Journal of Statistical Planning and Inference 205 (2020) 115–128

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A single-index model with multiple-links

Hyung Park [a,*], Eva Petkova [a], Thaddeus Tarpey [a], R. Todd Ogden [b]

[a] Department of Population Health, New York University, New York, NY 10016, USA
[b] Department of Biostatistics, Columbia University, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 16 August 2018
Received in revised form 16 January 2019
Accepted 27 May 2019
Available online 4 July 2019

Keywords:
Single-index models
Treatment effect modifier
Biosignature

ABSTRACT

In a regression model for treatment outcome in a randomized clinical trial, a treatment effect modifier is a covariate that has an interaction with the treatment variable, implying that the treatment efficacies vary across values of such a covariate. In this paper, we present a method for determining a composite variable from a set of baseline covariates, that can have a nonlinear association with the treatment outcome, and acts as a composite treatment effect modifier. We introduce a parsimonious generalization of the *single-index models* that targets the effect of the interaction between the treatment conditions and the vector of covariates on the outcome, a *single-index model with multiple-links* (SIMML) that estimates a single linear combination of the covariates (i.e., a single-index), with treatment-specific nonparametric *link* functions. The approach emphasizes a focus on the treatment-by-covariates interaction effects on the treatment outcome that are relevant for making optimal treatment decisions. Asymptotic results for estimator are obtained under possible model misspecification. A treatment decision rule based on the derived single-index is defined, and it is compared to other methods for estimating optimal treatment decision rules. An application to a clinical trial for the treatment of depression is presented.

1. Introduction

In precision medicine, a critical concern is to identify baseline measures that have distinct relationships with the outcome from different treatments so that patient-specific treatment decisions can be made (Murphy, 2003; Robins, 2004). Such variables are called treatment effect modifiers, and these can be useful in determining a treatment decision rule that will select a treatment for a patient based on observations made at baseline. There is a growing need to extract treatment effect modifiers from (usually noisy) baseline patient data that, more and more commonly, consist of a large number of clinical and biological characteristics.

Typically, treatment effect modifiers (or, "moderators") are identified either one by one, using one model for each potential predictor, or from a large model which includes all potential predictors and their (two-way) interactions with treatment, and then testing for significance of the interaction terms, almost exclusively using linear models. In the linear model context, Petkova et al. (2016) proposed a model using a linear combination (i.e., an index) of patients' characteristics, termed a generated effect modifier (GEM) constructed to optimize the interaction with a treatment indicator. Such a composite variable approach is especially appealing for complex diseases such as psychiatric diseases, in which each baseline characteristic may only have a small treatment modifying effect. In such settings, it is not common to find variables that are individually strong moderators of treatment effects.

* Corresponding author.
E-mail address: hyung.park@nyumc.org (H. Park).

https://doi.org/10.1016/j.jspi.2019.05.008
0378-3758/

Here we present novel flexible methods for determining composite variables that permit non-linear association with the outcome. In particular, the proposed methods allow the conditional expectation of the outcomes to have a flexible treatment-specific link function with an index. We define the index to be a one-dimensional linear combination of the covariates. This approach is related to *single-index models* (Brillinger, 1982; Stoker, 1986; Powell et al., 1989; Hardle et al., 1993; Xia and Li, 1999; Horowitz, 2009; Antoniadis et al., 2004), as well as to single-index model generalizations such as projection pursuit regression (Friedman and Stuetzle, 1981) and *multiple-index* models (Xia, 2008; Yuan, 2011). We employ a single projection of the covariates (i.e., an index) to summarize the variability of the baseline covariates, and multiple link functions to connect the derived single-index to the treatment-specific mean responses; we call these *single-index models with multiple-links* (SIMML). This single-index model with multiple-links provides a parsimonious extension of the single-index model in modeling the effect of the interaction between a categorical treatment variable and a vector-valued covariate. The dependence of treatment-specific outcomes on a common single-index improves the interpretability, and helps in determining treatment decision rules. This approach generalizes the notion of a composite "treatment effect modifier" from the linear model setting, to a nonparametric context, to define a nonparametric generated effect modifier.

2. A single-index model with multiple-links (SIMML)

Let $X = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ denote the set of covariates. Let T denote the categorical (treatment assignment) variable of interest, taking values in $\{1, \ldots, K\}$ with nonzero probabilities $(\pi_1, \ldots \pi_K)$ that sum to one. Let $Y \in \mathbb{R}$ denote an outcome variable; without loss of generality, we assume that a higher value of Y is preferred. We focus on data arising from a randomized experiment, however, the method can be extended to observational studies.

A common approach to interrogate the effect of the interaction between X and the treatment indicator T on an outcome is to fit a regression model separately for each of the K treatment groups, as functions of X. For instance, a single-index model can be fitted separately for each treatment group t, resulting in K indices, $\boldsymbol{\beta}_t^\top X$, $t \in \{1, \ldots, K\}$. We refer to this as a K-index model; it has the form

$$\mathbb{E}\,(Y \mid T = t, X = x) = g_t(\boldsymbol{\beta}_t^\top x) \quad (t = 1, \ldots, K), \tag{1}$$

where both the treatment-specific nonparametric link functions $g_t(\cdot)$, and the treatment-specific index vectors $\boldsymbol{\beta}_t \in \mathbb{R}^p$, need to be estimated for each group t. (The vectors $\boldsymbol{\beta}_t$ need to satisfy some identifiability condition (Lin and Kulasekera, 2007).) While this is a reasonable approach, the K indices of model (1) lack useful interpretation as effect modifiers and often lead to over-parametrization.

For parsimony and insight, the SIMML constrains the $\boldsymbol{\beta}_t$ in (1) to be equal, and it requires separate nonparametrically defined curves for each treatment t as a function of a single index $\boldsymbol{\alpha}^\top X$ common for all t:

$$\mathbb{E}\,(Y \mid T = t, X = x) = g_t(\boldsymbol{\alpha}^\top x) \quad (t = 1, \ldots, K), \tag{2}$$

where both the links g_t and the vector $\boldsymbol{\alpha}$ need to be estimated. The SIMML (2) provides a single parsimonious biosignature, $\boldsymbol{\alpha}^\top X \in \mathbb{R}$. Due to the nonparametric nature of g_t, the scale of $\boldsymbol{\alpha}$ is not identifiable in (2) and to address this we restrict $\boldsymbol{\alpha}$ to be in $\Theta = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top | \sum_{j=1}^p \alpha_j^2 = 1, \alpha_p > 0\}$, i.e., to be in the upper hemisphere of the unit sphere.

If the true model for the treatment-specific outcome Y_t is not a SIMML, then the SIMML can be regarded as the L^2 projection of the treatment specific mean outcome $m_t(X) = \mathbb{E}(Y_t \mid X)$ on the single index $u = \boldsymbol{\alpha}^\top X$,

$$g_t(u) = \mathbb{E}(m_t(X) \mid \boldsymbol{\alpha}^\top X = u) \quad (t = 1, \ldots, K), \tag{3}$$

for each given $\boldsymbol{\alpha}$. Specifically, suppose the true treatment-specific model can be expressed as

$$Y_t = m_t(X) + \sigma_t(X)\epsilon \quad (t = 1, \ldots, K), \tag{4}$$

in which $\mathbb{E}(\epsilon \mid X) = 0$, $\mathbb{E}(\epsilon^2 \mid X) = 1$. Let $R(\boldsymbol{\alpha}) = \sum_{t=1}^K \pi_t \mathbb{E}\left(Y_t - g_t(\boldsymbol{\alpha}^\top X)\right)^2$, where g_t is defined in (3) and let

$$\boldsymbol{\alpha}_0 := \underset{\boldsymbol{\alpha} \in \Theta}{\arg\min} \quad R(\boldsymbol{\alpha}). \tag{5}$$

Then $\boldsymbol{\alpha}_0$ can be shown to be the minimizer of the cross-entropy (e.g., Mackay, 2003) between the SIMML (2) and the general model (4) under the Gaussian noise assumption. Here, the cross-entropy of an arbitrary distribution with probability density f, with respect to another reference distribution $\mathcal{P}$ is defined as $\mathbb{E}_{\mathcal{P}}(-\log f)$, where the expectation is take with respect to the distribution $\mathcal{P}$. Model (3) evaluated at $\boldsymbol{\alpha}_0$ can be viewed as the "projection" (in the sense of the closest point) of the true distribution $\mathcal{P}$ (4) onto the space Θ of the SIMML distribution, using the Kullback–Leibler divergence as a distance measure.

The SIMML (2) allows a visualization useful for characterizing differential treatment effects, varying with the single-index $\boldsymbol{\alpha}^\top X$. As $X \in \mathbb{R}^p$ varies, the mean response of model (2) changes only in the specific direction $\boldsymbol{\alpha} \in \Theta$, and the effect of varying X, described by the link functions g_t, is different for each treatment condition $t \in \{1, \ldots, K\}$. Therefore, the single-index can be viewed as a useful biosignature for describing differential treatment effects, provided that $g_t \neq g_{t'}$ for at least one pair $t, t' \in \{1, \ldots, K\}$.

3. Estimation

While any smoothing technique can be used to approximate the unspecified smooth links $g_t(\cdot)$ in (2), in this paper, we will focus on cubic B-splines. Specifically, $g_t(u) \approx \eta_t^\top Z_t(u)$, for some coefficients $\eta_t \in \mathbb{R}^{d_t}$. Here, $Z_t(u) = \left[B_1(u), \quad \ldots, \quad B_{d_t}(u)\right]^\top \in \mathbb{R}^{d_t}$ consists of a set of d_t normalized cubic B-spline basis functions (de Boor, 2001). Let n_t be the sample size for the tth treatment group and $n = \sum_{t=1}^K n_t$ denote the total sample size. Note, d_t depends on n_t (see Assumption 5 and Wang and Yang, 2009). For a given $\boldsymbol{\alpha}$, let $\mathbb{Z}_{\boldsymbol{\alpha},t}$ denote the B-spline evaluation matrix $(n_t \times d_t)$, so that the ith row is $Z_t(\boldsymbol{\alpha}^\top X_{ti})^\top$, which is the B-spline evaluation of the ith individual from the tth treatment group. The subscript $\boldsymbol{\alpha}$ in the matrix $\mathbb{Z}_{\boldsymbol{\alpha},t}$ highlights its dependence on $\boldsymbol{\alpha}$. Without loss of generality we assume that the outcome and the covariates are all centered at zero for each treatment group t, so that the model does not involve any intercept terms.

For sample data, SIMML (2) can be represented by

$$\left[\boldsymbol{Y}\right]_{n\times 1} = \left[\mathbb{Z}_{\boldsymbol{\alpha}}\right]_{n\times(\sum_{t=1}^K d_t)} \left[\boldsymbol{\eta}\right]_{(\sum_{t=1}^K d_t)\times 1} + \left[\boldsymbol{\epsilon}\right]_{n\times 1}, \tag{6}$$

where $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_K^\top\right]^\top$ is the observed response vector with $\boldsymbol{Y}_t \in \mathbb{R}^{n_t}$, $\mathbb{Z}_{\boldsymbol{\alpha}}$ is $n \times (\sum_{t=1}^K d_t)$ block-diagonal B-spline design matrix of the $\mathbb{Z}_{\boldsymbol{\alpha},t}$'s, $\boldsymbol{\eta} = \left[\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_K^\top\right]^\top$ is the B-spline coefficient vector, and $\boldsymbol{\epsilon} = \left[\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_K^\top\right]^\top$ is a mean zero noise vector with covariance matrix $\sigma^2 \boldsymbol{I}_n$.

For a given $\boldsymbol{\alpha}$, we define the $n \times n$ single-index projection matrix to be $\mathbb{S}_{\boldsymbol{\alpha}} = \mathbb{Z}_{\boldsymbol{\alpha}} \left(\mathbb{Z}_{\boldsymbol{\alpha}}^T \mathbb{Z}_{\boldsymbol{\alpha}}\right)^{-1} \mathbb{Z}_{\boldsymbol{\alpha}}^T$. Assuming Gaussian noise and treating $\boldsymbol{\eta}$ as a nuisance parameter, the negative "profile" loglikelihood of $\boldsymbol{\alpha}$, up to a constant multiplier, is

$$Q(\boldsymbol{\alpha}) = \|\boldsymbol{Y} - \mathbb{S}_{\boldsymbol{\alpha}} \boldsymbol{Y}\|^2. \tag{7}$$

We define the profile likelihood estimator of the index parameter $\boldsymbol{\alpha}$ as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}\in\Theta}{\arg\min} \quad Q(\boldsymbol{\alpha}). \tag{8}$$

Each link function $g_t(\cdot)$ in (2) can be estimated by

$$\hat{g}_t(u) = Z_t(u)^\top \left(\mathbb{Z}_{\hat{\boldsymbol{\alpha}},t}^T \mathbb{Z}_{\hat{\boldsymbol{\alpha}},t}\right)^{-1} \mathbb{Z}_{\hat{\boldsymbol{\alpha}},t}^T \boldsymbol{Y}_t \quad (t = 1, \ldots, K), \tag{9}$$

where $\mathbb{Z}_{\hat{\boldsymbol{\alpha}},t}$ is $\mathbb{Z}_{\boldsymbol{\alpha},t}$ evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$.

To solve (8), we can perform a procedure that alternates between the following two steps: first, for a fixed $\boldsymbol{\alpha}$, estimate each link function $g_t(\cdot)$ in (2) by (9), where $\hat{\boldsymbol{\alpha}}$ is taken at $\boldsymbol{\alpha}$; second, for a fixed $\hat{g}_t(u)$, perform iteratively reweighted least squares (IRLS) to approximately solve (8) for $\boldsymbol{\alpha}$. These two steps can be iterated until convergence.

4. Asymptotic theory

In this section, we establish the asymptotic results of the profile estimator $\hat{\boldsymbol{\alpha}}$ in (8) under possible misspecification, when the true model is assumed to be (4). Let us denote the pth component of the vector $\boldsymbol{\alpha}_0$ in (5) by $\alpha_{0,p} (> 0$, since $\boldsymbol{\alpha}_0 \in \Theta)$. By the completeness property of $\mathbb{R}$, we can always find some $c > 0$ such that $\alpha_{0,p} \geq c$, and therefore, without loss of generality, we may assume that $\boldsymbol{\alpha}_0$ is in a compact set $\Theta_c = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top \in \mathbb{R}^p | \sum_{j=1}^p \alpha_j^2 = 1, \alpha_p \geq c\}$, with an appropriate choice of small $c > 0$. Further, to avoid the complication from the restricted parameter space Θ_c, we can consider instead the "pth component removed" $R(\boldsymbol{\alpha})$ in (5), as follows:

$$R(\boldsymbol{\alpha}_{-p}) = R\left(\alpha_1, \ldots, \alpha_{p-1}, \sqrt{1 - (\alpha_1^2 + \cdots + \alpha_{p-1}^2)}\right), \tag{10}$$

where a vector $\boldsymbol{\alpha}_{-p} = (\alpha_1, \alpha_2, \ldots, \alpha_{p-1}) \in \mathbb{R}^{p-1}$ lives inside the unit *ball*. Let the "pth component removed" value of $\boldsymbol{\alpha}_0$ in (5) be denoted by $\boldsymbol{\alpha}_{0,-p} \in \mathbb{R}^{p-1}$. Similarly, let the "$p$th component removed" value of the corresponding profile estimator $\hat{\boldsymbol{\alpha}}$ in (8) be denoted by $\hat{\boldsymbol{\alpha}}_{-p} \in \mathbb{R}^{p-1}$. The following conditions are assumed for the asymptotic results.

Assumption 1. The objective function $R(\boldsymbol{\alpha}_{-p})$ in (10) is locally convex at $\boldsymbol{\alpha}_{0,-p}$, and its Hessian function, $H(\boldsymbol{\alpha}_{-p})$ evaluated at $\boldsymbol{\alpha}_{-p} = \boldsymbol{\alpha}_{0,-p}$, is positive definite, with bounded eigenvalues.

Assumption 2. The underlying mean functions $m_t(X)$ in (4) are in $C^{(4)}(B_a^p)$, $t \in \{1, \ldots, K\}$ for some finite $a > 0$, where B_a^p is the p-dimensional ball with center 0 and radius a and $C^{(q)}(B_a^p) = \left\{f \mid \text{the } q\text{th order partial derivatives of } f \text{ are continuous in } B_a^p\right\}$.

Assumption 3. The probability density function of $X, f_X(x) \in C^{(4)}(B_a^p)$, and there exist constants $0 < c_f < C_f$ such that $c_f/\mathrm{Vol}_p(B_a^p) \leq f_X(x) \leq C_f/\mathrm{Vol}_p(B_a^p)$, if $x \in B_a^p$, and $f_X(x) = 0$, if $x \notin B_a^p$.

Assumption 4. The underlying noise ϵ in (4) satisfies $\mathbb{E}(\epsilon \mid X) = 0$ with $\mathbb{E}(\epsilon^2 \mid X) = 1$, and there exists a constant $C_\epsilon > 0$, such that $\sup_{x \in B_a^p} \mathbb{E}(|\epsilon|^3 \mid X = x) < C_\epsilon$. For each group $t \in \{1, \ldots, K\}$, the standard deviation function $\sigma_t(x)$ is continuous in B_a^p, with $0 < c_{\sigma_t} \leq \inf_{x \in B_a^p} \sigma_t(x) \leq \inf_{x \in B_a^p} \sigma_t(x) \leq C_{\sigma_t} < \infty$, for some constants $0 < c_{\sigma_t} < C_{\sigma_t}$.

Assumption 5. The number of interior knots, $N_t (= d_t - 4)$, in the cubic B-spline approximation of the link function $g_t(\cdot)$ for the tth treatment group satisfies: $n_t^{1/6} \ll N_t \ll n_t^{1/5} (\log(n_t))^{-(2/5)}$, $t \in \{1, \ldots, K\}$.

The first theorem establishes consistency of the estimator (8) and the second theorem establishes asymptotic normality of the estimator $\hat{\boldsymbol{\alpha}}_{-p}$ for $\boldsymbol{\alpha}_{0,-p}$.

Theorem 1 (*Consistency*). *Under Assumptions 1 to 5, $\hat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}_0$ almost surely, where $\boldsymbol{\alpha}_0$ is defined in (5).*

Theorem 2 (*Asymptotic Normality*). *Under Assumptions 1 to 5, $\sqrt{n}(\hat{\boldsymbol{\alpha}}_{-p} - \boldsymbol{\alpha}_{0,-p}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{0,-p}})$ in distribution, with asymptotic covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{0,-p}} = \boldsymbol{H}^{-1}_{\boldsymbol{\alpha}_{0,-p}} \boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}} \boldsymbol{H}^{-1}_{\boldsymbol{\alpha}_{0,-p}}$, where the matrix $\boldsymbol{H}_{\boldsymbol{\alpha}_{0,-p}}$ is the Hessian matrix $\boldsymbol{H}(\boldsymbol{\alpha}_{-p}) = \frac{\partial^2}{\partial \boldsymbol{\alpha}_{-p} \partial \boldsymbol{\alpha}_{-p}^T} R(\boldsymbol{\alpha}_{-p})$ evaluated at $\boldsymbol{\alpha}_{-p} = \boldsymbol{\alpha}_{0,-p}$, and the matrix $\boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}}$ is defined in Appendix.*

The proofs of the theorems are given in Appendix.

5. Simulation illustrations

5.1. Performance on estimating treatment decision rules

A treatment decision function, $\mathcal{D}(X) : \mathbb{R}^p \mapsto \{1, \ldots, K\}$, mapping a subject's baseline characteristics $X \in \mathbb{R}^p$ to one of K available treatments, defines a treatment decision rule for the single decision time point (Murphy, 2003; Robins, 2004; Zhang et al., 2012; Cai et al., 2011; Qian and Murphy, 2011). Given covariates X, a treatment decision rule based on SIMML is $\mathcal{D}(X) = \arg\max_{t \in \{1,\ldots,K\}} g_t(\boldsymbol{\alpha}^\top X)$. We investigate the performance of the estimated treatment decision rules of the form $\mathcal{D}(X) = \arg\max_{t \in \{1,\ldots,K\}} \mathbb{E}\big(Y \mid X, T = t\big)$, where the conditional expectation is obtained from various modeling procedures.

In our simulation settings, the baseline covariate vector $X = (x_1, \ldots, x_p)^\top \sim \mathcal{N}(0, \Psi_X)$, with Ψ_X having $1's$ on the diagonal and 0.1 everywhere else. We consider $K = 2$ with different noise levels for the two treatment groups: $\epsilon_1 \sim \mathcal{N}(0, 0.4^2)$, $\epsilon_2 \sim \mathcal{N}(0, 0.2^2)$. The outcome data are generated under the following fairly broad model

$$Y_t = \delta M(\boldsymbol{\mu}^\top X; \nu) + C_t(\boldsymbol{\alpha}^\top X; \omega) + \epsilon_t \quad (t = 1, 2). \tag{11}$$

As a function of the index $\boldsymbol{\mu}^\top X$, M is referred to as the "main effect" of X. As functions of the other index $\boldsymbol{\alpha}^\top X$, the C_t's are referred to as the "contrast" functions that define the treatment-by-X interaction. Here, we will use the parameters ν and ω to control the degree of non-linearity of M and C_t's, respectively.

An optimal treatment decision rule depends only on the C_t's, not on M or the ϵ_t's. The parameter δ in (11) controls the relative contribution of the "signal" component $C_t's$ to the variance in the outcomes, and is calibrated to obtain a relative contribution of 0.35. The contrast functions C_t's in (11) are set to

$$C_t(u; \omega) = \begin{cases} C_1(u; \omega) = +1 - \cos\big(0.5\pi\omega u\big) + 0.5(u - \omega) \\ C_2(u; \omega) = -1 + \cos\big(0.5\pi\omega u\big) - 0.5(u - \omega), \end{cases} \tag{12}$$

where, if $\omega = 0$, then the C_t's are linear functions; and they are more nonlinear for larger values of ω. We considered three cases, corresponding to *linear* ($\omega = 0$), *moderately nonlinear* ($\omega = 0.5$), and *highly nonlinear* ($\omega = 1$) C_t's, respectively, illustrated in the first three panels of Fig. 1. We set the main effect function M in (11) to be

$$M(u; \nu) = 0.5u - \sin(0.5\pi\nu u),$$

where, as ν increases, the degree of nonlinearity in the main effect function M increases. We considered two cases, $\nu = 0$, corresponding to a *linear* M; and $\nu = 1$, corresponding to a *nonlinear* M, illustrated in the fourth and the fifth panel of Fig. 1. We set $p = 5$ and $p = 10$ with $\boldsymbol{\alpha} = (1, \ldots, 5)^\top$ and $\boldsymbol{\alpha} = (1, \ldots, 10)^\top$, respectively, each standardized to have norm one. We set μ to be proportional to a vector of 1's, standardized to have norm one. Two treatment groups were considered, with equal sample sizes $n_1 = n_2 = 40$. We used $d_1 = d_2 = 5$ B-spline basis functions to approximate the link functions. The treatment decision rules were based on the following regression models: (i) SIMML (2) estimated from maximizing the profile likelihood; (ii) the K-Index model (1) fitted separately for each treatment group by the B-spline approach of Wang and Yang (2009), denoted as K-Index; (iii) the linear GEM model (Petkova et al., 2016) estimated under the criterion of maximizing the difference in the treatment-specific slope, denoted as linGEM; and (iv) linear regression models fitted separately for each treatment group under the least squares criterion, denoted as K-LR. For each scenario, using the outcome Y from a simulated test set (of size 10^5), we computed the proportion of correct decisions (PCD) of the treatment decision rules estimated from each method and the methods were compared in terms of PCD using boxplots from 200 training datasets.

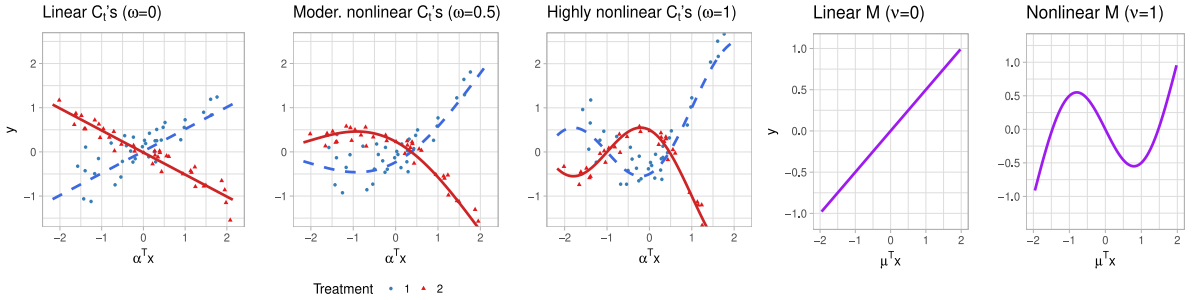

Fig. 1. The first panel shows the *linear* contrast C_t's ($\omega = 0$), the second panel the *moderately nonlinear* contrast C_t's ($\omega = 0.5$), and the third panel displays *highly nonlinear* contrast C_t's ($\omega = 1$). Data points are generated from model (11) with $\delta = 0$ and $p = 5$. The fourth and the fifth panels show the *linear* ($\nu = 0$) and the *nonlinear* main effect M ($\nu = 1$), respectively.

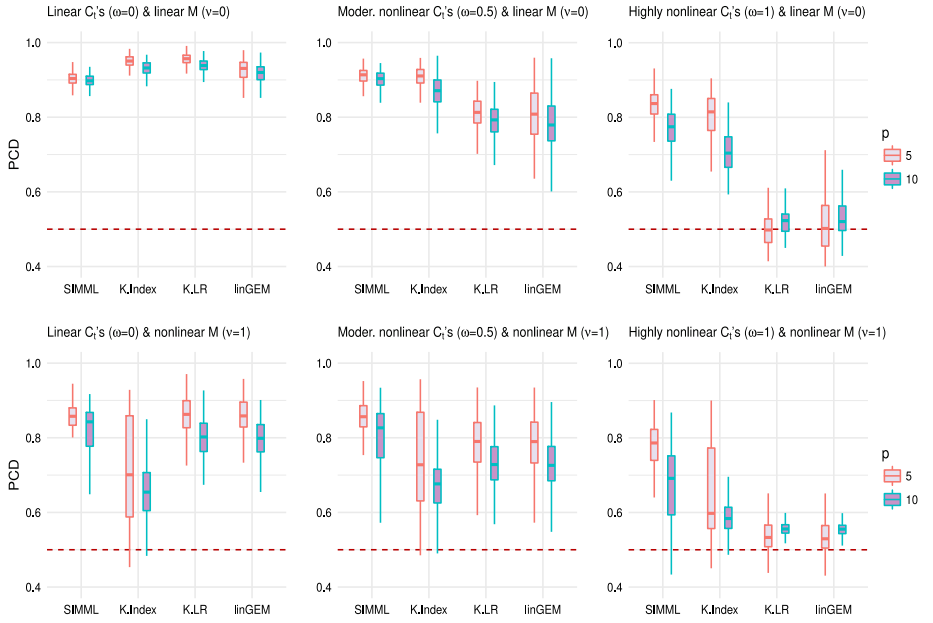

Fig. 2. Boxplots of the proportion of correct decisions (PCD) of the treatment decision rules obtained from 200 training datasets for each of the four methods. Each panel corresponds to one of the six combinations of $\omega \in \{0, 0.5, 1\}$ and $\nu \in \{0, 1\}$: the shape of the contrast functions C_t's controlled by ω; the shape of the main effect function M controlled by ν; the number of predictors $p \in \{5, 10\}$. The sample sizes are $n_1 = n_2 = 40$.

Fig. 2 shows that SIMML outperforms all other methods, except for the case under the *linear* M and C_t's in which all 4 approaches perform well. The K-Index model is clearly second best, under the linear M ($\nu = 0$) (the top panels) with the nonlinear C_t's ($\omega = 0.5$ and $\omega = 1$). However, with a more complex M function ($\nu = 1$) (the bottom panels), the performance of the K-index approach is considerably worse compared to SIMML. Given a relatively small sample size and under the complex main effect, the SIMML that emphasizes the treatment contrasts through the common single-index is more effective in estimating optimal treatment decisions than the K-Index model. As would be expected, additional complexity in the contrasts C_t's ($\omega = 0.5$ and $\omega = 1$) has a greater effect on the performance of the more restrictive models (linGEM and K-LR) than it does on the flexible models (SIMML and K-index). The number of covariates, p, also has a clear impact on the performance of all methods. As p changes from 5 (red) to 10 (blue), the deterioration in performance is more pronounced for the K-Index model that requires separate fits for each treatment and thus involves estimation of more parameters ($K(p-1)+Kd$), compared to the more parsimonious SIMML with a fewer number of parameters ($p-1+Kd$) to be estimated.

5.2. Coverage probability of asymptotic 95% confidence intervals

The next simulation experiment assesses the coverage probability of the asymptotic confidence intervals derived from Theorem 2. The data were generated under model (11) with $\delta = 0$ (i.e., no main effect M) with $p = 5$ covariates. We set the SIMML index vector $\boldsymbol{\alpha}(= \boldsymbol{\alpha}_0)$ to be stepwise increasing: $(1, \ldots, 5)^\top$, normalized to have unit L^2 norm. The associated

Table 1
Depression randomized clinical trial: Description of the $p = 9$ baseline covariates (means and SDs); the estimated values ("Indiv. Value") of treatment decision rules from each individual covariate, using either the B-spline regression ("Nonpar.", in the third column) or the linear regression ("Linear", in the fourth column); the estimated single-index coefficients (in the last three columns), and the values of the associated treatment decision rules from the three methods (in the bottom row).

(Label) Baseline	Mean	Indiv. value		Coefficients α_j's, $j \in \{1, \ldots, 9\}$		
Patient characteristics	(SD)	Nonpar.	Linear	SIMML*	SIMML	linGEM
(x_1) Age at evaluation	38.00 (13.84)	8.56	8.24	−0.53	−0.50	−0.43
(x_2) Severity of depression	18.80 (4.29)	6.85	7.07	−0.07	−0.13	−0.37
(x_3) Dur. MDD (month)	38.19 (53.17)	7.42	7.33	0.08	−0.18	0.20
(x_4) Age at MDD	16.46 (6.09)	6.29	6.95	0.23	0.05	0.31
(x_5) Axis II	3.92 (1.43)	7.16	7.11	0.23	0.20	0.17
(x_6) Word Fluency	37.42 (11.68)	7.64	7.11	0.11	0.09	0.27
(x_7) Flanker RT	59.51 (26.63)	8.19	8.39	0.12	0.23	−0.18
(x_8) Post-conflict adjus.	0.07 (0.12)	6.73	7.23	−0.30	−0.29	−0.18
(x_9) Flanker Accuracy	0.22 (0.15)	7.89	8.37	0.70	0.70	0.59
Value from single-index model				9.34	8.72	8.22

contrast functions, C_t's, are given by (12). As in Section 5.1, we consider three levels of the curvature of the contrasts, corresponding to *linear* ($\omega = 0$), *moderately nonlinear* ($\omega = 0.5$), and *highly nonlinear* ($\omega = 1$) contrasts (see Fig. 1). In (11), the standard deviations of the noise ϵ_t were set to 0.5. We set the sample size $n = n_1 + n_2$ with $n_1 = n_2$. With varying $n \in \{50, 100, 200, 400, 800, 1600, 3200\}$, the number of interior knots used in the B-spline approximation, N_t, was determined to be $N_t = \left[n_t^{1/5.5}\right]$, as recommended by Wang and Yang (2009) ($[v]$ denotes the integer part of v). Two hundred datasets were generated for all combinations of n and ω. For each (i.e., the jth) component α_j of $\boldsymbol{\alpha}$, the proportion of times the 95% asymptotic confidence interval contains the true value of α_j was recorded in Table C.2 in Appendix. Notice that the 5th (i.e., the pth) element is estimated to satisfy the constraint $\boldsymbol{\alpha} \in \Theta$ in Theorem 2. To obtain the confidence intervals for the 5th component, we applied Theorem 2 with the 4th component removed (without loss of generality), and obtained the confidence intervals for the 5th component.

We note that the choice of $N_t = \left[n_t^{1/5.5}\right]$ is an approximation to the N_t of Assumption 5 which requires $n_t^{1/6} \ll N_t \ll n_t^{1/5}(\log(n_t))^{-(2/5)}$, as such N_t can only be obtained for a very large n_t. Nevertheless, in Table C.2 in the Appendix, as the sample size $n(= n_1 + n_2)$ increases, the "actual" coverage probability gets closer to the "nominal" coverage probability, with better coverage results for the *linear* and the *moderately nonlinear* contrasts ($\omega \in \{0, 0.5\}$) compared to the *highly nonlinear* contrasts ($\omega = 1$).

6. Application to data from a randomized clinical trial

Major depressive disorder afflicts millions and, according to the World Health Organization, it is the leading cause of disability worldwide. It is a highly heterogeneous disorder, however, no individual biological or clinical marker has demonstrated sufficient ability to match individuals to efficacious treatment. Here we illustrate the utility of the proposed SIMML method for estimating a composite biomarker and treatment decision rules, with an application to data from a randomized clinical trial comparing an antidepressant and placebo for treating depression.

Of the 166 subjects, 88 were randomized to placebo and 78 to the antidepressant. In addition to standard clinical assessments, patients underwent neuropsychiatric testing prior to treatments. Table 1 summarizes the information on $p = 9$ baseline patient characteristics, $X = (x_1, \ldots, x_9)^\top$. These baseline covariates were considered as potential treatment effect modifiers, and standardized to have unit variance. The treatment outcome Y was the improvement in symptom severity from week 0 (baseline) to week 8 and thus larger values of the outcome were better.

Fig. 3 shows the treatment outcomes Y against each of the 9 baseline covariates, for the placebo group (blue) and the active drug group (red). The estimated B-spline approximated curves for each individual covariate are shown with the associated 95% confidence bands: the solid blue curves for the placebo group and the dotted red curves for the active drug group. In Fig. 3, each individual covariate has at most a small treatment modifying effect, as its treatment-specific curves do not differ much.

One natural measure for the effectiveness of a treatment decision rule $\mathcal{D}$ is called the "value" (V) of a treatment decision rule $\mathcal{D}$ (Qian and Murphy, 2011), which is defined as the expected mean outcome if everyone in the population receives treatment according to that rule:

$$\mathrm{V}(\mathcal{D}) = \mathbb{E}_X\left(\mathbb{E}_{Y|X}\left(Y \mid X, T = \mathcal{D}(X)\right)\right). \tag{13}$$

In the third and the fourth column of Table 1, "Indiv. Value" refers to the estimated "value" of a decision rule $\mathcal{D}$ estimated from each of the 9 individual covariates, using the following two approaches for estimating $\mathcal{D}$: the B-spline regressions of the treatment-specific outcome on each individual covariate ("Nonpar." in the third column of Table 1) as suggested by the overlaid curves in Fig. 3, and the linear regressions of the treatment-specific outcome on each individual covariate

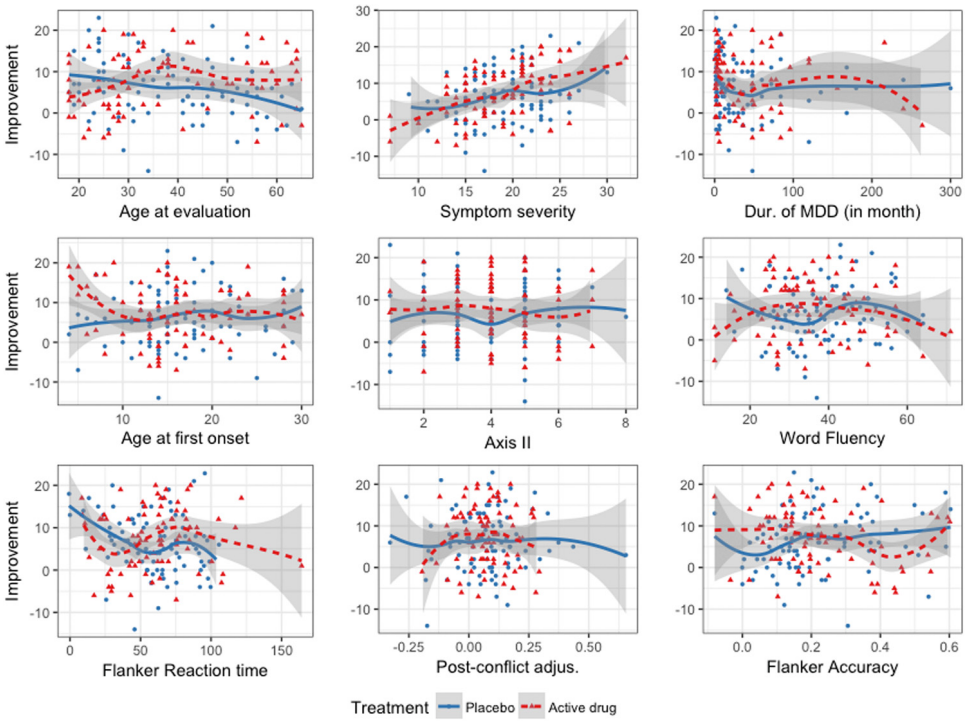

Fig. 3. Depression randomized clinical trial: For each of the 9 baseline covariates individually, treatment-specific spline approximated regression curves with 5 basis functions are overlaid on to the data points; the placebo group is the blue solid curve and the active drug group is the red dotted curve. The associated 95% confidence bands of the regression curves were also plotted.

("Linear" in the fourth column of Table 1). The value (13) of $\mathcal{D}$ can be estimated by the inverse probability weighted estimator (Murphy, 2005):

$$\hat{V}(\mathcal{D}) = \sum_{i=1}^{\tilde{n}} Y_i I_{T_i=\mathcal{D}(X_i)} / \sum_{i=1}^{\tilde{n}} I_{T_i=\mathcal{D}(X_i)}, \tag{14}$$

using a testing set, say, $\{(Y_i, X_i, T_i), i = 1, \ldots, \tilde{n}\}$, where, if one uses only the jth covariate for estimating $\mathcal{D}$, then $X_i = x_{ij}$. The data were randomly split into a training set and a testing set with a ratio of 10 to 1. This splitting was performed 500 times, each time estimating $\mathcal{D}$ on the training set and computing (14) from the testing set. Values (14) are averaged over the 500 splits.

The SIMML can be made more efficient by incorporating a main effect component $\boldsymbol{\beta}^\top D(X)$ in the model, i.e., we can consider $\mathbb{E}(Y \mid T = t, X = x) = \boldsymbol{\beta}^T D(x) + g_t(\boldsymbol{\alpha}^\top x)$, for an appropriate vector-valued function $D(X)$. If the $n \times q$ matrix $\mathbb{D}$ denotes the evaluation of $D(X)$ on the sample data, then for each $\boldsymbol{\alpha}$, the negative "profile" loglikelihood (7) under this extended model (with Gaussian outcome), up to constants, is $Q^*(\boldsymbol{\alpha}) = \|\tilde{\boldsymbol{Y}} - \mathbb{S}_{\boldsymbol{\alpha}}\tilde{\boldsymbol{Y}}\|^2$, where $\tilde{\boldsymbol{Y}} = \left(\boldsymbol{I}_n - (\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}})\,\mathbb{D}\left(\mathbb{D}^T\mathbb{D}\right)^{-1}\mathbb{D}^T\right)\boldsymbol{Y}$. In this analysis, we took $D(X) = X$. We refer to this approach as "main effect adjusted" profile likelihood SIMML and denote it by SIMML*.

In Table 1, the last three columns show the estimated single-index coefficients $\boldsymbol{\alpha}$ obtained by two different SIMMLs (SIMML* and SIMML) and the linear GEM (linGEM) which restricts the link $g_t(\cdot)$ to be a linear function. In Fig. 4, the estimated pairs of link functions are plotted against the approach-specific single-index $\boldsymbol{\alpha}^\top X$, obtained from applying the two SIMML approaches and the linear GEM approach. From Figs. 3 and 4, it appears that the index $\boldsymbol{\alpha}^\top X$ exhibits a stronger moderating effect of treatment than the individual covariates. Also, the shapes of the regression curves from the SIMML approaches appear to capture a nonlinear treatment-by-index interaction effect, especially due to some non-monotone relationship between the index and the outcome in the active drug group.

In Fig. 5, we illustrate the single-index coefficient estimates from each of the methods, and the associated 95% confidence intervals obtained from a bias-corrected and accelerated (BC_a, DiCiccio and Efron, 1996) bootstrap with 500 replications. The coverage of the asymptotic-based confidence intervals for this sample size is not expected to be very good (based on the simulation results in Section 5.2) and thus instead we used bootstrap confidence intervals. The magnitude of the estimated coefficients $\alpha_1, \ldots, \alpha_9$ reflects the relative importance of the covariates $x_1, \ldots, x_9$ in determining a composite treatment effect modifier $\boldsymbol{\alpha}^\top X$.

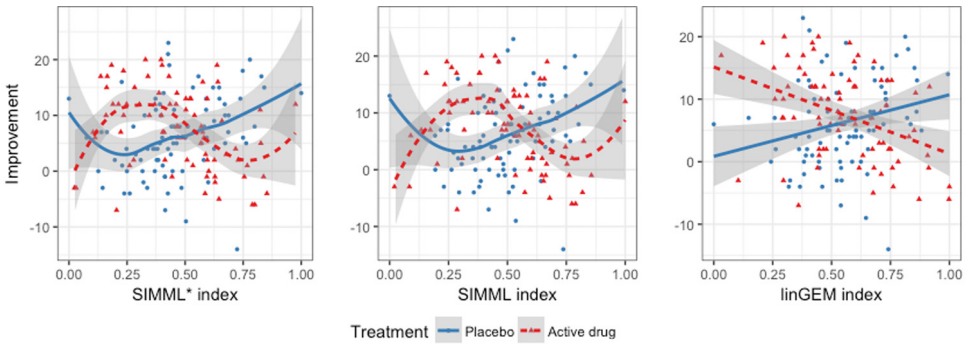

Fig. 4. Depression randomized clinical trial: Pair of estimated link functions (g_1 and g_2) obtained from SIMML with the "main effect adjusted" profile likelihood (first panel), SIMML with the (main effect un-adjusted) profile likelihood (second panel), and the linear GEM model estimated under the criterion maximizing the difference in the linear regression slopes (third panel), respectively, for the placebo group (blue solid curves) and the active drug group (red dotted curves). The 95% confidence bands were constructed conditioning on the single-index coefficient $\boldsymbol{\alpha}$. For each treatment group, the observed outcomes are plotted against the estimated single-index.

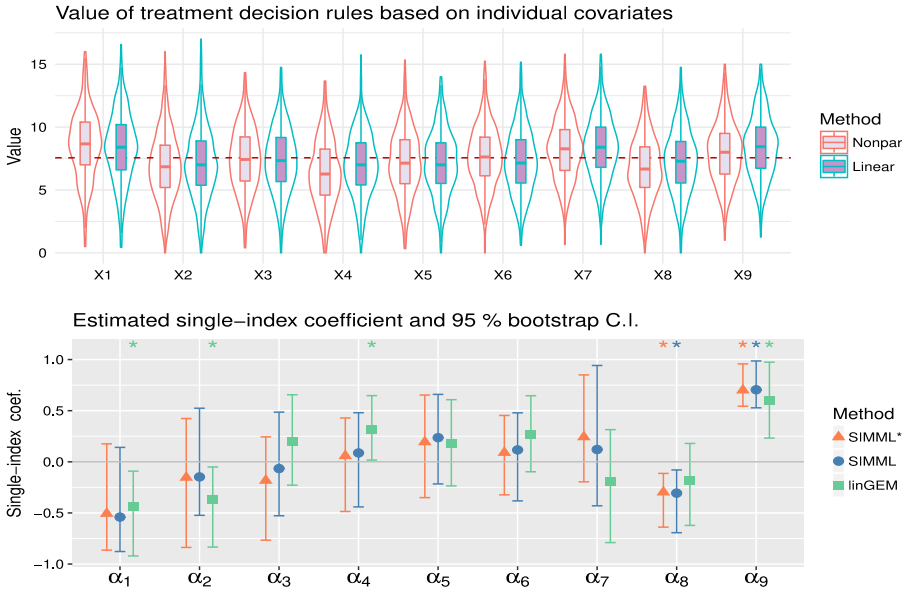

Fig. 5. Depression randomized clinical trial: Top panel: Violin plots of the estimated values of treatment decision rules based on each of the individual covariates $x_1, \ldots, x_9$, determined from univariate nonparametric and linear regressions, respectively, obtained from 500 randomly split testing sets (with higher values preferred). Bottom panel: The estimated index coefficients $\alpha_1, \ldots, \alpha_9$, associated with the covariates $x_1, \ldots, x_9$, and the 95% confidence intervals for each of the three methods, obtained from BC_a bootstrap with 500 replications. An estimated significant coefficient is marked with $*$ on top of each confidence interval.

In this analysis, the incorporation of the "main effect" component improved the value of treatment decision rules determined from the proposed SIMML method, as illustrated in the boxplots in Fig. 6; we compared the two SIMML approaches (SIMML* and SIMML); the linear GEM (linGEM) and the two approaches based on separate regression models for each treatment group (K-Index and K-LR), with respect to the estimated values (14) of the treatment decision rules. For comparison, we also included the decision to treat everyone with placebo (All PBO), and the decision to treat everyone with the active drug (All DRG). The results are summarized in Fig. 6.

In Fig. 6, in terms of the averaged estimated values (14) estimated from the aforementioned 500 randomly split testing sets, the proposed SIMML approaches outperform all other methods. The visualization (see Fig. 4) indicates that the superiority of the active drug over placebo does not linearly decrease with the index, but rather, it appears to remain relatively constant to the left of the crossing point, exhibiting some nonlinear patterns. Finally, we note that the value of the treatment decision rule All PBO was lower than the value of the treatment decision rule All DRG, and that all treatment decision rules that took patient characteristics into account outperformed the decision of treating everyone with the drug (which is standard current clinical practice). In particular, the superiority the treatment decision rule SIMML* over treating everyone with the drug in terms of value was of similar magnitude of the superiority of the decision to treat everyone with

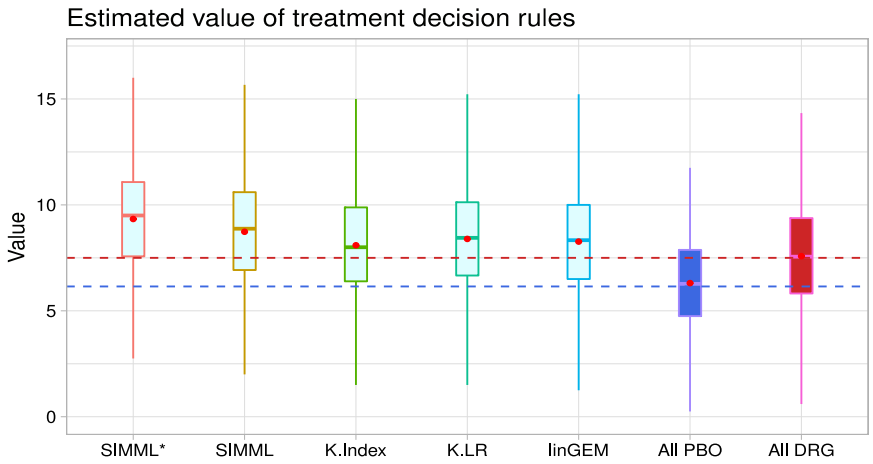

Fig. 6. Depression randomized clinical trial: Boxplots of the estimated values of treatment decision rules, obtained from the 500 randomly split testing sets (higher values are preferred). The estimated values (and the standard deviations) are given as follow. SIMML*: 9.34 (2.68); SIMML: 8.72 (2.68); K-Index: 8.04 (2.69); K-LR: 8.36 (2.69); linear GEM (linGEM): 8.22 (2.67); All placebo (PBO): 6.17 (2.63); All drug (DRG): 7.57 (2.67).

the drug versus treating everyone with placebo. This is a clear indication that patient characteristics can help treatment decisions for patients with depression, and the more flexible SIMML methods are well suited for developing treatment decision rules. Particularly, the proposed methods show that combining patient characteristics with little moderating effects of a treatment can result in a strong treatment effect modifier which exhibits nonlinear association with the outcome that can help with making treatment decisions.

7. Discussion

The SIMML model (6) can be extended in various ways, for example, by allowing treatment-specific noise variances σ_t^2. Under a Gaussian noise assumption, the B-spline approximated profile log likelihood of $\boldsymbol{\alpha}$, that profiles out the nuisance parameters σ_t^2 and $\boldsymbol{\eta}_t$, up to constants, is $\sum_{t=1}^K n_t \log Q_t(\boldsymbol{\alpha})$, in which $Q_t(\boldsymbol{\alpha}) = \|\boldsymbol{Y}_t - \mathbb{S}_{\boldsymbol{\alpha},t}\boldsymbol{Y}_t\|^2/n_t$. The corresponding profile estimator of $\boldsymbol{\alpha}$ is $\arg\min_{\boldsymbol{\alpha}\in\Theta} \sum_{t=1}^K n_t \log Q_t(\boldsymbol{\alpha})$. The estimation can be performed similarly as in the estimation of $\hat{\boldsymbol{\alpha}}$ in (8), but the criterion function $Q(\boldsymbol{\alpha})$ will be replaced by $\sum_{t=1}^K n_t \log Q_t(\boldsymbol{\alpha})$.

The SIMML can also be extended to generalized linear models (GLM) in which the outcome variable is a member of the exponential family. The standard form of the density is $f_Y(Y;\theta,\phi) = \exp\{(Y\theta - b(\theta))/a(\phi) + c(Y,\phi)\}$, with a canonical link function $h(\cdot)$. We can extend the SIMML approach to the GLM setting with treatment-specific natural parameters θ_t, $t \in \{1,\ldots,K\}$ by modeling the treatment-specific outcomes as a function of a single-index $\boldsymbol{\alpha}^\top X$: $\theta_t(x) = h^{-1}(\mathbb{E}(Y \mid T = t, X = x)) = g_t(\boldsymbol{\alpha}^\top x)$, $t \in \{1,\ldots,K\}$; $g_t(\cdot)$, hence $\theta_t(x) \in \mathbb{R}$, can be approximated, for example, by B-splines. The approximates can be denoted by $\tilde{\theta}_t(x) = \boldsymbol{\eta}_t^\top Z_t(\boldsymbol{\alpha}^\top x)$ for some $\boldsymbol{\eta}_t \in \mathbb{R}^{d_t}$. As in Section 3, the general strategy of nonlinear maximization of the "profile" likelihood over $\boldsymbol{\alpha} \in \Theta$, where we profile out $\boldsymbol{\eta}_t$ for each value of $\boldsymbol{\alpha}$, can be employed. The dispersion parameter ϕ can also be profiled out. Other potential extensions involve incorporating variable selection in high-dimensional covariate settings using a regularization method and incorporating functional-valued data objects (such as images) as patient covariates.

An important extension to the SIMML model is to factor out baseline effects common to all treatment groups, by allowing an unspecified main-effect term $\mu(X)$ (e.g., Tian et al., 2014) in the model. Generally, this can be handled by an "orthogonalization" approach, and the estimation can be performed under the framework of A-learning (Murphy, 2003; Lu et al., 2011; Shi et al., 2016, 2018; Jeng et al., 2018). To elaborate, consider the following extension of model (2),

$$\mathbb{E}(Y \mid T = t, X = x) = \mu(x) + g_t(\boldsymbol{\alpha}^\top x) \quad (t = 1,\ldots,K), \tag{15}$$

where we impose a structural constraint, $\mathbb{E}_T\left(g_T(\boldsymbol{\alpha}^\top X) \mid X\right) = \sum_{t=1}^K \pi_t g_t(\boldsymbol{\alpha}^\top X) = 0$, which is a sufficient condition for orthogonality between the SIMML, $g_T(\boldsymbol{\alpha}^\top X)$, and the unspecified main effect, $\mu(X)$ in (15), as in Jeng et al. (2018). Optimization of model (15) can be achieved by constrained least squares under this orthogonality constraint and A-learning can be employed for estimating an optimal treatment decision rule, focusing on estimating the interactions in the presence of the unspecified main effect $\mu(X)$. The technicalities of this adjustment are treated in a separate work.

Acknowledgment

This work was supported by National Institute of Health (NIH) grant 5 R01 MH099003.

Appendix A. The asymptotic covariance matrix in Theorem 2

Define $R_t(\boldsymbol{\alpha}) = \mathbb{E}_{Y,X|T=t}\left(Y - g_t(\boldsymbol{\alpha}^\top X)\right)^2$, $t \in \{1,\ldots,K\}$. In Theorem 2, the asymptotic covariance matrix is given as $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{0,-p}} = \boldsymbol{H}^{-1}_{\boldsymbol{\alpha}_{0,-p}} \boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}} \boldsymbol{H}^{-1}_{\boldsymbol{\alpha}_{0,-p}}$. Here, the Hessian matrix $\boldsymbol{H}_{\boldsymbol{\alpha}_{0,-p}} = \left[H_{j,q}\right]_{j,q=1}^{p-1}$ evaluated at $\boldsymbol{\alpha}_{-p} = \boldsymbol{\alpha}_{0,-p}$ has its (j,q)th element given by

$$H_{j,q} = \sum_{t=1}^{K} \pi_t \Bigg[\frac{\partial^2}{\partial\alpha_j \partial\alpha_q} R_t(\boldsymbol{\alpha}) - \frac{\alpha_j}{\alpha_p} \frac{\partial^2}{\partial\alpha_p \partial\alpha_q} R_t(\boldsymbol{\alpha}) - \frac{\alpha_q}{\alpha_p} \frac{\partial^2}{\partial\alpha_p \partial\alpha_j} R_t(\boldsymbol{\alpha}) - \frac{\alpha_j \alpha_q}{\alpha_p^3} \frac{\partial}{\partial\alpha_p} R_t(\boldsymbol{\alpha}) + \frac{\alpha_j \alpha_q}{\alpha_p^2} \frac{\partial^2}{\partial\alpha_p^2} R_t(\boldsymbol{\alpha}) \Bigg] \Bigg|_{\alpha=\alpha_0}. \tag{A.1}$$

The matrix $\boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}} = \left[W_{j,q}\right]_{j,q=1}^{p-1}$ evaluated at $\boldsymbol{\alpha}_{-p} = \boldsymbol{\alpha}_{0,-p}$ has its (j,q)th element given by

$$W_{j,q} = \sum_{t=1}^{K} \pi_t \mathbb{E}_{Y,X|T=t} \Bigg(\left\{ 2\left(g_t(u_{\boldsymbol{\alpha}}) - Y\right) \left(\frac{\partial}{\partial\alpha_j} g_t(u_{\boldsymbol{\alpha}}) - \frac{\alpha_j}{\alpha_p} \frac{\partial}{\partial\alpha_p} g_t(u_{\boldsymbol{\alpha}}) \right) + \frac{\partial}{\partial\alpha_j} R_t(\boldsymbol{\alpha}) - \frac{\alpha_j}{\alpha_p} \frac{\partial}{\partial\alpha_p} R_t(\boldsymbol{\alpha}) \right\} \times \left\{ 2\left(g_t(u_{\boldsymbol{\alpha}}) - Y\right) \left(\frac{\partial}{\partial\alpha_q} g_t(u_{\boldsymbol{\alpha}}) - \frac{\alpha_q}{\alpha_p} \frac{\partial}{\partial\alpha_p} g_t(u_{\boldsymbol{\alpha}}) \right) + \frac{\partial}{\partial\alpha_q} R_t(\boldsymbol{\alpha}) - \frac{\alpha_q}{\alpha_p} \frac{\partial}{\partial\alpha_p} R_t(\boldsymbol{\alpha}) \right\} \Bigg) \Bigg|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} \tag{A.2}$$

where $u_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^\top X$.

Appendix B. Proof

B.1. Proof of Theorem 1

Proof. Let us write $Q_t(\boldsymbol{\alpha}) = \|\boldsymbol{Y}_t - \mathbb{S}_{\boldsymbol{\alpha},t}\boldsymbol{Y}_t\|^2/n_t$ and $Q(\boldsymbol{\alpha}) = \sum_{t=1}^K n_t Q_t(\boldsymbol{\alpha})/n$. Under Assumptions 2–4, by the results from A.14 of Wang and Yang (2007), we have

$$\sup_{\boldsymbol{\alpha}\in\Theta_c} |Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})| \le O((n_t^{-1/2} h_t^{-1/2} \log n_t)^2 + (h_t^4)^2) + O(n_t^{-1/2} \log n_t h_t^{-1/2} + h^4)$$

almost surely, where $h_t = \frac{1}{N_t+1}$ is the distance between knot points, and N_t (note, $N_t = d_t - 4$) is the number of interior knots on $[0,1]$. Since we choose N_t such that $n_t^{1/6} \ll N_t \ll n_t^{1/5}(\log(n_t))^{-(2/5)}$ for all $t \in \{1,\ldots,K\}$, under Assumption 5,

$$\sup_{\boldsymbol{\alpha}\in\Theta_c} |Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})| \to 0 \quad t \in \{1,\ldots,K\},$$

almost surely. By the continuous mapping theorem,

$$\sup_{\boldsymbol{\alpha}\in\Theta_c} \left| \sum_{t=1}^K \frac{n_t}{n} Q_t(\boldsymbol{\alpha}) - \sum_{t=1}^K \pi_t R_t(\boldsymbol{\alpha}) \right| \le \sup_{\boldsymbol{\alpha}\in\Theta_c} \sum_{t=1}^K \left| \frac{n_t}{n} Q_t(\boldsymbol{\alpha}) - \pi_t R_t(\boldsymbol{\alpha}) \right| \to 0$$

almost surely, therefore, we have

$$\sup_{\boldsymbol{\alpha}\in\Theta_c} |Q(\boldsymbol{\alpha}) - R(\boldsymbol{\alpha})| \to 0, \tag{B.1}$$

almost surely. Denote by $(\Omega, \mathcal{F}, \mathcal{P})$ the probability space on which all $\{Y_i, T_i, X_i^\top\}_{i=1}^\infty$ are defined. By (B.1), for any $\delta > 0$, $\omega \in \Omega$, there is an integer $n^*(\omega)$, such that $Q(\boldsymbol{\alpha}_0, \omega) - R(\boldsymbol{\alpha}_0) < \delta/2$, whenever $n > n^*(\omega)$. Since $\hat{\boldsymbol{\alpha}}(\omega)$ is the minimizer of $Q(\boldsymbol{\alpha}, \omega)$, we have $Q(\hat{\boldsymbol{\alpha}}(\omega), \omega) - R(\boldsymbol{\alpha}_0) < \delta/2$. Also, by (B.1), there exists an integer $n^{**}(\omega)$, such that $R(\hat{\boldsymbol{\alpha}}(\omega), \omega) - Q(\hat{\boldsymbol{\alpha}}(\omega), \omega) < \delta/2$, whenever $n > n^{**}(\omega)$. Therefore, whenever $n > \max(n^*(\omega), n^{**}(\omega))$, we have $R(\hat{\boldsymbol{\alpha}}(\omega), \omega) - R(\boldsymbol{\alpha}_0) < \delta$. The strong consistency $\hat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}_0$ follows from the local convexity of Assumption 1. □

B.2. Proof of Theorem 2

Proof. We first derive the expression (A.1) from the Appendix for the Hessian matrix. We can write $R(\boldsymbol{\alpha}_{-p}) = \sum_{t=1}^K \pi_t R_t(\boldsymbol{\alpha}_{-p})$, where the "$p$th component removed" function corresponding to the tth treatment is $R_t(\boldsymbol{\alpha}_{-p}) = R_t\left(\alpha_1, \ldots, \alpha_{p-1}, \sqrt{1 - (\alpha_1^2 + \cdots + \alpha_{p-1}^2)}\right)$. Applying the chain rule for taking the derivative of $R_t(\boldsymbol{\alpha}_{-p})$ with respect to α_j, we obtain

$$\frac{\partial}{\partial\alpha_j} R_t(\boldsymbol{\alpha}_{-p}) = \frac{\partial}{\partial\alpha_j} R_t(\boldsymbol{\alpha}) - \frac{\alpha_j}{\alpha_p} \frac{\partial}{\partial\alpha_p} R_t(\boldsymbol{\alpha}) \tag{B.2}$$

for each $j \in \{1, \ldots, p-1\}$. Taking another derivative of (B.2) with respect to α_q, for each $q \in \{1, \ldots, p-1\}$, again by applications of the chain rule,

$$\frac{\partial^2}{\partial\alpha_q\partial\alpha_j}R_t(\boldsymbol{\alpha}_{-p}) = \frac{\partial^2}{\partial\alpha_q\partial\alpha_j}R_t(\boldsymbol{\alpha}) - \frac{\alpha_q}{\alpha_p}\frac{\partial^2}{\partial\alpha_p\partial\alpha_j}R_t(\boldsymbol{\alpha}) - \frac{\alpha_j}{\alpha_p}\frac{\partial^2}{\partial\alpha_q\partial\alpha_p}R_t(\boldsymbol{\alpha}) - \frac{\partial}{\partial\alpha_q}\left(\frac{\alpha_j}{\alpha_p}\right)\frac{\partial}{\partial\alpha_p}R_t(\boldsymbol{\alpha}) + \frac{\alpha_q\alpha_j}{\alpha_p^2}\frac{\partial^2}{\partial\alpha_p\partial\alpha_p}R_t(\boldsymbol{\alpha}). \tag{B.3}$$

After summing (B.3) over the groups $t \in \{1, \ldots, K\}$, weighted by the group probabilities $\pi_1, \ldots, \pi_K$, evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, we obtain (A.1).

Next, we examine the asymptotics of the profile estimator $\hat{\boldsymbol{\alpha}}$. From *A*.15 of Wang and Yang (2007) and under Assumptions 2–5, we have

$$\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\frac{\partial}{\partial\alpha_j}\{Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})\} - \frac{1}{n_t}\sum_{i=1}^{n_t}\xi_{\boldsymbol{\alpha},i,j,t}\right| = o\left(n_t^{-1/2}\right) \tag{B.4}$$

almost surely, with $\xi_{\boldsymbol{\alpha},i,j,t} = 2\{g_t(u_{\boldsymbol{\alpha},ti}) - Y_{ti}\}\frac{\partial}{\partial\alpha_j}g_t(u_{\boldsymbol{\alpha},ti}) - \frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha})$, where $u_{\boldsymbol{\alpha},ti} = \boldsymbol{\alpha}^\top X_{ti}$, and furthermore

$$\begin{aligned}&\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\frac{\partial}{\partial\alpha_j}\{Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})\}\right| = o(1),\\ &\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le q,j\le p}\left|\frac{\partial^2}{\partial\alpha_q\partial\alpha_j}\{Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})\}\right| = o(1),\end{aligned} \tag{B.5}$$

almost surely, for each group $t \in \{1, \ldots, K\}$.

Now, we will prove that the estimated score of $Q(\boldsymbol{\alpha}_{-p}) = \sum_{t=1}^K \hat{\pi}_t Q_t(\boldsymbol{\alpha}_{-p})$, where $\hat{\pi}_t = \sum_{i=1}^n I(T_i = t)/n$, evaluated at $\boldsymbol{\alpha}_{-p} = \boldsymbol{\alpha}_{0,-p}$, is represented up to $o(n^{-1/2})$ almost surely, by a sum of mean-zero independent random variables, which we denote by $\boldsymbol{\eta}_i \in \mathbb{R}^{p-1}$, $i \in \{1, \ldots, n\}$, where $n = \sum_{t=1}^K n_t$. Let us denote the estimated score function by $\hat{\Psi}(\boldsymbol{\alpha}_{-p}) = \frac{\partial}{\partial\boldsymbol{\alpha}_{-p}^\top}Q(\boldsymbol{\alpha}_{-p})$, where $\boldsymbol{\alpha}_{-p} \in \mathbb{R}^{p-1}$. We will show

$$\sup_{1\le j\le p-1}\left|\hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p}) - \frac{1}{n}\sum_{i=1}^n \eta_{i,j}\right| = o(n^{-1/2}), \tag{B.6}$$

almost surely, where $\hat{\Psi}_j(\boldsymbol{\alpha}_{-p}) \in \mathbb{R}$ is the jth component of the score function $\hat{\Psi}(\boldsymbol{\alpha}_{-p})$ and $\eta_{i,j} \in \mathbb{R}$ is the jth component of the random variable $\boldsymbol{\eta}_i$. In order to employ the result (B.4), we first consider the score function defined on the set Θ_c, i.e., the score function $\hat{\Psi}_j(\boldsymbol{\alpha})$, instead of the "$p$th component removed" score function defined on $\mathbb{R}^{p-1}$, i.e., $\hat{\Psi}_j(\boldsymbol{\alpha}_{-p})$. We will show that, for some mean-zero independent random variables, which we denote by $\xi^*_{\boldsymbol{\alpha},i,j}$, $i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}$,

$$\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\frac{\partial}{\partial\alpha_j}\{Q(\boldsymbol{\alpha}) - R(\boldsymbol{\alpha})\} - \frac{1}{n}\sum_{i=1}^n \xi^*_{\boldsymbol{\alpha},i,j}\right| = o(n^{-1/2}) \tag{B.7}$$

is satisfied almost surely. Let us set the desired mean-zero independent random variable $\xi^*_{\boldsymbol{\alpha},i,j}$ to be $\xi^*_{\boldsymbol{\alpha},i,j} = \sum_{t=1}^K \xi^*_{\boldsymbol{\alpha},i,j,t}$, where

$$\xi^*_{\boldsymbol{\alpha},i,j,t} = \left[2\{g_t(\boldsymbol{\alpha}^\top X_i) - Y_i\}\frac{\partial}{\partial\alpha_j}g_t(\boldsymbol{\alpha}^\top X_i) - \frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha})\right] I(T_i = t),$$

which must satisfy the following:

$$\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\sum_{t=1}^K \pi_t\left[\frac{\partial}{\partial\alpha_j}Q_t(\boldsymbol{\alpha}) - \frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha})\right] - \frac{1}{n}\sum_{i=1}^n\sum_{t=1}^K \xi^*_{\boldsymbol{\alpha},i,j,t}\right| = o(n^{-1/2}). \tag{B.8}$$

We can write

$$\begin{aligned}&\left|\sum_{t=1}^K \pi_t\left[\frac{\partial}{\partial\alpha_j}Q_t(\boldsymbol{\alpha}) - \frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha})\right] - \frac{1}{n}\sum_{t=1}^K\sum_{i=1}^n \xi^*_{\boldsymbol{\alpha},i,j,t}\right|\\ &= \left|\sum_{t=1}^K \pi_t\left[\frac{\partial}{\partial\alpha_j}Q_t(\boldsymbol{\alpha}) - \frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha}) - \frac{1}{\pi_t}\frac{n_t}{n}\frac{1}{n_t}\sum_{i=1}^{n_t}\xi_{\boldsymbol{\alpha},i,j,t}\right]\right|,\end{aligned}$$

where $\xi_{\boldsymbol{\alpha},i,j,t}$ is defined in (B.4). Therefore, applying the continuous mapping theorem and Slutsky's theorem to (B.4) leads to the desired result (B.8).

Next, we will show (B.6), the result corresponding to the "pth component removed" estimated score function, $\hat{\Psi}(\boldsymbol{\alpha}_{-p})$ on $\mathbb{R}^{p-1}$. Considering the linear operator $\frac{\partial}{\partial\alpha_j} - \frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}$, we note that by the chain rule,

$$\left(\frac{\partial}{\partial\alpha_j} - \frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}\right)\{Q(\boldsymbol{\alpha}) - Q(\boldsymbol{\alpha})\} = \hat{\Psi}_j(\boldsymbol{\alpha}_{-p}) - \Psi_j(\boldsymbol{\alpha}_{-p}),$$

for $j \in \{1, \ldots, p-1\}$, where $\Psi_j(\boldsymbol{\alpha}_{-p})$ denotes the jth component of the gradient of $R(\boldsymbol{\alpha}_{-p})$. If we set the approximation variable $\eta_{i,j}$ of (B.6) to be

$$\begin{aligned}\eta_{i,j} &= \xi^*_{\boldsymbol{\alpha},i,j} - \frac{\alpha_j}{\alpha_p}\xi^*_{\boldsymbol{\alpha},i,p} \\ &= \sum_{t=1}^{K}\Bigg[2\{g_t(\boldsymbol{\alpha}^\top X_i) - Y_i\}\left\{\frac{\partial}{\partial\alpha_j}g_t(\boldsymbol{\alpha}^\top X_i) - \frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}g_t(\boldsymbol{\alpha}^\top X_i)\right\} \\ &\qquad + \frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha}) - \frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}R_t(\boldsymbol{\alpha})\Bigg] I(T_i = t),\end{aligned} \tag{B.9}$$

then we can show

$$\begin{aligned}&\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p-1}\left|\left(\frac{\partial}{\partial\alpha_j} - \frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}\right)\{Q(\boldsymbol{\alpha}) - R(\boldsymbol{\alpha})\} - \frac{1}{n}\sum_{i=1}^{n}\eta_{i,j}\right| \\ &\le \sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p-1}\left|\frac{\partial}{\partial\alpha_j}\big(Q(\boldsymbol{\alpha}) - R(\boldsymbol{\alpha})\big) - \frac{1}{n}\sum_{i=1}^{n}\xi^*_{\boldsymbol{\alpha},i,j}\right| \\ &+ \sup_{\boldsymbol{\alpha}\in\Theta_c}\frac{\alpha_j}{\alpha_p}\left|\frac{\partial}{\partial\alpha_p}\big(Q(\boldsymbol{\alpha}) - R(\boldsymbol{\alpha})\big) - \frac{1}{n}\sum_{i=1}^{n}\xi^*_{\boldsymbol{\alpha},i,p}\right| = o(n^{-1/2}),\end{aligned} \tag{B.10}$$

by the triangle inequality and the result of (B.7). Since $\Psi_j(\boldsymbol{\alpha}_{-p})$ is evaluated at the minimum $\boldsymbol{\alpha}_{0,-p}$, we have

$$\Psi_j(\boldsymbol{\alpha}_{0,-p}) = \left(\frac{\partial}{\partial\alpha_j} - \frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}\right)\{Q(\boldsymbol{\alpha})\}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} = 0, \tag{B.11}$$

by the local convexity under Assumption 1. Then we obtain the desired result of (B.6), by (B.10) and (B.11).

The uniform consistency of the observed Hessian, $\hat{\boldsymbol{H}}(\boldsymbol{\alpha}) = \frac{\partial^2}{\partial\boldsymbol{\alpha}_{-p}\partial\boldsymbol{\alpha}_{-p}^\top}Q(\boldsymbol{\alpha}_{-p})$, to the population Hessian $\boldsymbol{H}(\boldsymbol{\alpha}_{-p})$ of (A.1) follows directly from the results of (B.5) under Assumptions 2–5, with applications of the continuous mapping theorem.

Finally, we prove the main result. Consider the random variable $\hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p})$ introduced in (B.6), and the following parametrization: for each component $j \in \{1, \ldots, p-1\}$

$$f_j(s) = \hat{\Psi}_j\big(s\hat{\boldsymbol{\alpha}}_{-p} + (1-s)\boldsymbol{\alpha}_{0,-p}\big), \quad s \in [0,1].$$

Taking the derivative with respect to t, we have by the chain rule

$$\frac{d}{dt}f_j(s) = \sum_{m=1}^{p-1}\frac{\partial}{\partial\alpha_m}\hat{\Psi}_j\big(s\hat{\boldsymbol{\alpha}}_{-p} + (1-s)\boldsymbol{\alpha}_{0,-p}\big)\big(\hat{\alpha}_m - \alpha_{0,m}\big).$$

Since $\hat{\Psi}_j(\hat{\boldsymbol{\alpha}}_{-p}) = 0$ by the definition of $\hat{\boldsymbol{\alpha}}_{-p}$, it follows that $f_j(1) - f_j(0) = \hat{\Psi}_j(\hat{\boldsymbol{\alpha}}_{-p}) - \hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p}) = -\hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p})$. Therefore, for any particular $j = 1, \ldots, p-1$, there exists $s_j^* \in [0,1]$ by the mean value theorem, such that

$$\begin{aligned}-\hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p}) = \Bigg[&\frac{\partial}{\partial\alpha_1}\hat{\Psi}_j\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p} + (1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big), \\ &\ldots, \frac{\partial}{\partial\alpha_{p-1}}\hat{\Psi}_j\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p} + (1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big)\Bigg]\Big[\hat{\boldsymbol{\alpha}}_{-p} - \boldsymbol{\alpha}_{0,-p}\Big],\end{aligned}$$

which is just

$$\begin{aligned}\Bigg[&\frac{\partial^2}{\partial\alpha_1\partial\alpha_j}\hat{Q}\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p} + (1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big), \\ &\ldots, \frac{\partial^2}{\partial\alpha_{p-1}\partial\alpha_j}\hat{Q}\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p} + (1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big)\Bigg]\Big[\hat{\boldsymbol{\alpha}}_{-p} - \boldsymbol{\alpha}_{0,-p}\Big],\end{aligned} \tag{B.12}$$

Table C.2
The proportion of time ("Coverage") that the asymptotic 95% confidence interval contains the true value of $\alpha_j, j \in \{1, \ldots, 5\}$, for varying $\omega \in \{0, 0.5, 1\}$, corresponding to *linear*, *moderately nonlinear*, and *highly nonlinear* contrasts, respectively, with varying n ($= n_1 + n_2$, where $n_1 = n_2$).

n	$\omega = 0$ (*linear*)					$\omega = 0.5$ (*moderate nonlinear*)					$\omega = 1$ (*highly nonlinear*)				
	α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
50	0.36	0.45	0.43	0.44	0.42	0.49	0.46	0.45	0.46	0.40	0.59	0.58	0.57	0.55	0.52
100	0.64	0.67	0.72	0.68	0.64	0.76	0.75	0.80	0.72	0.76	0.89	0.82	0.84	0.75	0.73
200	0.77	0.77	0.79	0.78	0.73	0.88	0.83	0.82	0.85	0.79	0.92	0.88	0.84	0.78	0.81
400	0.85	0.90	0.87	0.87	0.85	0.88	0.88	0.88	0.82	0.85	0.95	0.88	0.84	0.79	0.78
800	0.95	0.92	0.91	0.89	0.88	0.92	0.89	0.92	0.89	0.87	0.92	0.91	0.83	0.78	0.81
1600	0.93	0.93	0.92	0.93	0.92	0.94	0.94	0.91	0.93	0.91	0.93	0.90	0.87	0.84	0.81
3200	0.94	0.95	0.94	0.94	0.94	0.96	0.94	0.90	0.92	0.90	0.93	0.92	0.87	0.90	0.85

where $\left[\hat{\boldsymbol{\alpha}}_{-p} - \boldsymbol{\alpha}_{0,-p}\right]$ is a $p-1$ dimensional random vector. Writing (B.12) in matrix notation, we have

$$-\hat{\Psi}(\boldsymbol{\alpha}_{0,-p}) = \left[\frac{\partial^2}{\partial\alpha_q \partial\alpha_j}\hat{Q}\left(s_j^*\hat{\boldsymbol{\alpha}}_{-p} + (1-s_j^*)\boldsymbol{\alpha}_{0,-p}\right)\right]_{j,q=1}^{p-1}\left[\hat{\boldsymbol{\alpha}}_{-p} - \boldsymbol{\alpha}_{0,-p}\right]. \tag{B.13}$$

Then, by (B.13) one can write

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_{-p} - \boldsymbol{\alpha}_{0,-p}) = -\left\{\left[\frac{\partial^2}{\partial\alpha_q \partial\alpha_j}\hat{Q}\left(s_j^*\hat{\boldsymbol{\alpha}}_{-p} + (1-s_j^*)\boldsymbol{\alpha}_{0,-p}\right)\right]_{j,q=1}^{p-1}\right\}^{-1}\sqrt{n}\hat{\Psi}(\boldsymbol{\alpha}_{0,-p}). \tag{B.14}$$

Meanwhile, by (B.6), for each component $j \in \{1, \ldots, p-1\}$ of $\hat{\Psi}(\boldsymbol{\alpha}_{0,-p})$, we can write

$$\hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p}) = \frac{1}{n}\sum_{i=1}^{n}\eta_{i,j} + o(n^{-1/2}), \tag{B.15}$$

almost surely with $\mathbb{E}(\eta_{i,j}) = 0$. The variance–covariance matrix of the random vector $\boldsymbol{\eta}_i = \left[\eta_{i,1}, \ldots, \eta_{i,p-1}\right]^\top \in \mathbb{R}^{p-1}$ evaluated at $\boldsymbol{\alpha}_{-p} = \boldsymbol{\alpha}_{0,-p}$, where $\eta_{i,j}$ are specified in (B.9), is given in (A.2), where it is denoted by $\boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}}$. From (B.15), the central limit theorem ensures that $\sqrt{n}\hat{\Psi}(\boldsymbol{\alpha}_{0,-p}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}})$ in distribution. Now, by the representation of (B.14) together with an application of Slutsky's theorem on the observed Hessian, we obtain $\sqrt{n}(\hat{\boldsymbol{\alpha}}_{0,-p} - \boldsymbol{\alpha}_{0,-p}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{0,-p}})$ in distribution, where $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{0,-p}} = \boldsymbol{H}^{-1}_{\boldsymbol{\alpha}_{0,-p}}\boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}}\boldsymbol{H}^{-1}_{\boldsymbol{\alpha}_{0,-p}}$, which is the desired result of Theorem 2. □

Appendix C. Table for Section 5.2 coverage probability of asymptotic 95% confidence intervals

See Table C.2.

References

Antoniadis, A., Gregoire, G., McKeague, I., 2004. Bayesian estimation in single-index models. Statist. Sinica 14, 1147–1164.
de Boor, C., 2001. A Practical Guide to Splines. Springer-Verlag, New York.
Brillinger, R.D., 1982. A generalized linear model with "Gaussian" regressor variables. In: Bickel, P., Doksum, K., Hodges, J. (Eds.), In A Festschrift for Erich L. Lehman. Wadsworth, New York.
Cai, T., Tian, L., Wong, P.H., Wei, L.J., 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics 12, 270–282.
DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Statist. Sci. 11, 189–228.
Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression. J. Amer. Statist. Assoc. 76, 817–823.
Hardle, W., Hall, P., Ichimura, H., 1993. Optimal smoothing in single-index models. Ann. Statist. 21, 157–178.
Horowitz, J.L., 2009. Semiparametric and Nonparametric Methods in Econometrics. Springer.
Jeng, X., Lu, W., Peng, H., 2018. High-dimensional inference for personalized treatment decision. Electron. J. Stat. 12, 2074–2089.
Lin, W., Kulasekera, K.B., 2007. Uniqueness of a single index model. Biometrika 94, 496–501.
Lu, W., Zhang, H., Zeng, D., 2011. Variable selection for optimal treatment decision. Stat. Methods Med. Res. 22, 493–504.
Mackay, D.J., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.
Murphy, S.A., 2003. Optimal dynamic treatment regimes. J. R. Stat. Soc. Ser. B Stat. Methodol. 65, 331–355.
Murphy, S.A., 2005. A generalization error for Q-learning. J. Mach. Learn. 6, 1073–1097.
Petkova, E., Tarpey, T., Su, Z., Ogden, R.T., 2016. Generated effect modifiers in randomized clinical trials. Biostatistics 18, 105–118.
Powell, J., Stock, J., Stoker, T., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.
Qian, M., Murphy, S.A., 2011. Performance guarantees for individualized treatment rules. Ann. Statist. 39, 1180–1210.
Robins, J., 2004. Optimal structural nested models for optimal sequential decisions. In: Proceedings of the Second Seattle Symposium on Biostatistics. Springer, New York.
Shi, C., Fan, A., Song, R., Lu, W., 2018. High-dimensional A-learning for optimal dynamic treatment regimes. Ann. Statist. 46 (3), 925–957..
Shi, C., Song, R., Lu, W., 2016. Robust learning for optimal treatment decision with np-dimensionality. Electron. J. Stat. 10, 2894–2921.

Stoker, T.M., 1986. Consistent estimation of scaled coefficients. Econometrica 54, 1461–1481.
Tian, L., Alizadeh, A., Gentles, A., Tibshrani, R., 2014. A simple method for estimating interactions between a treatment and a large number of covariates. J. Amer. Statist. Assoc. 109 (508), 1517–1532.
Wang, L., Yang, L., 2007. Spline Single-Index Prediction Model. Technical Report, https://arxiv.org/abs/0704.0302.
Wang, L., Yang, L., 2009. Spline estimation of single-index models. Statist. Sinica 19, 765–783.
Xia, Y., 2008. A multiple-index model and dimension reduction. J. Amer. Statist. Assoc. 103, 1631–1640.
Xia, Y., Li, W., 1999. On single index coefficient regression models. J. Amer. Statist. Assoc. 94, 1275–1285.
Yuan, M., 2011. On the identifiabliity of additive index models. Statist. Sinica 21, 1901–1911.
Zhang, B., Tsiatis, A.A., Laber, E.B., Davidian, M., 2012. A robust method for estimating optimal treatment regimes. Biometrics 68, 1010–1018.