{g}, \widehat{h}_1, \ldots, \widehat{h}_d)$, $\widehat{\beta}$ is consistent and β_0 is an interior point of $\mathcal{B}$. We next show that $b_{jk}^\dagger$ $(a_k^\dagger)$ belongs to $\mathbf{H}_{\widetilde{c}_j}^{r_j}[0,1]$ $(\mathbf{H}_{\widetilde{c}_0}^{r_0}[l_v, u_v])$ for some $0 < \widetilde{c}_j < \infty$ and $j = 0, 1, \ldots, d$ such that there exists a $b_{jkn}^\dagger \in \mathcal{H}_{jn}$ $(a_{kn}^\dagger \in \mathcal{G}_n)$ satisfying

$$\|b_{jk}^\dagger - b_{jkn}^\dagger\|_\infty = O(n^{-r_j/(2r_j+1)}) \quad \text{(A.17)}$$

$$\|a_{kn}^\dagger - a_k^\dagger\|_\infty = O(n^{-r_0/(2r_0+1)}) \quad \text{(A.18)}$$

by (6) and the assumption that $K_j \asymp n^{1/(2r_j+1)}$. Since $\mathbb{P}_n\dot{\ell}_{\widehat{h}_j}[b_{jkn}] = 0$ and $\mathbb{P}_n\dot{\ell}_{\widehat{g}}[a_{kn}] = 0$ for any $b_{jkn} \in \mathcal{H}_{jn}$ and $a_{kn} \in \mathcal{G}_n$, it remains to show

$$\mathbb{P}_n\left\{\dot{\ell}_{\widehat{h}_j}[b_{jkn}^\dagger] - \dot{\ell}_{\widehat{h}_j}[b_{jk}^\dagger]\right\} = o_P(n^{-1/2}), \quad \text{(A.19)}$$

$$\mathbb{P}_n\left\{\dot{\ell}_{\widehat{g}}[a_{kn}^\dagger] - \dot{\ell}_{\widehat{g}}[a_k^\dagger]\right\} = o_P(n^{-1/2}) \quad \text{(A.20)}$$

for verifying Condition B1.

Now we show $b_{jk}^\dagger \in \mathbf{H}_{\widetilde{c}_j}^{r_j}[0,1]$ and (A.19). Following the analysis in Page 2282 of Ma & Kosorok (2005a), we can write, with $\bar{a}_I^\dagger(v) = \int_{l_v}^v \exp(g_0(s))\bar{a}^\dagger(s)ds$,

$$\begin{aligned}\bar{b}_j^\dagger &= \Pi_j D(v,w) - \Pi_j \bar{a}_I^\dagger(v) - \sum_{i\neq j}\Pi_j \bar{b}_i^\dagger \\ &= \Pi_j D(v,w) - \int_{l_v}^{u_v} \bar{a}_I^\dagger(v) Sf(v,w_j)dv - \sum_{i\neq j}\int_0^1 \bar{b}_i^\dagger(w_i) Tf(w_i,w_j)dw_i.\end{aligned}$$

According to Lemma 5 and dominated convergence theorem, we know that $b_{jk}^\dagger(w_j) \in \mathbf{H}_{\widetilde{c}_j}^{r_j}[0,1]$ under Condition M5, $b_{jk}^\dagger \in L_2^0(w_j)$ and $a_k^\dagger \in L_2(H)$ (thus $a_{Ik}^\dagger$ is uniformly bounded) for some

$0 < \widetilde{c}_j < \infty$. As for (A.19), we first decompose its left hand side as $I_{1n} + I_{2n}$, where

$$\begin{aligned} I_{1n} &= P\left\{\dot{\ell}_{\widehat{h}_j}[b^\dagger_{jkn} - b^\dagger_{jk}] - \dot{\ell}_{h_{j0}}[b^\dagger_{jkn} - b^\dagger_{jk}]\right\}, \\ I_{2n} &= (\mathbb{P}_n - P)\left\{\dot{\ell}_{\widehat{h}_j}[b^\dagger_{jkn} - b^\dagger_{jk}]\right\}. \end{aligned}$$

By Cauchy-Schwartz Inequality, we have $I_{1n} \lesssim \|b^\dagger_{kjn} - b^\dagger_{kj}\|_\infty \|\widehat{\theta} - \theta_0\|_2$ based on Conditions M4(a), P1 & P2. Thus, (A.13) and (A.17) imply that $I_{1n} = O_P(n^{-2r/(2r+1)}) = o_P(n^{-1/2})$ since $r > 1/2$.

To show $I_{2n} = o_P(n^{-1/2})$, we need to make use of Lemma 3.4.2 in Van de Geer & Wellner (1996). We first construct the following class of functions:

$$\mathcal{I}_n = \left\{ f_{\theta, b_{jkn}}(x) = \dot{\ell}_{h_j}[b_{jkn} - b^\dagger_{jk}](x;\alpha) : \alpha \in \mathcal{A}_n(n^{\frac{-r}{2r+1}}) \text{ and } b_{jkn} \in \mathcal{H}'_{jn}\left(n^{\frac{-r_j}{2r_j+1}}\right)\right\},$$

where $\mathcal{A}_n(\delta) \equiv \{\alpha \in \mathcal{A}_n : d(\alpha, \alpha_0) \leq C_1\delta\}$ and $\mathcal{H}'_{jn}(\delta) \equiv \{b_{jkn} \in \mathcal{H}_{jn} : \|b_{jkn} - b^\dagger_{jk}\|_\infty \leq C_2\delta\}$ for some $0 < C_1, C_2 < \infty$. Let $\Theta_n(\delta) = \{\beta' z + H(v) + \sum_{j=1}^d h_j(w_j) : \alpha \in \mathcal{A}_n(\delta)\}$. It is easy to verify that, for every x,

$$|f_{\theta_1, b_{jkn1}}(x) - f_{\theta_2, b_{jkn2}}(x)| \lesssim \|\theta_1 - \theta_2\|_\infty + \|b_{jkn1} - b_{jkn2}\|_\infty, \tag{A.21}$$

where $\theta_j \in \Theta_n(n^{-r/(2r+1)})$ for $j = 1, 2$. Let $\theta^1, \ldots, \theta^{N(\epsilon, \Theta_n(n^{-r/(2r+1)}), \|\cdot\|_\infty)}$ and

$$b^1_{jkn}, \ldots, b^{N(\epsilon, \mathcal{H}'_{jn}(n^{-r_j/(2r_j+1)}), \|\cdot\|_\infty)}_{jkn}$$

be the ϵ-cover for $\Theta_n(n^{-r/(2r+1)})$ and $\mathcal{H}'_{jn}(n^{-r_j/(2r_j+1)})$, respectively. Thus, we can construct the bracket $[f_{\theta^i, b^l_{jkn}} - 2C\epsilon, f_{\theta^i, b^l_{jkn}} + 2C\epsilon]$ covering $\mathcal{I}_n$. The bracket size is $4C\epsilon$. Hence, we obtain

$$\begin{aligned} & H_B(\epsilon, \mathcal{I}_n, L_2(P_X)) \\ \leq\ & H(\epsilon/(4C), \Theta_n(n^{\frac{-r}{2r+1}}), \|\cdot\|_\infty) + H(\epsilon/(4C), \mathcal{H}'_{jn}(n^{\frac{-r_j}{2r_j+1}}), \|\cdot\|_\infty) \\ \lesssim\ & \max_{0 \leq j \leq d}\{K_j\} \log(1 + n^{-r/(2r+1)}/\epsilon) \end{aligned}$$

based on Lemma 2. The corresponding δ-bracketing entropy integral is calculated as

$$J_B(\delta, \mathcal{I}_n, L_2(P_X)) \equiv \int_0^\delta \sqrt{1 + H_B(\epsilon, \mathcal{I}_n, L_2(P_X))} \lesssim \max_{0 \leq j \leq d}\{\sqrt{K_j}\} n^{-\frac{r}{4r+2}} \delta^{1/2}. \tag{A.22}$$

Now, it is ready to apply Lemma 3.4.2 in Van de Geer & Wellner (1996) to show $E\|\mathbb{G}_n\|_{\mathcal{I}_n} = o(1)$ implying $I_{2n} = o_P(n^{-1/2})$. Note that $\|f\|_2 \lesssim \|b_{jkn} - bjk^\dagger\|_2$ and $\|f\|_\infty \leq \|b_{jkn} - bjk^\dagger\|_\infty$ for any $f \in \mathcal{I}_n$, and thus δ and M in Lemma 3.4.2 of Van de Geer & Wellner (1996) are both

chosen as $K_j^{-r_j}$, i.e., $n^{-r_j/(2r_j+1)}$. Then, by Lemma 3.4.2 of Van de Geer & Wellner (1996) and (A.22), we have that

$$E\|\mathbb{G}_n\|_{\mathcal{I}_n} = O\left(n^{-\left(\frac{r-1}{4r+2}+\frac{r_j}{4r_j+2}\right)} \vee n^{-\frac{4r-1}{4r+2}}\right) = o(1).$$

This completes the proof of (A.19).

We next show (A.20) by similar arguments. Similarly, we have

$$\bar{a}_I^\dagger(v) = \Pi_a D(v,w) - \sum_{j=1}^{d}\int_0^1 \bar{b}_j^\dagger(w_j) U f(w_j, v) dw_j.$$

Recall that $\bar{a}_I^\dagger(v) = \int_{l_v}^v \exp(g_0(s))\bar{a}^\dagger(s)ds$. Under Condition M6 and the assumption that $g_0 \in \mathbf{H}_{c_0}^{r_0}[l_v, u_v]$, we can show that $a_{Ik}^\dagger \in \mathbf{H}_{\widetilde{c}_0}^{r_0+1}[l_v, u_v]$, which implies that $a_k^\dagger \in \mathbf{H}_{\widetilde{c}_0}^{r_0}[l_v, u_v]$ for some $0 < \widetilde{c}_0 < \infty$, based on Lemma 5. We next show that $I'_{1n} = o_P(n^{-1/2})$ and $I'_{2n} = o_P(n^{-1/2})$, where

$$\begin{aligned} I'_{1n} &= P\left\{\dot{\ell}_{\widehat{g}}[a_{kn}^\dagger - a_k^\dagger] - \dot{\ell}_{g_0}[a_{kn}^\dagger - a_k^\dagger]\right\}, \\ I'_{2n} &= (\mathbb{P}_n - P)\left\{\dot{\ell}_{\widehat{g}}[a_{kn}^\dagger - a_k^\dagger]\right\}. \end{aligned}$$

Similarly, by Cauchy-Schwartz Inequality, we can show that

$$\begin{aligned} I'_{1n} &\lesssim \|a_{kn}^\dagger - a_k^\dagger\|_\infty \|\widehat{\theta} - \theta_0\|_2 + P\left[\int_{l_v}^v (\exp(\widehat{g}) - \exp(g_0))(s)(a_{kn}^\dagger - a_k^\dagger)(s)ds\right] \\ &\lesssim \|a_{kn}^\dagger - a_k^\dagger\|_\infty \left(\|\widehat{\theta} - \theta_0\|_2 + \|\widehat{H} - H_0\|_2\right) \\ &\lesssim O_P(n^{-r/(2r+1)}) = o_P(n^{-1/2}) \end{aligned}$$

by choosing $K_j \asymp n^{1/(2r_j+1)}$. Following similar arguments in analyzing I_{2n}, we can show that $I'_{2n} = o_P(n^{-1/2})$. Thus, we have verified Condition B1 in Lemma 4. We again apply Lemma 3.4.2 of Van de Geer & Wellner (1996) to verify Assumption B2. The details are skipped due to the similarity of the previous analysis.

It remains to verify Assumption B3. This can be easily established using the Taylor expansion in Banach space. However, we first need to reparameterize the efficient score function $\widetilde{\ell}_\beta(X;\alpha)$ as

$$\begin{aligned} \widetilde{\ell}_\beta(X;\alpha^*) &= ZQ_\theta(X) - \left[\int_{l_v}^V \bar{a}^\dagger(s)dH(s) + \sum_{j=1}^{d}\bar{b}_j^\dagger(W_j)\right] Q_\theta(X) \\ &\equiv \dot{\ell}_\beta(X;\alpha^*) - \dot{\ell}_\eta[\bar{c}^\dagger](X;\alpha^*), \end{aligned}$$

where $\alpha^* = (\beta, H, h_1, \ldots, h_d)$, $\eta = (H, h_1, \ldots, h_d)$ and $\bar{c}^\dagger = (\bar{a}^\dagger, \bar{b}_1^\dagger, \ldots, \bar{b}_d^\dagger)$. We first derive two useful equalities (A.26)-(A.27). Let E_{α^*} be the expectation corresponding to the reparametrized likelihood under the parameter α^*. Since $E_{\alpha^*}\widetilde{\ell}_\beta(X;\alpha^*) = 0$, we have

$$\frac{\partial}{\partial t}|_{t=0} E_{\alpha_t^*}\widetilde{\ell}_\beta(X;\alpha_t^*) = 0, \tag{A.23}$$

where $\alpha_t^* = \alpha_0^* + t\epsilon$. Define $\widetilde{\ell}_{\beta,\beta}$ and $\widetilde{\ell}_{\beta,\eta}[c]$ as the first derivative of $\widetilde{\ell}_\beta$ w.r.t. β and η (along the direction c), respectively. By setting $\epsilon = (\epsilon'_\beta, 0, \ldots, 0)'$ and $\epsilon = (0, e)' = (0, \Delta H, b_1, \ldots, b_d)'$, respectively, some calculations reveal that

$$\begin{aligned} E\left\{\widetilde{\ell}_{\beta,\beta}(X;\alpha_0^*)\epsilon_\beta\right\} + E\left\{\widetilde{\ell}_\beta(X;\alpha_0^*)\dot{\ell}'_\beta(X;\alpha_0^*)\epsilon_\beta\right\} &= 0, & \text{(A.24)}\\ E\left\{\widetilde{\ell}_{\beta,\eta}[e](X;\alpha_0^*)\right\} + E\left\{\widetilde{\ell}_\beta(X;\alpha_0^*)\dot{\ell}'_\eta[e](X;\alpha_0^*)\right\} &= 0 & \text{(A.25)}\end{aligned}$$

based on (A.23). By considering the orthogonal property of $\widetilde{\ell}_{\beta_0}$ and the above reparametrization, we obtain the following two useful facts:

$$\widetilde{I}_0 = -E\left\{\widetilde{\ell}_{\beta,\beta}(X;\alpha_0^*)\right\}, \tag{A.26}$$

$$E\left\{\widetilde{\ell}_{\beta,\eta}[e](X;\alpha_0^*)\right\} = 0 \tag{A.27}$$

based on (A.24) and (A.25).

Define $\widetilde{\ell}_{\beta,\alpha^*,\alpha^*}[h_1,h_2](X;\alpha^*)$ as the second order Fréchet derivative of $\widetilde{\ell}_\beta$ w.r.t. α^* along the direction $[h_1,h_2]$ at the point α^*. The same notation rule applies to $\dot{\ell}_{\beta,\alpha^*,\alpha^*}[h_1,h_2](X;\alpha^*)$ and $\dot{\ell}_{\eta,\alpha^*,\alpha^*}[h_1,h_2,h_3](X;\alpha^*)$. Now we are ready to express the Taylor expansion as follows.

$$\begin{aligned} & E[\widetilde{\ell}_\beta(X;\alpha) - \widetilde{\ell}_\beta(X;\alpha_0)]\\ = \;& E[\widetilde{\ell}_\beta(X;\alpha^*) - \widetilde{\ell}_\beta(X;\alpha_0^*)]\\ = \;& E\left\{\widetilde{\ell}_{\beta,\beta}(X;\alpha_0^*)\right\}(\beta-\beta_0) + E\left\{\widetilde{\ell}_{\beta,\eta}[\eta-\eta_0](X;\alpha_0^*)\right\}\\ & +\frac{1}{2}E\left\{\widetilde{\ell}_{\beta,\alpha^*,\alpha^*}[\Delta\alpha^*,\Delta\alpha^*](X;\widetilde{\alpha}^*)\right\}\\ = \;& -\widetilde{I}_0(\beta-\beta_0)\\ & +\frac{1}{2}E\left\{\dot{\ell}_{\beta,\alpha^*,\alpha^*}[\Delta\alpha^*,\Delta\alpha^*](X;\widetilde{\alpha}^*) - \dot{\ell}_{\eta,\alpha^*,\alpha^*}[\bar{c}^\dagger,\Delta\alpha^*,\Delta\alpha^*](X;\widetilde{\alpha}^*)\right\}, \end{aligned}$$

where $\Delta\alpha^* = \alpha^* - \alpha_0^*$ and $\widetilde{\alpha}^*$ lies between α^* and α_0^*. The last equation in the above follows from (A.26) & (A.27). Now we only need to show that the second term in the last equation is of the order

$$o(\|\beta-\beta_0\|) + o(n^{-1/2}).$$

Let $\Delta H = H - H_0$ and $\Delta h_j = h_j - h_{j0}$. After some algebra, we obtain

$$\begin{aligned}
& \dot{\ell}_{\beta,\alpha^*,\alpha^*}[\Delta\alpha^*, \Delta\alpha^*](X; \widetilde{\alpha}^*) \\
= \;& Z\ddot{Q}_{\widetilde{\theta}} \left[Z'(\beta - \beta_0) + \Delta H(V) + \sum_{j=1}^{d} \Delta h_j(W_j) \right]^2, \\
& \dot{\ell}_{\eta,\alpha^*,\alpha^*}[\bar{c}^\dagger, \Delta\alpha^*, \Delta\alpha^*](X; \widetilde{\alpha}^*) \\
= \;& \left[\int_{l_v}^{V} \bar{a}^\dagger(s) dH(s) + \sum_{j=1}^{d} \bar{b}_j^\dagger(W_j) \right] \ddot{Q}_{\widetilde{\theta}} \left[Z'(\beta - \beta_0) + \Delta H(V) + \sum_{j=1}^{d} \Delta h_j(W_j) \right]^2 \\
& + 2 \left[\int_{l_v}^{V} \bar{a}^\dagger(s) d\Delta H(s) \right] \dot{Q}_{\widetilde{\theta}} \left[Z'(\beta - \beta_0) + \Delta H(V) + \sum_{j=1}^{d} \Delta h_j(W_j) \right],
\end{aligned}$$

where $\widetilde{\theta}$ lies between θ and θ_0. Considering the assumption that $d(\alpha, \alpha_0) \leq C_1 n^{-r/(2r+1)}$ and the previously shown result that $a_k^\dagger$ and $b_{jk}^\dagger$ are both uniformly bounded, we can verify Assumption B3 based on the above expressions. This completes the proof of Theorem 2. □

Proof of Theorem 3

To facilitate the understanding, we first provide the roadmap of our proof here. For simplicity, we write $S_k(X; \alpha_0, w_k)$ and $S_k(X; \widehat{\alpha}, w_k)$ as $S_k^0[w_k]$ and $\widehat{S}_k[w_k]$, respectively. Based on the definitions of $\widetilde{I}_0$ and (19), we know their (k, k')-th entry can be written as

$$\widetilde{I}_0(k, k') = E S_k^0[w_k^\dagger] S_{k'}^0[w_{k'}^\dagger], \tag{A.28}$$
$$\widehat{I}(k, k') = \mathbb{P}_n \widehat{S}_k[\widehat{w}_k^\dagger] \widehat{S}_{k'}[\widehat{w}_{k'}^\dagger], \tag{A.29}$$

where $w_k^\dagger = (a_k^\dagger, b_{1k}^\dagger, \ldots, b_{dk}^\dagger)$ and $\widehat{w}_k^\dagger = ((\gamma_{0k}^\dagger)'\mathbf{B}_0, (\gamma_{1k}^\dagger)'\mathbf{B}_1, \ldots, (\gamma_{dk}^\dagger)'\mathbf{B}_d)$. Recall that $\mathcal{W}_n = \mathcal{G}_n \times \prod_{j=1}^{d} \mathcal{H}_{jn}$. Define $\widetilde{w}_k^\dagger \equiv \arg\min_{w_k \in \mathcal{W}_n} E\{S_k^0[w_k]\}^2$. To establish $\widehat{I} \xrightarrow{P} \widetilde{I}_0$, we need to establish the following three equations step by step:

$$\widehat{I}(k, k') = E S_k^0[\widehat{w}_k^\dagger] S_{k'}^0[\widehat{w}_{k'}^\dagger] + o_P(1), \tag{A.30}$$
$$E S_k^0[\widehat{w}_k^\dagger] S_{k'}^0[\widehat{w}_{k'}^\dagger] - E S_k^0[\widetilde{w}_k^\dagger] S_{k'}^0[\widetilde{w}_{k'}^\dagger] = o_P(1), \tag{A.31}$$
$$E S_k^0[\widetilde{w}_k^\dagger] S_{k'}^0[\widetilde{w}_{k'}^\dagger] - \widetilde{I}_0(k, k') = o(1). \tag{A.32}$$

We first consider (A.30). It is easy to show that

$$E \left[\sup_{\alpha \in \mathcal{N}_0, w_k \in \mathcal{W}_n} |S_k(X; \alpha, w_k)|^2 \right] \leq \text{const.} < \infty \tag{A.33}$$

since $\mathcal{A}$ and $\mathcal{W}_k$ are both assumed to be compact. Note that (A.33) implies that $\{S_k(x;\alpha,w_k) : \alpha \in \mathcal{N}_0, w_k \in \mathcal{W}_n\}$ is P-Glivenko-Cantelli. Then, we know that, uniformly over $w_k, w_{k'} \in \mathcal{W}_n$,

$$\mathbb{P}_n \widehat{S}_k[w_k]\widehat{S}_{k'}[w_{k'}] = E\widehat{S}_k[w_k]\widehat{S}_{k'}[w_{k'}] + o_P(1) \tag{A.34}$$

by considering Corollary 9.27 of Kosorok (2008). Uniformly over $w_k, w_{k'} \in \mathcal{W}_n$, we have

$$\begin{aligned} & \left| E\widehat{S}_k[w_k]\widehat{S}_{k'}[w_{k'}] - ES_k^0[w_k]S_{k'}^0[w_{k'}] \right| \\ \leq\ & E\left|\widehat{S}_k[w_k](\widehat{S}_{k'}[w_{k'}] - S_{k'}^0[w_{k'}])\right| + E\left|S_{k'}^0[w_{k'}](\widehat{S}_k[w_k] - S_k^0[w_k])\right| \\ \leq\ & \|\widehat{S}_k^2[w_k]\|_2 \|\widehat{S}_{k'}[w_{k'}] - S_{k'}^0[w_{k'}]\|_2 + \|S_{k'}^0[w_{k'}]\|_2 \|\widehat{S}_k[w_k] - S_k^0[w_k]\|_2 \\ \leq\ & o_P(1), \end{aligned} \tag{A.35}$$

where the last inequality follows from Lemma 6 (together with the consistency of $\widehat{\alpha}$) & (A.33). Combining (A.34) and (A.35), we have obtained that

$$\sup_{w_k, w_{k'} \in \mathcal{W}_n} \left| \mathbb{P}_n \widehat{S}_k[w_k]\widehat{S}_{k'}[w_{k'}] - ES_k^0[w_k]S_{k'}^0[w_{k'}] \right| = o_P(1), \tag{A.36}$$

which implies (A.30).

We next consider (A.31). By similar analysis applied to (A.35), we know that (A.31) holds if $\|S_k^0[\widetilde{w}_k^\dagger] - S_k^0[\widehat{w}_k^\dagger]\|_2 = o_P(1)$. Denote $M_n(w)$ and $M(w)$ as $\mathbb{P}_n \widehat{S}_k^2[w]$ and $\|S_k^0[w]\|_2^2$, respectively. The definition of $\widetilde{w}_k^\dagger$ further implies that

$$\begin{aligned} \|S_k^0[\widetilde{w}_k^\dagger] - S_k^0[\widehat{w}_k^\dagger]\|_2^2 &= \|S_k^0[\widehat{w}_k^\dagger]\|_2^2 - \|S_k^0[\widetilde{w}_k^\dagger]\|_2^2, \\ &= \mathbb{P}_n \widehat{S}_k^2[\widehat{w}_k^\dagger] - \|S_k^0[\widetilde{w}_k^\dagger]\|_2^2 + o_p(1), \\ &= M_n(\widehat{w}_k^\dagger) - M(\widetilde{w}_k^\dagger) + o_P(1), \end{aligned}$$

where the second equality follows from (A.36). By the definitions of $\widehat{w}_k^\dagger$ and $\widetilde{w}_k^\dagger$, we have

$$M_n(\widehat{w}_k^\dagger) - M(\widehat{w}_k^\dagger) \leq M_n(\widehat{w}_k^\dagger) - M(\widetilde{w}_k^\dagger) \leq M_n(\widetilde{w}_k^\dagger) - M(\widetilde{w}_k^\dagger).$$

Therefore, we conclude the proof of (A.31) by applying (A.36) to the above inequality.

In the end, we consider (A.32). Again, by the form of $\widetilde{I}_0(k,k')$ given in (A.28) and similar analysis in (A.31), we only need to show $\|S_k^0[\widetilde{w}_k^\dagger] - S_k^0[w_k^\dagger]\|_2 = o(1)$. By the definitions of $\widetilde{w}_k^\dagger$

and $w_k^\dagger$, we have

$$\begin{aligned}
\|S_k^0[\widetilde{w}_k^\dagger] - S_k^0[w_k^\dagger]\|_2^2 &= \inf_{w_k \in \mathcal{W}_n} E\left[\dot{\ell}_{g_0}[a_k^\dagger] - \dot{\ell}_{g_0}[a_k] + \sum_{j=1}^{d}(\dot{\ell}_{h_{j0}}[b_{jk}^\dagger] - \dot{\ell}_{h_{j0}}[b_{jk}])\right]^2 \\
&\lesssim \inf_{w_k \in \mathcal{W}_n}\left\{\|\dot{\ell}_{g_0}[a_k^\dagger] - \dot{\ell}_{g_0}[a_k]\|_2^2 + \sum_{j=1}^{d}\|\dot{\ell}_{h_{j0}}[b_{jk}^\dagger] - \dot{\ell}_{h_{j0}}[b_{jk}]\|_2^2\right\} \\
&\lesssim \inf_{a_k \in \mathcal{G}_n}\|\dot{\ell}_{g_0}[a_k^\dagger] - \dot{\ell}_{g_0}[a_k]\|_2^2 + \sum_{j=1}^{d}\inf_{b_{jk} \in \mathcal{H}_{jn}}\|\dot{\ell}_{h_{j0}}[b_{jk}^\dagger] - \dot{\ell}_{h_{j0}}[b_{jk}]\|_2^2 \\
&\lesssim \inf_{a_k \in \mathcal{G}_n}\|a_k^\dagger - a_k\|_\infty^2 + \sum_{j=1}^{d}\left\{\inf_{b_{jk} \in \mathcal{H}_{jn}}\|b_{jk}^\dagger - b_{jk}\|_\infty^2\right\},
\end{aligned}$$

where the last inequality trivially follows from the form of $\dot{\ell}_g[a]$ and $\dot{\ell}_{h_j}[b_j]$. In the proof of Theorem 2, we show that $a_k^\dagger \in H_{\widetilde{c}_0}^{r_0}[l_v, u_v]$ and $b_{jk}^\dagger \in H_{\widetilde{c}_j}^{r_j}[0,1]$. Thus, we have $\|S_k^0[\widetilde{w}_k^\dagger] - S_k^0[w_k^\dagger]\|_2 \to 0$ based on the last inequality in the above. This completes the whole proof. □

References

BANERJEE, M., BISWAS, P. AND GHOSH, D. (2006). A Semiparametric Binary Regression Model involving Monotonicity Constraints. *Scandinavian Journal of Statistics*, **33** 673–697.

BANERJEE, M., MUKHERJEE, D. AND MISHRA, S. (2009). Semiparametric Binary Regression Models under Shape Constraints with an Application to Indian Schooling Data. *Journal of Econometrics*. **149** 101-117.

BICKEL, P., KLAASSEN, C.A., RITOV, Y. AND WELLNER, J.A. (1993). Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Univ. Press

CHEN, K. AND TONG, X. (2010). Varying Coefficient Transformation Models with Censored Data. *Biometrika*, **97** 969–976.

CHEN, X. AND SHEN, X. (1998). Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica*, **66** 289–314.

CHENG, G. AND HUANG, J. (2010). Bootstrap Consistency for General Semiparametric M-estimation. *Annals of Statistics*, 38, 2884-2915.

DABROWSKA, D.M. AND DOKSUM, K.A. (1988). Partial Likelihood in Transformation Models with Censored Data. *Scandinavian Journal of Statistics*, **15** 1–23.

DABROWSKA, D.M. AND DOKSUM, K.A. (1988). Estimation and Testing in a Two-Sample Generalized Odds Rate Model. *Journal of American Statistical Association*, **83** 744–749.

DOKSUM, K.A. AND GASKO, M. (1990). On a Correspondence between Models in Binary Regression Analysis and in Survival Analysis. *Journal of American Statistical Association*, **83** 744–749.

HUANG, J. AND ROSSINI, A.J. (1997). Sieve Estimation for the Proportional-Odds Failure-Time Regression Model with Interval Censoring. *Journal of American Statistical Association*, **92** 960–967.

HUANG, J. (1999). Efficient Estimation of the Partly Linear Additive Cox Model. *Annals of Statistics*, **27** 1536–1563.

KALBFLEISCH, J.D. AND PRENTICE, R.L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, New York.

KOSOROK, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

LAM, K.F. AND XUE, H. (2005). A semiparametric regression cure model with current status data. *Biometrika*, **92**, 573-586.

MA, S. (2009). Cure model with current status data. *Statistica Sinica*, **19**, 233-249.

MA, S. AND KOSOROK, M.R. (2005a). Penalized Log-likelihood Estimator for Partly Linear Transformation Models with Current Status Data. *Annals of Statistics*, **33** 2256–2290.

MA, S. AND KOSOROK, M.R. (2005b). Robust Semiparametric M-estimation and the Weighted Bootstrap. *Journal of Multivariate Statistics*, **96**, 190-217.

RADCHENKO, P. (2008). Mixed-Rates Asymptotics. *Annals of Statistics*, **36** 287–309.

SASIENI, P. (1992). Nan-orthogonal Projections and Their Application to Calculating the Information in a Partly Linear Cox Model. *Scandinavian Journal of Statistics* **19** 215–233.

SHEN, X. (1998). Proportional Odds Regression and Sieve Maximum Likelihood Estimation. *Biometrika*, **85** 165–177.

SHEN, X. (2000). Linear Regression with Current Status Data. *Journal of American Statistical Association*, **95** 842–852.

STONE, C. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *Annals of Statistics*, **10** 1040–1053.

STONE, C. (1985). Additive Regression and Other Nonparametric Models. *Annals of Statistics*, **13** 689–705.

VAN DE GEER, S. (2000). Empirical Processes in M-Estimation. Cambridge University Press.

VAN DER VAART, A. W., AND WELLNER, J. A. (1996). Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York

XUE, H., LAM, K.F., AND LI, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association*, **99**, 346-356.

YU, A.K.F., KWAN, K.Y.W., CHAN, D.H.Y., AND FONG, D.Y.T. (2001). Clinical features of 46 eyes with calcified hydrogel intraocular lenses. *Journal of Cataract and Refractive Surgery*, **27**, 1596-1606.

ZHANG, Y., HUA, L. AND HUANG, J. (2010). A Spline-based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data, *Scandinavian Journal of Statistics* 37 338-354.