Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Estimation of Optimal Individualized Treatment Rules Using a Covariate-Specific Treatment Effect Curve With High-Dimensional Covariates

Wenchuan Guo, Xiao-Hua Zhou & Shujie Ma

To cite this article: Wenchuan Guo, Xiao-Hua Zhou & Shujie Ma (2021) Estimation of Optimal Individualized Treatment Rules Using a Covariate-Specific Treatment Effect Curve With High-Dimensional Covariates, Journal of the American Statistical Association, 116:533, 309-321, DOI: 10.1080/01621459.2020.1865167

To link to this article: https://doi.org/10.1080/01621459.2020.1865167

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 533, 309–321: Theory and Methods Special Issue on Precision Medicine and Individualized Policy Discovery
https://doi.org/10.1080/01621459.2020.1865167

Estimation of Optimal Individualized Treatment Rules Using a Covariate-Specific Treatment Effect Curve With High-Dimensional Covariates

Wenchuan Guo[a,b], Xiao-Hua Zhou[c], and Shujie Ma[a]

[a]Department of Statistics, University of California Riverside, Riverside, CA; [b]Global Biometric Sciences, Bristol-Myers Squibb, Pennington, NJ; [c]Beijing International Center for Mathematical Research, and Department of Biostatistics, Peking University, Beijing, China

ABSTRACT

With a large number of baseline covariates, we propose a new semiparametric modeling strategy for heterogeneous treatment effect estimation and individualized treatment selection, which are two major goals in personalized medicine. We achieve the first goal through estimating a covariate-specific treatment effect (CSTE) curve modeled as an unknown function of a weighted linear combination of all baseline covariates. The weight or the coefficient for each covariate is estimated by fitting a sparse semiparametric logistic single-index coefficient model. The CSTE curve is estimated by a spline-backfitted kernel procedure, which enables us to further construct a simultaneous confidence band (SCB) for the CSTE curve under a desired confidence level. Based on the SCB, we find the subgroups of patients that benefit from each treatment, so that we can make individualized treatment selection. The innovations of the proposed method are 3-fold. First, the proposed method can quantify variability associated with the estimated optimal individualized treatment rule with high-dimensional covariates. Second, the proposed method is very flexible to depict both local and global associations between the treatment and baseline covariates in the presence of high-dimensional covariates, and thus it enjoys flexibility while achieving dimensionality reduction. Third, the SCB achieves the nominal confidence level asymptotically, and it provides a uniform inferential tool in making individualized treatment decisions. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2019
Accepted September 2020

KEYWORDS
High-dimensional covariates;
Optimal treatment selection;
Personalized medicine;
Semiparametric model

1. Introduction

Personalized medicine aims to tailor medical treatments according to patient characteristics, and it has gained much attention in modern biomedical research. The success of personalized medicine crucially depends on the development of reliable statistical tools for estimating an optimal treatment regime given the data collected from clinical trials or observational studies (Kosorok and Laber 2019). In the literature, there are two general statistical approaches and their hybrids for deriving an optimal individualized treatment rule (ITR) based on clinical trial data. The first general approach targets at direct optimization of the population average outcome under an ITR. Several independent research groups have proposed a new framework of deriving the ITR that recasts the problem of maximizing treatment benefit as a weighted classification problem (Rubin and van der Laan 2012; Zhang et al. 2012; Zhao et al. 2012; Chen, Zeng, and Kosorok 2016; Zhou et al. 2017). Several semiparametric and nonparametric methods have been further proposed under this framework (Zhao et al. 2012; Huang and Fong 2014; Huang 2015; Laber and Zhao 2015; Zhu, Huang, and Zhou 2018), and they are more robust against model misspecification than the parametric approaches. Under this framework, the convergence rates of the resulting estimators were thoroughly studied in Zhao et al. (2012) and Zhang et al. (2018), among others, but it is generally difficult to build statistical inference upon the estimated ITRs (Laber and Qian 2019). To solve this problem, Jiang et al. (2019) proposed an entropy learning method, and Wager and Athey (2018) established asymptotic normality of a causal forest estimator.

The second general approach is to model the difference in average outcomes between two treatment groups conditional on predictive biomarkers that provide information about the effect of a therapeutic intervention. There are hybrids from the two approaches using mean-modeling (Zhang et al. 2012; Taylor, Cheng, and Foster 2015; Zhang and Zhang 2018; Luckett et al. 2020), but the hybrid methods also inherit the difficulty of developing inferential tools from the first general approach. In this article, we focus on the idea of the second approach. With a single predictive biomarker, several authors have proposed to plot the estimated conditional treatment difference against the biomarker's values or its percentiles (e.g., a covariate-specific treatment effect (CSTE) curve) obtained from nonparametric smoothing, along with its confidence bands, to derive an optimal ITR (Zhou and Ma 2012; Janes et al. 2014; Ma and Zhou 2014; Han, Zhou, and Liu 2017). It provides a direct and effective visual way of using the predictive biomarker for deriving an ITR, but the nonparametric smoothing method suffers

CONTACT Xiao-Hua Zhou azhou@math.pku.edu.cn Beijing International Center for Mathematical Research, and Department of Biostatistics, Peking University, Beijing 100871, China; Shujie Ma shujie.ma@ucr.edu Department of Statistics, University of California Riverside, Riverside, CA 92521.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

from "curse of dimensionality." With multiple biomarkers, Qian and Murphy (2011) considered a parametric conditional mean model that involves higher order terms, and employed a penalized method for estimation. Cai et al. (2011) proposed a two-step method, and Kang, Janes, and Huang (2014) devised a boosting iterative algorithm to reduce the bias caused by the model misspecification of a working model. Moreover, Shi et al. (2018) proposed a penalized multi-stage A-learning, and Zhu, Zeng, and Song (2019) developed a high-dimensional Q-learning method to simultaneously estimate the optimal dynamic treatment regimes and select important variables. The existing methods try to either provide a flexible modeling strategy by relaxing the restrictive structural assumption embodied in parametric models, or handle high-dimensional covariates, but not both.

In our article, we aim at developing a new method that estimates the outcome model in a flexible way as well as selecting important variables simultaneously, when a large number of baseline characteristics such as genetic variables are present. Moreover, we would like to provide a uniform inferential procedure that takes into account uncertainty associated with the estimated optimal ITR, and graphically explores heterogeneity of treatment effects based on varying levels of biomarkers. To achieve these goals, we propose a generalization of the CSTE curve suitable for high dimensional baseline covariates. The CSTE curve represents the predictive ability of covariates in evaluating whether a patient responds better to one treatment over another. It depicts the heterogeneous treatment effects on the outcome through an unknown function of covariates, and thus it can visually represent the magnitude of the predictive ability of the covariates.

We propose a penalized semiparametric modeling approach for estimating the CSTE curve and selecting variables simultaneously. The proposed semiparametric model enjoys flexibility while achieving dimensionality reduction. It is motivated by the logistic varying-coefficient model considered in Han, Zhou, and Liu (2017) for treatment selection with one covariate, and it meets the immediate needs from modern biomedical studies which can have a large number of baseline covariates. To make use of all covariates, one simple but effective way is to derive a weighted linear combination of all covariates as a summary predictor, and model the CSTE curve as an unknown function of this summary predictor. The weight or the coefficient for each covariate represents how important the covariate is for the prediction of the outcome. As a result, our proposed model involves two sets of high-dimensional coefficients fed into additive unknown functions for the two treatment groups. The development of the estimation procedure and the associated statistical properties is challenging and it needs different tools from the high-dimensional single-index model (Radchenko 2015). Moreover, we establish different convergence rates for the estimators of the high-dimensional coefficients and the estimators for the unknown additive functions, respectively. This new theoretical result makes it possible to further construct a simultaneous confidence band (SCB) for the CSTE curve based on the asymptotic extreme value distribution of a spline-backfitted kernel estimator (Wang and Yang 2007; Liu, Yang, and Härdle 2013; Zheng et al. 2016) for the unknown function of interest, while Radchenko (2015) and other related works only provide a nonseparable convergence rate for the estimators of the coefficients and the unknown function in single-index models. Based on the SCB, we identify the subgroups of patients that benefit from each treatment, and the proposed method is flexible enough to depict both local and global associations between the treatment and baseline covariates.

The rest of the article is organized as follows. In Section 2, we introduce the CSTE curve and the proposed semiparametric logistic single-index coefficient model. Section 3 presents the estimation procedure and the asymptotic properties of the proposed estimators. Section 4 illustrates the application of the CSTE curves and the SCBs for treatment selection. In Section 5, we evaluate the finite sample properties of the proposed method via simulation studies, while Section 6 illustrates the usefulness of the proposed method through the analysis of a real data example. A discussion is given in Section 7. All technical proofs are relegated to the online supplementary materials.

2. Methodology

We consider a sample of n subjects, a binary treatment, denoted by $Z_i = 1$ if the subject i is assigned to treatment and $Z_i = 0$ otherwise, a p-dimensional vector of covariates, denoted by X_i, and binary-valued outcomes, denoted by Y_i. Let $\left(Y_i, Z_i, X_i^\top\right), i = 1, \ldots, n$, be independent and identically distributed (iid) copies of (Y, Z, X). The goal is to estimate the optimal treatment regime using the observed data. We consider the CSTE curve given in Han, Zhou, and Liu (2017), which has the following form:

$$\mathrm{CSTE}(X) = \mathrm{logit}(E(Y(1)|X)) - \mathrm{logit}(E(Y(0)|X)),$$

where $\mathrm{logit}(u) = \log(u) - \log(1-u)$, and $Y(1)$ and $Y(0)$ denote the potential outcomes if the active treatment and the control treatment are received, respectively (Rubin 2005). Moreover, $Y \equiv ZY(1) + (1 - Z)Y(0)$. Under the unconfoundeness assumption such that $(Y(0), Y(1)) \perp Z|X$, the CSTE curve can be re-expressed as

$$\mathrm{CSTE}(X) = \mathrm{logit}(E(Y|X, Z = 1)) - \mathrm{logit}(E(Y|X, Z = 0)).$$

Denoting $\mu(X, Z) = E(Y|X, Z)$, we model the logarithm of odds ratio as

$$\mathrm{logit}(\mu(X, Z)) = g_1(X^\top \beta_1) \cdot Z + g_2(X^\top \beta_2), \qquad (1)$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are unknown single-valued functions of p variables, $\beta_1 = (\beta_{11}, \ldots, \beta_{1p})$ and $\beta_2 = (\beta_{21}, \ldots, \beta_{2p})$ are two p-vectors of unknown parameters. We see

$$\mathrm{CSTE}(X) = \mathrm{logit}(\mu(X, 1)) - \mathrm{logit}(\mu(X, 0)) = g_1(X^\top \beta_1).$$

The two sets of high-dimensional coefficients and the two unknown functions in (1) are chosen to simultaneously maximize the log-likelihood function of the binomial distribution. The proposed model is a flexible semiparametric model and is robust against model misspecification. We call it sparse logistic single index coefficient model (SLSICM). Our SLSICM contains the varying coefficient (VC) model considered in Han, Zhou, and Liu (2017) as a special case. We have the same form when

$p = 1$. As an extension of the VC model, the varying-index coefficient model considered in Ma and Song (2015) can be applied to the cases with several covariates and continuous responses. Moreover, an important and related work, Song et al. (2017), proposed a semiparametric model in which g_1 has a single-index structure and g_2 is a pure nonparametric function of X. This model suffers from the curse of dimensionality when the number of covariates is large. In our article, we allow the number of covariates to be much larger than sample size. As a result, our proposed SLSICM involves two sets of high-dimensional unknown coefficients built into additive unknown functions with a nonlinear link function, and it is considered as a hybrid of the high-dimensional single index model (HSIM) and the nonparametric additive model. Developing the estimation procedure and the statistical theories for SLSICM is challenging, and it needs different tools from HSIM. It is worth noting that estimation of HSIM has been studied in the past several years, and most existing works use a sliced inverse regression approach to estimate the index coefficients in HSIM under a linearity condition on the covariates; see Jiang and Liu (2014) and Neykov, Liu, and Cai (2016), among others. This method is not applicable to our setting, and it does not directly estimate the unknown function. Radchenko (2015) proposed to estimate the coefficients and the unknown function in HSIM jointly, but they only provide a nonseparable convergence rate for the resulting estimators of the coefficients and the unknown function. In our proposed SLSICM, we establish different convergence rates for the estimators of the coefficients β_1 and β_2 and the estimators of the unknown functions g_1 and g_2, respectively. This new theoretical result makes it possible to further construct the asymptotic SCB based on the asymptotic extreme value distribution of a spline-backfitted kernel estimator for the CSTE curve.

Unlike parametric models, the parameter vectors β_k are not identified without further assumptions. For the purpose of model identification, we assume that β_k for $k = 1, 2$ belong to the parameter space $\Theta = \{\beta = (\beta_1^\top, \beta_2^\top)^\top : \|\beta_k\| = 1, \beta_{k1} > 0, \beta_k \in R^p, k = 1, 2\}$, where $\|\cdot\|$ denotes the l_2 norm of a vector. This restriction together with Assumption 2 given in Section 3.2 will make the model (1) identifiable. We will provide a formal proof of the model identifiability in Section 1.1 of the supplementary materials. Based on the constraint that $\|\beta_k\| = 1$, we eliminate the first component in β_k and obtain the resulting parameter space: $\Theta_{-1} = \{\beta_{k,-1} = (\beta_{k2}, \ldots, \beta_{kp})^\top : \sum_{j>1} \beta_{kj}^2 < 1, k = 1, 2\}$. Let $\beta_{k1} = \sqrt{1 - \sum_{j>1} \beta_{kj}^2}$. The derivative with respect to the coefficients $(\beta_{k2}, \ldots, \beta_{kp})^\top$ can be easily obtained using the chain rule under the above parameter space. For high-dimensional problems (with a large number of covariates), p can be much larger than n but only a small number of covariates are important or relevant for treatment selection. To this end, we assume that the number of nonzero elements increases as n increases, but it is much smaller than n . Without loss of generality, we assume that only the first $s_k = s_{kn}$ components of β_k are nonzeros, that is, we can write the true values as $\beta_k = (\beta_{k1}, \ldots, \beta_{ks_k}, 0, \ldots, 0)^\top$. For model identifiability, we require one component in β_k to be nonzero. Without loss of generality, we let the first component β_{k1} to be nonzero.

3. Estimation and Theory

3.1. Algorithm

In this subsection, we discuss the estimation of model (1). We minimize the negative log-likelihood function simultaneously with respect to the parameters β_k and the functions g_k for $(k = 1, 2)$. The SLSICM has the form $\text{logit}(\mu(X, Z)) = g_1(X^\top \beta_1) \cdot Z + g_2(X^\top \beta_2)$. Therefore, we seek the minimizer of the following negative log-likelihood function given (X_i, Y_i, Z_i), $i = 1, \ldots, n$:

$$\frac{1}{n}\sum_{i=1}^{n} \log\{1 + \exp(g_1(X_i^\top \beta_1)Z_i + g_2(X_i^\top \beta_2))\} - \frac{1}{n}\sum_{i=1}^{n} Y_i\{g_1(X_i^\top \beta_1)Z_i + g_2(X_i^\top \beta_2)\}. \quad (2)$$

To overcome the problem of high-dimensional covariates, we exploit the sparsity through parameter regularization. With the sparsity constraint of β_k's, we minimize the following penalized negative log-likelihood

$$\frac{1}{n}\sum_{i=1}^{n} \log\{1 + \exp(g_1(X_i^\top \beta_1)Z_i + g_2(X_i^\top \beta_2))\} - \frac{1}{n}\sum_{i=1}^{n} Y_i\{g_1(X_i^\top \beta_1)Z_i + g_2(X_i^\top \beta_2)\} + \sum_{k=1}^{2}\sum_{j=2}^{p} p(\beta_{kj}, \lambda), \quad (3)$$

where $p(\cdot)$ is a penalty function with a tuning parameter λ that controls the level of sparsity in β_k, $k = 1, 2$.

The functions $g_k(\cdot)$, $k = 1, 2$, are unspecified and are estimated using B-splines regression. Next, we introduce the B-splines that will be used to approximate the unknown functions. For $k = 1, 2$, we assume the support of $g_k(\cdot)$ is $[\inf_X(X^\top \beta_k), \sup_X(X^\top \beta_k)] = [a_k, b_k]$. Let $a_k = t_{0,0} < t_{1,k} < \cdots < t_{N_k,k} < b_k = t_{N_k+1}$ be an equally spaced partition of $[a_k, b_k]$, called interior knots. Then, $[a_k, b_k]$ is divided into subintervals $I_{\ell,k} = [t_{\ell,k}, t_{\ell+1,k})$, $0 \le \ell \le N_k - 1$ and $I_{N_k} = [t_{N_k}, t_{N_k+1}]$, satisfying $\max_{0\le\ell\le N_k} |t_{\ell+1,k} - t_{\ell,l}| / \min_{0\le\ell\le N_k} |t_{\ell+1,k} - t_{\ell,k}| \le M$ uniformly in n for some constant $0 < M < \infty$, where $N_k \equiv N_{k,n}$ increases with the sample size n . We write the normalized B spline basis of this space (de Boor 2001) as $B_k(u_k) = \{B_{\ell,k}(u_k) : 1 \le \ell \le L_{n,k}\}^\top$, where the number of spline basis functions is $L_{n,k} = L_k = N_k + q_k$, and q_k is the spline order. For computational convenience, we let $N_k = N$ and $q_k = q$ so that $L_k = L$. In practice, cubic splines with order $q = 4$ are often used. By the result in de Boor (2001), the nonparametric function can be approximated well by a spline function such that $g_k(X^\top \beta_k) \approx B_k(X^\top \beta_k)^\top \delta_k$ for some $\delta_k \in R^L$, $k = 1, 2$. For notational simplicity, we write $B_k(X^\top \beta_k)$ as $B(X^\top \beta_k)$. Therefore, the estimates of the unknown index parameters β_k and the spline coefficients δ_k are the minimizers of

$$\frac{1}{n}\sum_{i=1}^{n} \log\{1 + \exp(B_1(X_i^\top \beta_1)^\top \delta_1 Z_i + B_2(X_i^\top \beta_2)^\top \delta_2)\}$$

$$-\frac{1}{n}\sum_{i=1}^{n} Y_i\{B_1(X_i^\top\beta_1)^\top\delta_1 z_i + B_2(X_i^\top\beta_2)^\top\delta_2\} + \sum_{k=1}^{2}\sum_{j=2}^{p} p(\beta_{kj},\lambda). \quad (4)$$

Denote $U_k = X^\top\beta_k$ for $k = 1,2$. For notational simplicity, we use U to denote U_k by suppressing k. To allow for possibly unbounded support $\mathcal{U}$ of U, according to the method given in Chen and Christensen (2015), we can weigh the B-spline basis functions by a sequence of nonnegative weighting functions $w_n : \mathcal{U} \to \{0,1\}$ given by $w_n(u) = 1$ if $u \in \mathcal{D}_n$ and $w_n(u) = 0$ otherwise, where $\mathcal{D}_n \subseteq \mathcal{U}$ is compact, convex and has nonempty interior and $\mathcal{D}_n \subseteq \mathcal{D}_{n+1}$ for all n. For the choices of $\mathcal{D}_n$, we refer to Chen and Christensen (2015) for the discussions. When $\mathcal{U}$ is compact, we simply set $w_n(u) = 1$ for all $u \in \mathcal{U}$. Then we obtain a set of new B-spline basis functions: $B_k^w(u_k) = \{B_{\ell,k}(u_k)w_n(u_k) : 1 \le \ell \le L_{n,k}\}^\top$, and replace $B_k(u_k)$ by $B_k^w(u_k)$ in the above minimization problem for obtaining the estimators $\widehat{\beta}_k$ and $\widehat{\delta}_k$. As given in Chen and Christensen (2015), the asymptotic properties including consistency and asymptotic distribution for the estimator of the unknown function $g_1(u)$ still hold for $u \in \mathcal{D}_n$.

In principle, we would like to obtain the estimates of β_k and δ_k by minimizing the penalized negative log-likelihood given in (4). In practice, this minimization is different to achieve, and thus an iterative algorithm (Watson and Engle 1983) is often applied. To this end, we obtain the estimates of $\widehat{\beta}_k$ and $\widehat{\delta}_k$ through an iterative algorithm described as follows, although our theoretical properties given in Section 3.2 are established for the estimators which minimize (4).

Step 1: Given β_k, the solution of δ_k is easily obtained. Reexpressing the model gives

$$\begin{aligned}\text{logit}(\mu(X,Z)) &\approx B(X^\top\beta_k)^\top\delta_1 Z + B(X^\top\beta_k)^\top\delta_2 \\ &= \left(ZB(X^\top\beta_k)^\top, B(X^\top\beta_k)^\top\right)\begin{pmatrix}\delta_1\\ \delta_2\end{pmatrix}.\end{aligned}$$

This can be viewed as a logistic regression model using $\left(ZB(X^\top\beta_k)^\top, B(X^\top\beta_k)^\top\right)^\top$ as the regressors without intercept term.

Step 2: Given δ_k, it remains to find the solution that minimizes (3) with respect to β_k. Let β_k^{old} and $\beta_{k,-1}^{\text{old}}$ be the current estimates for β_k and $\beta_{k,-1}$, respectively. Let $\widetilde{g}_k(X^\top\beta_k) = B_k(X^\top\beta_k)^\top\delta_k$. We approximate

$$\widetilde{g}_k(X^\top\beta_k) \approx \widetilde{g}_k(X^\top\beta_k^{\text{old}}) + \widetilde{g}_k'(X^\top\beta_k^{\text{old}})X^\top J(\beta_k^{\text{old}})(\beta_{k,-1} - \beta_{k,-1}^{\text{old}}),$$

where $J(\beta_k) = \partial\beta_k/\partial\beta_{k,-1} = (-\beta_{k,-1}/\sqrt{1-\|\beta_{k,-1}\|_2^2}, I_{p-1})^\top$ is the Jacobian matrix of size p by $p-1$. To obtain the sparse estimates of β_k, we carry out a regularized logistic regression with $\left[\{Z\widetilde{g}_1'(X^\top\beta_1^{\text{old}})J(\beta_1^{\text{old}})^\top X\}^\top, \{\widetilde{g}_2'(X^\top\beta_2^{\text{old}})J(\beta_2^{\text{old}})^\top X\}^\top\right]^\top$ as the regressors with a known intercept term given as

$$\begin{aligned}&Z\widetilde{g}_1(X^\top\beta_1^{\text{old}}) + \widetilde{g}_2(X^\top\beta_2^{\text{old}}) - Z\widetilde{g}_1'(X^\top\beta_1^{\text{old}})X^\top J(\beta_1^{\text{old}})\beta_{1,-1}^{\text{old}} \\ &\quad - \widetilde{g}_2'(X^\top\beta_2^{\text{old}})X^\top J(\beta_2^{\text{old}})\beta_{2,-1}^{\text{old}}.\end{aligned}$$

This produces an updated vector $\beta_{k,-1}^{\text{new}}$. Then we set $\beta_k^{\text{new}} = (\sqrt{1-\|\beta_{k,-1}^{\text{new}}\|^2}, (\beta_{k,-1}^{\text{new}})^\top)^\top$, for $k = 1,2$. Steps 1 and 2 are repeated until convergence.

We obtain the initial value of β_k through fitting a regularized logistic regression by assuming that $g_k(X^\top\beta_k) = X^\top\beta_k$. In Step 2, we use the coordinate descent algorithm (Breheny and Huang 2011) to fit the regularized regression. Moreover, we choose to use the nonconvex penalties such as MCP and SCAD which induce nearly unbiased estimators. The MCP (Zhang 2010) has the form $p_\gamma(t,\lambda) = \lambda\int_0^t(1 - x/(\gamma\lambda))_+dx, \gamma > 1$ and the SCAD (Fan and Li 2001) penalty is $p_\gamma(t,\lambda) = \lambda\int_0^t \min\{1, (\gamma - x/\lambda)_+/(\gamma-1)\}dx, \gamma > 2$, where γ is a parameter that controls the concavity of the penalty functions. In particular, both penalties converge to the L_1 penalty as $\gamma \to \infty$. We put γ in the subscript to indicate the dependence of these penalty functions on it. In practice, we treat γ as a fixed constant. The B-spline basis functions and their derivatives are calculated using the `bsplineS` function in R package `fda`.

From the above algorithm, we obtain the spline estimators of the functions $g_1(\cdot)$ and $g_2(\cdot)$. However, the spline estimator only has convergence rates but its asymptotic distribution is not available in the additive model settings with multiple unknown functions (Stone 1985), so no measures of confidence can be assigned to the estimators for conducting statistical inference (Wang and Yang 2007). The spline-backfitted kernel (SBK) estimator is designed to overcome this issue for generalized additive models (Liu, Yang, and Härdle 2013), which combines the strengths of kernel and spline smoothing, is easy to implement and has asymptotic distributions. Denote the SBK estimator of $g_1(\cdot)$ as $\widehat{g}_{1,\text{sbk}}(\cdot)$. We obtain $\widehat{g}_{1,\text{sbk}}(X_0^\top\widehat{\beta}_1)$ for a new input vector X_0 as follow:

Step 3: Given the spline estimate $\widehat{g}_2(X_i^\top\widehat{\beta}_2)$, the loss criterion for a local linear logistic regression can be expressed as the following negative quasi-likelihood function:

$$\begin{aligned}l(a,b,X) = -\frac{1}{n}\sum_i [Y_i(aZ_i + \widehat{g}_2(X_i^\top\widehat{\beta}_2)) \\ - \log(1+\exp(aZ_i + \widehat{g}_2(X_i^\top\widehat{\beta}_2)))]K_h(X_i^\top\widehat{\beta}_1 - X^\top\widehat{\beta}_1),\end{aligned}$$

where $K_h(\cdot)$ is a kernel function with bandwidth h. We obtain the estimate $\widehat{a}(X_0)$ by minimizing the above loss function. Then the predicted CSTE value at X_0 is

$$\text{CSTE}(X_0) \approx \widehat{g}_{1,\text{sbk}}(X_0^\top\widehat{\beta}_1) = \widehat{a}(X_0). \quad (5)$$

Based on the above SBK estimator of CSTE, we can construct a SCB which is used for optimal treatment selection. The details will be discussed in Section 4.

3.2. Asymptotic Analysis

For any positive sequences $\{a_n\}$ and $\{b_n\}$, let $a_n \asymp b_n$ denote $\lim_{n\to\infty} a_n b_n^{-1} = C$ for a constant $0 < C < \infty$ and $a_n \ll b_n$ denote $\lim_{n\to\infty} a_n b_n^{-1} = 0$. For a vector $a = (a_1,\ldots,a_p)^\top \in R^p$, denote $\|a\| = (\sum_{l=1}^p a_l^2)^{1/2}$ and $\|a\|_\infty = \max_l |a_l|$. For a matrix $A = (A_{ij})$, denote $\|A\| = \max_{\|\zeta\|=1}\|A\zeta\|$, $\|A\|_\infty = \max_i \sum_j |A_{ij}|$, and $\|A\|_{2,\infty} = \max_i ||A_i||$, where A_i is the ith row. For a symmetric matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of A, respectively. We assume

that the nonsparsity size $s = \max(s_1, s_2) \ll n$ and the dimensionality satisfies $\log p = O(n^\alpha)$ for some $\alpha \in (0,1)$. Denote $\beta_{k1} = (\beta_{k1}, \ldots, \beta_{ks_k})^\top, \beta_{k1,-1} = (\beta_{k2}, \ldots, \beta_{ks_k})^\top, \beta_{k2} = (\beta_{k(s_k+1)}, \ldots, \beta_{kp})^\top$. Then we write $\beta_{(1)} = (\beta_{11}^\top, \beta_{21}^\top)^\top, \beta_{(2)} = (\beta_{12}^\top, \beta_{22}^\top)^\top, \beta_{(1),-1} = (\beta_{11,-1}^\top, \beta_{21,-1}^\top)^\top$. Denote the Jacobian matrix as $J(\beta_{k1}) = \partial\beta_{k1}/\beta_{k1,-1}, k = 1,2$, and $J(\beta_{(1)}) = \partial\beta_{(1)}/\partial\beta_{(1),-1} = \text{diag}(J(\beta_{11}), J(\beta_{21}))$, which is a block diagonal matrix. We use the superscript "0" to represent the true values.

Denote the first $s_k (1 \le k \le 2)$ components of X_i as $X_{i,k1} = (x_{ij}, 1 \le j \le s_k)^\top$ and the last $p - s_k$ components as $X_{i,k2} = (x_{ij}, s_k < j \le p)^\top$. Denote $S(x) = (1 + e^{-x})^{-1}$ as the sigmoid function. Then the true expected value of the response given (X, Z) is

$$E(Y|X,Z) = \mu(X,Z) = S(g_1(X^\top\beta_1)Z + g_2(X^\top\beta_2)).$$

Define the space $\mathcal{M}$ as a collection of functions with finite L_2 norm on $\mathscr{C} \times \{0,1\}$ by

$$\mathcal{M} = \left\{ g(x,z) = g_1(x^\top\beta_1^0)z + g_2(x^\top\beta_2^0), E\{g_k(X^\top\beta_k^0)\}^2 < \infty \right\}.$$

For a given random variable U, define its projection onto the space $\mathcal{M}$ as

$$\mathbb{P}_{\mathcal{M}}(U) = \arg\min_{g\in\mathcal{M}} E\left[w^0\{U - g(X,Z)\}^2\right],$$

where $w^0 = \pi^0(1-\pi^0)$ and $\pi^0 = \mu(X,Z)$. For a vector $U = \{U_1, \ldots, U_d\}$, let

$$\mathbb{P}_{\mathcal{M}}(U) = \{\mathbb{P}_{\mathcal{M}}(U_1), \ldots, \mathbb{P}_{\mathcal{M}}(U_d)\}^\top.$$

Moreover, we define

$$\Omega_n = \frac{1}{n}\sum_{i=1}^n J(\beta_{(1)}^0)^\top \begin{pmatrix} g_1'(X_{i,11}^\top\beta_{11}^0)\widetilde{X}_{i,11}Z_i \\ g_2'(X_{i,21}^\top\beta_{21}^0)\widetilde{X}_{i,12} \end{pmatrix} \begin{pmatrix} g_1'(X_{i,11}^\top\beta_{11}^0)\widetilde{X}_{i,11}Z_i \\ g_2'(X_{i,21}^\top\beta_{21}^0)\widetilde{X}_{i,12} \end{pmatrix}^\top J(\beta_{(1)}^0), \tag{6}$$

where $\widetilde{X}_{i,kv} = X_{i,kv} - \mathbb{P}_{\mathcal{M}}(X_{i,kv})$ for $k, v = 1, 2$, and

$$\Phi_n = \frac{1}{n}\sum_{i=1}^n \sigma^2(X_i,Z_i) J(\beta_{(1)}^0)^\top \begin{pmatrix} g_1'(X_{i,11}^\top\beta_{11}^0)\widetilde{X}_{i,11}Z_i \\ g_2'(X_{i,21}^\top\beta_{21}^0)\widetilde{X}_{i,21} \end{pmatrix} \begin{pmatrix} g_1'(X_{i,11}^\top\beta_{11}^0)\widetilde{X}_{i,11}Z_i \\ g_2'(X_{i,21}^\top\beta_{21}^0)\widetilde{X}_{i,21} \end{pmatrix}^\top J(\beta_{(1)}^0),$$

where $\sigma^2(X,Z) = E[\{Y - S(g(X,Z))\}^2|X,Z]$.

To establish asymptotic properties, we need the following regularity conditions.

Assumption 1. The penalty function $p_\gamma(t,\lambda)$ is a nondecreasing symmetric function and concave on $[0,\infty)$. For some constant $a > 0$, $\rho(t) = \lambda^{-1}p_\gamma(t,\lambda)$ is a constant for all $t \ge a\lambda$ and $\rho(0) = 0$. $\rho'(t)$ exists and is continuous except for a finite number of t and $\rho'(0+) = 1$.

Assumption 2. For $k = 1,2$, for any given $\beta_k \in \Theta$, g_k is a nonconstant function on $\{x \in \mathcal{X}\}$, where $\mathcal{X}$ is compact, convex and has nonempty interior, and $g_k \in \mathcal{H}_r$ for some $r > 1$, where $\mathcal{H}_r$ is the collection of all functions such that the qth order derivative satisfies the Hölder condition of order γ with $r \equiv q + \gamma$ and $q \ge 1$, that is, for any $\phi \in \mathcal{H}_r$, there is a $C_0 \in (0,\infty)$ such that $|\phi^{(q)}(u_1) - \phi^{(q)}(u_2)| \le C_0|u_1 - u_2|^\gamma$ for all u_1 and u_2 in the support of g_k.

Assumption 3. For $1 \le j \le s_k$, $E(X_j^{2+2(\varkappa+1)}) \le C_{\varkappa,k}$ for some constant $\varkappa \in (0,1)$. For $s_k \le j \le p - s_k$, $E(|X_j|^{2+\varrho}) \le C_{\varrho,k}$, for some $\varrho > (8/3)(1-\alpha)^{-1} - 2$ and $\varrho \ge 2$.

Assumption 4. There exist $c, c', c_k \in (0,\infty)$ such that $\lambda_{\min}(\Omega_n) \ge c$, $\lambda_{\min}(\Phi_n) \ge c'$ almost surely, and $\|E(\widetilde{X}_{k2}\widetilde{X}_{k1}^\top)\|_{2,\infty} \le c_k$, where $\widetilde{X}_{kv} = X_{kv} - \mathbb{P}_{\mathcal{M}}(X_{kv})$, for $k, v = 1, 2$.

Assumption 5. Let $w_n = 2^{-1}\min\{|\beta_{kj}^0| : 2 \le j \le s, k = 1,2\}$. Assume that $\max((s/n)^{1/2}, L^{1-r} + L^{3/2}n^{\alpha/2-1/2}) \ll \lambda \ll w_n$.

Assumption 1 is a typical condition on the penalty function, see Fan and Lv (2011). The concave penalties such as SCAD and MCP satisfy Assumption 1. Assumption 2 is a typical smoothness condition on the unknown nonparametric function, see, for instance, Condition (C3) in Ma and He (2016). Moreover, we require that the unknown functions be nonconstant functions to identify the index parameters β_k. Assumption 3 is required for the covariates, see Condition (A5) in Ma and Yang (2011). Moreover, the design matrix needs to satisfy Assumption 4. A similar condition can be found in Fan and Lv (2011). Assumption 5 assumes that half of the minimum nonzero signal in β_k^0 is bounded by some thresholding value, which is allowed to go to zero as $n \to \infty$. This assumption is needed for variable selection consistency established in Theorem 1.

Denote $\beta_{-1} = (\beta_{11,-1}^\top, \beta_{12}^\top, \beta_{21,-1}^\top, \beta_{22}^\top)^\top$. Let $s = \max(s_1, s_2)$. Theorem 1 establishes the consistency for the parameters in model (1).

Theorem 1. Under Assumptions (A1)–(A5), and $\alpha \in (0, (2r-5/2)/(2r-1))$, $L^{1-r}s^{1/2} = o(1)$, $n^{\alpha-1}(L^3 + s) = o(1)$, $n^{1/2}L^{1-2r} = o(1)$, $\log(n)(s^{1/2} + L^{1/2})L^{3/2}n^{-1/2} = o(1)$, there exists a strict local minimizer $\widehat{\beta}_{-1} = (\widehat{\beta}_{11,-1}^\top, \widehat{\beta}_{12}^\top, \widehat{\beta}_{21,-1}^\top, \widehat{\beta}_{22}^\top)^\top$ of the loss function given in (3) such that $\widehat{\beta}_{k2} = 0$ for $1 \le k \le 2$ with probability approaching 1 as $n \to \infty$, and $\|\widehat{\beta}_{-1} - \beta_{-1}^0\| = O_p(\sqrt{s/n})$.

Remark. Based on the assumption given in Theorem 1, the number of spline basis functions needs to satisfy $n^{1/\{2(2r-1)\}} \ll L \ll \min\{n^{1/4}\{\log(n)\}^{-1}, n^{(1-\alpha)/3}\}$.

The following Theorem presents the convergence rate for the spline estimator of the unknown functions.

Theorem 2. Under conditions given in Theorem 1, we have $n^{-1}\sum_{i=1}^n\{\widehat{g}_k(X_i^\top\widehat{\beta}_k) - g_k(X_i^\top\beta_k^0)\}^2 = O_p(L^{-2r} + L/n + s/n)$, for $k = 1, 2$.

To estimate the SCB, we need the following assumptions, see Assumptions (A4) and (A6) in Zheng et al. (2016).

Assumption 6. Let $r = 2$. The kernel function K is symmetric probability density function supported on $[-1,1]$ and has bounded derivative. The bandwidth h satisfies $h = h_n = o(n^{-1/5}(\log n)^{-1/5})$ and $h^{-1} = O(n^{1/5}(\log n)^\delta)$ for some constant $\delta > 1/5$.

Assumption 7. The joint density of $X^\top\beta_1^0$ and $X^\top\beta_2^0$ is a bounded and continuous function. The marginal probability density functions have continuous derivatives and the same bound as the joint density. When $\beta_1^0 = \beta_2^0 = \beta^0$, the probability density function of $X^\top\beta^0$ is bounded, continuous, and has continuous derivative.

We borrow some notations in Zheng et al. (2016):

$$\sigma^2(u) = E[\mu(u,Z)(1-\mu(u,Z))|Z=1], D(u) = f(u)\sigma^2(u),$$
$$v^2(u) = ||K||_2^2 f(u)\sigma^2(u), \tag{7}$$

where $\mu(u,Z) = S(g_1(u)Z + g_2(X^\top\beta_2^0))$ and $f(u)$ is the density function of $X^\top\beta_1^0$. Define the quantile function

$$Q_h(\alpha) = a_h + a_h^{-1}[\log(\sqrt{C_K}/(2\pi)) - \log(-\log\sqrt{1-\alpha})]$$

for any $\alpha \in (0,1)$, where $a_h = \sqrt{-2\log h}$ and $C_K = ||K'||_2^2/||K||_2^2$. Without loss of generality, assume $x^\top\beta_1^0$ is in the range $[0,1]$ and let $\mathscr{C}$ be a set of x such that $\mathscr{C} = \{x : x^\top\beta_1^0 \in [h, 1-h], x \in R^p\}$. Theorem 3 is an adaptation from Theorem 1 in Zheng et al. (2016). It provides a method to construct the SCB for g_1.

Theorem 3. Under Assumptions 1–7, and $\alpha \in (0, 2/5)$, $sn^{-1/10}(\log n)^{-3/5} = O(1)$ and $n^{1/5} \ll L \ll \min\{n^{1/4}\{\log(n)\}^{-1}, n^{(1-\alpha)/3}\}$, we have

$$\lim_{n\to\infty} P\left(\sup_{x\in\mathscr{C}} \left|\frac{\widehat{g}_{1,\text{sbk}}(x^\top\widehat{\beta}_1) - g_1(x^\top\beta_1^0)}{\sigma_n(x^\top\widehat{\beta}_1)}\right| \le Q_h(\alpha)\right) = 1-\alpha,$$

where $\sigma_n(x^\top\beta_1) = n^{-0.5}h^{-0.5}v(x^\top\beta_1)/D(x^\top\beta_1)$.

Remark 1. Based on the proof for Theorem 3, we can readily obtain that $\sigma_n(x^\top\widehat{\beta}_1)^{-1}\{\widehat{g}_{1,\text{sbk}}(x^\top\widehat{\beta}_1) - g_1(x^\top\beta_1^0)\} \to \mathcal{N}(0,1)$ in distribution, for given $x \in \mathscr{C}$. Thus, the $100(1-\alpha)\%$ pointwise confidence interval for $g_1(x^\top\beta_1^0)$ is $\widehat{g}_{1,\text{sbk}}(x^\top\widehat{\beta}_1) \pm Z_{1-\alpha/2}\sigma_n(x^\top\widehat{\beta}_1)$, where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)100\%$ quantile of the standard normal distribution.

Remark 2. For details of implementations of kernel and spline estimation for functions in (7), we refer to Zheng et al. (2016). As suggested in Zheng et al. (2016), we use a data-driven under smoothing bandwidth $h = h_{\text{opt}}(\log n)^{-1/4}$ and h_{opt} is given in Zheng et al. (2016). We let the number of spline interior knots be $\lfloor n^{1/5}(\log n)\rfloor + 1$. The $100(1-\alpha)\%$ SCB for $g_1(x^\top\beta_1^0)$ is $\widehat{g}_{1,\text{sbk}}(x^\top\widehat{\beta}_1) \pm \sigma_n(x^\top\widehat{\beta}_1)Q_h(\alpha)$.

Remark 3. From Theorem 3, we see that the width of the SCB has the order of $\sqrt{\log(h)n/h}$. The width of the pointwise confidence interval given in Remark 1 has the order of $\sqrt{n/h}$. Based on the assumption for the bandwidth h, the SCB is wider than the pointwsie confidence interval asymptotically. In deriving the asymptotic distribution of the SBK estimator for the unknown function, it also involves the estimation error from the parameter estimation, which is given in Theorem 1. We assume that s increases with n in a slow rate, so that this estimation error is negligible in the construction of the asymptotic SCB. Our interest focuses on constructing the SCB for the CSTE curve, based on which we make treatment recommendations to a group of new patients. For conducting inference for the parameters, it can be a future interesting research topic to explore.

The theoretical result in Theorem 3 is obtained from the asymptotic extreme value distribution of the SBK estimator for the CSTE curve. This is the key result for obtaining our asymptotic SCB, which achieves the nominal confidence level asymptotically. It is worth noting that we provide different convergence rates for the estimators of the high-dimensional coefficients and the estimators for the unknown functions in Theorems 1 and 2, respectively. This theoretical result enables us to further derive the asymptotic distribution of $\widehat{g}_{1,\text{sbk}}$, when the number of important covariates satisfies certain order given in Theorem 3. The SCB provides an important uniform inferential tool for identifying subgroups of patients that benefit from each treatment and make treatment recommendations to the subgroups, while the pointwise confidence interval can only provide treatment selection for a given patient.

4. Treatment Selection via Confidence Bands

In this section, attention is focused on making individualized treatment decision rule for patients. We provide an example to illustrate how to select the optimal treatment based on the CSTE curve and its SCBs. The goal of treatment selection is to find which group of patients will benefit from new treatment based on their covariates. By the definition of CSTE curve, if we assume that the outcome of interest is death, the CSTE curve is the odds ratio of the treatment effect in reducing the probability of death. That is, a positive CSTE(X) value means that the patients will not benefit from new treatment since they may have a higher death rate than patients who receive old treatment. We define the cutoff points as the places where the upper and lower confidence bands equal to 0. Based on these cutoff points, we are able to identify the regions with the positive and negative values of CSTE(X), respectively, so that it will guide us to select the best treatment for a future patient. To summarize, this treatment selection method consists of the following steps:

Step 1. Use the existing dataset to obtain the SBK estimator $\widehat{g}_{1,\text{sbk}}(x^\top\widehat{\beta}_1)$ of the CSTE curve and the corresponding SCBs $\widehat{g}_{1,\text{sbk}}(x^\top\widehat{\beta}_1) \pm \sigma_n(x^\top\widehat{\beta}_1)Q_h(\alpha)$ at a given confidence level $100(1-\alpha)\%$.

Step 2. Identify the cutoff points for $x^\top\widehat{\beta}_1$ and the regions of positive and negative values of the constructed SCBs, as illustrated in Figure 1.

Step 3. For a new patient, we first calculate $\tilde{x}^\top\widehat{\beta}_1$, where $\tilde{x}$ is the observed value of the covariates for this new patient. Then we make treatment recommendation for this patient based on what region the value of $x^\top\widehat{\beta}$ falls into.

We use the following example to illustrate the method of using the SCBs to select optimal treatment for patients. We assume that $X^\top\hat{\beta}_1$ has a range of $(-4, 2)$. In Figure 1, the solid line represents a CSTE curve and dashed lines above and below the curve are the corresponding 95% SCBs. We assume that the outcome variable Y is the indicator of death. As shown in Figure 1, the CSTE is decreasing when the value of $X^\top\hat{\beta}_1$ is from -4 to -1.6 and it is increasing when the value of $X^\top\hat{\beta}_1$ is from -1.6 to 2. In general, when the $X^\top\hat{\beta}_1$ value of a patient

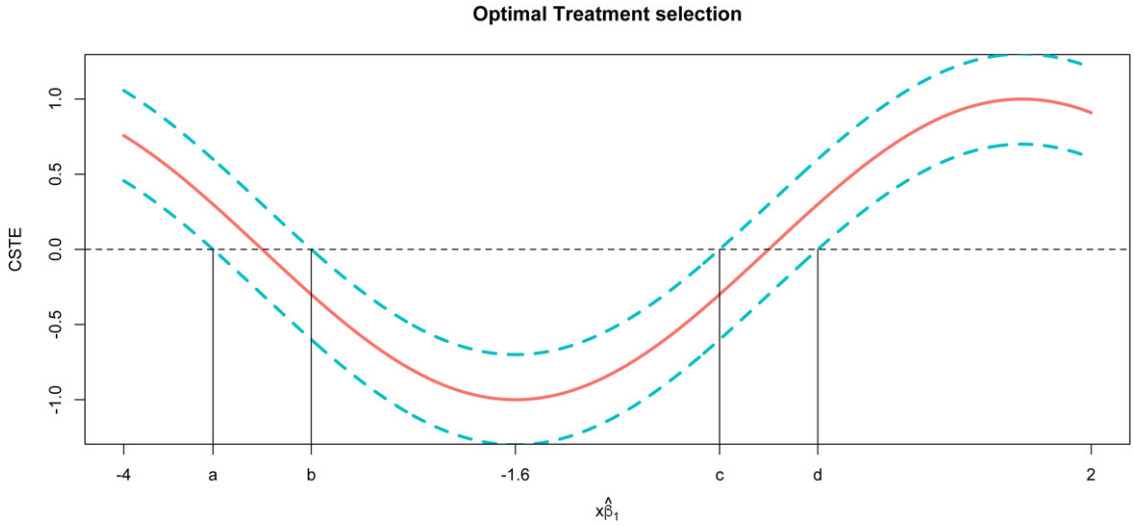


Figure 1. A simulated example, demonstrating CSTE estimation, SCBs, and cutoff points. The solid curve is the estimate of the CSTE function. The dashed curves are the corresponding SCBs of the CSTE curve. Four vertical lines indicate the locations of the cutoff points.

is within the range of $(-4, -1.6)$, a larger value implies that the patient more likely benefits from the new treatment than from the old treatment. On the other hand, if the patient's $X^\top\hat{\beta}_1$ value falls into $(-1.6, 2)$, the new treatment is more beneficial when a smaller value of $X^\top\hat{\beta}_1$ is observed. Moreover, the results in Figure 1 also imply that we are 95% confident that a new patient with the $X^\top\hat{\beta}_1$ value falling into the region $[-4, a]$ or $[d, 2]$ will benefit from the old treatment, since the lower bands are all above zero in this region. If $X_0^\top\hat{\beta}_1$ falls into region $[b, c]$, then we have 95% confidence that the patient should receive the new treatment, as the upper bands are all below zero in this region. For the intervals $[a, b]$ and $[c, d]$, zero is covered in the confidence band indicating that there is no significant difference between the effects of the new and old treatments.

It is worth nothing that the regions of indifference can also be constructed based on a parametric model using the limiting distribution of the parametric estimators; see more discussions in Wu (2016). However, this approach is only directly applicable to the settings with low-dimensional covariates and to each given patient instead of a group of patients. Han, Zhou, and Liu (2017) defined a modified version of the CSTE curve using the quantile of the covariate to compare covariate's capacities for predicting responses to a treatment. Then they use the "best" covariate as a guidance to select treatment. This may be time-consuming when there is a large number of covariates. Our method selects relevant covariates automatically in the estimation procedure and combines information from all covariates through a weighted combination of those covariates with the weights estimated from the data. The weight reflects how important the corresponding covariate is for predicting the response. As a result, our method is more convenient and flexible in selecting optimal treatment.

5. Simulation Study

In this section, we investigate the finite-sample performance of our proposed method via simulated datasets. We run all simulations in R in a linux cluster. We consider the following examples:

Example 1. $\text{logit}(\mu(X,Z)) = X^\top\beta_1(1 - X^\top\beta_1)Z + \exp(X^\top\beta_2)$.

Example 2. $\text{logit}(\mu(X,Z)) = (X^\top\beta_1)^2 Z + \sin(\pi X^\top\beta_2/2)$.

Example 3. $\text{logit}(\mu(X,Z)) = -\exp(X^\top\beta_1)Z/1.5 + (X^\top\beta_2)^2$.

Example 4. $\text{logit}(\mu(X,Z)) = X^\top\beta_1(1-X^\top\beta_1)Z + \exp(X^\top\beta_2) + c \cdot \sin(\pi X^\top\beta_3/2)$.

Example 5. $\text{logit}(\mu(X,Z)) = [X^\top\beta_1(1 - X^\top\beta_1) + c \cdot \sin(\pi X^\top\beta_3/2)]Z + \exp(X^\top\beta_2)$.

The simulated data are generated as follows: the outcome Y is sampled from a binomial distribution with probability of success equals to $\mu(x, z)$; the covariates X are generated from a truncated multivariate normal distribution with mean vector 0, covariance matrix with $\Sigma_{ij} = 0.5^{|i-j|}$, and each covariate is truncated by $(-2, 2)$; the binary scale covariate Z is sampled from Binomial$(1, 0.5)$, which means that each subject is randomly assigned to either control or treatment group. We set $\beta_1 = (1, 1, 1, 0, \ldots, 0)'/\sqrt{3}$, $\beta_2 = (1, -2, 0, \ldots, 0)'/\sqrt{5}$, and $\beta_3 = (1, 1, 1, 0, \ldots, 0)'/\sqrt{3}$. For Examples 1–3, we set $p = 10, 50, 100, 500$, and sample size $n = 500, 750, 1000$. For Examples 4 and 5, we consider that g_2 and g_1 are misspecified respectively. We choose $p = 50$, $n = 1000$, and $c = 0.1, 0.2, 0.5$. For each pair of n and p, we repeat the simulations $J = 300$ times.

To obtain sparse estimates of β_k for high-dimensional cases, we choose SCAD as the penalty function and let $\gamma = 3.7$ (Fan and Li 2001). The optimal tuning parameter λ is chosen from a geometrically increasing sequence of 30 parameters by minimizing the modified Bayesian information criterion (Wang, Li, and Leng 2009): $\text{BIC}_\lambda = -2\cdot\log(\text{Loss}) + df_\lambda \cdot \log(n) \cdot C(p)$, where Loss is the loss function in (3), df_λ is the number of nonzero elements in β_k, $C(p) = C(p) = \log\log(p)$.

We evaluate the following metrics for the nonparametric function $g_1(\cdot)$ based on 200 equally spaced grid points on the

Table 1. Simulation results for Examples 1–3.

	Model 1			Model 2			Model 3		
(n, p)	MSE	MAE	CP	MSE	MAE	CP	MSE	MAE	CP
(500, Oracle)	0.230	0.364	0.880	0.217	0.336	0.861	0.389	0.400	0.840
(750, Oracle)	0.125	0.276	0.925	0.113	0.251	0.905	0.143	0.268	0.905
(1000, Oracle)	0.095	0.23	0.947	0.088	0.230	0.935	0.083	0.220	0.940
(500, 10)	0.280	0.494	0.874	0.325	0.428	0.889	0.405	0.461	0.858
(750, 10)	0.169	0.318	0.897	0.151	0.293	0.924	0.329	0.472	0.885
(1000, 10)	0.112	0.258	0.939	0.095	0.243	0.929	0.099	0.241	0.931
(500, 50)	0.530	0.461	0.871	0.506	0.576	0.870	0.679	0.621	0.863
(750, 50)	0.258	0.371	0.930	0.331	0.426	0.903	0.280	0.393	0.920
(1000, 50)	0.100	0.248	0.957	0.129	0.280	0.955	0.099	0.242	0.959
(500, 100)	0.698	0.594	0.836	0.525	0.566	0.826	0.640	0.660	0.862
(750, 100)	0.153	0.305	0.906	0.452	0.477	0.896	0.141	0.291	0.907
(1000, 100)	0.108	0.230	0.947	0.188	0.338	0.920	0.093	0.237	0.940
(500, 500)	0.718	0.789	0.826	0.642	0.777	0.835	0.683	0.519	0.831
(750, 500)	0.413	0.503	0.857	0.434	0.453	0.860	0.524	0.444	0.854
(1000, 500)	0.123	0.249	0.930	0.174	0.319	0.925	0.143	0.287	0.912

NOTE: Oracle estimator is obtained when true index sets are given. True functions are $x(1-x), x^2, -\exp(x)/1.5$. MSE and MAE represent the mean squared error and the mean absolute error of the estimator for $g_1(\cdot)$, and CP represents the coverage probability of its SCBs.

Table 2. Simulation results for Examples 4 and 5 (misspecified cases) with $p = 50$ and $n = 1000$.

	Model 4			Model 5		
	MSE	MAE	CP	MSE	MAE	CP
$c = 0$	0.100	0.248	0.957	0.100	0.248	0.957
$c = 0.1$	0.138	0.212	0.933	0.161	0.293	0.938
$c = 0.2$	0.136	0.235	0.931	0.266	0.338	0.914
$c = 0.5$	0.208	0.341	0.904	0.290	0.415	0.895

NOTE: MSE and MAE represent the mean squared error and the mean absolute error of the estimator for $g_1(\cdot)$, and CP represents the coverage probability of its SCBs.

range of $\hat{\eta}_1 = X^\top \hat{\beta}_1$: the mean squared error (MSE) of its estimator; the mean absolute error (MAE) of its estimator; the average coverage probability (CP) of its SCBs. Since the models in Example 1–3 are correctly specified, we compute the average number of parameters that are incorrectly estimated to be nonzero, the average number of parameters that are incorrectly estimated as zero; the proportions that all relevant covariates are correctly selected and the proportions that some relevant covariates are not selected.

We define that the oracle estimator of $g_1(\cdot)$ is obtained when the true indexes of nonzero components in β_1 and β_2 are given in Examples 1–3. The results are summarized in Tables 1 and 2. In Table 1, we see that the coverage probabilities are slightly less than 95% but close to 90% when the sample size is 750. When the sample is 1000, the empirical coverage is close to the nominal 95% confidence level. The mean square error and mean absolute error also decrease as sample size n increases. This indicates that the estimates of confidence bands become more accurate as the sample size increases. Table 1 shows that the proposed asymptotic SCB performs well when the sample size is relatively large. However, with small sample size, a bootstrap procedure is recommended to construct the SCB. This can be a future research topic to explore. In Table 2, when c is small (i.e., the degree of misspecification is small), the numerical results are comparable to those for $c = 0$ with no misspecification. However, the performance deteriorates when c becomes large. Moreover, the method is more robust to the misspecification of g_2 than that of g_1. The model selection results are summarized in Table 3. The false positive and false negative are decreasing when n increases. It is worth noting that in practice we often need to make a reliable treatment recommendation strategy to a group of patients instead of a specific given patient, and the SCBs can serve the former purpose well. In Table 4, we also compare the performance of the pointwise confidence interval (CI) with that of the SCB for a group of randomly generated new patients with group size 1, 10, 100, 200, 500, respectively. The CI and SCB are constructed based on the data with sample $n = 1000, p = 10$. We see that the pointwise CI has low empirical coverage rates when the group size is greater than 1. The pointwise CI can work well when we only have one patient. However, when we have a group of patients with different covariate values, we can make an incorrect treatment decision based on the pointwise CI for these patients.

6. A Real Data Example

In this section, we present and discuss the results of applying the procedure described in previous sections to a real dataset. We illustrate the applications of the CSTE curve in a real-world example and provide some insight into the interpretation of the results. The goal is to analyze the effect of Zhengtianwan in the treatment of migraines. This is a multicenter, randomized, double-blind, placebo-controlled trial on the effectiveness of Zhengtian pill on treating patients with migraines. A migraine is a common neurological disorder characterized by recurrent headache attacks. Migraine treatment involves acute and prophylactic therapy. The objective of this study is to evaluate the efficacy and safety of Zhengtian Pill for migraine prophylaxis. Zhengtian Pill, a Chinese Patented Medicine approved by the State Food and Drug Administration of China in 1987, has been used in clinical practice for more than 20 years in China to stimulate blood circulation, dredge collaterals, alleviate pains, and nourish the liver. To evaluate the effectiveness of Zhengtian Pill in preventing the onset of migraine attacks, a large-scale, randomized and prospective clinical study was conducted. Eligible patients were monitored during a baseline period of four weeks, during which the headache characteristics were recorded as baseline data. In this period, any use of migraine preventive medications was prohibited. After the baseline period, a

Table 3. Model selection results for Examples 1–3.

	Model 1				Model 2				Model 3			
(n, p)	FPR	FNR	C	IC	FPR	FNR	C	IC	FPR	FNR	C	IC
(500, 10)	0.282	0.025	0.885	0.115	0.274	0.019	0.904	0.095	0.381	0.015	0.909	0.090
(750, 10)	0.206	0.008	0.956	0.043	0.170	0.007	0.960	0.039	0.280	0.018	0.941	0.058
(1000, 10)	0.162	0.000	1.000	0.000	0.110	0.005	0.994	0.005	0.217	0.002	0.989	0.010
(500, 50)	0.271	0.042	0.785	0.214	0.172	0.181	0.564	0.435	0.165	0.236	0.535	0.465
(750, 50)	0.108	0.041	0.823	0.176	0.178	0.025	0.892	0.107	0.094	0.125	0.563	0.436
(1000, 50)	0.070	0.009	0.952	0.047	0.111	0.005	0.975	0.025	0.082	0.065	0.788	0.211
(500, 100)	0.063	0.093	0.625	0.375	0.107	0.084	0.605	0.395	0.093	0.087	0.670	0.330
(750, 100)	0.055	0.065	0.703	0.296	0.098	0.044	0.688	0.311	0.040	0.106	0.681	0.318
(1000, 100)	0.031	0.032	0.845	0.154	0.053	0.037	0.830	0.170	0.021	0.061	0.803	0.196
(500, 500)	0.114	0.138	0.471	0.519	0.119	0.164	0.223	0.777	0.104	0.166	0.243	0.757
(750, 500)	0.974	0.096	0.540	0.460	0.069	0.138	0.294	0.706	0.059	0.132	0.302	0.698
(1000, 500)	0.030	0.062	0.609	0.391	0.039	0.102	0.413	0.587	0.049	0.094	0.449	0.551

NOTE: FP (false positive): zero is estimated as nonzero; FN (false negative): nonzero is estimated as zero; C (correctly selected): all relevant covariates are selected; IC (incorrectly selected): some relevant covariates are not selected.

Table 4. Comparisons between the pointwise confidence interval and the SCBs for the empirical coverage rate of a group of randomly generated new patients with group size 1, 10, 100, 200, and 500, respectively.

	Group size	1	10	100	200	500
Model 1	CI	0.962	0.806	0.621	0.561	0.554
	SCBs	1.000	0.991	0.952	0.948	0.935
Model 2	CI	0.959	0.911	0.835	0.771	0.773
	SCBs	1.000	1.000	0.963	0.951	0.954
Model 3	CI	0.961	0.853	0.694	0.661	0.668
	SCBs	1.000	0.981	0.952	0.934	0.937

12-week treatment period and four-week follow-up period were carried out. Patients were requested to keep a headache diary throughout the whole study period, from which investigators were able to extract detailed information of migraine attacks including migraine days, frequency, duration, and intensity as well as the use of acute medication during the study period. The outcome measures were evaluated at 4, 8, and 12 weeks, and during the follow-up period. The patients who met the inclusion criteria were randomly assigned into the experimental group and control group in a 1:1 ratio using a computer-generated stochastic system.

In our analysis, the response variable Y is a binary outcome indicating if the number of days that headaches occur has decreased 8 weeks after patients were treated. Z is another indicator: $Z = 1$ means the subject is assigned in the experimental group; $Z = 0$ means in the control group. The covariates x_1 to x_3 are gender (0 for male, 1 for female), height and body weight, x_4 to x_{12} are overall scores of Traditional Chinese Medicine Syndromes (Huozheng, Fengzheng, Xueyu, Tanshi, Qixu, Yuzheng, Xuexu, Yinxu, and Yangxu) for migraine. Those scores, with a range of 0–20, are commonly used to describe the severity of migraine symptoms in Traditional Chinese Medicine. The higher of the score means that the syndrome is more severe. All covariates are centered and standardized as input. We have 204 observations where 99 are in the experimental group and 105 are in the control group. The purpose of this exercise is to model the odds ratio as a function of those covariates. The number of covariates $p = 12$ might be regarded as small, so we estimate the CSTE curve using the algorithm in Section 3.1 with and without model selection. We use SCAD as our penalty function and the optimal tuning parameter is selected via the modified BIC criterion. The corresponding SCBs for the CSTE curve are calculated. The results are not intended as definitive analyses of these data.

Table 5 summarizes the point estimates of β_1 and the corresponding standard errors. Figure 2 shows two estimated CSTE curve and their SCBs: (a) using all 12 variables; (b) using all variables except x_7 and x_8 (not selected). To aid in interpretation for each covariate, we depict each covariate versus g_1 where other covariates are projected onto their mean values in Figure 3. As we can see, Huozheng has a monotonic dependence on the CSTE, but the corresponding relationships for other overall scores are highly nonlinear; most of them have a quadratic appearance. On the basis of these figures, we can conclude that the odds ratio does depend on the linear combination of the covariates in a nonlinear manner.

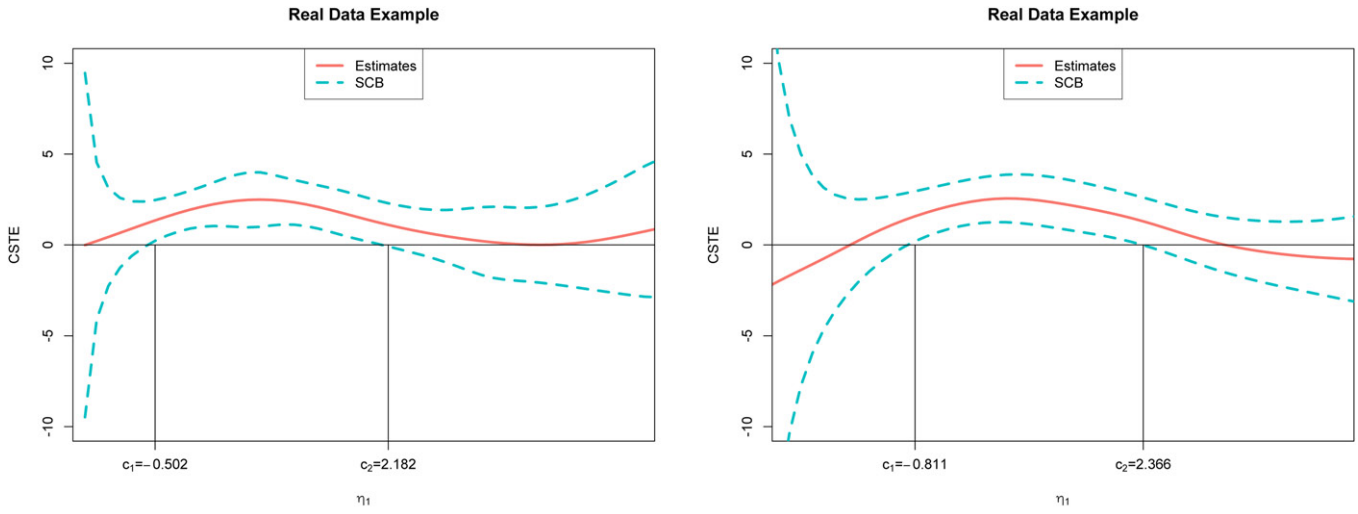
When all variables are used to estimate the curve, the two cutoff points are $c_1 = -0.502, c_2 = 2.182$. The estimates of the overall rating of biomarker values are $x^\top \hat{\beta}_1$ where the majority of the points (>95%) fall into the interval $(-1.2, 5)$. Two cutoff points divide this interval into three parts. Since the response variable $y = 1$ represents headache improvements after 8 weeks treatment, higher CSTE value means that the patient more likely benefits from the treatment. Suppose a new patient with biomarker value x_0. When $x^\top \hat{\beta}_1$ falls into $[-1.2, c_1]$ and $[c_2, 5]$, the treatment does not have a significant effect. When it falls into (c_1, c_2), then this treatment can improve the migraine of this patient. In the Zhengtianwan data, there are 37.7% of patients fall into the interval. On the other hand, when x_7, x_8 are removed from the model, we obtain a new estimate $\hat{\beta}_{1,ms}$ of β_1. As we can see from Figure 2(b), the positive region of the CSTE curve becomes wider than that in (a), and the two cutoff points become -0.811 and 2.366. The majority of $x^\top \hat{\beta}_{1,ms}$ falls into $-(2.5, 5)$. When $x^\top \hat{\beta}_{1,ms}$ is in the range of $[-0.811, 2.366]$, the patient should receive Zhengtianwan treatment and 50.98% of all subjects in the training dataset fall into this interval.

The number of covariates is not very large in this dataset, so it is possible to fit a sparse linear model with high order terms and make treatment recommendation based on the parametric estimate of the CSTE curve. To this end, we model g_k for $k = 1, 2$ as a multivariate polynomial function of the covariates X with degree of two, that is, $g_k(X) = \alpha_{0,k} + \sum_j \alpha_{j,k} X_j + \sum_{j \le j'} \alpha_{jj',k} X_j X_{j'}$. We obtain the estimates $(\hat{\alpha}_{0,k}, \hat{\alpha}_{j,k}, \hat{\alpha}_{jj',k})$ of the unknown coefficients $(\alpha_{0,k}, \alpha_{j,k}, \alpha_{jj',k})$ by fitting regularized L_1 logistic regression

Table 5. Estimates of β_1 and the corresponding standard errors (in parentheses).

Variable	Gender	Height	Weight	Huozheng	Fengzheng	Xueyu	Tanshi	Qixu	Yuzheng	Xuexu	Yinxu	Yangxu
β_1	0.88	0.10	−0.04	−0.06	0.08	−0.09	−0.08	0.04	0.18	−0.24	0.30	−0.03
	(0.23)	(0.01)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.11)	(0.04)	(0.05)	(0.07)	(0.02)
Penalized β_1	0.70	0.11	−0.05	−0.04	0.14	−0.15	0.00	0.00	0.21	−0.38	0.49	−0.14
	(0.31)	(0.02)	(0.01)	(0.02)	(0.03)	(0.01)	(0.00)	(0.00)	(0.05)	(0.07)	(0.09)	(0.06)

NOTE: First row: Using the unpenalized model; second row: using the penalized model.



(a) Without variable selection (b) With variable selection

Figure 2. Real data example. Red curve is the CSTE curve; blue dashed curves are the SCBs. (a) Two cutoff points are −0.502 and 2.182; (b) two cutoff points are −0.811 and 2.366.

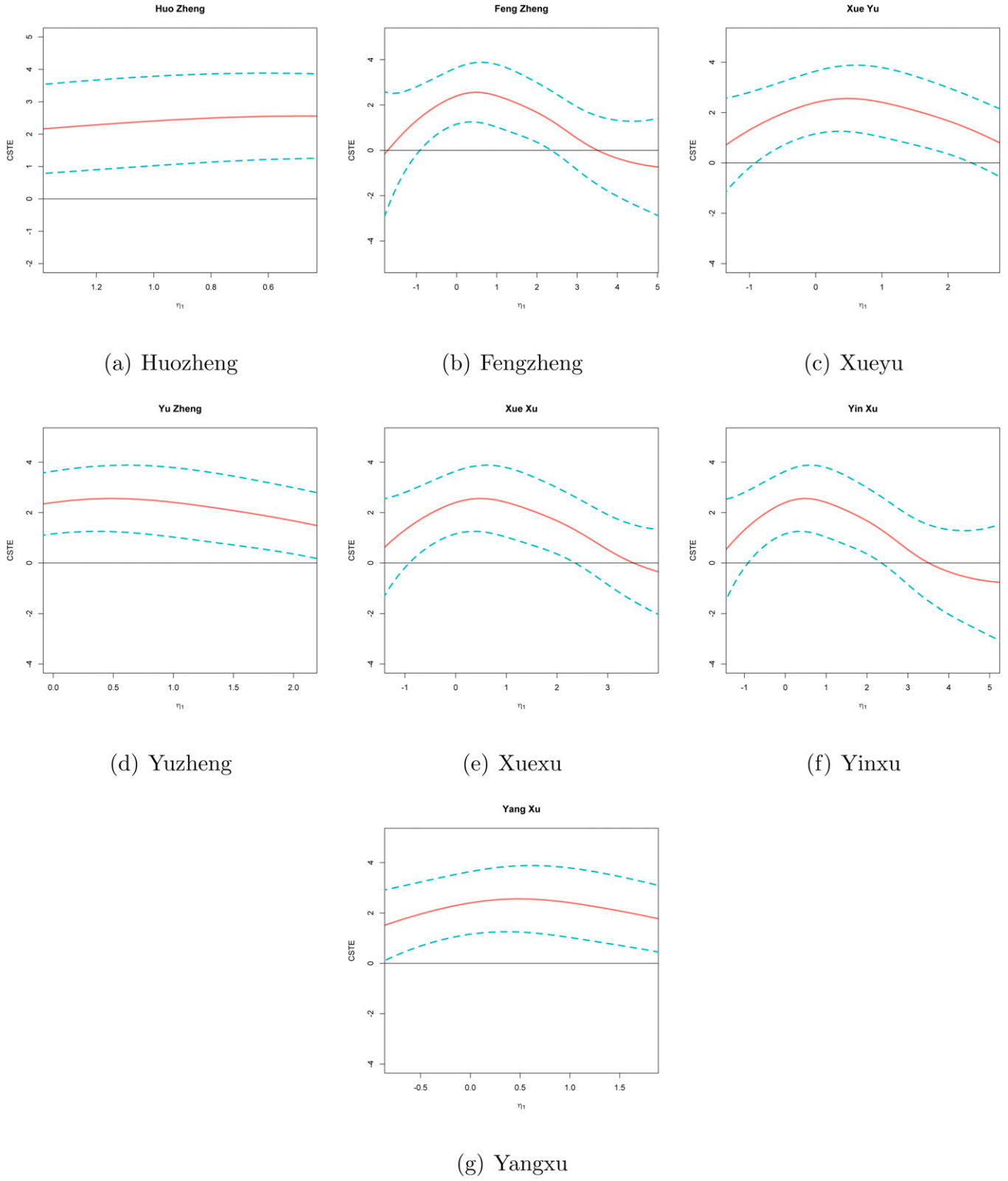
with the tuning parameter selected by 3-fold cross-validation. For a new patient with the observed covariates $\tilde{x}$, we obtain the predicted value for the CSTE curve as $\hat{g}_1(\tilde{x}) = \hat{\alpha}_{0,1} + \sum_j \hat{\alpha}_{j,1}\tilde{x}_j + \sum_{j\le j'} \hat{\alpha}_{jj',1}\tilde{x}_j\tilde{x}_{j'}$. For this parametric modeling method, we can only use the predicted value for g_1 to make treatment recommendation. Thus, based on the rule that a subject would benefit from the treatment if $\hat{g}_1$ is positive, 67.64% of the subjects should receive the treatment. We see that this percentage is higher than that obtained by our proposed method, as the parametric method only uses the point estimate for treatment selection. The construction of confidence intervals for g_1 by the parametric method is very challenging, as it involves finding the joint asymptotic distribution of $(\hat{\alpha}_{0,k},\hat{\alpha}_{j,k},\hat{\alpha}_{jj',k})$ after model selection, which is a difficult problem. Moreover, SCBs are not available based on the parametric modeling method. As discussed and shown in the simulation studies given in Section 5, SCBs are more reliable than the pointwise CIs for treatment recommendation for a group of patients.

7. Discussion

Both the simulation and real-world studies in Sections 5 and 6 suggest that the modeling procedure for CSTE curve can successfully detect and model complicated nonlinear relationships between binary response and high-dimensional covariates. In practice, the nonlinear dependencies we suggest are not characteristic of all situations. We adapt the spline-backfitted kernel smoothing to construct the SCBs for the nonlinear functions to choose the optimal treatment. Moreover, the SCBs can be used to verify the presence of nonlinear relationships as well.

Our model is motivated by the desire to provide an individualized decision rule for patients along with the ability to deal with high-dimensional covariates when the outcome is binary. The semiparametric modeling approach can be viewed as a generalization of the CSTE curve with one covariate proposed in Han, Zhou, and Liu (2017) in the sense that the odds ratio depends on a weighted linear combination of all covariates. Although we consider a single decision with two treatment options, our model can be readily generalized to multiple treatment arms.

To identify model (1), we require that the unknown functions $g_1(\cdot)$ and $g_2(\cdot)$ be nonconstant functions on their supports. Otherwise, the index parameters β_1 and β_2 are not identified. That is, if $g_k(\beta_k^\top x) = a_k$ for all $x \in \mathcal{X}$, then β_k can take an arbitrary value. Our treatment recommendation method is proposed based on the prior information that at least one baseline covariate is useful, and the treatment effect is not zero for some values of the covariates. Based on this prior information, our goal is to estimate the optimal treatment regimes and provide recommendations for individual patients based on their observed baseline covariates. For example, Figures 1 and 2 show that the estimated values of $g_1(\cdot)$ are nonzero in some regions, so that one can apply our method to identify those regions and then make treatment recommendations to patients. In practice, if we conjecture that all covariates are irrelevant, then for an arbitrarily given value of β_k, $g_k(\beta_k^\top x) = a_k$ for all $x \in \mathcal{X}$. We



(a) Huozheng (b) Fengzheng (c) Xueyu

(d) Yuzheng (e) Xuexu (f) Yinxu

(g) Yangxu

Figure 3. Plot one covariate against CSTE curve with other covariates fixed on their mean values.

can conduct a test for this conjecture by choosing a value for β_k, estimating the unknown function $g_k\left(\beta_k^\top x\right)$ and then testing whether the function is a constant or not. Likewise, we can use this approach to testing whether $g_1\left(\beta_1^\top x\right) = 0$ for all $x \in \mathcal{X}$, that is, whether the treatment has effect at all. If they are rejected, next we can fit our model to more precisely find the optimal value of β_k. Since this article focuses on making individualized treatment recommendations to patients based on their available covariates, we leave this interesting topic as a future work to explore.

Supplementary Materials

Supplementary materials include the technical proofs for all the theoretical results.

Acknowledgments

The authors are grateful to the editor, the associate editor, and two anonymous reviewers for their constructive comments that helped us improve the article substantially.

Funding

Guo and Ma's research was supported by NSF grants DMS-1712558 and DMS-2014221, NIH grant R01 ES024732-03 and UCR Academic Senate CoR grant. Zhou's research was supported by NSFC 81773546.

References

Breheny, P., and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications to Biological Feature Selection," *The Annals of Applied Statistics*, 5, 232–253. [312]

Cai, T., Tian, L., Wong, P. H., and Wei, L. J. (2011), "Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selections," *Biostatistics*, 12, 270–282. [310]

Chen, G., Zeng, D., and Kosorok, M. R. (2016), "Personalized Dose Finding Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 111, 1509–1547. [309]

Chen, X., and Christensen, T. M. (2015), "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions," *Journal of Econometrics*, 188, 447–465. [312]

de Boor, C. (2001), *A Practical Guide to Splines*, Applied Mathematical Sciences, New York: Springer-Verlag. [311]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [312,315]

Fan, J., and Lv, J. (2011), "Nonconcave Penalized Likelihood With NP-Dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484. [313]

Han, K., Zhou, X.-H., and Liu, B. (2017), "CSTE Curve for Selection the Optimal Treatment When Outcome Is Binary," *Scientia Sinica Mathematica*, 47, 497–514. [309,310,315,318]

Huang, Y. (2015), "Identifying Optimal Biomarker Combinations for Treatment Selection Through Randomized Controlled Trials," *Clinical Trials*, 12, 348–56. [309]

Huang, Y., and Fong, Y. (2014), "Identifying Optimal Biomarker Combinations for Treatment Selection via a Robust Kernel Method," *Biometrics*, 70, 891–901. [309]

Janes, H., Brown, M. D., Huang, Y., and Pepe, M. S. (2014), "An Approach to Evaluating and Comparing Biomarkers for Patient Treatment Selection," *The International Journal of Biostatistics*, 10, 99–121. [309]

Jiang, B., and Liu, J. (2014), "Variable Selection for General Index Models via Sliced Inverse Regression," *The Annals of Statistics*, 42, 1751–1786. [311]

Jiang, B., Song, R., Li, J., and Zeng, D. (2019), "Entropy Learning for Dynamic Treatment Regimes," *Statistica Sinica*, 29, 1633–1655. [309]

Kang, C., Janes, H., and Huang, Y. (2014), "Combining Biomarkers to Optimize Patient Treatment Recommendations," *Biometrics*, 70, 695–707. [310]

Kosorok, M. R., and Laber, E. B. (2019), "Precision Medicine," *Annual Review of Statistics and Its Application*, 6, 263–286. [309]

Laber, E. B., and Qian, M. (2019), "Generalization Error for Decision Problems," *Wiley StatsRef: Statistics Reference Online* (accepted). [309]

Laber, E. B., and Zhao, Y. Q. (2015), "Tree-Based Methods for Individualized Treatment Regimes," *Biometrika*, 102, 501–514. [309]

Liu, R., Yang, L., and Härdle, W. K. (2013), "Oracally Efficient Two-Step Estimation of Generalized Additive Model," *Journal of the American Statistical Association*, 108, 619–631. [310,312]

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020), "Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning," *Journal of the American Statistical Association*, 115, 692–706. [309]

Ma, S., and He, X. (2016), "Inference for Single-Index Quantile Regression Models With Profile Optimization," *The Annals of Statistics*, 44, 1234–1268. [313]

Ma, S., and Song, P. X.-K. (2015), "Varying Index Coefficient Models," *Journal of the American Statistical Association*, 110, 341–356. [311]

Ma, S., and Yang, L. (2011), "A Jump-Detecting Procedure Based on Spline Estimation," *Journal of Nonparametric Statistics*, 23, 67–81. [313]

Ma, Y., and Zhou, X.-H. (2014), "Treatment Selection in a Randomized Clinical Trial via Covariate-Specific Treatment Effect Curves," *Statistical Methods in Medical Research*, 26, 124–141. [309]

Neykov, M., Liu, J., and Cai, T. (2016), "L_1-Regularized Least Squares for Support Recovery of High Dimensional Single Index Models With Gaussian Designs," *Journal of Machine Learning Research*, 17, 1–37. [311]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *The Annals of Statistics*, 39, 1180–1210. [310]

Radchenko, P. (2015), "High Dimensional Single Index Models," *Journal of Multivariate Analysis*, 139, 266–282. [310,311]

Rubin, D. B. (2005), "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, 100, 322–331. [310]

Rubin, D. B., and van der Laan, M. J. (2012), "Statistical Issues and Limitations in Personalized Medicine Research With Clinical Trials," *The International Journal of Biostatistics*, 8, 18. [309]

Shi, C., Fan, A., Song, R., and Lu, W. (2018), "High-Dimensional *a*-Learning for Optimal Dynamic Treatment Regimes," *The Annals of Statistics*, 46, 925–957. [310]

Song, R., Luo, S., Zeng, D., Zhang, H., Lu, W., and Li, Z. (2017), "Semiparametric Single-Index Model for Estimating Optimal Individualized Treatment Strategy," *Electronic Journal of Statistics*, 11, 364–384. [311]

Stone, C. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [312]

Taylor, J. M. G., Cheng, W., and Foster, J. C. (2015), "Reader Reaction to 'a Robust Method for Estimating Optimal Treatment Regimes' by Zhang et al. (2012)," *Biometrics*, 71, 267–273. [309]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [309]

Wang, H., Li, B., and Leng, C. (2009), "Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters," *Journal of the Royal Statistical Society*, Series B, 71, 671–683. [315]

Wang, L., and Yang, L. (2007), "Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model," *The Annals of Statistics*, 35, 2474–2503. [310,312]

Watson, M. W., and Engle, R. F. (1983), "Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models," *Journal of Econometrics*, 23, 385–400. [312]

Wu, T. (2016), "Set Valued Dynamic Treatment Regimes," Ph.D. thesis, The University of Michigan. [315]

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012), "Estimating Optimal Treatment Regimes From a Classification Perspective," *Stat*, 1, 103–114. [309]

Zhang, B., and Zhang, M. (2018), "C-Learning: A New Classification Framework to Estimate Optimal Dynamic Treatment Regimes," *Biometrics*, 74, 891–899. [309]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [312]

Zhang, Y., Laber, E. B., Davidian, M., and Tsiatis, A. A. (2018), "Interpretable Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 113, 1541–1549. [309]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [309]

Zheng, S., Liu, R., Yang, L., and Härdle, W. K. (2016), "Statistical Inference for Generalized Additive Models: Simultaneous Confidence Corridors and Variable Selection," *TEST*, 25, 607–626. [310,313,314]

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017), "Residual Weighted Learning for Estimating Individualized Treatment Rules," *Journal of the American Statistical Association*, 112, 169–187. [309]

Zhou, X.-H., and Ma, Y. B. (2012), "BATE Curve in Assessment of Clinical Utility of Predictive Biomarkers," *Science China Mathematics*, 55, 1529–1552. [309]

Zhu, K., Huang, Y., and Zhou, X.-H. (2018), "Tree-Based Ensemble Methods for Individualized Treatment Rules," *Biostatistics & Epidemiology*, 2, 61–83. [309]

Zhu, W., Zeng, D., and Song, R. (2019), "Proper Inference for Value Function in High-Dimensional Q-Learning for Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 114, 1404–1417. [310]