Flexible Regression Models for Estimating Interactions between a Treatment and Scalar/Functional Predictors

Hyung G. Park

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Flexible Regression Models for Estimating Interactions between a Treatment and Scalar/Functional Predictors

Hyung G. Park

In this dissertation, we develop regression models for estimating interactions between a treatment variable and a set of baseline predictors in their effect on the outcome in a randomized trial, without restriction to a linear relationship. The proposed semiparametric/nonparametic regression approaches for representing interactions generalize the notion of an interaction between a categorical treatment variable and a set of predictors on the outcome, from a linear model context.

In Chapter 2, we develop a model for determining a composite predictor from a set of baseline predictors that can have a nonlinear interaction with the treatment indicator, implying that the treatment efficacies can vary across values of such a predictor without a linearity restriction. 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 predictors on the outcome. A common approach to interrogate such treatment-by-predictor interaction is to fit a regression curve as a function of the predictors separately for each treatment group. For parsimony and insight, we propose a single-index model with multiple-links that estimates a single linear combination of the predictors (i.e., a single-index), with treatment-specific nonparametrically-defined link functions. The approach emphasizes a focus on the treatment-by-predictors 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 to illustrate the proposed approach for deriving treatment decision rules.

In Chapter 3, we allow the proposed single-index model with multiple-links to have an unspecified main effect of the predictors on the outcome. This extension greatly increases the utility of the proposed regression approach for estimating the treatment-by-predictors interactions. By obviating the need to model the main effect, the proposed method extends the modified covariate approach of [Tian *et al.*, 2014] into a semiparametric regression framework. Also, the approach extends [Tian *et al.*, 2014] into general K treatment arms.

In Chapter 4, we introduce a regularization method to deal with the potential high dimensionality of the predictor space and to simultaneously select relevant treatment effect modifiers exhibiting possibly nonlinear associations with the outcome. We present a set of extensive simulations to illustrate the performance of the treatment decision rules estimated from the proposed method. An application to a clinical trial for the treatment of depression is presented to illustrate the proposed approach for deriving treatment decision rules.

In Chapter 5, we develop a novel additive regression model for estimating interactions between a treatment and a potentially large number of functional/scalar predictor. If the main effect of baseline predictors is misspecified or high-dimensional (or, infinite dimensional), any standard nonparametric or semiparametric approach for estimating the treatment-by-predictors interactions tends to be not satisfactory because it is prone to (possibly severe) inconsistency and poor approximation to the true treatment-by-predictors interaction effect. To deal with this problem, we impose a constraint on the model space, giving the orthogonality between the main and the interaction effects. This modeling method is particularly appealing in the functional regression context, since a functional predictor, due to its infinite dimensional nature, must go through some sort of dimension reduction, which essentially involves a main effect model misspecification. The main effect and the interaction effect can be estimated separately due to the orthogonality between the two effects, which side-steps the issue of misspecification of the main effect. The proposed approach extends the modified covariate approach of [Tian *et al.*, 2014] into an additive regression model framework. We impose a concave penalty in estimation, and the method simultaneously selects functional/scalar treatment effect modifiers that exhibit possibly nonlinear interaction effects with the treatment indicator. The dissertation concludes in Chapter 6.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to give my sincere thanks to my adviser, Dr. Todd Ogden, for his guidance, encouragement, and assistance. Looking back upon my time at Columbia, I feel so fortunate to have known and worked with him as his student. Not only he taught me valuable statistical insights but also he served as a role model who I hope to emulate.

I would also like to deeply thank the chair of my dissertation committee, Dr. Ian McKeague, for his advice and encouragement. I would also like to give my sincere thanks to the other members of my committee: Dr. Eva Petkova, Dr. Min Qian, and Dr. Seonjoo Lee, for their time and insightful comments.

I am thankful to my fellow doctoral students who took the classes together and the professors who taught me. I gratefully acknowledge the fellowship I received from the department. I would also like to thank Dr. Eva Petkova and Dr. Thad Tarpey for their guidance and invaluable support on our projects together, and thank Dr. Seonjoo Lee for the opportunity to serve as a research assistant and the project with her. I give my sincere thanks to Dr. Nan Laird, Dr. L.J. Wei, Dr. Francesca Dominici, and Dr. Armin Schwartzman, for their encouragements when applying to my doctoral study. I give my special thanks to Dr. Gail Gong, who first taught me how to do research work. I also express my thanks to Dr. Mingue Park and Dr. Hyungjun Cho, for their invaluable supports.

My deepest thanks go to my mother and grandmother. Thank you for your unwavering belief in me and for all the love and the goodness that I received from you, everywhere in my life.

Just as a tree takes the sunlight to synthesize organic molecules, I feel that I live my life out of the love and the support from the people surrounding me.

For my family

Chapter 1

Introduction

Precision medicine represents a powerful and effective general approach for disease treatment and prevention that takes into account individual variability in genetic structure, environment, and lifestyle for each person. Its growth is not only helped by technological advances in detecting and measuring a wide range of biomedical information, such as brain imaging (structure, function, connectivity), molecular, genomic, cellular, clinical, behavioral, physiological, and environmental characteristics, but also helped by the increasing pace of developing treatment options. The most daunting challenge for precision medicine is discovery of the treatment implications of the available complex and large-scale biological information.

To develop strategies for precision medicine, it is important to identify treatment and predictor interactions ([Royston and Sauerbrei, 2008], [Tian *et al.*, 2014]) particularly in the setting of randomized clinical trials (RCT). There are many RCTs dedicated to discovering the treatment implications based on individual patient's characteristics. Just in major depressive disorder (MDD), for example, recent large-scale studies include iSPOT-D: International Study to Predict Optimized Treatment for Depression, PReDICT: Predictors of Remission in Depression to Individual and Combined Treatments, and EMBARC: Establishing Moderators and Biosignatures of Antidepressant Response for clinical Care, among others.

Recent breakthroughs in biotechnology allows a vast amount of data available for exploring for potential interaction effect with the treatment and assisting in the optimal treatment decision for individual patients ([Tian *et al.*, 2014]). For example, data from modern medical experiments include more and more commonly not only traditional clinical measures, but also increasingly complex information such as genetic information (e.g., [van't Veer and Bernards, 2008]) and brain structure or functions, measured from neuroimaging modalities such as magnetic resonance imaging (MRI), functional MRI (fMRI), electroencephalogram (EEG), among others. This motivates the need for developing an efficient and also flexible statistical method for discovery of biomarkers from high-dimensional data, specifically designed to estimate the interactions between a treatment and high-dimensional pretreatment predictors.

In particular, development of individualized treatment decisions rules (ITRs) based on patient characteristic data measured at baseline is an increasingly important topic in precision medicine. Much research has been done since the seminal papers of [Murphy, 2003] and [Robins, 2004]. Regression-based methodologies are intended to optimize the ITRs by estimating treatment-specific mean responses (e.g., [Qian and Murphy, 2011], [Zhang *et al.*, 2012], [Gunter *et al.*, 2011], [Lu *et al.*, 2011], among others), while seeking robustness with respect to model misspecification. Extensions that allow functional data objects to be incorporated as baseline predictors have also been developed (e.g., [McKeague and Qian, 2014], [Ciarleglio *et al.*, 2015a]). Machine learning approaches for developing ITRs originate from computer science literature, and can often be framed in the context of classification problems ([Zhang *et al.*, 2012], [Zhang *et al.*, 2012]), for example, the outcome weighted learning (OWL) (e.g., [Zhao *et al.*, 2012], [Zhao *et al.*, 2015], [Song *et al.*, 2015]) based on support vector machines, tree-based classification (e.g. [Laber and Zhao, 2015]), and the [Kang *et al.*, 2014] method based on adaptive boosting, among others. In these settings of optimizing ITRs, a major challenge is in the discovery of biomarkers that exhibit interaction effects with the treatment indicator when large amount of patient characteristics are available.

Suppose we are given pre-treatment predictors $X \in \mathcal{X}$, a treatment variable T that takes a value in a finite, discrete treatment space, say, $\mathcal{T} = \{1, \cdots, K\}$, and a real-valued response variable Y. We assume that a larger Y is preferred, without loss of generality. Let the distribution of (Y, T, X) be denoted by $\mathcal{P}$. An ITR $\mathcal{D} : \mathcal{X} \to \mathcal{T}$ is a deterministic decision rule that maps $\mathcal{X}$ into the treatment space $\mathcal{T}$. For any fixed ITR $\mathcal{D}$, let $\mathcal{P}^{\mathcal{D}}$ denote the distribution of (Y, T, X) conditioning on $T = \mathcal{D}(X)$, i.e., the treatments are chosen according to the rule $\mathcal{D}$. Let $\mathbb{E}^{\mathcal{D}}$ denote the expectation with respect to $\mathcal{P}^{\mathcal{D}}$. A natural measure for the effectiveness of $\mathcal{D}$ is the expected outcome that would have resulted if $\mathcal{D}$ had been used to choose treatment for the entire study population

$$V(\mathcal{D}) = \mathbb{E}^{\mathcal{D}}(Y), \tag{1.1}$$

which is often called the "value" associated with $\mathcal{D}$ ([Murphy, 2005], [Qian and Murphy, 2011]). A larger value of (1.1) is preferred. Therefore, an ITR that maximizes the function $\mathcal{D} \to V(\mathcal{D})$ over all $\mathcal{D}$ is called optimal. It can be easily verified that any $\mathcal{D}_0(X)$ with

$$\mathcal{D}_0(X) \ \in \ \arg\max_{t \in \mathcal{T}} \mathbb{E}(Y \mid X, T = t), \quad X \in \mathcal{X} \tag{1.2}$$

is optimal ([Murphy, 2005], [McKeague and Qian, 2014]), where $\mathbb{E}$ denotes expectation under $\mathcal{P}$.

A first natural approach to estimate the optimal ITR is then to maximize an empirical version of the mean response (1.1) (or its surrogate) over a class of ITRs, in a classification context, for example, as in OWL ([Zhao *et al.*, 2012]). Although the classification approaches can be appealing in many settings, in this dissertation, we will focus on a regression approach that estimates the conditional expectations $\mathbb{E}(Y \mid X, T = t)$, $t \in \mathcal{T}$ in (1.2), as the regression-based approaches are most frequently utilized in practice, and often come with great interpretability. We will employ a two-step procedure (e.g., [Qian and Murphy, 2011]) that first estimates the conditional expectation $\mathbb{E}(Y \mid X, T)$ using a regression model and then from this estimated conditional expectation derives the estimated treatment.

In (1.2), if the conditional expectation is modeled correctly, then the two-step procedure consistently estimates the optimal ITR. [Qian and Murphy, 2011] derived several finite sample upper bounds on the difference between the mean response (1.1) to the optimal ITR and the mean response to the estimated ITR. If the part of the model for the conditional expectation involving the treatment-by-predictors interaction effect is correct, then the upper bounds imply that, although a surrogate two-step procedure is used, the estimated ITR is consistent. The upper bound of [Qian and Murphy, 2011] is an improvement over that of ([Murphy, 2005]), in the sense that the upper bound depends only on how well we approximate the interaction effect term, and not on the main effect term of the conditional expectation. These upper bounds guarantee that if the T-by-X interaction effect is consistently estimated (for example, estimated under the squared error loss), then the value (1.1) of the estimated ITR will converge to the optimal value.

However, if the approximation space for the interaction effect does not provide an interaction effect term close to the true interaction effect term of the conditional expectation, then the two-step procedure does not provide the best value of the considered ITRs in the approximation space ([Qian and Murphy, 2011]). This is due to the "mismatch" ([Murphy, 2005]) between the loss functions (weighted 0-1 loss for directly maximizing the value and the squared error loss for approximating the conditional expectation). In other words, if the interaction effect model is misspecified, then the ITR obtained from the two-step procedure may not be the best ITR within the class of ITRs defined by the model. [McKeague and Qian, 2014] noted that this issue also arises when a smooth surrogate of the empirical value function is maximized ([Zhao *et al.*, 2012]).

The primary focus of this dissertation is on developing flexible regression approaches to accurately approximate the interaction effect term of the conditional expectation. The semiparametric/nonparametric regression approaches that we develop in this dissertation for estimating the interaction effect term will reduce the concerns regarding the mismatch between the two loss functions, that occur from the misspecification of the interaction effect term in the model.

[Qian and Murphy, 2011] approximated the conditional expectation using L^1 penalized least squares with a rich linear model. However, the approach is generally not robust to the main effect model misspecification, and is restricted to a parametric regression model. The presence of main effect, which often have much bigger effect on the outcome than the treatment interactions, makes the consistent estimation of the interaction effect very difficult, if the main effect model is misspecified or is high-dimensional. [Tian *et al.*, 2014] proposed a novel approach to consistently estimate the covariates and treatment interactions without the need for modeling main effects. Their method modified the covariates in a simple way, and then fit a standard model using the modified covariates and no main effects. However, the approach is limited to a linear regression framework. In realistic situations, the knowledge of the true functional forms for models of interactions is often lacking, and a linear model is generally restrictive.

The main contribution of this dissertation is in generalizing the work of [Tian *et al.*, 2014] to a single-index model framework in Chapter 3, and to an additive model framework in Chapter 5. These extensions provide robust and flexible regression approach to developing ITRs in many situations, particularly when we deal with a large number of baseline predictors, that includes multiple functional predictors.

The thesis is organized as follows. In Chapter 2, we introduce a flexible model for determining composite predictors that permit nonlinear association with the outcome. In Chapter 3, we will consider a more general model where we assume an unspecified structure for the main effect component. In Chapter 4, we use a L^1 regularization to consider a large model for the treatment effect modification. In Chapter 5, we develop a sparse additive regression model for estimating interactions between a treatment and a large number of functional/scalar predictors. The thesis concludes in Chapter 6.

Chapter 2

A Single-index model with multiple-links

2.1 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 uncommon to find variables that are individually strong moderators of treatment effects.

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) models. This single-index models with multiple-links provides a parsimonious generalization 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 determining ITRs. This approach extends the notion of a "treatment effect modifier" from the linear model setting, to a single-index model framework, to define a nonparametric generated effect modifier.

2.2 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 probabilities $(\pi_1, \ldots \pi_K)$ that sum to one. Let $Y \in \mathbb{R}$ denote an outcome variable, where 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 conventional approach to study 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, $\boldsymbol{\beta}_t^\top X$, $t \in \{1, \dots, K\}$. We refer to this as a K-index model; it has the form

$$\mathbb{E}\left(Y \mid T = t, X = x\right) = g_t(\boldsymbol{\beta}_t^\top x) \quad (t = 1, \dots, K), \tag{2.1}$$

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. [Wu and Rolling, 2016] proposed this model for dimension reduction in optimizing ITRs. (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 (2.1) lack useful interpretation as effect modifiers and often lead to over-parametrization.

The SIMML constrains the $\boldsymbol{\beta}_t$ in (2.1) to be equal, and it requires separate nonparametrically defiend curves for each treatment t as a function of a single index $\boldsymbol{\alpha}^\top X$ common for all t:

$$\mathbb{E}\left(Y \mid T = t, X = x\right) = g_t(\boldsymbol{\alpha}^\top x) \quad (t = 1, \dots, K), \tag{2.2}$$

where both the links g_t and the vector $\boldsymbol{\alpha}$ need to be estimated. Due to the nonparametric nature of g_t, the scale of $\boldsymbol{\alpha}$ is not identifiable in (2.2) and to address this we restrict $\boldsymbol{\alpha}$ to be in $\Theta = \{\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^\top | \sum_{j=1}^p \alpha_j^2 = 1, \alpha_p > 0\}$, 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 $u = \boldsymbol{\alpha}^\top X$,

$$g_t(u) = \mathbb{E}(m_t(X) \mid \boldsymbol{\alpha}^\top X = u) \quad (t = 1, \dots, K), \tag{2.3}$$

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

$$Y_t = m_t(X) + \sigma_t(X)\epsilon \quad (t = 1, \dots, K), \tag{2.4}$$

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 (2.3) and let

$$\boldsymbol{\alpha}_0 := \underset{\boldsymbol{\alpha} \in \Theta}{\arg\min} \quad R(\boldsymbol{\alpha}). \tag{2.5}$$

Then $\boldsymbol{\alpha}_0$ can be shown to be the minimizer of the cross-entropy (e.g., [Mackay, 2003]) between the SIMML (2.2) and the general model (2.4) under the Gaussian noise assumption. Here, the cross-entropy of an arbitrary distribution with the probability density (or, mass) 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 (2.3) evaluated at $\boldsymbol{\alpha}_0$ can be viewed as the "projection" (in the sense of the closest point) of the true distribution $\mathcal{P}$ (2.4) onto the space Θ of the SIMML distribution, using the Kullback-Leibler divergence as a distance measure.

The SIMML (2.2) 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 (2.2) 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 that can describe differential treatment effects, provided that $g_t \neq g_{t'}$ for at least one pair $t, t' \in \{1, \ldots, K\}$.

2.3 Criteria for estimation

2.3.1 Profile likelihood maximization

While any nonparametric smoother can be employed to approximate the unspecified smooth link functions $g_t(\cdot)$ in (2.2), in this chapter, we will apply cubic splines. Specifically, $g_t(u) \approx \boldsymbol{\eta}_t^\top Z(u)$, for some $\boldsymbol{\eta}_t \in \mathbb{R}^d$. Here, $Z(u) = \begin{bmatrix} B_1(u), & \ldots, & B_d(u) \end{bmatrix}^\top \in \mathbb{R}^d$ consists of a set of d normalized cubic B-spline basis functions [de Boor, 2001]. For ease of notation, the number of basis functions, d, is taken to be the same across treatments but in practice d may vary by treatment. 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. For a given $\boldsymbol{\alpha}$, let $\mathbb{Z}_{\boldsymbol{\alpha},t}$ denote the B-spline evaluation matrix ($n_t \times d$), so that the ith row is $Z(\boldsymbol{\alpha}^\top X_{ti})^\top$, 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 $\boldsymbol{\alpha}$.

For sample data, SIMML (2.2) can be represented by

$$\begin{bmatrix}\boldsymbol{Y}\end{bmatrix}_{n\times 1} = \begin{bmatrix}\mathbb{Z}_{\boldsymbol{\alpha}}\end{bmatrix}_{n\times Kd} \begin{bmatrix}\boldsymbol{\eta}\end{bmatrix}_{Kd\times 1} + \begin{bmatrix}\boldsymbol{\epsilon}\end{bmatrix}_{n\times 1}, \tag{2.6}$$

where $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_K^\top\right]^\top$ is the observed response vector in which $\boldsymbol{Y}_t \in \mathbb{R}^{n_t}$, $\mathbb{Z}_{\boldsymbol{\alpha}}$ is $n \times Kd$ 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 $\sigma^2 \boldsymbol{I}_n$.

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

$$Q(\boldsymbol{\alpha}) = \|\boldsymbol{Y} - \mathbb{S}_{\boldsymbol{\alpha}} \boldsymbol{Y}\|^2. \tag{2.7}$$

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

$$\hat{\boldsymbol{\alpha}} \quad = \quad \underset{\boldsymbol{\alpha} \in \Theta}{\arg\min} \quad Q(\boldsymbol{\alpha}). \tag{2.8}$$

Each link functions $g_t(\cdot)$ in (2.2) can be estimated by

$$\hat{g}_t(u) = Z(u)^\top \left(\mathbb{Z}_{\hat{\boldsymbol{\alpha}},t}^T \mathbb{Z}_{\hat{\boldsymbol{\alpha}},t}\right)^{-1} \mathbb{Z}_{\hat{\boldsymbol{\alpha}},t}^T \boldsymbol{Y}_t \quad (t = 1, \ldots, K), \tag{2.9}$$

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

2.3.2 Maximizing L^2 distance between two link functions

A natural criterion for choosing $\boldsymbol{\alpha}$ in the SIMML (2.2) in terms of moderator analysis is to maximize an interaction effect. In the special case of linear link functions in the SIMML

$$g_t(\boldsymbol{\alpha}^\top X) = \gamma_{t0} + \gamma_t \boldsymbol{\alpha}^\top X \quad (t = 1, \ldots, K). \tag{2.10}$$

[Petkova *et al.*, 2016] proposed estimating $\boldsymbol{\alpha}$ to maximize the variability of the GEM slopes γ_t's, weighted by their respective probabilities π_t; this was called the "numerator" criterion because it corresponds to maximizing the numerator of a F-test statistic for significance of an interaction effect.

Analogously, for nonlinear $g_t(\cdot)$, $\boldsymbol{\alpha}$ can be chosen to maximize the variance of the distance between any two link functions, e.g., $g_1(\cdot)$ and $g_2(\cdot)$ of (2.2). Assuming that the outcome has been centered at 0 for each treatment group and that the observations are independent, maximizing the variance corresponds to maximizing the L^2 distance between the two link functions over $\boldsymbol{\alpha} \in \Theta$, which simplifies to

$$\mathbb{E}_X \left(g_1(\boldsymbol{\alpha}^\top X) - g_2(\boldsymbol{\alpha}^\top X) \right)^2 = \int g_1^2(\boldsymbol{\alpha}^\top x) f_X(x) dx + \int g_2^2(\boldsymbol{\alpha}^\top x) f_X(x) dx, \qquad (2.11)$$

where $f_X(\cdot)$ is the density of X (assumed equal across treatment groups due to randomization). Given $\boldsymbol{\alpha}$, define the $n_t \times n_t$ matrix $\mathbb{S}_{\boldsymbol{\alpha},t} = \mathbb{Z}_{\boldsymbol{\alpha},t} \left(\mathbb{Z}_{\boldsymbol{\alpha},t}^\top \mathbb{Z}_{\boldsymbol{\alpha},t} \right)^{-1} \mathbb{Z}_{\boldsymbol{\alpha},t}^\top$ for $t = 1, 2$. Then the index coefficients $\boldsymbol{\alpha}$ can be chosen to maximize the corresponding empirical approximation of the L^2 distance (2.11)

$$W(\boldsymbol{\alpha}) = \boldsymbol{Y}_1^T \mathbb{S}_{\boldsymbol{\alpha},1} \boldsymbol{Y}_1 / n_1 + \boldsymbol{Y}_2^T \mathbb{S}_{\boldsymbol{\alpha},2} \boldsymbol{Y}_2 / n_2, \qquad (2.12)$$

and

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \Theta}{\arg\max} \quad W(\boldsymbol{\alpha}). \qquad (2.13)$$

The associated link functions $\hat{g}_t(\cdot)$ can be obtained by (2.9).

2.4 Estimation

Suppose we have a set of observations $\{(Y_{ti}, X_{ti})_{i=1}^{n_t}\}_{t=1}^K$. For each candidate $\boldsymbol{\alpha}$, [Wang and Yang, 2009] suggested to take an integral transformation of each candidate index variable $\boldsymbol{\alpha}^\top X_{ti}$, $\boldsymbol{\alpha} \in \Theta$, to $u_{\boldsymbol{\alpha},ti} = F_p(\boldsymbol{\alpha}^\top X_{ti})$, where F_p is a re-scaled centered Beta$\{\ (p+1)/2, (p+1)/2\ \}$ cumulative distribution function

$$F_p(\nu) = \int_{-1}^{\nu/R} (1-t^2)^{(p-1)/2} \Gamma(p+1) / (\Gamma\{(p+1)/2\}^2 \, 2^p) dt, \quad \nu \in [-R, R],$$

in which $R = \max_{t,i} |\boldsymbol{\alpha}^\top X_{ti}|$. For any fixed $\boldsymbol{\alpha}$, this (transformed) index has a quasi-uniform $[0,1]$ distribution, and it is reasonable to use equally-spaced knots on $[0,1]$ when applying spline smoothing ([Wang and Yang, 2009]).

In an attempt to avoid being trapped in local minima in solving (2.8), we employ a variant of a gradient descent algorithm; we adopt elements of the Cuckoo search ([Yang and Deb, 2009], [Yang and Deb, 2014]). The Cuckoo search considers multiple (say, $C > 1$) evolving candidate solutions, where each of the C candidates takes independent random walks, specified by the step size, $s > 0$ which follows a heavy-tailed distribution (we take the absolute value of the standard Cauchy distribution), with some direction in $\mathbb{R}^p$ (we take the negative of the gradient of the criterion function $Q(\boldsymbol{\alpha})$, which we denote by $-\boldsymbol{\nabla} \in \mathbb{R}^p$). From the rth candidate at the hth update, $\hat{\boldsymbol{\alpha}}_{(r)}^{(h)} \in \mathbb{R}^p$, a new position $\hat{\boldsymbol{\alpha}}_{(r)}^{(h+1)} \in \mathbb{R}^p$ is generated by

$$\hat{\boldsymbol{\alpha}}_{(r)}^{(h+1)} = \hat{\boldsymbol{\alpha}}_{(r)}^{(h)} - s_{(r)}^{(h)} l_{(r)}^{(h)} \boldsymbol{\nabla}_{(r)}^{(h)}, \tag{2.14}$$

where $s_{(r)}^{(h)}$ is the random step size; $\boldsymbol{\nabla}_{(r)}^{(h)} \in \mathbb{R}^p$ is the gradient $\boldsymbol{\nabla}$ evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_{(r)}^{(h)}$; $l_{(r)}^{(h)} \in \mathbb{R}$ is an adaptive multiplier which is [Ong, 2014]

$$l_{(r)}^{(h)} = \left\{ \begin{array}{ll} l_L + (l_U - l_L)(Q_{(r)}^{(h)} - Q_{\min}^{(h)})/(Q_{\text{avg}}^{(h)} - Q_{\min}^{(h)}), & \text{if} \quad Q_{(r)}^{(h)} < Q_{\text{avg}}^{(h)} \\ l_U, & \text{if} \quad Q_{(r)}^{(h)} \geq Q_{\text{avg}}^{(h)} \end{array} \right\}, \tag{2.15}$$

where $l_L > 0$ and $l_U \geq l_L$ are pre-specified minimum and maximum step sizes, respectively, depending on the scale of the problem. In (2.15), $Q_{(r)}^{(h)}$ corresponds to the criterion value (2.7); $Q_{\min}^{(h)}$ is the minimum, and $Q_{\text{avg}}^{(h)}$ is the average of the criterion values at the hth update, respectively, computed from C candidate solutions. Intuitively, when a candidate solution is close to the minimum, the algorithm focuses more on the local search. The jth component of the gradient $\boldsymbol{\nabla}$ is

$$\nabla_j(\boldsymbol{\alpha}) = 2\boldsymbol{Y}^\top(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}})\dot{\mathbb{Z}}_{\boldsymbol{\alpha}}^{(j)}(\mathbb{Z}_{\boldsymbol{\alpha}}^\top \mathbb{Z}_{\boldsymbol{\alpha}})^{-1}\mathbb{Z}_{\boldsymbol{\alpha}}^\top \boldsymbol{Y}, \quad j \in \{1, \ldots, p\}, \tag{2.16}$$

where $\dot{\mathbb{Z}}_{\boldsymbol{\alpha}}^{(j)}$ is an $n \times dK$ block diagonal matrix with the blocks equal to

$$\dot{\mathbb{Z}}_{\boldsymbol{\alpha},t}^{(j)} = \Big[B_1'(F_p(\boldsymbol{X}_t\boldsymbol{\alpha})) * F_p'(\boldsymbol{X}_t\boldsymbol{\alpha}) * \boldsymbol{X}_{t,j} \quad \ldots \quad B_d'(F_p(\boldsymbol{X}_t\boldsymbol{\alpha})) * F_p'(\boldsymbol{X}_t\boldsymbol{\alpha}) * \boldsymbol{X}_{t,j} \Big]_{n_t \times d}$$

for $t = 1, \ldots, K$. Here, $\boldsymbol{X}_t$ is $n_t \times p$ covariate matrix of the tth treatment group, the jth column $\boldsymbol{X}_{t,j}$ corresponds to the jth covariate from the tth treatment, and $*$ denotes element-wise multiplication.

To obtain the L^2 distance maximizer of Section 2.3.2, instead of the profile likelihood maximizer, we simply change the objective function to be the negative of the criterion function of (2.12), and $\boldsymbol{\nabla}_{(r)}^{(h)}$ to be the negative gradient of (2.12), evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_{(r)}^{(h)}$.

The algorithm reduces to an ordinary gradient descent, if $s_{(r)}^{(h)}$ in (2.14) is set to be non-random and $C = 1$. In such cases, we can take as an initial solution $\hat{\boldsymbol{\alpha}}^{(h=1)}$ of gradient descent the leading eigenvector of the weighted "between-group" covariance matrix of the ordinary least-square regression coefficient vector (weighted by group assignment probabilities π_t), i.e., the solution that maximizes the variability of the GEM slopes from the linear model of [Petkova *et al.*, 2016].

Algorithm 1 An algorithm for estimating the single index coefficient $\boldsymbol{\alpha}$

```
Center $Y_t$ at $\mathbf{0}$ and center and scale $X_t$ for each $t = 1, \ldots, K$.
Set $h \leftarrow 1$. Generate $C$ candidate solutions, $\boldsymbol{A}^{(h)} = \left[\hat{\boldsymbol{\alpha}}_{(1)}, \ldots, \hat{\boldsymbol{\alpha}}_{(C)}\right]^\top$, where $\hat{\boldsymbol{\alpha}}_{(r)} \in \Theta$.
Evaluate (2.7) for each $\hat{\boldsymbol{\alpha}}_{(r)}$, and obtain $Q^{(h)} = \left[Q(\hat{\boldsymbol{\alpha}}_{(1)}), \ldots, Q(\hat{\boldsymbol{\alpha}}_{(C)})\right]^\top$.
for $h = 1, 2, \ldots$ until convergence do
    Set $h \leftarrow h + 1$.
    Set $\boldsymbol{A}^{(h)} \leftarrow \boldsymbol{A}^{(h-1)}$, and $Q^{(h)} \leftarrow Q^{(h-1)}$.
    Compute $Q^{(h)}_{\text{avg}}$ and $Q^{(h)}_{\min}$ from $Q^{(h)}$, obtain the multipliers $\left[l^{(h)}_{(1)}, \ldots, l^{(h)}_{(C)}\right]^\top$ in (2.15).
    Compute the $C \times p$ gradient matrix $\left[\nabla(\hat{\boldsymbol{\alpha}}^{(h)}_{(1)}), \ldots, \nabla(\hat{\boldsymbol{\alpha}}^{(h)}_{(C)})\right]^\top$ by using (2.16).
    for $r = 1, \ldots, C$ do
        Generate a new solution $\boldsymbol{\alpha}^{\text{temp},(h)}_{(r)}$ via (2.14) from $\boldsymbol{\alpha}^{(h)}_{(r)}$. Set $\boldsymbol{\alpha}^{\text{temp},(h)}_{(r)}$ to be in $\Theta$.
        Evaluate (2.7) at $\boldsymbol{\alpha}^{\text{temp},(h)}_{(r)}$, and denote the evaluated value by $Q^{\text{temp},(h)}_{(r)}$.
        if $Q^{\text{temp},(h)}_{(r)} < Q^{(h)}_{(r)}$ then $\boldsymbol{\alpha}^{(h)}_{(r)} \leftarrow \boldsymbol{\alpha}^{\text{temp},(h)}_{(r)}$ and $Q^{(h)}_{(r)} \leftarrow Q^{\text{temp},(h)}_{(r)}$.
    end for
end for
Output $\hat{\boldsymbol{\alpha}}$ that corresponds to $Q^{(h)}_{\min}$.
```

Once $\hat{\boldsymbol{\alpha}}$ is obtained, the estimates of the link functions $\hat{g}_t(\cdot)$ can be obtained by (2.9). The fitted SIMML is then $\hat{g}_t(F_p(\hat{\boldsymbol{\alpha}}^\top X))$, $t \in \{1, \ldots, K\}$.

2.5 Asymptotic theory

In this section, we establish the asymptotic results of the profile estimator $\hat{\boldsymbol{\alpha}}$ in (2.8), under possible misspecification, when the true model is assumed to be (2.4). We assume that the data consist of n random vectors $\{(Y_i, T_i, X_i), i = 1, \ldots, n\}$ on an underling probability space $(\Omega, \mathcal{F}, \mathcal{P})$, in which $\{(Y_i, T_i, X_i) : T_i = t\}$ are n_t identically distributed random vectors for $t = 1, \ldots, K$, with the ratio n_t/n converging to a constant, $\pi_t \in (0, 1)$ almost surely, as

$n \to \infty$. Observations between groups are assumed to be independent.

Let us denote the pth component of the vector $\boldsymbol{\alpha}_0$ by $\boldsymbol{\alpha}_{0,p}(> 0$, since $\boldsymbol{\alpha}_0 \in \Theta)$. By the completeness property of $\mathbb{R}$, we can always find some $c > 0$ such that $\boldsymbol{\alpha}_{0,p} \geq c$, and therefore, without loss of generality, we may assume that $\boldsymbol{\alpha}_0$ is in a compact set $\Theta_c = \{\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^\top \in \mathbb{R}^p | \sum_{j=1}^p \alpha_j^2 = 1, \alpha_p \geq c\}$, with an appropriate choice of small $c > 0$.

Theorem 1. *(Consistency) Under Assumption 1 to 5 in the Appendix, $\hat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}_0$ almost surely.*

The proof of **Theorem** 1 is given in the Appendix. To avoid the complication from the restricted parameter space Θ_c, we will consider instead the "pth component removed" $R(\boldsymbol{\alpha})$ in (2.5) as follows

$$R(\boldsymbol{\alpha}_{-p}) = R\left(\alpha_1, \dots, \alpha_{p-1}, \sqrt{1 - (\alpha_1^2 + \dots + \alpha_{p-1}^2)}\right), \tag{2.17}$$

where a vector $\boldsymbol{\alpha}_{-p} = (\alpha_1, \alpha_2, \dots, \alpha_{p-1}) \in \mathbb{R}^{p-1}$ lives inside the unit ball. Let the "pth component removed" value of the optimal $\boldsymbol{\alpha}_0$ be denoted by $\boldsymbol{\alpha}_{0,-p} \in \mathbb{R}^{p-1}$. Similarly, let the corresponding profile estimator $\hat{\boldsymbol{\alpha}}$ in (2.8) be denoted by $\hat{\boldsymbol{\alpha}}_{-p} \in \mathbb{R}^{p-1}$.

Theorem 2. *(Asymptotic Normality) Under Assumption 1 to 5 in the Appendix, $\sqrt{n}(\hat{\boldsymbol{\alpha}}_{-p} - \boldsymbol{\alpha}_{0,-p}) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{0,-p}})$ 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 the Appendix.*

The proof of **Theorem** 2 is given in the Appendix. The asymptotic confidence intervals for the index coefficients can be constructed using the asymptotic covariance in **Theorem** 2.

2.6 Simulation studies

2.6.1 Precision of estimators

We investigate the precision of the $\boldsymbol{\alpha}$ estimators in the simple case of $K = 2$ treatment groups and $p = 2$ baseline covariates. The covariates $X = (x_1, x_2)^\top$ are generated from

$\mathcal{N}(0, \Psi_X)$, where $\Psi_X = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$. The outcomes in the two treatment groups are simulated under $Y_t = m_k(X; \delta) + \epsilon_t$ with $\epsilon_k \sim \mathcal{N}(0, 0.1^2)(t = 1, 2)$, where

$$\begin{aligned} m_1(X; \delta) &= 0.5(\boldsymbol{\alpha}_1^\top X)^2 - \delta/3(\boldsymbol{\alpha}_2^\top X)^2 \\ m_2(X; \delta) &= 1 - 0.5(\boldsymbol{\alpha}_1^\top X)^2 + \delta/3(\boldsymbol{\alpha}_2^\top X)^2, \end{aligned} \tag{2.18}$$

in which $\boldsymbol{\alpha}_1 = (1, 1)^\top/\sqrt{2}$ and $\boldsymbol{\alpha}_2 = (-1, 1)^\top/\sqrt{2}$. If $\delta = 0$, the data generation model (2.18) is a "genuine" SIMML (2.2), and the single index of the SIMML is $\boldsymbol{\alpha}_1^\top X$. However, if $\delta \neq 0$, model (2.18) is not in the class of SIMML since it involves two indices, $\boldsymbol{\alpha}_1^\top X$ and $\boldsymbol{\alpha}_2^\top X$, and the approximate SIMML is $g_t(\boldsymbol{\alpha}^\top X) = E[m_t(X; \delta) \mid \boldsymbol{\alpha}^\top X]$, $t = 1, 2$, for some $\boldsymbol{\alpha} \in \Theta$. The parameter δ in (2.18) controls the relative influence of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. We consider two cases in the simulations: $\delta = 0$ ("genuine" SIMML) and $\delta = 1$, for which $\boldsymbol{\alpha}_1$ has a stronger influence than $\boldsymbol{\alpha}_2$ (and so $\boldsymbol{\alpha}_1^\top X$ can be considered as the more important composite covariate that modifies the effect of treatments).

We investigate how the estimator of the index coefficient $\boldsymbol{\alpha} = (a_1, a_2)^\top$ of (approximated) SIMML (2.2) obtained via maximizing the SIMML profile likelihood (i.e., $\hat{\boldsymbol{\alpha}}$ in (2.8)) compares to the estimator obtained by maximizing the L^2 distance between the two link functions from (2.13). For the purposes of visualization we express $\boldsymbol{\alpha} \in \Theta \subset \mathbb{R}^2$ in polar coordinates. If Cartesian coordinates are transformed into polar coordinates, $(a_1, a_2)^\top$ on the unit half circle Θ can be represented by a single parameter, θ for $0 \leq \theta < \pi$, where θ is an angle in radians. Then $\boldsymbol{\alpha}_1$ corresponds to $\theta_1 = \pi/4$ and $\boldsymbol{\alpha}_2$ corresponds to $\theta_2 = 3\pi/4$. As a function of the angle θ, the criterion function of an unbiased estimator would have a peak at θ_1 and a smaller one at θ_2.

We simulated 100 data sets under the above described setup (2.18), and averaged the values of the criterion functions (the profile likelihood and the L^2 distance, respectively) for each $\theta \in [0, \pi]$. The resulting averaged criterion functions are shown on the third and the fourth panels of Figure 2.2, for the case of $\delta = 0$ and $\delta = 1$, respectively. Both the profile likelihood and L^2 distance maximizers have a global peak at $\theta_1 = \pi/4$ and a smaller local peak at $\theta_2 = 3\pi/4$. The profile likelihood maximizer, however, shows a sharper peak around the true value θ_1 compared to the L^2 distance maximizer, indicating better efficiency in estimation, for both $\delta \in \{0, 1\}$.

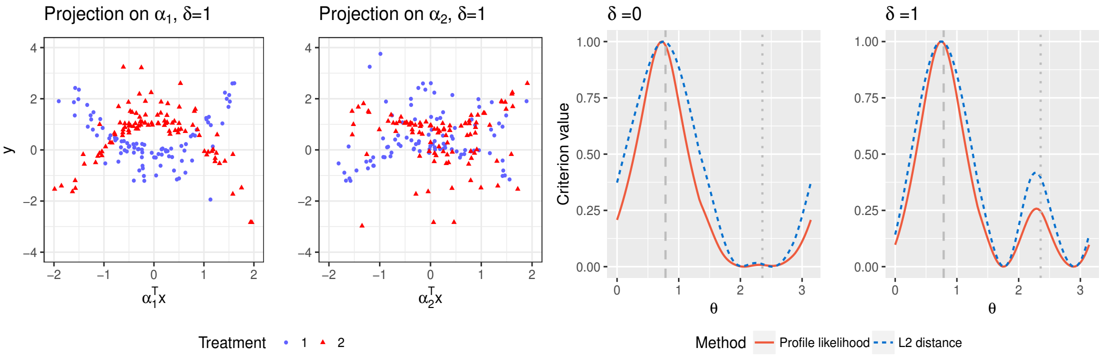

Figure 2.1: The first two panels: the outcomes simulated under model (2.18) when $\delta = 1$, plotted against the "first" index $\boldsymbol{\alpha}_1^\top X$ in the first panel, and against the "second" index $\boldsymbol{\alpha}_2^\top X$ in the second panel, for the two treatment groups (blue dots and red triangles respectively). The third ($\delta = 0$ case) and the fourth ($\delta = 1$ case) panels: the criterion functions of the profile likelihood maximizer (the red solid curve) and the L^2 distance maximizer (the blue dotted curve), averaged over 100 simulated datasets, each scaled to have height 1. The dashed grey vertical line indicates the angle $\theta_1 = \pi/4$ that corresponds to $\boldsymbol{\alpha}_1$, and the vertical dotted grey line indicates $\theta_2 = 3\pi/4$ that corresponds to $\boldsymbol{\alpha}_2$.

2.6.2 ITR performance

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 an ITR for the single decision time point ([Murphy, 2003], [Robins, 2004], [Cai *et al.*, 2011], [Qian and Murphy, 2011]). Given covariates X, an ITR based on SIMML is $\mathcal{D}(X) = \arg\max_{t \in \{1,\ldots,K\}} g_t(\alpha^\top X)$. We investigate the performance of the estimated ITRs of the form $\mathcal{D}(X) = \arg\max_{t \in \{1,\ldots,K\}} \mathbb{E}[Y \mid X, T = t]$, where the conditional expectation is obtained from various modeling procedures.

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

$$Y_t = \delta M(\boldsymbol{\mu}^\top X; \nu) + C_t(\boldsymbol{\alpha}^\top X; \omega) + \epsilon_t \quad (t = 1, 2). \tag{2.19}$$

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 (2.19) controls the relative contribution of the $C_t's$ to the variance in the outcomes, and is calibrated to obtain the relative contribution of 0.35. The contrast functions C_t's in (2.19) are set to

$$C_t(u; \omega) = \begin{cases} C_1(u; \omega) = +1 - \cos\big(0.5\pi\omega u\big) + 0.5(u - \omega) \\ C_2(u; \omega) = -1 + \cos\big(0.5\pi\omega u\big) - 0.5(u - \omega), \end{cases} \tag{2.20}$$

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 Figure 2.2. We set the main effect function M in (2.19) to be

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

where, as ν increases, the degree of nonlinearity in M increases. We considered two cases, $\nu = 0$, corresponding to a *linear* M; and $\nu = 1$, corresponding to a *nonlinear* M, illustrated

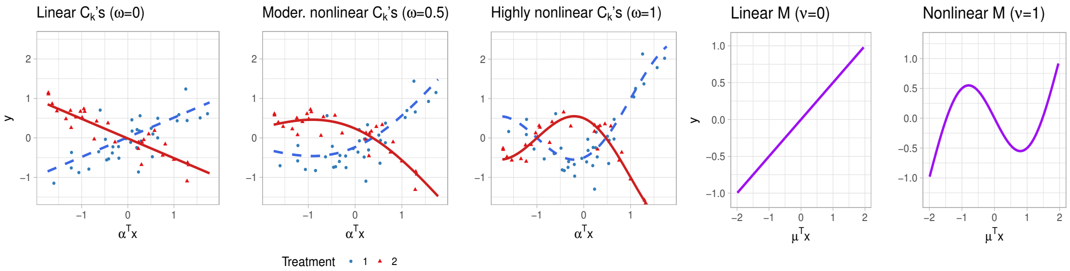

Figure 2.2: The first panel shows the *linear* contrast C_t's ($\omega = 0$), the second panel the *moderately nonlinear* contrast C_t's ($\omega = 0.5$), and the third panel displays *highly nonlinear* contrast C_t's ($\omega = 1$). Data points are generated from model (2.19) with $\delta = 0$ and $p = 5$. The fourth and the fifth panel shows the *linear* ($\nu = 0$) and the *nonlinear* main effect M ($\nu = 1$), respectively.

in the fourth and the fifth panels of Figure 2.2. We considered $p = 5$ 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 unequal sample sizes: $n_1 = 40$ and $n_2 = 30$. We used $d = 5$ B-spline basis functions. We compared the treatment decision rules determined based on the following regression models: (i) SIMML (2.2) estimated from maximizing the profile likelihood; (ii) the K-Index model (2.1) 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. We reported the boxplots of PCDs obtained from 200 training datasets.

Figure 2.3 shows that SIMML outperformed all other methods, except for the case under the *linear* M and C_t's in which all 4 approaches performed well. The K-Index method was 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, for more complex M function ($\nu = 1$) (the bottom panels), the performance of the K-index model was considerably worse compared

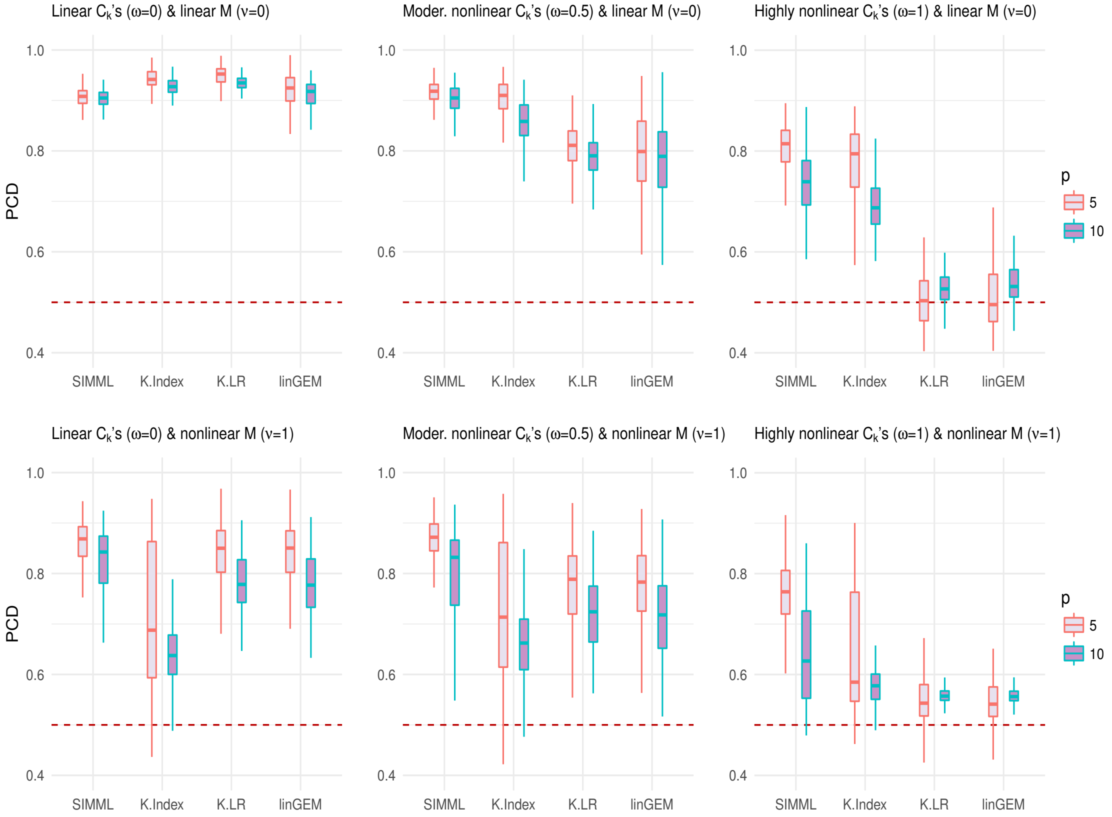

Figure 2.3: Boxplots of the PCDs of the treatment decision rules obtained from 200 training datasets for each of the four methods. Each panel corresponds to one of the six combinations of $\omega \in \{0, 0.5, 1\}$ and $\nu \in \{0, 1\}$: the shape of the contrast functions C_t's controlled by ω; the shape of the main effect function M controlled by ν; the number of predictors $p \in \{5, 10\}$. The sample sizes are $n_1 = 40$, $n_2 = 30$.

to SIMML. When the underlying model is complex, given a relatively small sample size, the SIMML in which the treatment contrast was emphasized through the common single-index was more effective in estimating treatment decision rules than the K-Index model. As would be expected, additional complexity in C_t's ($\omega = 0.5$ and $\omega = 1$) had a greater effect on the performance of the more restrictive models (linGEM and K-LR) than it did on the flexible models (SIMML and K-index). It was clear that the number of covariates p also had a major impact on the performance of all methods. As p changed from 5 (red) to 10 (blue), the deterioration in performance was more pronounced for the K-Index model that requires separate fits for each treatment and thus involve estimation of more parameters ($K(p-1) + Kd$), than the more parsimonious SIMML with a fewer number of parameters ($p - 1 + Kd$) to be estimated.

2.6.3 Coverage probability of asymptotic 95% confidence intervals

We present a simulation experiment that assesses the coverage probability of the asymptotic confidence intervals derived from **Theorem** 2. The data were generated under model (2.19) with $\delta = 0$ (i.e., no main effect M) with $p = 5$ predictors. 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 contrast functions, C_t's, are given by (2.20), and two levels of the curvature of the contrasts are considered, corresponding to a single and multiple-crossings cases, $\omega \in \{0, 1\}$, respectively (see Figure 2.2). To set signal to noise ratio at 1 for both scenarios, the noise standard deviations were set to 0.64 and 0.89, corresponding to $\omega = 0$ and $\omega = 1$ respectively. We considered unequal sample sizes for the $K = 2$ groups by setting $n = n_1 + n_2$ where $2n_1 = 3n_2$. With varying $n \in \{100, 200, 400, 800, 1600, 3200, 6400\}$, the number of interior knots used in the B-spline approximation, was determined to be $N = \left[n_1^{1/5.5}\right]$ as recommended by [Wang and Yang, 2009]. Five hundred data sets 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 the table in the Appendix. The 5th (i.e., the pth) element is estimated to satisfy the constraint $\boldsymbol{\alpha} \in \Theta$ in Theorem 2.

As the the sample size increases, the "actual" coverage probability gets closer to the "nominal" coverage probability, with better coverage results for the single-crossing scenario ($\omega = 0$) compared to the multiple-crossing scenario ($\omega = 1$).

2.7 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 determining ITRs 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. Patients were tested Flanker [Flanker and Eriksen, 1974] and A not B Working Memory (AnotB; [Herrera-Guzman *et al.*, 2009]), for which reaction time (RT) and accuracy were assessed. In addition, RT was recorded for a choice task [Deary *et al.*, 2011]. Four baseline clinical and demographic characteristics were also assessed: (i) current patient age; (ii) severity of depressive symptoms measured by the Hamilton Rating Scale for Depression (HRSD); (iii) duration of the current major depressive episode; and (iv) age of onset of first major depressive episode. Table 2.1 summarizes the information on the $p = 9$ baseline patient characteristics, $X = (x_1, \dots, x_9)^\top$, starting with the means and standard deviations of the original (untransformed) covariates. The treatment outcome Y was the improvement in symptom severity from baseline to week 8 and thus larger values of the outcome were better.

Figure 2.4 shows the outcome Y against each of the 9 baseline covariates for placebo (blue) and active drug (red). The estimated B-spline approximated curves 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. From the figure, we can see that each individual covariate marginally has at most a small modifying treatment effect.

One natural measure for the effectiveness of an ITR ($\mathcal{D}$) is the expected mean outcome

(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

Table 2.1: 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.") or the linear regression ("linear"); the estimated singe-index of the three (single-index based) methods, with the estimated values of the treatment decision rules.

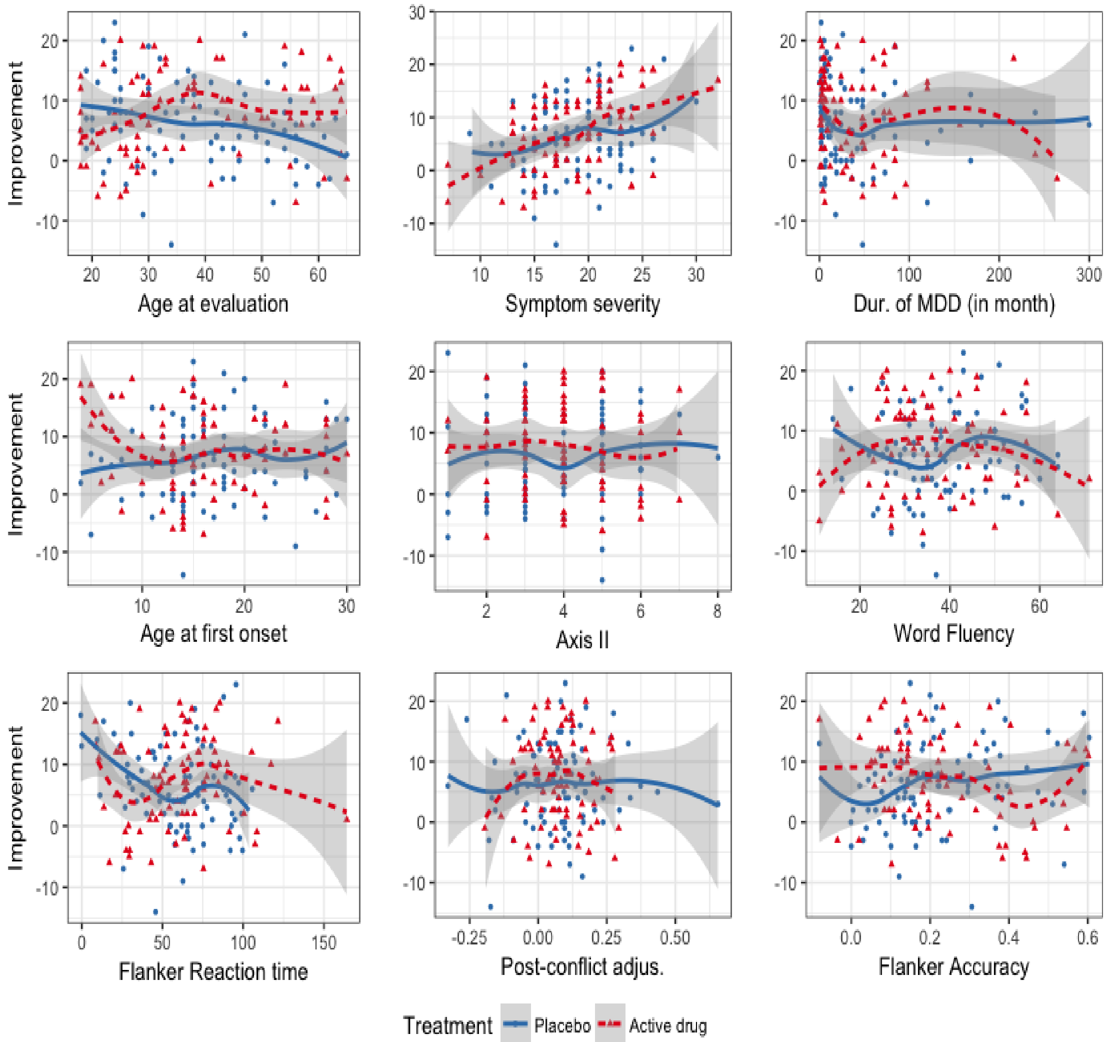

Figure 2.4: For each of the 9 baseline covariates individually, treatment-specific spline approximated regression curves with 5 basis functions are overlaid on to the data points; the placebo group is the blue solid curve and the active drug group is the red dotted curve. The associated 95% confidence bands of the regression curves were also plotted.

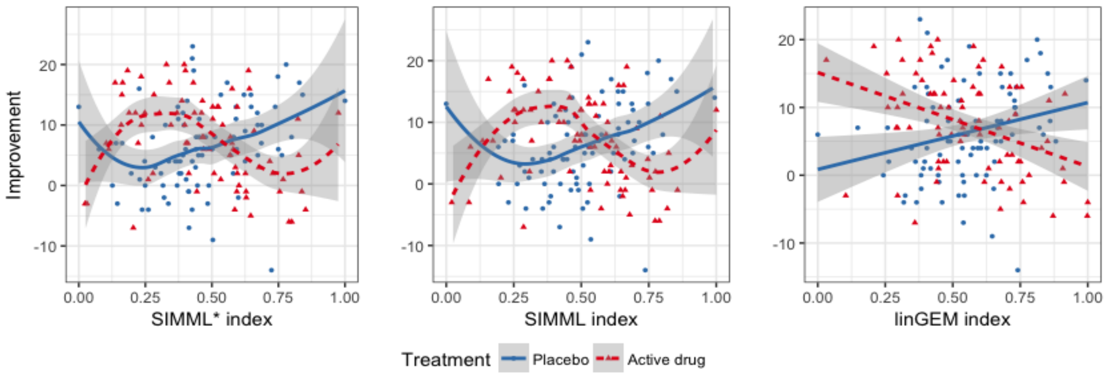

Figure 2.5: Pair of estimated link functions (g_1 and g_2) obtained from SIMML with the "main effect adjusted" profile likelihood (first panel), SIMML with the (main effect unadjusted) profile likelihood (second panel), and the linear GEM model estimated under the criterion maximizing the difference 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 group, observed values of the outcomes are plotted against the estimated index.

if everyone in the population receives treatment according to that rule, the "value" (V) (1.1) of a decision rule $\mathcal{D}$ ([Qian and Murphy, 2011]):

$$\mathrm{V}(\mathcal{D}) = \mathbb{E}_X \left[\mathbb{E}_{Y|X}[Y \mid X, T = \mathcal{D}(X)] \right]. \tag{2.21}$$

In Table 2.1, "Indiv. Value" refers to the estimated value of an ITR based on each individual covariate, based on two approaches for determining ITRs: the aforementioned B-spline approximated regressions of the outcome on a single covariate (nonpar. links) as suggested by the overlaied curves in Figure 2.4, and linear regressions of the outcome on a single covariate (linear links). The value (2.21) of an ITR $\mathcal{D}$ can be estimated by the inverse probability weighted estimator ([Murphy, 2005]):

$$\hat{V}(\mathcal{D}) = \sum_{i=1}^{\tilde{n}} Y_i I_{T_i = \mathcal{D}(X_i)} / \sum_{i=1}^{\tilde{n}} I_{T_i = \mathcal{D}(X_i)}, \tag{2.22}$$

using a testing set, say, $\{(Y_i, X_i, T_i), i = 1, \ldots, \tilde{n}\}$, in which, if we use only each individual covariate, then $X_i = x_{ij}$, for each $j = 1, \ldots, 9$. 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

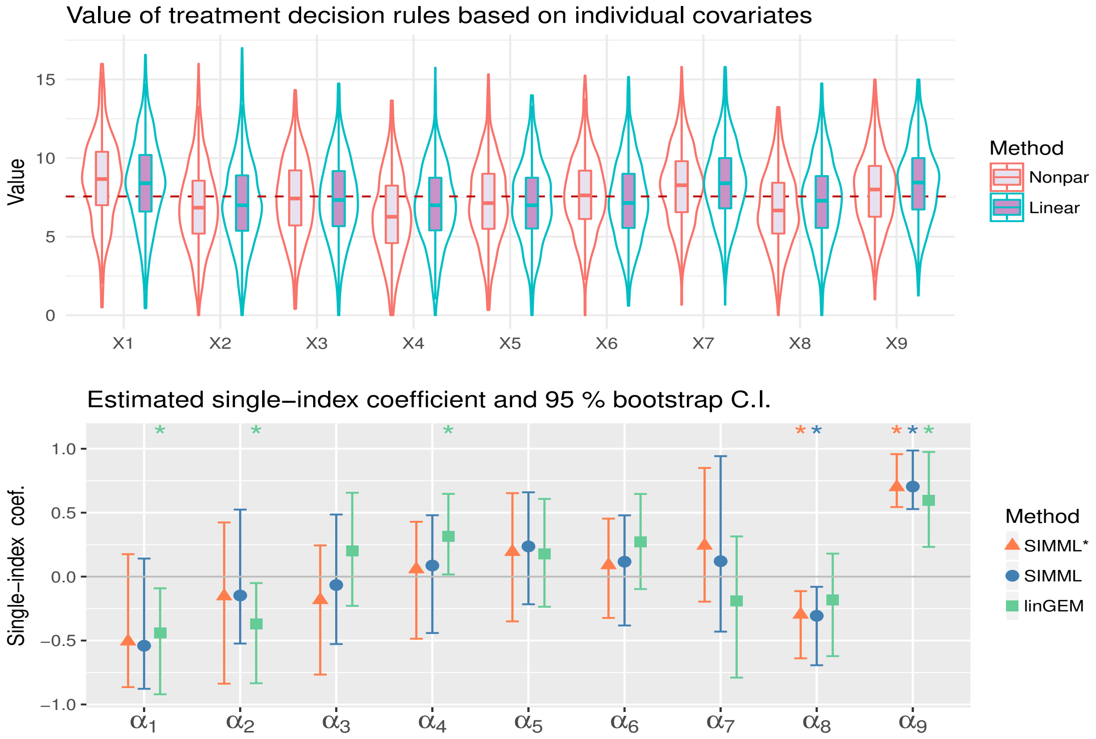

Figure 2.6: Top row: Violin plots of the estimated values of ITRs based on each of the individual predictors $x_1, \ldots, x_9$, determined from univariate nonparametric and linear regressions, respectively, obtained from 500 randomly split testing sets (with higher values preferred). Bottom row: The estimated single-index coefficients $\alpha_1, \ldots, \alpha_9$, associated with the covariates $x_1, \ldots, x_9$. The associated 95% confidence intervals obtained from BC$_a$ bootstrap with 500 replications are illustrated. Estimated significant coefficients are marked with $*$ on the top.

time estimating $\mathcal{D}$ on the training set and computing (2.22) from the testing set. We reported the averaged values. The last three columns of Table 2.1 show the estimated index coefficients ($\boldsymbol{\alpha}$) obtained by two different SIMMLs and the linear GEM (linGEM) method.

The SIMML can be made more efficient by incorporating a main effect component $\boldsymbol{\beta}^\top D(X)$ in the model, i.e., we consider $\mathbb{E}\left(Y \mid T=t, X=x\right) = \boldsymbol{\beta}^T D(x) + g_t(\boldsymbol{\alpha}^\top x)$, for an appropriate vector-valued function $D(X)$. If the $n \times q$ matrix $\mathbb{D}$ is the evaluation of $D(X)$ on the sample data, then for each $\boldsymbol{\alpha}$, the profile loglikelihood under this extended model (with Gaussian outcome), up to constants, is $Q^*(\boldsymbol{\alpha}) = \|(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}})\tilde{\boldsymbol{Y}}\|^2$, 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 Figure 2.5, the estimated pairs of link functions are plotted against the approach-specific single index variable, obtained from applying the two SIMML approaches and the linear GEM approach. The shapes of the regression curves capture a nonlinear treatment-by-index interaction effect, especially due to some non-monotone relationship between the index in the outcome in the drug group. In Figure 2.6, the coefficient estimates from each of those single index-based methods, and the associated 95% confidence intervals obtained from a bias-corrected and accelerated (BC$_a$, [DiCiccio and Efron, 1996]) bootstrap with 500 replications are presented. The estimated single index coefficients reflect the relative importance of the baseline covariates $x_1, \ldots, x_9$ in determining a composite treatment effect modifier, $\boldsymbol{\alpha}^\top X$, that is used for defining the ITRs.

In this analysis, the incorporation of the "main effect" component improved the value of ITRs determined from the proposed SIMML method, as illustrated in the boxplots in Figure 2.7; 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 (2.22) of the ITRs. 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 Figure 2.7.

The proposed SIMML approaches, in terms of the averaged estimated values (2.22) estimated from the aforementioned 500 randomly split testing sets, appeared to outperform

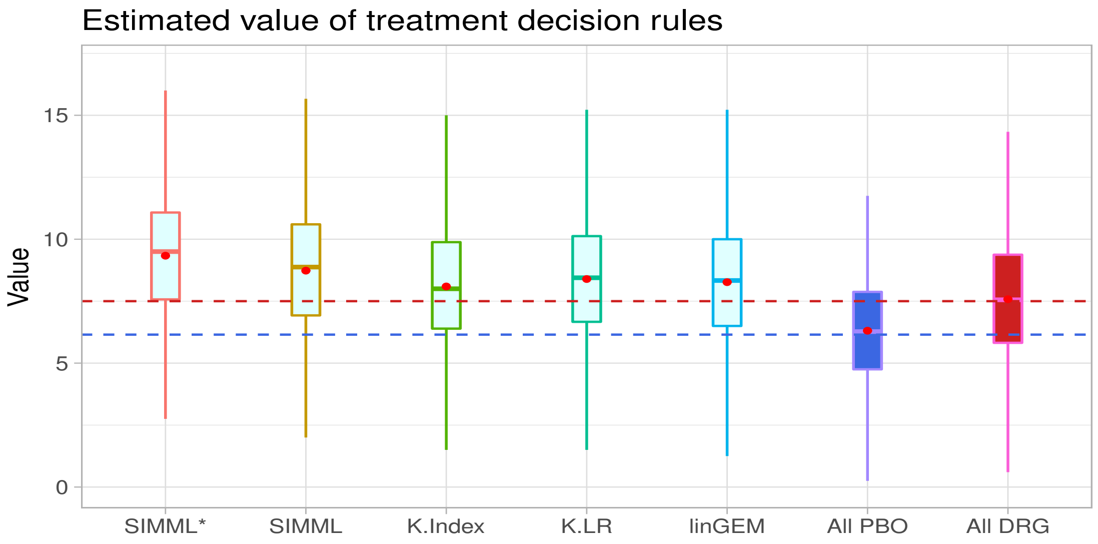

Figure 2.7: Boxplots of the estimated values of ITRs obtained from the 500 randomly split testing sets (higher values are preferred). The estimated values (and the standard deviations) are given as follows: SIMML*: 9.34 (2.68); SIMML: 8.72 (2.68); K-Index: 8.04 (2.69); K-LR: 8.36 (2.69); linGEM: 8.22 (2.67); All PBO: 6.17 (2.63); All DRG: 7.57 (2.67).

K-Index, exceeding the value estimated for the policy of assigning everyone to receive the active drug, while also outperforming linGEM and K-LR. The visualization (see Figure 2.5) 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 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 ITRs.

2.8 Discussion

Model (2.6) can be extended 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 $Q_t(\boldsymbol{\alpha}) = \|(\boldsymbol{I}_{n_t} - \mathbb{S}_{\boldsymbol{\alpha},t})\boldsymbol{Y}_t\|^2/n_t$. The corresponding profile estimator of $\boldsymbol{\alpha}$ is $\underset{\boldsymbol{\alpha}\in\Theta}{\arg\min} \sum_{t=1}^{K} n_t \log Q_t(\boldsymbol{\alpha})$. The estimation can be performed similarly as in the estimation of $\hat{\boldsymbol{\alpha}}$ in (2.8), 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 $f_Y(Y;\theta,\phi) = \exp\{(Y\theta - b(\theta))/a(\phi) + c(Y,\phi)\}$, given a canonical link function $h(\cdot)$. We can extend the SIMML approach to the GLM setting with treatment-specific natural parameters θ_k, $k \in \{1,\dots,K\}$ by modeling the treatment-specific outcomes as a function of a single index $\boldsymbol{\alpha}^\top X$: $\theta_t(x) = h^{-1}(\mathbb{E}(Y \mid T = t, X = x)) = g_t(\boldsymbol{\alpha}^\top x)$, $t \in \{1,\dots,K\}$; $g_t(\cdot)$ and hence $\theta_t(x)$ will be approximated by cubic B-splines. The approximates can be denoted by $\tilde{\theta}_t(x) = \boldsymbol{\eta}_t^\top Z(\boldsymbol{\alpha}^\top x)$ for some $\boldsymbol{\eta}_t \in \mathbb{R}^d$, as in Section 2.4. The general strategy of nonlinear maximization of the profile likelihood over $\boldsymbol{\eta}_t$, for each $\boldsymbol{\alpha}$ can be employed. The dispersion parameter ϕ can also be profiled out. Other potential extensions involve performing variable selection in high-dimensional covariate settings and incorporating functional-valued data objects (such as images) as patient covariates.

Chapter 3

A Constrained single-index model with multiple-links for interactions

3.1 Introduction

A major challenge in constructing ITRs from a dataset with (Y_i, T_i, X_i), $i = 1, \ldots, n$, lies in the detection of relatively small treatment effect-modification-related variations (i.e., the T-by-X interaction effect on the outcome) against relatively large non-treatment-related variations (i.e., the main effect of X on the outcome). In this chapter, we target the question of estimating $\boldsymbol{\alpha}^\top X$ in the SIMML model (2.2) that captures the treatment effect modification-related variabilities in the dataset, in the presence of relatively large, unspecified main effect-related variabilities.

We propose a semiparametric regression approach specifically designed to model the interactions between a treatment indicator and predictors, without the need to model the main effects. By obviating the need to model the main effect, the proposed approach extends the modified covariate approach of [Tian *et al.*, 2014] into a semiparametric regression framework, that uses treatment-specific nonparametrically-defined link functions. The method provides a valuable regression approach that allows nonlinear interaction effects in developing ITRs.

[Tian *et al.*, 2014] proposed a simple and elegant approach to estimate the potentially large number of covariates and treatment interactions, without the need to model the main effects in analyzing data from a RCT. The method termed the modified covariate approach (MCA) simply codes the treatment variable as ±0.5 and then includes the products of this variable with each covariate under a linear model framework for estimating interactions. When the dimension of covariates, p, is high, MCA can also incorporate appropriate regularization procedures to select treatment effect modifiers that interact with the treatment.

In this chapter, we will extend SIMML (2.2) to have an unspecified main effect component on the outcome, and this approach takes MCA as a special case under a linear link restriction. This generalization is analogous to the extension from a classical linear model into a single-index model ([Brillinger, 1982], [Stoker, 1986], [Powell *et al.*, 1989], [Hardle *et al.*, 1993], [Xia and Li, 1999], [Horowitz, 2009], [Antoniadis *et al.*, 2004]), which employs a flexible data-driven link function to estimate the mean response, in the context of estimating the treatment-by-predictors interactions.

In Chapter 2, we introduced a flexible approach for determining composite variables that permit nonlinear association with the outcome Y, using a SIMML,

$$\mathbb{E}(Y \mid X, T) = g_T(\boldsymbol{\alpha}^\top X), \quad T \in \{1, \ldots, K\}. \tag{3.1}$$

which employs a single projection $\boldsymbol{\alpha}^\top X$ to summarize the variability of the baseline predicdtors, and multiple (treatment-specific) nonparametrically-defined link functions, $g_T(\cdot)$, $T \in \{1, \ldots, K\}$, to connect the derived single-index to the treatment-specific mean responses.

Although model (3.1) provides a parsimonious generalization of SIM for modeling the interaction effect between T and X, it assumes a fairly restrictive model for a main effect, since both the main effect of X and the T-by-X interaction effect are restricted to be functions of the common single-index $\boldsymbol{\alpha}^\top X$.

Consider a more general model as a true model, for instance, $\mathbb{E}(Y \mid X, T) = \mu(\boldsymbol{\mu}^\top X) + g_T(\boldsymbol{\alpha}^\top X)$, for some distinct vectors $\boldsymbol{\mu} \neq \boldsymbol{\alpha}$ and some (smooth) main effect function $\mu(\cdot)$, and the interaction effect functions $g_T(\cdot)$, $T \in \{1, \ldots, K\}$. If one tries to fit SIMML (3.1) to the data when the variance of the main effect $\mu(\boldsymbol{\mu}^\top X)$ is larger than that of the interaction effect $g_T(\boldsymbol{\alpha}^\top X)$, then a standard least squares estimate of $\boldsymbol{\alpha}$ would look more like $\boldsymbol{\mu}$, rather than the $\boldsymbol{\alpha}$ involved in the T-by-X interaction effect functions. Consequently, the estimate of $\boldsymbol{\alpha}$ obtained from model (3.1) would not be very informative for estimating the interactions.

In the following section, we will introduce an extended version of SIMML that includes an unspecified main effect component, which can solve the limitation of the SIMML (3.1).

3.2 Models

In this chapter, we will assume a completely unspecified structure, say $\mu(X)$, for the main effect component, and consider an extension from model (3.1)

$$\mathbb{E}(Y \mid X, T) = \underbrace{\mu(X)}_{\text{“main” effect}} + \underbrace{g_T(\boldsymbol{\alpha}^\top X)}_{\text{“interaction” effect}}, \quad T \in \{1, \dots, K\}, \tag{3.2}$$

and develop a methodology to estimate $g_T(\boldsymbol{\alpha}^\top X)$. Model (3.2) removes the single-index restriction of the main effect model. More precisely, we will develop a methodology that obviates the need to model the main effect $\mu(X)$ when estimating the interaction effect. As in Chapter 2, $\boldsymbol{\alpha} \in \mathbb{R}^p$ corresponds to a direction that we project X into, hence is restricted to have a unit norm $\|\boldsymbol{\alpha}\| = 1$, and the set of link functions $g_T(\cdot)$, $T \in \{1, \dots, K\}$, models the nonlinear treatment T-by-X interaction effects, as general smooth functions of the single-index $\boldsymbol{\alpha}^\top X$.

MCA, which has been shown to be an effective methodology under a linear model to estimate the interaction effects between a binary treatment indicator and covariates, assumes

$$\mathbb{E}(Y \mid X, T) = \mu(X) + \boldsymbol{\alpha}^\top X (-1)^T / 2, \tag{3.3}$$

where $T \in \{1, 2\}$ with an equal probability 0.5. The form of the main effect $\mu(X)$ is left unspecified. If we take $g_T(u) = u(-1)^T/2$ and the number of treatment options $K = 2$, SIMML (3.2) reduces to MCA (3.3). In both (3.2) and (3.3), without loss of generality, we take the centered Y and centered/scaled $X \in \mathbb{R}^p$ within each treatment. We note that MCA, as introduced by [Tian *et al.*, 2014], is limited to the case of the binary treatment indicator (i.e., two treatment groups). (3.2) extends the method to the context of a general K treatment group case.

In the (extended) SIMML (3.2), a necessary and sufficient condition for orthogonality between the main effect component and the interaction effect component

$$\mu(X) \perp g_T(\boldsymbol{\alpha}^\top X), \quad \text{a.s.,} \tag{3.4}$$

is given by (if a nontrivial main effect)

$$\mathbb{E}_T\left(g_T(\boldsymbol{\alpha}^\top X) \mid X\right) = 0, \quad \text{a.s.}, \tag{3.5}$$

i.e., the link function $g_T(\cdot)$ has mean zero with respect to the treatment indicator T. The condition (3.5) for the orthogonality can be easily verified, using the law of iterated expectations and independence between T and X, as in the case of an RCT.

For model (3.2), the orthogonality (3.4) is attractive, since the main effect, $\mu(X)$, and the interaction effect, $g_T(\boldsymbol{\alpha}^\top X)$, can be estimated separately. When our interest is in interactions, this suggests a simpler working model than (3.2), using the interaction effect component only and no main effect

$$\mathbb{E}(Y \mid X, T) = g_T(\boldsymbol{\alpha}^\top X), \quad T \in \{1, \ldots, K\}, \tag{3.6}$$

subject to (3.5), which gives the orthogonality (3.4). We will call model (3.6) a constrained SIMML (a SIMML, constrained by the orthogonality condition (3.5)). Working with a constrained SIMML is appealing, since we do not have to specify the form of the main effects, side-stepping issues with misspecification of $\mu(X)$, which is potentially a complicated function. In the following, we propose the criterion for optimizing model (3.2) by using the working model (3.6).

3.3 Criterion

To optimize the interaction effect component $g_T\big(\boldsymbol{\alpha}^\top X\big)$ in model (3.2), we propose the following constrained least squares criterion

$$\begin{aligned} &\underset{\boldsymbol{\alpha}, g_T}{\text{minimize}} \quad \mathbb{E}\Big(Y - g_T\big(\boldsymbol{\alpha}^\top X\big)\Big)^2 \\ &\text{subject to} \quad \mathbb{E}_T\left(g_T\right) = 0. \end{aligned} \tag{3.7}$$

Additionally, we need an identifiability condition for $\boldsymbol{\alpha}$ due to the nonparametric nature of $g_T(\cdot)$, for example, $\|\boldsymbol{\alpha}\| = 1$ with its first component $\alpha_1 > 0$. For a fixed $\boldsymbol{\alpha}$, we can write (3.7) in the penalized Lagrangian form

$$\mathbb{E}\Big(Y - g_T\big(\boldsymbol{\alpha}^\top X\big)\Big)^2 + \lambda \mathbb{E}_T\left(g_T\right), \tag{3.8}$$

where $\lambda > 0$ is the Lagrange multiplier. The minimizer function g_T of (3.8), for each $\boldsymbol{\alpha}$, has a closed-form expression in the population setting.

Theorem 3. *Given $\boldsymbol{\alpha}$, the minimizer g_T of (3.8) satisfies*

$$g_T(\boldsymbol{\alpha}^\top X) = \mathbb{E}\left(Y \mid \boldsymbol{\alpha}^\top X, T\right) - \mathbb{E}\left(Y \mid \boldsymbol{\alpha}^\top X\right), \quad a.s.. \tag{3.9}$$

The proof of Theorem 3 is in the Appendix. Theorem 3 suggests that the optimization problem (3.7) in the population setting can be split into two iterative steps. First, given $\boldsymbol{\alpha}$, the link g_T can be found by (3.9). Second, given the link g_T, $\boldsymbol{\alpha}$ can be found by minimizing

$$\mathbb{E}\Big(Y - g_T(\boldsymbol{\alpha}^\top X)\Big)^2, \tag{3.10}$$

subject to $\|\boldsymbol{\alpha}\| = 1$, with $\alpha_1 > 0$, for model identifiability. These two steps can be iterated until convergence, to obtain a population solution of (3.7). To obtain a sample counterpart of the population solution, we can insert sample estimates into the population algorithm, as in fitting generalized additive models ([Hastie and Tibshirani, 1999]).

3.4 Estimation

Suppose that we are given data (Y_i, T_i, X_i), $i = 1, \ldots, n$, where $X_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, the treatment indicator T_i takes a value $t \in \{1, \ldots, K\}$, and $n = \sum_{t=1}^{K} n_t$ is the total sample size, in which n_t denotes the sample size for the tth treatment group, i.e., $\{i \mid T_i = t, i = 1, \ldots, n\}$.

Let us write the $n \times 1$ vector $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_K^\top\right)^\top$, in which the $n_t \times 1$ vector $\boldsymbol{Y}_t = (Y_1, \ldots, Y_{n_t})^\top$ is the observed response vector that corresponds to the tth treatment group. For the regression function $g_T(\boldsymbol{\alpha}^\top X)$, let us write the $n \times 1$ stacked-up vector, $\boldsymbol{g}_{\boldsymbol{\alpha}} = \left(\boldsymbol{g}_{\boldsymbol{\alpha},1}^\top, \ldots, \boldsymbol{g}_{\boldsymbol{\alpha},K}^\top\right)^\top$, where the $n_t \times 1$ vector $\boldsymbol{g}_{\boldsymbol{\alpha},t} = \left(g_t(\boldsymbol{\alpha}^\top X_{1_t}), \ldots, g_t(\boldsymbol{\alpha}^\top X_{n_t})\right)^\top$ is the evaluation vector of the observations from the tth treatment group, for each $t \in \{1, \ldots, K\}$. Both the link functions $g_t(\cdot)$'s and the projection vector $\boldsymbol{\alpha}$ need to be estimated.

3.4.1 Algorithm

We need to approximate the conditional expectations in (3.9) to obtain a sample estimate of the population solution. Any nonparametric smoothers can be used to approximate

(3.9), for example, B-splines [de Boor, 2001] and local kernel regression ([Ruppert and Wand, 1994], [Hardle and Muller, 2012]). For each candidate $\boldsymbol{\alpha}$, let $\mathbb{S}_{\boldsymbol{\alpha}}^{(**)}$ denote a suitable nonparametric smoother for approximating the bivariate conditional expectation $\mathbb{E}(Y \mid \boldsymbol{\alpha}^\top X, T)$ in (3.9). Similarly, let $\mathbb{S}_{\boldsymbol{\alpha}}^{(*)}$ denote a suitable nonparametric smoother for the univariate conditional expectation $\mathbb{E}(Y \mid \boldsymbol{\alpha}^\top X)$ in (3.9). Then we can define a smoother $\mathbb{S}_{\boldsymbol{\alpha}}$ for approximating g_T in (3.9) that smooths the response vector $\boldsymbol{Y}$

$$\mathbb{S}_{\boldsymbol{\alpha}} = \mathbb{S}_{\boldsymbol{\alpha}}^{(**)} - \mathbb{S}_{\boldsymbol{\alpha}}^{(*)} \tag{3.11}$$

$$\hat{\boldsymbol{g}}_{\boldsymbol{\alpha}} = \mathbb{S}_{\boldsymbol{\alpha}}\boldsymbol{Y}, \tag{3.12}$$

and obtain the link estimates $\hat{g}_T$, $T \in \{1, \ldots, K\}$. For given $\hat{g}_T$'s, a sample counterpart for (3.10) is

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \sum_{i=1}^{n} \big(Y_i - \hat{g}_{T_i}(\boldsymbol{\alpha}^\top X_i)\big)^2 \quad \big(= \|\boldsymbol{Y} - \hat{\boldsymbol{g}}_{\boldsymbol{\alpha}}\|^2\big), \\ &\text{subject to} \quad \|\boldsymbol{\alpha}\| = 1, \alpha_1 > 0, \end{aligned} \tag{3.13}$$

which can be solved for $\boldsymbol{\alpha}$, for example, by iteratively weighted least squares (IWLS) (e.g., [Nelder and Wedderburn, 1972]) fixing the link estimates $\hat{g}_T$'s. Expressions (3.12) and (3.13) suggest an iterative procedure for approximately solving (3.7) over the link function g_T and $\boldsymbol{\alpha}$. We summarize below an algorithm that alternates between updating the link functions (i.e., Step 1), and updating $\boldsymbol{\alpha}$ (i.e., Step 2).

Algorithm 2 Estimation of constrained SIMML

```
1: Initialize $\boldsymbol{\alpha}$ (e.g., using an MCA estimate)
2: for iteration until convergence do
3:     Fix $\boldsymbol{\alpha}$, compute $\mathbb{S}_{\boldsymbol{\alpha}}$ in (3.11), and obtain $\hat{\boldsymbol{g}}_{\boldsymbol{\alpha}}$ by (3.12) and $\hat{g}_T$'s.
4:     Fix $\hat{g}_T$'s, and obtain $\boldsymbol{\alpha}$ by solving (3.13).
5: end for
```

Remark 1. *For the $K = 2$ case, if the link function $g_T(u)$ is restricted to be linear (i.e., $g_T(u) = \gamma_T u$ for some scalar γ_T, for $T \in \{1, 2\}$), then the constrained SIMML estimate of $\boldsymbol{\alpha}$ reduces to the MCA estimate of $\boldsymbol{\alpha}$ in (3.3) up to a constant multiplier. A justification for this is provided in the Appendix.*

3.4.2 Main effect augmentation

In this section, we will describe how to estimate the extended SIMML (3.2) when we have a working main effect model for $\mu(X)$, say, $\mu(X) = \boldsymbol{\mu}^\top D(X)$, with an appropriate vector-valued function $D(X) \in \mathbb{R}^q$ and an unknown $\boldsymbol{\mu} \in \mathbb{R}^q$. In the population setting, to optimize the working model (3.6) with a main effect, we suggest the following constrained criterion, which extends (3.7)

$$\begin{aligned} &\underset{\boldsymbol{\mu},\boldsymbol{\alpha},g_T}{\text{minimize}} \quad \mathbb{E}\Big(Y - \boldsymbol{\mu}^\top D(X) - g_T\big(\boldsymbol{\alpha}^\top X\big)\Big)^2 \\ &\text{subject to} \quad \mathbb{E}_T\left(g_T\right) = 0, \end{aligned} \tag{3.14}$$

where $\|\boldsymbol{\alpha}\| = 1$ and $\alpha_1 > 0$ for model identifiability. Given $\boldsymbol{\alpha}$, the minimizer g_T of (3.14) satisfies

$$g_T(\boldsymbol{\alpha}^\top X) = \mathbb{E}(R \mid \boldsymbol{\alpha}^\top X, T) - \mathbb{E}(R \mid \boldsymbol{\alpha}^\top X), \quad \text{a.s.,} \tag{3.15}$$

where $R = Y - \boldsymbol{\mu}^\top D(X)$; and the minimizer $\boldsymbol{\mu}$ of (3.14) satisfies

$$\mathbb{E}\big(D(X)^\top D(X)\big)\boldsymbol{\mu} = \mathbb{E}\big(D(X)^\top\big(Y - g_T(\boldsymbol{\alpha}^\top X)\big)\big). \tag{3.16}$$

For sample data, based on (3.15) and (3.16), estimation of g_T and $\boldsymbol{\mu}$, for each fixed $\boldsymbol{\alpha}$, can be alternated until convergence as in Algorithm 2. If necessary, an appropriate regularization (e.g., the Lasso ([Tibshirani, 1996])) can be employed to approximately solve (3.16) based on the sample data. A good choice of a working model $\boldsymbol{\mu}^\top D(X)$ for the main effect $\mu(X)$ can increase the efficiency of the estimator, analogous to the efficiency augmentation in [Tian *et al.*, 2014].

3.4.3 Details for estimating the projection vector

In this subsection, we describe how to estimate the projection vector $\boldsymbol{\alpha}$. In implementing the algorithm 2, the link estimates $\hat{g}_T$'s can be approximated, for example, by a cubic polynomial. Based on such approximated link functions, IWLS can be performed to approximately minimize (3.13) over $\boldsymbol{\alpha}$. The estimated $\hat{\boldsymbol{\alpha}}$ can be rescaled to have unit L^2 norm (with $\hat{\alpha}_1 > 0$). To implement the IWLS, we provide an expression for the gradient vectors, as follows. For each fixed $\boldsymbol{\alpha}$, let $R_i = Y_i - \hat{g}_{T_i}(\boldsymbol{\alpha}^\top X_i)$ (or, if we account for the working main

effect, set $R_i = Y_i - \hat{\boldsymbol{\mu}}^\top D(X_i) - \hat{g}_{T_i}(\boldsymbol{\alpha}^\top X_i))$, and the residual vector $\boldsymbol{R} = (R_1, \ldots, R_n)^\top$. Let us write the $n \times 1$ vector of the first derivative of the (approximated) link functions $\hat{g}'_{T_i}(\boldsymbol{\alpha}^\top X_i)$ evaluated at $\{(T_i, X_i)_{i=1}^n\}$ by
$\hat{\boldsymbol{g}}'_{\boldsymbol{\alpha}} = (\hat{g}'_1(\boldsymbol{\alpha}^\top X_{1_1}), \ldots, \hat{g}'_1(\boldsymbol{\alpha}^\top X_{n_1}), \ldots, \hat{g}'_K(\boldsymbol{\alpha}^\top X_{1_K}), \ldots, \hat{g}'_K(\boldsymbol{\alpha}^\top X_{n_K}))^\top$. Then the jth component of the gradient of the residual sum of squares with respect to $\boldsymbol{\alpha}$ is

$$\nabla_j = -\boldsymbol{R}^\top \left(\hat{\boldsymbol{g}}'_{\boldsymbol{\alpha}} * \boldsymbol{X}_j \right), \quad j = 1, \ldots, p, \tag{3.17}$$

where $\boldsymbol{X}_j = (X_{1,j}, \ldots, X_{n,j})^\top$ is the observed $n \times 1$ vector for the jth predictor, and $*$ denotes element-wise multiplication of vectors.

3.5 Connection to the modified covariate approach

In this section, we will describe connections between the MCA and the methods of using treatment-specific link functions, for example, SIMML (3.2). In comparison to the MCA models (3.3), SIMML (3.2) permits a nonlinear interaction between T and a linear combination of the predictors, $\boldsymbol{\alpha}^\top X$, on the outcome. [Petkova *et al.*, 2016] called the composite predictor $\boldsymbol{\alpha}^\top X$ a generated effect modifier (GEM), however, their treatment-specific link function was restricted to be a linear function. An equivalence between a GEM model under the orthogonality constraint and the MCA will be described in this section.

3.5.1 Sufficient reduction

As a function of X, let us consider the individualized treatment effect contrast, defined as

$$C(X) := \sum_{t=1}^{K} c_t \mathbb{E}\left(Y \mid X, T = t\right) \tag{3.18}$$

where the "contrast" vector $(c_1, \ldots, c_K)^\top$ is under a zero-sum constraint $\sum_{t=1}^K c_t = 0$; for example, if $K = 2$, we can consider $c_1 = 1$ and $c_2 = -1$. The contrast (3.18) makes the individualized treatment efficacy comparison across the K treatment conditions as a function of the pretreatment predictors X. Note, the main effect of X cancels out in (3.18) due to the zero sum constraint in $(c_1, \ldots, c_K)^\top$. For now, let us assume a classical linear model for the interaction effect term in the conditional expectation, $\mathbb{E}(Y \mid X, T = t)$, which is sometimes called the "quality" of treatment t at observation X ([Qian and Murphy, 2011])

$$\mathbb{E}(Y \mid X, T = t) = \mu(X) + \boldsymbol{\beta}_t^\top X, \tag{3.19}$$

with distinct $\boldsymbol{\beta}_t$'s. Assuming $p > K - 1$, let us introduce the $p \times p$ "between" group dispersion matrix of the coefficients $\boldsymbol{\beta}_t$'s in (3.19), defined by

$$\boldsymbol{B} = \sum_{t=1}^{K} \pi_t (\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})^\top, \tag{3.20}$$

where π_t is probability that T takes a value $t \in \{1, \ldots, K\}$; the $p \times 1$ vector $\bar{\boldsymbol{\beta}} = \sum_{t=1}^{K} \pi_t \boldsymbol{\beta}_t$ is the weighted average of $\boldsymbol{\beta}_t$'s. Let us define the $p \times (K-1)$ orthogonal matrix $\boldsymbol{\Phi} = [\boldsymbol{\xi}_1; \ldots; \boldsymbol{\xi}_{K-1}]$, that consists of the $K-1$ (normalized) leading eigenvectors $\boldsymbol{\xi}_m$'s of the matrix $\boldsymbol{B}$, associated with the nonzero eigenvalues (there are at most $K-1$ nonzero eigenvalues). In the terminology of [Cook, 2007] and [Adragni and Cook, 2009], the span($\boldsymbol{\Phi}$) produces a sufficient reduction for representing the individualized treatment effect contrast (3.18).

Lemma 1. *Under model (3.19), $C(X) = C(\boldsymbol{\Phi}^\top X)$, i.e., the span($\boldsymbol{\Phi}$) is sufficient for representing $C(X)$.*

The proof of Lemma 1 is in the Appendix. In the following subsections, the estimation of the leading eigenvector $\boldsymbol{\xi}_1$ of $\boldsymbol{B}$ in (3.20) in a linear regression framework will be tied to the MCA.

3.5.2 Linear GEM models

The "numerator" method of [Petkova *et al.*, 2016] considers the following class of models

$$\mathbb{E}(Y \mid X, T = t) = \mu(X) + \boldsymbol{\xi}_1^\top X \gamma_t, \quad t \in \{1, \ldots, K\}, \tag{3.21}$$

where $\boldsymbol{\xi}_1$ is the leading eigenvector of $\boldsymbol{B}$ in (3.20), and the treatment t-specific slopes $\gamma_t \in \mathbb{R}$, $t \in \{1, \ldots, K\}$, attached to the single-index $\boldsymbol{\xi}_1^\top X$, describe the differential treatment response over the different treatment conditions. In [Petkova *et al.*, 2016], the composite covariate $\boldsymbol{\xi}_1^\top X \in \mathbb{R}$ was called a generated (treatment) effect modifier (GEM).

Let us reparametrize model (3.21) by setting

$$\gamma_t^* := \gamma_t - \gamma_0, \quad t \in \{1, \ldots, K\}, \tag{3.22}$$

where $\gamma_0 := \sum_{t=1}^{K} \pi_t \gamma_t$. Then, model (3.21) is rewritten as

$$\mathbb{E}(Y \mid X, T = t) = \underbrace{\mu(X) + \boldsymbol{\xi}_1^\top X \gamma_0}_{\text{“main” effect}} \quad + \quad \underbrace{\boldsymbol{\xi}_1^\top X \gamma_t^*}_{\text{“interaction” effect}}, \tag{3.23}$$

where the term $\mu(X) + \boldsymbol{\xi}_1^\top X \gamma_0$ models the main effect of X on the outcome, and the second term $\boldsymbol{\xi}_1^\top X \gamma_t^*$ models the treatment t-specific effect as a linear function of X (i.e., the T-by-X interaction), where

$$\sum_{t=1}^{K} \pi_t \gamma_t^* = 0, \tag{3.24}$$

acting as the identifiability constraint of the representation (3.23).

The constraint (3.24) suffices to give the orthogonality between the two subspaces

$$\operatorname{span}(\mu(X), X\gamma_0) \quad \perp \quad \operatorname{span}(X\gamma_T^*), \tag{3.25}$$

due to the randomization on $T \in \{1, \ldots, K\}$ independent of X, in which the associated inner product between arbitrary two variables is defined as the covariance between the two variables. In estimation, the orthogonality (3.25) allows us to asymptotically separate the interaction effect component from the main effect component in model (3.23). For estimating the coefficients $\boldsymbol{\xi}_1 \gamma_t^*$, this leads us to consider the following working model

$$\mathbb{E}(Y \mid X, T = t) = X^\top \boldsymbol{\xi}_1 \gamma_t^*, \quad t \in \{1, \ldots, K\}, \tag{3.26}$$

under the constraint (3.24), without specifying the form of the main effect.

Lemma 2. *Assuming (3.19), the population constrained least square solution of γ_t^* of model (3.26), subject to the constraint (3.24), is given by $\gamma_t^* = \boldsymbol{\xi}_1^\top(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})$, $t \in \{1, \ldots, K\}$.*

The proof of Lemma 2 is in the Appendix. Lemma 2 implies that, in (3.26), the coefficient $\boldsymbol{\xi}_1 \gamma_t^* \in \mathbb{R}^p$ attached to X for the treatment group t has a closed form constrained least squares solution $\boldsymbol{\xi}_1 \gamma_t^* = \boldsymbol{\xi}_1 \boldsymbol{\xi}_1^\top(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})$, for each $t \in \{1, \ldots, K\}$, which is simply the projection of $(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) \in \mathbb{R}^p$ onto the span of $\boldsymbol{\xi}_1$,

If $K = 2$, there is only one eigenvector $\boldsymbol{\xi}_1$ of the matrix $\boldsymbol{B}$ associated with a nonzero eigenvalue. In particular, $\boldsymbol{\xi}_1$ has a closed form, given by $\boldsymbol{\xi}_1 = (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)/\sqrt{\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2}$, where, without loss of generality, we took $\pi_1 = \pi_2 = 1/2$ for the simplicity. This leads us to the equivalence between optimizing model (3.26) under the constraint (3.24) and the MCA.

Lemma 3. *Under (3.19), if $K = 2$ with $\pi_1 = \pi_2 = 1/2$, optimizing model (3.26) using the constrained least squares criterion subject to (3.24) is equivalent to optimizing model (3.3) (without the main effect) using the least squares criterion, i.e., the MCA.*

Proof. If $K = 2$, $\boldsymbol{\xi}_1 = (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)/\sqrt{\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2}$. Therefore, by Lemma 2, model (3.26) under the constraint (3.24) can be written in terms of $\boldsymbol{\beta}_t$

$$\begin{aligned}
\mathbb{E}(Y \mid X, T = t) &= X^\top \boldsymbol{\xi}_1 \boldsymbol{\xi}_1^\top (\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}), \quad t = 1, 2 \\
&= X^\top \boldsymbol{\xi}_1 \boldsymbol{\xi}_1^\top (\boldsymbol{\beta}_t - \frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2}), \quad t = 1, 2 \\
&= \begin{cases} X^\top (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1) \frac{(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top}{\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2} \frac{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}{2} = -\frac{1}{2} X^\top (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1), & t = 1, \\ X^\top (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1) \frac{(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top}{\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2} \frac{(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)}{2} = +\frac{1}{2} X^\top (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1), & t = 2. \end{cases}
\end{aligned} \tag{3.27}$$

Setting $\boldsymbol{\alpha} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ leads to the MCA working model, $\mathbb{E}(Y \mid X, T = t) = X^\top \boldsymbol{\alpha} \frac{1}{2}(-1)^t$, $t = 1, 2$. □

From Lemma 3, MCA can be viewed as a special approach that estimates the sufficient single-dimension reduction vector $\boldsymbol{\xi}_1$ of model (3.26), if we restrict our attention to the $K = 2$ case, that can be estimated by optimizing model (3.26) under an appropriate centering of γ_t^*, i.e., subject to the constraint (3.24).

3.5.3 $K \geq 3$ case

The method of "undetermined" coefficients with γ_t^*, $t \in \{1, \dots, K\}$ in (3.26), in which we impose the orthogonality between the main effect and the interaction effect by the constraint (3.24), sheds light on a simple approach to modeling interactions when $K \geq 3$. The projection vector $\boldsymbol{\xi}_1$ in (3.26), the leading eigenvector of $\boldsymbol{B}$, can be computed from sample estimates of the $\boldsymbol{\beta}_t$'s, which can be estimated by linear regression (via the Lasso, for example) based on observations from the tth treatment group. Then the γ_t^*'s in (3.26) can be computed using Lemma 2.

3.5.4 Extension to a semiparametric model

The equivalence between the MCA working model and model (3.26) under the constraint (3.24) gives an insight on generalizing the MCA to a semiparametric regression framework.

A semiparametric counterpart of the linear GEM model (3.21) is the SIMML (3.2), in which a set of nonparametrically-defined link functions g_t, $t \in \{1, \ldots, K\}$ replaces the set of treatment-specific scalar coefficients γ_t, $t \in \{1, \ldots, K\}$. We will present the followings for completeness, although materials overlap with those from Section 3.3. Given an arbitrary $\boldsymbol{\alpha} \in \mathbb{R}^p$ and a set of arbitrary (smooth) functions g_t, $t \in \{1, \ldots, K\}$, let us introduce, analogous to the reparametrization (3.22)

$$g_t^*(\boldsymbol{\alpha}^\top X) := g_t(\boldsymbol{\alpha}^\top X) - g_0(\boldsymbol{\alpha}^\top X), \tag{3.28}$$

where $g_0(\boldsymbol{\alpha}^\top X) := \sum_{t=1}^K \pi_t g_t(\boldsymbol{\alpha}^\top X)$. Then model (3.2) can be rewritten by

$$\mathbb{E}\left(Y \mid X, T = t\right) = \underbrace{\mu(X) + g_0\big(\boldsymbol{\alpha}^\top X\big)}_{\text{“main” effect}} \quad + \quad \underbrace{g_t^*\big(\boldsymbol{\alpha}^\top X\big)}_{\text{“interaction” effect}} \quad , \tag{3.29}$$

for $t \in \{1, \ldots, K\}$. By definition (3.28), we have

$$\sum_{t=1}^K \pi_t g_t^*(\boldsymbol{\alpha}^\top X) = 0, \tag{3.30}$$

acting as the identifiability condition of the representation (3.29). Due to the randomization of $T \in \{1, \ldots, K\}$ independent of X, the condition (3.30) implies the orthogonality between the two subspaces

$$\text{span}\left(\mu(X), g_0\big(\boldsymbol{\alpha}^\top X\big)\right) \quad \perp \quad \text{span}\left(g_T^*(\boldsymbol{\alpha}^\top X)\right). \tag{3.31}$$

Under the constraint (3.30), the estimation of $g_T^*(\boldsymbol{\alpha}^\top X)$ can be asymptotically separated from estimating the main effect $\mu(X) + g_0\big(\boldsymbol{\alpha}^\top X\big)$. Thus, the empirical version of the expected constrained squared error

$$\mathbb{E}\left(Y - g_T^*(\boldsymbol{\alpha}^\top X)\right)^2, \tag{3.32}$$

subject to (3.30), i.e., $\mathbb{E}_T(g_T^*) = 0$, can be minimized. In Theorem 3, we have the population constrained least squares solution of g_T^* in terms of the conditional expectations (3.9). We note that Lemma 2 is a special case of Theorem 3, in which the link g_T^* is restricted to be a linear function. Using an orthogonalized model (3.29), the optimization (3.13) for $\boldsymbol{\alpha}$ can be performed.

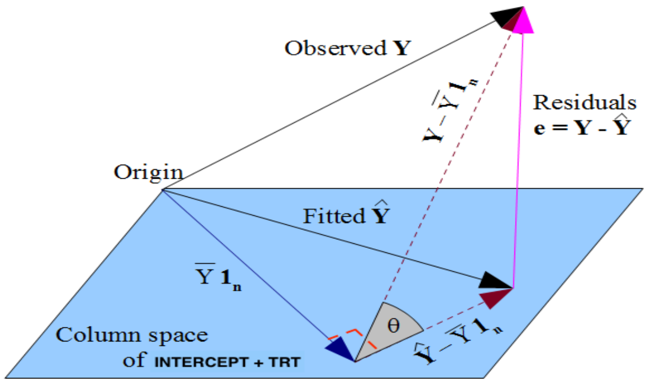

Figure 3.1: In the simple linear regression $\mathbb{E}(Y \mid T)$ of the outcomes on the treatments, the fitted $\hat{\boldsymbol{Y}}$ for the model $\mathbb{E}(Y \mid T)$ is the orthogonal projection of the observed $\boldsymbol{Y}$ onto the plane of the column space spanned by the intercept and the treatments. The fitted vector for the intercept-only model $\mathbb{E}(Y \mid 1)$ is $\bar{Y}\mathbf{1}_n$. In the picture, the magnitude of the "effect" of the intercept (i.e., averaging), which gets modified by the treatment (i.e., treatment-specific averaging), can be quantified by the squared length of $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$.

3.5.5 Some geometric intuition

In this section, we will provide some geometric intuition of optimizing a SIMML by the criterion (3.32). For analogy, we take the simple regression of Y on the treatment indicator T with no covariate X. In this case, g_T^* in (3.9) is simply

$$g_T^* = \mathbb{E}(Y \mid 1, T) - \mathbb{E}(Y \mid 1). \tag{3.33}$$

Suppose we are given a dataset $(Y_i, T_i)_{i=1}^n$. In Figure 3.1, the projections $\mathbb{E}(Y \mid 1, T)$ and $\mathbb{E}(Y \mid 1)$ can be represented by $\hat{\boldsymbol{Y}}$ and $\bar{Y}\mathbf{1}_n$, respectively. Here, $\hat{\boldsymbol{Y}}$ is the $n \times 1$ fitted vector of the model $\mathbb{E}(Y \mid 1, T)$ (the vector of the treatment group-specific averages), $\boldsymbol{Y}$ is the $n \times 1$ observed vector of the responses $(Y_1, \ldots, Y_n)^\top$, $\mathbf{1}_n$ is the $n \times 1$ vector of ones, and $\bar{Y} = \sum_{i=1}^n Y_i/n$, denotes the grand average. Then, in Figure 3.1, g_T^* in (3.33) is represented by the side $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$.

Given each g_T^* that satisfies $\mathbb{E}_T(g_T^*) = 0$ (3.30), minimizing the criterion (3.32)

$$\mathbb{E}\left(Y - g_T^*(\boldsymbol{\alpha}^\top X)\right)^2 = \mathbb{E}\left(Y^2\right) - \mathbb{E}\left(g_T^*(\boldsymbol{\alpha}^\top X)\right)^2$$

over $\boldsymbol{\alpha}$ corresponds to maximizing $\mathbb{E}\left(g_T^*(\boldsymbol{\alpha}^\top X)\right)^2$ over $\boldsymbol{\alpha}$. In Figure 3.1, this maximization is analogous to maximizing the squared length of the adjacent $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$, i.e., $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2$, of the right-angled triangle defined by the hypotenuse $\boldsymbol{Y} - \bar{Y}\mathbf{1}_n$. Note, the adjacent $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$ is always orthogonal to the "main effect" vector $\bar{Y}\mathbf{1}_n$.

In Figure 3.1, $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2$, corresponding to a sample version of $\mathbb{E}\,(g_T^*)^2$ where g_T^* is defined in (3.33), gives the magnitude of how much the "effect" of 1 (i.e., the simple averaging) gets modified by T (i.e., the T-specific averaging), hence it quantifies the intensity of the interaction effect between T and 1. Analogously, in our problem with covariates X, the quantity $\mathbb{E}\left(g_T^*(\boldsymbol{\alpha}^\top X)\right)^2$ gives the magnitude of the interaction effect between T and the single-index $\boldsymbol{\alpha}^\top X$, which is to be maximized over $(\boldsymbol{\alpha}, g_t^*, t \in \{1, \ldots, K\})$ subject to (3.30). This can be achieved by optimizing (3.32) subject to (3.30).

This maximization corresponds to optimizing the space of functions defined on $(\boldsymbol{\alpha}^\top X, T)$ that is represented by the blue plane in Figure 3.1, by choosing an optimal $\boldsymbol{\alpha}$ that gives the minimal angle θ formed by the hypotenuse $\boldsymbol{Y} - \bar{Y}\mathbf{1}_n$ and the adjacent $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$ (i.e., the two dashed lines in Figure 3.1); or equivalently, maximizing the cosine of the angle θ, i.e., maximizing $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2$.

The two dashed lines in Figure 3.1 represent $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$ and $\boldsymbol{Y} - \bar{Y}\mathbf{1}_n$, corresponding to the fitted $(\hat{\boldsymbol{Y}})$ and the observed $(\boldsymbol{Y})$ vector, respectively, centered by the intercept vector $(\bar{Y}\mathbf{1}_n)$. Without the centering by the intercept, there is no orthogonal sum of squares decomposition

$$\|\boldsymbol{Y} - \bar{Y}\mathbf{1}_n\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2, \tag{3.34}$$

in which the second component $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2$ quantifies the T-by-1 interaction effect. Analogously, in orthogonalization (3.29), the "shifting" component $g_0(\boldsymbol{\alpha}^\top X)$ $\left(= \sum_{t=1}^K \pi_t g_t(\boldsymbol{\alpha}^\top X)\right)$ plays the role of an "intercept". Centered by $g_0(\boldsymbol{\alpha}^\top X)$ gives the following orthogonal decomposition, for any set of smooth functions g_t, $t \in \{1, \ldots, K\}$

$$\mathbb{E}\Big(Y - \mu(X) - g_0(\boldsymbol{\alpha}^\top X)\Big)^2 = \mathbb{E}\Big(Y - \mu(X) - g_T(\boldsymbol{\alpha}^\top X)\Big)^2 + \mathbb{E}\Big(g_T(\boldsymbol{\alpha}^\top X) - g_0(\boldsymbol{\alpha}^\top X)\Big)^2, \tag{3.35}$$

in which the second component $\mathbb{E}\left(g_T^*(\boldsymbol{\alpha}^\top X)\right)^2 \left(= \mathbb{E}\left(g_T(\boldsymbol{\alpha}^\top X) - g_0(\boldsymbol{\alpha}^\top X)\right)^2\right)$ quantifies the T-by-$\boldsymbol{\alpha}^\top X$ interaction effect, which is to be maximized over g_T and $\boldsymbol{\alpha}$, subject to some identifiability condition.

Finally, if we consider a single-index model (SIM)

$$\mathbb{E}(Y \mid X, T=t) = \mathbb{E}(Y \mid X) = g_0(\boldsymbol{\alpha}^\top X) \tag{3.36}$$

that is defined regardless of the treatment indicator T, then the SIM (3.36) is nested within the SIMML (3.1), with restriction $g_1 = \cdots = g_K (= g_0)$. Solving (3.7) for $\boldsymbol{\alpha}$ corresponds to choosing $\boldsymbol{\alpha}$ that maximizes the deviance (e.g., [Nelder and Wedderburn, 1972]) between the unconstrained SIMML (3.1) and the SIM (3.36), in which the both two models are separately optimized under the least squares criterion, for each fixed $\boldsymbol{\alpha}$. It follows that maximizing $\mathbb{E}\left(g_T(\boldsymbol{\alpha}^\top X) - g_0(\boldsymbol{\alpha}^\top X)\right)^2$ in (3.35) over $\boldsymbol{\alpha}$ leads to maximizing over $\boldsymbol{\alpha}$ the L^2 distance (the deviance, if we consider a Gaussian response), between the unconstrained SIMML (3.1) that permits T-by-X interactions versus the restricted SIM (3.36) that does not permit any interactions.

3.6 Simulation examples

In this section, we performed numerical studies to illustrate the performance of the proposed approach for estimating the SIMML and constructing associated ITRs.

3.6.1 Estimation criterion illustration

For the purpose of illustration, we first considered a simple case of $p = 2$ and $K = 2$. We generated $X_i = (x_{i1}, x_{i2})^\top$ from the independent bivariate Gaussian with unit variances and zero correlation. For the simulation set "A", we considered a highly nonlinear (a cosine function) contrast. We generated the outcomes $Y_i = \mu(X_i) + g_{T_i}(\boldsymbol{\alpha}^\top X_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 0.2^2)$, $i = 1, \ldots, n$, where we set

$$\begin{aligned} \mu(X_i) &= \delta \cos(\boldsymbol{\mu}^\top X_i) \\ g_{T_i}(\boldsymbol{\alpha}^\top X_i) &= T_i \left(\cos(\boldsymbol{\alpha}^\top X_i) - 0.5\right), \end{aligned} \tag{3.37}$$

where T_i takes a value in $\{-1, 1\}$ with equal probability 0.5, independently generated of X_i. For the simulation set "B", we considered a moderately nonlinear contrast. In (3.37), we took $g_{T_i}(\boldsymbol{\alpha}^\top X_i) = T_i \left(\sin(\boldsymbol{\alpha}^\top X_i) - \boldsymbol{\alpha}^\top X_i\right)$. Since the function $\sin(u)$ can be well approximated by u near 0, this setting gave an almost linear contrast. In the both settings,

$n = 200$. In generating the outcomes, there were two components: $\boldsymbol{\mu}^\top X$ and $\boldsymbol{\alpha}^\top X$. $\boldsymbol{\alpha}^\top X$ was a single-index associated with the T-by-X interactions, whereas $\boldsymbol{\mu}^\top X$ was a single-index associated with the main effects, hence, $\boldsymbol{\mu}^\top X$ was a "nuisance" component. We set $\boldsymbol{\alpha} := (1,1)^\top/\sqrt{2}$ and $\boldsymbol{\mu} := (-1,1)^\top/\sqrt{2}$. In (3.37), $\delta \in \{1,3,5\}$ was the main effect intensity parameter that controlled the contribution of $\mu(X)$ on the variance of Y. With the intensity parameter $\delta = 1$, $\delta = 3$, and $\delta = 5$, the contribution from the main effect $\mu(X)$ to the variance of Y was about 1, 8, and 20 times larger than that from $g_T(\boldsymbol{\alpha}^\top X)$, respectively, under the setting (3.37).

In this simulation example, the "constrained SIMML" refers to the SIMML that uses the criterion (3.7), and the "naïve SIMML" refers to the SIMML that uses (3.7) but without the "orthogonality" constraint $\mathbb{E}_T\left(g_T\right) = 0$. The naïve SIMML criterion corresponds to the profile likelihood criterion (2.7). For comparison, we included the least squares criterion of the MCA model (3.3). A main effect augmentation with the linear regression was implemented and regressed out before evaluating the least squares criterion.

If Cartesian coordinates are transformed into polar coordinates, then any vector $(c_1, c_2)^\top$ on the unit (i.e., radius 1) half circle can be represented by a single parameter $\theta \in [0,\pi]$, where θ is the angle in radians in polar coordinates. For the purposes of visualization we will express the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ in polar coordinates, so $\boldsymbol{\alpha}$ gets mapped to $\theta_1 = \pi/4$ and $\boldsymbol{\mu}$ gets mapped to $\theta_2 = 3\pi/4$.

We simulated 200 datasets under the above described setups, and averaged the values of the empirical version of the criterion function of the constrained SIMML, the naïve SIMML, and the MCA, respectively, for each value of $\theta \in [0,\pi]$ (evaluated on a dense grid). Then each of the averaged criterion functions were scaled to have height 1. In Figure 3.2, the resulting averaged criterion functions are displayed for the simulation set A.

In Figure 3.2, for all three cases of $\delta = 1$, $\delta = 3$, and $\delta = 5$, the constrained SIMML criterion had a "correct" global minimum at $\theta_1 = \pi/4$, implying that the minimization of the criterion function would lead to correctly identifying the T-by-X interaction effect component $\boldsymbol{\alpha}$. The naïve SIMML (the green dotted lines) had a correct minimum at $\theta_1 = \pi/4$ for the case $\delta = 1$ (i.e., when the main effect is relatively small), however, as the main effect intensity parameter increased from $\delta = 1$ to $\delta = 3$ and to $\delta = 5$, the criterion

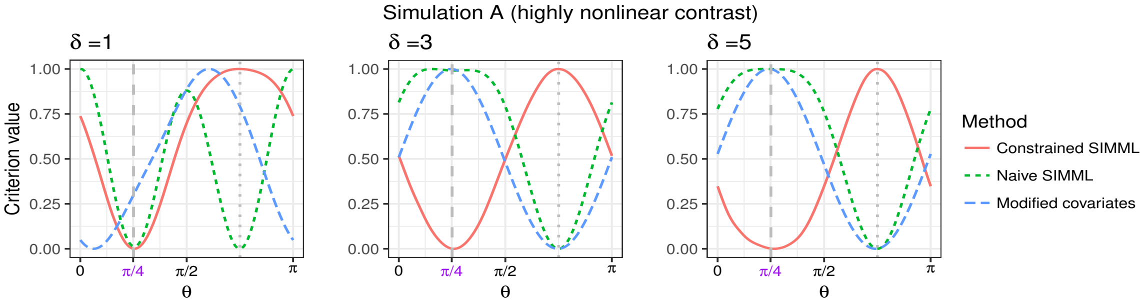

Figure 3.2: The empirical mean squared error criterion of the constrained SIMML, the naïve SIMML, and the MCA, respectively, averaged over 200 simulated datasets, for simulation set A. The vector $\boldsymbol{\alpha}$ corresponds to the angle $\theta_1 = \pi/4$, and the "nuisance" vector $\boldsymbol{\mu}$ corresponds to the angle $\theta_2 = 3\pi/4$. The grey dashed vertical line indicates the angle θ_1, corresponding to $\boldsymbol{\alpha}$, and the grey dotted vertical line indicates θ_2, corresponding to $\boldsymbol{\mu}$.

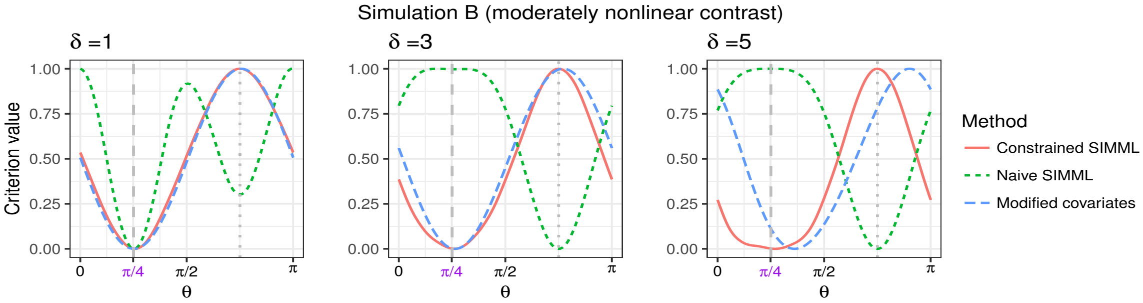

Figure 3.3: The empirical mean squared error criterion of the constrained SIMML, the naïve SIMML, and the MCA, respectively, averaged over 200 simulated datasets, for simulation set B. The vector $\boldsymbol{\alpha}$ corresponds to the angle $\theta_1 = \pi/4$, and the "nuisance" vector $\boldsymbol{\mu}$ corresponds to the angle $\theta_2 = 3\pi/4$. The grey dashed vertical line indicates the angle θ_1, corresponding to $\boldsymbol{\alpha}$, and the grey vertical dotted line indicates θ_2, corresponding to $\boldsymbol{\mu}$.

function took its global minimum at the nuisance component $\theta_2 = 3\pi/4$, implying that the minimization of the naïve SIMML criterion would lead to an estimate of $\boldsymbol{\alpha}$ of SIMML (2.2) that looks more like the "nuisance" vector $\boldsymbol{\mu}$, which would not be informative in developing ITRs. Under the highly nonlinear contrast, we note that the MCA criterion function did not provide a good prescriptive information in developing ITRs, as it behaved similarly to the naïve SIMML criterion.

In Figure 3.3, the results of the simulation set B are illustrated, where we investigated the estimation criterion comparison under a moderate nonlinear contrast. As in the simulation set A, the constrained SIMML criterion took a "correct" global minimum at $\theta_1 = \pi/4$ for all three cases of $\delta = 1$, $\delta = 3$, and $\delta = 5$. On the other hand, the naïve SIMML had its global minimum at the nuisance component $\theta_2 = 3\pi/4$, when $\delta = 3$ or $\delta = 5$ (i.e., when the main effect dominated the interaction effect). Since the contrast function g_T was almost linear in this simulation setting, the squared error criterion of the MCA behaved similarly to the constrained SIMML criterion function, and took its global minimum near $\theta_1 = \pi/4$, in all cases.

The results indicate that when a large main effect is present, it is essential to impose the "orthogonality" constraint when fitting the SIMML in order to capture the interaction-related variabilities.

3.6.2 ITR performance for $K = 2$ case

In this subsection and the next, we compared the performance of ITRs obtained from several methods. We note that regularization to deal with potential high dimensionality of X was not considered in this chapter. Methods with regularization will be considered in Chapter 4. We restricted our attention to ITRs of the form, $\mathcal{D}(X) = \underset{t\in\{1,\ldots,K\}}{\arg\max}\ \mathbb{E}\left(Y \mid X, T = t\right)$, where $\mathbb{E}\left(Y \mid X, T\right)$ was obtained by the following approaches:

SIMML Estimate the SIMML (3.6) subject to (3.5) (i.e., the constrained SIMML). An initial estimate for $\boldsymbol{\alpha}$ was obtained by the MCA (3.3) for the case $K = 2$, and by the linear GEM (3.26) for the case $K = 3$. The link function was approximated by cubic B-splines, with smoothing parameters selected by minimizing the generalized cross validation (GCV).

MCA Estimate the modified covariates model with efficiency augmentation (3.3) ([Tian *et al.*, 2014]). Model (3.3) was fitted by the ordinary least squares (OLS) for estimating interactions. Augmentation of main effect, $\mu(X) = \boldsymbol{\mu}^\top X$, was utilized, fitted by OLS. MCA is applicable only for $K = 2$, therefore considered only in the case of $K = 2$.

L.GEM Estimate the linear GEM model (3.26) as described in Section 3.5.3. The linear GEM is equivalent to the MCA in the population level when $K = 2$, therefore reported only for $K > 2$.

K.AM For each of the K groups separately, estimate an additive model (AM) ([Hastie and Tibshirani, 1999]), where the nonparametrically-defined component functions were approximated by cubic B-splines, with smoothing parameters selected by minimizing GCV.

K.LR For each of the K groups separately, estimate a linear regression (LR) model by OLS.

In this section and the next, the performance measure for an estimated ITR $\mathcal{D}$ was the proportion of correct decisions (PCD) of $\mathcal{D}$. Since we know the true data generating model for each simulation setting, we can calculate the PCD for each of the estimation methods, given each scenario. This measure was calculated from an independent testing set of size $n = 10000$.

In this section, we considered a relatively low dimensional X, i.e., $p \in \{5, 10\}$, since we did not consider regularizations, with a varying sample size $n \in \{200, 400\}$. As in the settings of Section 3.6.1, we generated the treatments T_i and the outcomes $Y_i = \mu(X_i) + g_{T_i}(\boldsymbol{\alpha}^\top X_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 0.2^2)$. We considered two cases for the treatment-specific link function $g_T(\nu)$: (1) a nonlinear contrast function $g_T(\nu) = (\cos(\nu) - 0.5)T$ that gave nonlinear T-by-X interaction effect, and (2) a linear contrast function $g_T(\nu) = 0.5\nu T$ that gave a linear T-by-X interaction effect, respectively.

For the main effect, we took $\mu(X) = \delta \cos(\boldsymbol{\mu}^\top X)$. The scaling parameter δ, taken at either $\delta = 1$ or $\delta = 2$, controlled the intensity of the main effect, representing a relatively small main effect case (about the same variance as the interaction effect) and a relatively large main effect case (about 3 times larger variance than that of the interaction effect),

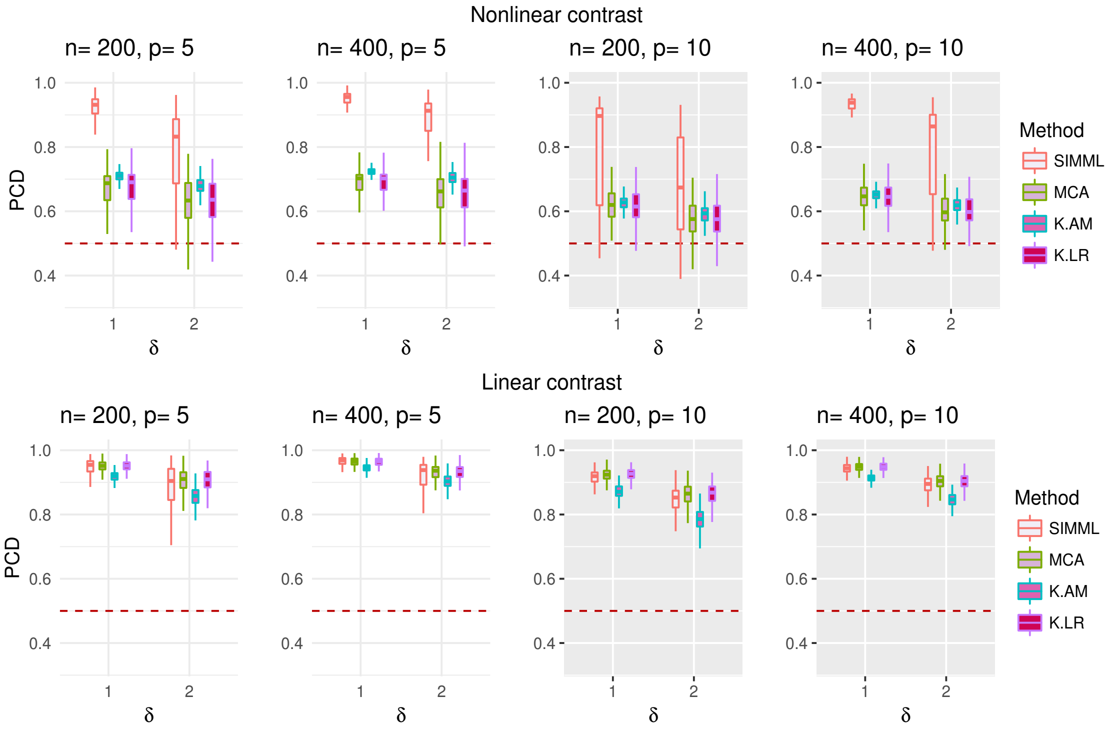

Figure 3.4: Top panels: boxplots of the PCDs of the ITRs estimated from the four methods (SIMML, MCA, K.AM, and K.LR) for the nonlinear contrast case. Lower panels: boxplots of the PCDs of the ITRs for the linear contrast case. For each case, $n \in \{200, 400\}$ and $p \in \{5, 10\}$ were considered.

respectively. We generated $X_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. For $p = 5$ and $p = 10$, we set $\boldsymbol{\alpha} = (1, 2, 3, 4, 5)^\top$ and $\boldsymbol{\alpha} = (-5, -4, -3, -2, -1, 1, 2, 3, 4, 5)^\top$, each standardized to have norm one. We set $\boldsymbol{\mu}$ to be proportional to a vector of 1's, standardized to have norm one. 200 training datasets were simulated for each scenario.

The results from the simulations are presented in Figure 3.4. For the nonlinear contrast, the performance of the SIMML was outstanding, particularly when $n = 400$ and $p = 5$. For the case with $n = 200$ and $p = 10$, the SIMML exhibited relatively large variabilities, indicating that a regularization might be necessary in estimation. Nevertheless, SIMML generally outperformed all other alternatives in all cases. Interestingly, the method significantly outperformed the K separate additive regressions (K.AM) which was also equipped with a set of flexible nonparametrically defined functions to model the nonlinear associations. This was because SIMML is more parsimonious than the K separate additive models. In addition, the SIMML estimates the interaction effects only, while the K separate additive regressions estimate both the main and the interaction effects, hence they tends to lose their efficiency in estimating the interaction effects given a limited sample size. Moreover, when the main effect is not an additive structure (as in this example), the K separate additive models suffer inconsistency in estimating the interaction effects, since the method is not made robust to the main effect model misspecification. For the nonlinear contrast, both the MCA and the K separate linear regressions (K.LR) were significantly outperformed by the SIMML that utilized the unspecified link functions to estimate the nonlinear interactions, indicating a clear benefit of fitting SIMML for estimating nonlinear interactions.

The case for the linear contrast is illustrated in the bottom panels of Figure 3.4. The proposed SIMML gave a similar performance level as the linear model based approaches (i.e., the MCA and K.LR) when $p = 5$. When $p = 10$, the MCA slightly outperformed the SIMML. This is expected, since the MCA is correctly specified in the linear contrast case, and is a special case of the SIMML but under a more parsimonious model for estimating the interactions. However, the K separate additive models were clearly outperformed by the MCA and K.LR in all cases, unlike the SIMML.

3.6.3 ITR performance for $K = 3$ case

In this subsection, we investigated the performance of the SIMML when the number of treatment groups $K = 3$. The treatments T_i that take values in $\{1, 2, 3\}$ with equal probability were generated, independently of X_i. The outcomes Y_i were generated as in the settings of Section 3.6.2, except that the interaction effect component, $g_T(\nu)$, was set at $\begin{cases} g_1(\nu) & = \nu^1(1-\nu)^4/B(2,5) \\ g_2(\nu) & = \nu^1(1-\nu)^1/B(2,4) \\ g_3(\nu) & = \nu^4(1-\nu)^0/B(5,1) \end{cases}$, in which $B(a, b) = (\Gamma(a)\Gamma(b))/\Gamma(a + b)$ is a Beta function, and $\nu = F(\boldsymbol{\alpha}^\top X)$, where F was the cumulative distribution function (CDF) of a re-scaled/centered $B((p^* + 1)/2, (p^* + 1)/2)$

$$F(u) = \int_{-1}^{u/R} \frac{\Gamma(p^* + 1)}{\Gamma\{(p^* + 1)/2\}^2 2^{p^*}} (1 - t^2)^{(p^*-1)/2} dt, \quad u \in [-R, R], \tag{3.38}$$

with p^* denoting the number of nonzero coefficients in $\boldsymbol{\alpha}$. In (3.38), R was the maximum of the absolute values of $\{(\boldsymbol{\alpha}^\top X_i)_{i=1}^n\}$. We used the transformation (3.38) because $\{F(\boldsymbol{\alpha}^\top X_i)_{i=1}^n\}$ is quasi-uniformly distributed on the interval $[0, 1]$ ([Wang and Yang, 2009]).

For the main effect, we set $\mu(X) = \delta \cos(\boldsymbol{\mu}^\top X)$ (as in the settings of Section 3.6.2). The intensity parameter $\delta \in \{1, 3\}$ controlled the intensity of the main effect, representing the relatively small (about the same variance as the interaction effect) and the relatively large main effect (about three times larger variance than the interaction effect) cases, respectively.

We generated $X_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ were set at the same as in the settings of Section 3.6.2. In the upper panel of Figure 3.5, we illustrated the functions $g_T(\nu)$, $\nu \in [0, 1]$, for each $T \in \{1, 2, 3\}$, with $n = 200$ data points generated from the model under $\delta = 0$ and $p = 5$. In the bottom panels of Figure 3.5, we displayed the boxplots of the PCDs of the ITRs estimated from the 4 different methods (SIMML, L.GEM, K.AM, and K.LR) (described in the beginning of Section 3.6.2), for each combination of $n \in \{200, 400\}$, $p \in \{5, 10\}$ and $\delta \in \{1, 3\}$.

The boxplots indicated that the proposed SIMML outperformed all other methods, in all cases. We note that the K separate additive regressions (K.AM) and the K separate linear regressions (K.LR) performed very badly when $n = 200$ and $p = 10$ particularly in the presence of a large main effect (i.e., when $\delta = 3$). These K separate regression methods

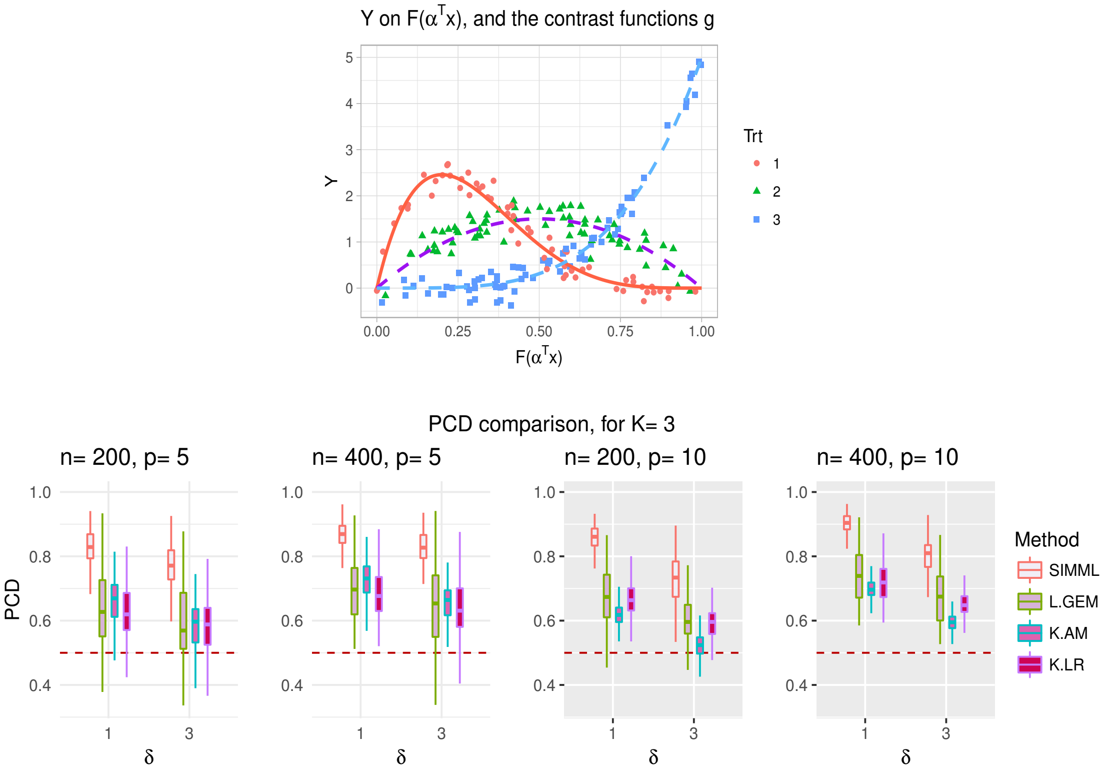

Figure 3.5: Upper panel: illustration of $g_t(\nu)$, $t \in \{1, 2, 3\}$, with simulated data points under $\delta = 0$. Lower panels: boxplots of the PCDs of the ITRs estimated from the four different methods (SIMML, L.GEM, L.AM, and K.LR) applied to 200 simulated datasets, for each combination of $n \in \{200, 400\}$, $p \in \{5, 10\}$ and $\delta \in \{1, 3\}$.

lack parsimony and interpretability, especially when $K > 2$, in comparison to the SIMML and the linear GEM method that estimate a single projection $\boldsymbol{\alpha}^\top X$, as opposed to the K separate projections. Moreover, the K separate regressions are generally not robust to the main effect misspecification. For this reason, when the main effect intensity increased, the K separate approaches hardly captured the T-by-X interaction-related variabilities useful for developing ITRs, resulting in a low-performing ITR in comparison to the SIMML.

3.7 Discussion

In this chapter, we presented a semiparametric regression model that uses a set of treatment-specific unspecified link functions defined on a single projection $\boldsymbol{\alpha}^\top X$, specifically intended to model the interaction effect between the treatment and a number of pretreatment predictors. The GEM models ([Petkova *et al.*, 2016]) and the MCA ([Tian *et al.*, 2014]) provide useful approaches to making ITRs by estimating an optimal linear combination of pretreatment predictors, under a linear model framework. In this chapter we proposed a semiparametric framework to model the interaction effect by estimating nonparametrically-defined treatment-specific link functions connected to a linear predictor $\boldsymbol{\alpha}^\top X$, without the need to specify the form of the main effect. This method is a special class of the projection pursuit regression ([Friedman and Stuetzle, 1981]), in which the "pursuit" of a linear projection $\boldsymbol{\alpha}^\top X$ is driven by optimizing the intensity of the interaction effect. The approach can be viewed as a general strategy of estimating possibly nonlinear interactions between a categorical variable and a possibly high dimensional vector-valued predictor on the outcomes, if an appropriate regularization to deal with the potential high dimensionality is employed. Such regularization method is described in Chapter 4.

Chapter 4

A Sparse constrained single-index model with multiple-links

4.1 Introduction

The search of biosignatures for treatment response that predict differential response to different treatment has been an active research topic. In a regression model for treatment outcome in a RCT, a treatment effect modifier is a covariate that has an interaction with the treatment indicator, implying that the treatment efficacies vary across values of such a covariate. Identification of such treatment effect modifiers that act as biosignatures of differential treatment response is crucial, as we move toward developing ITRs based on measurements made when a patient presents for treatment.

However, one challenging aspect of constructing ITRs from a RCT dataset is that there is relatively little clinical guidance on which baseline predictors might indicate better response to one treatment versus another, i.e., which predictors are "prescriptive" for assigning treatments. Identification of effective treatment effect modifiers from those individual baseline characteristics would help make better patient-specific treatment decisions. The primary aim of this chapter is to develop a methodology to select useful pretreatment measurements (i.e., treatment effect modifiers) from a potentially large number of baseline features, that will help predict patient response to treatment, and develop ITRs that will assign the treatment that is best for each patient. We will base our treatment effect modifier selection method on the constrained SIMML model (3.6), developed in Chapter 3. Estimating the SIMML (3.6) in a high-dimensional covariate space is likely to cause problems of overfitting. In this chapter we will employ an appropriate L^1 regularization that can avoid overfitting of the model, as well as can achieve simultaneous treatment effect modifier selection by obtaining a sparse estimate of $\boldsymbol{\alpha}$ in (3.6).

4.2 Treatment effect modifier selection

Although the SIMML formulation is a flexible and useful approach to estimating the treatment-by-X interactions, one shortcoming is that the linear projection $\boldsymbol{\alpha}^\top X$ is defined in terms of all the predictors in the model, i.e., model (3.6) forces all the predictors play a role in building an interaction term. However, there are often many pretreatment variables that may or may not be useful in constructing an optimal ITR. In such case, it would be advantageous to have a method that selects important predictors. Therefore, there is a need to develop an algorithm for treatment effect modifier selection.

It could be initially thought that such simultaneous treatment effect modifier selection could be easily performed, for example, using a L^1 penalized least squares for estimating $\boldsymbol{\alpha}$, as in the standard linear regression context (e.g., [Qian and Murphy, 2011]). However, only limited research has been conducted to extend the single index models to a high dimensional situation. This is mainly due to the nonconvexity of the sum of squared error function with respect to the single-index coefficient $\boldsymbol{\alpha}$, which complicates the estimation of a single index model as well as establishing theoretical properties of the estimators ([Radchanko, 2015]). [Wang and Yin, 2008] proposed an approach that introduces L^1 regularization into the minimum average variance estimation (MAVE) method of [Xia *et al.*, 2002], however, it is limited to a relatively low dimensional setting. For other examples, [Peng and Huang, 2011] estimate the single-index model by minimizing a penalized least squares criterion, performing simultaneous predictor selection, [Zhu *et al.*, 2011] use the adaptive lasso with kernel smoothing, and [Wang and Wang, 2015] use the smoothly clipped absolute deviation (SCAD) ([Fan and Li, 2001]) penalization under diverging number of parameters. However, [Radchanko, 2015] noted that such penalization approaches may be problematic in high-dimensional situations due to the non-convexity of the sum of squared error function.

4.2.1 A constrained L^1 regularization

Even if we fix $\hat{g}_T$ at the true link function g_T, the residual sum of squares function in (3.13) is not a convex function of the coefficient $\boldsymbol{\alpha}$. Such non-convexity is particularly problematic for performing a penalized estimation to regulate $\boldsymbol{\alpha}$. If we add a L^1 penalty, say $p_\lambda(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_1$, to the residual sum of squares $\sum_{i=1}^n \big(Y_i - \hat{g}_{T_i}(\boldsymbol{\alpha}^\top X_i)\big)^2$, then, due to the non-convexity of the residual sum of squares, the solution path of $\boldsymbol{\alpha}$ is generally not a continuous function of λ. For this reason, selecting an appropriate tuning parameter $\lambda > 0$ becomes extremely difficult.

However, [Radchanko, 2015] proposed a constrained L^1 regularization approach that handles such tuning parameter selection problem, which we will employ in estimating the SIMML (3.6). At the initialization step of the estimation, let us first compute component index, $\hat{a}$

$$\hat{a} \quad = \quad \underset{l\in\{1,\ldots,p\}}{\arg\min} \quad \sum_{i=1}^{n} \big(Y_i - \hat{g}_{T_i}(\boldsymbol{e}_l^\top X_i)\big)^2, \tag{4.1}$$

in which $\boldsymbol{e}_l = [0,\ldots,1,\ldots,0]^\top \in \mathbb{R}^p$, where the lth component equals 1, and all other components equal 0 for $l = 1,\ldots,p$ (i.e., the canonical basis of $\mathbb{R}^p$). In (4.1), $\hat{g}_t$, $t \in \{1,\ldots,K\}$, are (nonparametric) estimates of g_T in (3.9), with $\boldsymbol{\alpha} = \boldsymbol{e}_l$, $l = 1,\ldots,p$. Then $\hat{a}$ is component index that corresponds to the estimated best "signal" treatment effect modifier among the p predictors. (4.1) is a convex optimization problem. It is suggested in [Radchanko, 2015] to fix the $\hat{a}$th component of $\boldsymbol{\alpha}$ at 1 throughout the estimation procedure, i.e., fixing $\alpha_{\hat{a}} = 1$, to be used as the model identifiability constraint, instead of using the second line of (3.13), i.e., $\|\boldsymbol{\alpha}\| = 1$ with $\alpha_1 > 0$ for the identifiability constraint. Provided that the best individual treatment effect modifier (4.1) exists, these two identifiability constraints are equivalent.

It is suggested in [Radchanko, 2015] that the L^1 norm of $\boldsymbol{\alpha}$, $\|\boldsymbol{\alpha}\|_1$, say, $\lambda \geq 1$, be used as the sparsity tuning parameter for $\boldsymbol{\alpha}$. From $\lambda = 1$ at the beginning (the sparsest case) of the algorithm to some $\lambda = \lambda_{\max} > 1$, the L^1 norm $\|\boldsymbol{\alpha}\|_1 = \lambda$ can be increased gradually on a dense grid of $[1, \lambda_{\max}]$. For each given $\lambda \in [1, \lambda_{\max}]$, it is suggested to solve a constrained minimization problem, subject to the L^1 equality constraint, $\|\boldsymbol{\alpha}\|_1 = \lambda$, i.e.,

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \sum_{i=1}^{n} \big(Y_i - \hat{g}_{T_i}(\boldsymbol{\alpha}^\top X_i)\big)^2 \quad \big(= \|\boldsymbol{Y} - \hat{\boldsymbol{g}}_{\boldsymbol{\alpha}}\|^2\big), \\ &\text{subject to} \quad \|\boldsymbol{\alpha}\|_1 = \lambda, \end{aligned} \tag{4.2}$$

in which $\hat{g}_t$, $t \in \{1, \ldots, K\}$, are (nonparametric) estimates of g_T in (3.9) for each $\boldsymbol{\alpha}$. It is verified in [Radchanko, 2015] that solving the (L^1 equality constraint) optimization problem (4.2) for each and every $\lambda \in [1, \lambda_{\max}]$ constructs a solution path of $\boldsymbol{\alpha}$ that is continuous as a function of $\lambda \in [1, \lambda_{\max}]$. Therefore, the sparsity parameter λ can be reliably selected, for example, by minimizing the Akaike information criterion (AIC), the Bayesian information criterion (BIC), or cross-validated prediction error.

Say, $\epsilon > 0$ is a small number. For each given λ, the last computed $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_1 = \lambda - \epsilon$ can be used as s warm start in the search for the next one. Due to the continuity of the L^1 norm function $\|\cdot\|_1$, this search only needs to be conducted locally near the last computed $\boldsymbol{\alpha}$. Therefore a local (quadratic) approximation to the objective function in (4.2) near the last computed solution can be justified ([Radchanko, 2015]). This approach constructs a sequence of locally approximated convex objective functions near the last computed solutions, bypassing the issue of the noncontinuity of $\boldsymbol{\alpha}$ with respect to λ. The success of iterative algorithms depends on the initialization, i.e., solving (4.1). Due to the L^1 norm constraint $\|\boldsymbol{\alpha}\|_1 = \lambda$ in (4.2), we will implement a block coordinate descent (BCD) for solving (4.2), describe in Section 4.2.2.

4.2.2 Algorithm for treatment effect modifier selection

The sparsity parameter $\lambda \geq 1$ will be increased from 1 to next values, say, $\lambda + \epsilon$, on a dense grid in $[1, \lambda_{\max}]$. At each new $\lambda + \epsilon$, BCD can be performed to solve (4.2). This constructs a continuous path of solutions, $\boldsymbol{\alpha}^{(\lambda)}$'s, as a function of (increasing) λ. Following closely [Radchanko, 2015], the BCD algorithm for solving (4.2) for each given $\lambda + \epsilon$ is presented in the following.

Algorithm 3 Block-coordinate descent for optimizing $\boldsymbol{\alpha}$ subject to $\|\boldsymbol{\alpha}\|_1 = \lambda + \epsilon$, $\epsilon > 0$

```
Given: an initial estimate $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^p$ and $\hat{a}$ in (4.1)
Calculate $\boldsymbol{\nabla}$ (3.17) at $\boldsymbol{\alpha}^{(0)}$, and $L \leftarrow \arg\max_{j \in \{1,\dots,p\} \setminus \hat{a}} |\nabla_j|$.
Define $\mathcal{A} = \{\hat{a}, L\} \cup \{$indices of nonzero components of $\boldsymbol{\alpha}^{(0)}\}$.
Set $s_j = \text{sign}(\alpha_j)$; if $s_j = 0$, set $s_j = -\text{sign}(\nabla_j)$, $j = 1, \dots, p$.
$\alpha_L \leftarrow \alpha_L + s_L \epsilon$
for outer loop until convergence do
    $L \leftarrow \arg\max_{j \in \mathcal{A} \setminus \hat{a}} |\alpha_j|$.
    for inner loop $j \in \{\mathcal{A} \setminus \{\hat{a}, L\}\}$ do the $(j, L)$-block coordinate update:
        if $\alpha_j \neq 0$ or $(s_j \nabla_j \leq 0 \;\&\; |\nabla_j| \geq |\nabla_L|)$ then $\alpha_j \leftarrow \alpha_j + \Delta_j$, with $\Delta_j$ in (4.3).
        else $\Delta_j \leftarrow 0$.
        if $\alpha_j$ switches sign then $\Delta_j \leftarrow \Delta_j - \alpha_j$ and $\alpha_j \leftarrow 0$.
        $\alpha_L \leftarrow \alpha_L + \Delta_L$.
        if $\alpha_L$ switches sign then $\alpha_j \leftarrow \alpha_j + |\alpha_L| s_j$ and $\alpha_L \leftarrow 0$.
    end for
end for
Calculate $\boldsymbol{\nabla}$ (3.17) at $\boldsymbol{\alpha}$, and re-define $s_j$, $j = 1, \dots, p$.
if $\exists l \in \mathcal{A}^c$ for which $|\nabla_l| \geq |\nabla_L|$ then augment $\mathcal{A}$ with $l$, and repeat the above steps.
```

If we increase the tuning parameter λ to the next point on the grid, $\lambda^{(\text{new})} = \lambda + \epsilon$, $\epsilon > 0$, then the magnitude of some coefficients of $\boldsymbol{\alpha}$ needs to be increased from their current values to ensure the new L^1 norm constraint $\|\boldsymbol{\alpha}\|_1 = \lambda + \epsilon$ is satisfied. Let $L (\neq \hat{a})$ denote the component index of the gradient $\boldsymbol{\nabla}$ defined in (3.17) that corresponds to the largest absolute value. Without loss of generality, the increment ϵ can be brought to the Lth component, α_L, i.e., $\alpha_L \leftarrow \alpha_L + s_L \epsilon$, where the sign, s_L, is the opposite of the sign of the Lth component of the gradient, $-\text{sign}(\nabla_L)$, so that the change reduces the criterion function of (4.2).

It is suggested in [Radchanko, 2015] to use a block of size two that consists of $\{j, L\}$, where L is fixed for each outer loop, for performing the BCD. The algorithm cycles through the inner loop that optimizes individual blocks, until convergence of the outer loop. Within each block, a situation where a coefficient crosses zero is handled by setting that coefficient to exactly zero and correspondingly updating the other coefficient in the block. Here, we present an expression of Δ_j for the jth block's updating rule (see the Appendix for derivation), which is

$$\Delta_j = \frac{-\left(\nabla_j - S_{jL}\nabla_L\right)}{\|\hat{\boldsymbol{g}}'_{\boldsymbol{\alpha}} * (\boldsymbol{X}_j - S_{jL}\boldsymbol{X}_L)\|^2}, \tag{4.3}$$

where $S_{jL} := \text{sign}(\alpha_L\alpha_j) - \text{sign}(\alpha_L\nabla_j)I_{\{\alpha_j=0\}}$. To explain S_{jL}, it indicates the sign of $\alpha_L\alpha_j$, but, when $\alpha_j = 0$, the "sign" of α_j is the sign of $-\nabla_j$. The value of Δ_L is determined through the relationship

$$\Delta_L = -\Delta_j S_{jL}, \tag{4.4}$$

which makes the L^1 norm preserved (at $\|\boldsymbol{\alpha}\|_1 = \lambda + \epsilon$) for each within-block update.

4.3 Simulation examples

In Section 4.3.1 and Section 4.3.2, we performed numerical studies to illustrate the performance of the ITRs obtained from the SIMML with L^1 regularization, and were compared to several other alternatives, in a relatively high dimensional predictor space setting. In Section 4.3.3, we compared the treatment effect modifier selection performance of the L^1 regularized SIMML with the MCA.

4.3.1 ITR performance for $K = 2$ case

We first considered the case with the number of treatment groups $K = 2$. To simulate the data, we used the same settings as in Section 3.6.2, except that we increased the dimensionality of X, to relatively high dimensions $p = 100$ and $p = 200$. When estimating the ITRs, each of the methods considered was equipped with an appropriate regularization to deal with the high dimensionality, as described in the following.

SIMML Fit the constrained SIMML (3.2), estimated by the constrained L^1 regularization procedure described in Section 4.2.2, with the sparsity parameter λ selected by minimizing the AIC. The link g_T was approximated by cubic B-splines, with smoothing parameters selected by minimizing the GCV. An initial estimate for $\boldsymbol{\alpha}$ was obtained by the MCA (3.3) for the case $K = 2$, and by the linear GEM (3.26) for the case $K = 3$, both estimated via the Lasso, as described below.

MCA Estimate the modified covariates model (3.3) with efficiency augmentation ([Tian *et al.*, 2014]). Model (3.3) was fitted by the Lasso with a 10-fold cross validation for estimating interactions. Augmentation of main effect, $\mu(X) = \boldsymbol{\mu}^\top X$, was utilized, fitted by the Lasso with a 10-fold cross validation. MCA is applicable only for $K = 2$, therefore considered in the case of $K = 2$ only.

L.GEM Estimate the linear GEM model (3.26) as described in Section 3.5.3, where a Lasso penalized linear regression with a 10-fold cross validation was performed to estimate the linear model coefficients β_t, based on observations from the tth treatment group. The linear GEM is equivalent to the MCA in the population level when $K = 2$, therefore reported only for $K > 2$.

K.AM For each of the K groups separately, estimate a sparse additive model (SAM) ([Ravikumar *et al.*, 2009]), where the nonparametrically-defined component functions were approximated by cubic B-splines, with smoothing parameters selected by minimizing the GCV. The sparsity parameter was selected by minimizing the AIC.

K.LR For each of the K groups separately, estimate a linear regression (LR) model by the Lasso with a 10-fold cross validation.

The data generation settings are the same as in the settings of Section 3.6.2, except that we considered sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$. We set $\boldsymbol{\alpha} = (\underbrace{1, 0.5, 0.25, 0.125}_{\text{4 nonzeros}}, \underbrace{0, \ldots, 0}_{p-4})^\top$, and $\boldsymbol{\mu} = (\underbrace{0, \ldots, 0}_{p-4}, \underbrace{1, 0.5, 0.25, 0.125}_{\text{4 nonzeros}})^\top$, thus there were only 4 "signal" predictors that exhibited interactions with the treatment T.

In Figure 4.1, we present the boxplots of the PCDs (computed from testing datasets) of the estimated ITRs for each combination of $n \in \{200, 400\}$, $p \in \{100, 200\}$, and the main effect intensity $\delta \in \{1, 2\}$, obtained from 200 simulation runs, for the nonlinear contrast cases in the top panels and the linear contrast cases in the bottom panels. Under the linear contrast functions, the MCA slightly outperformed the SIMML approach. This is expected,

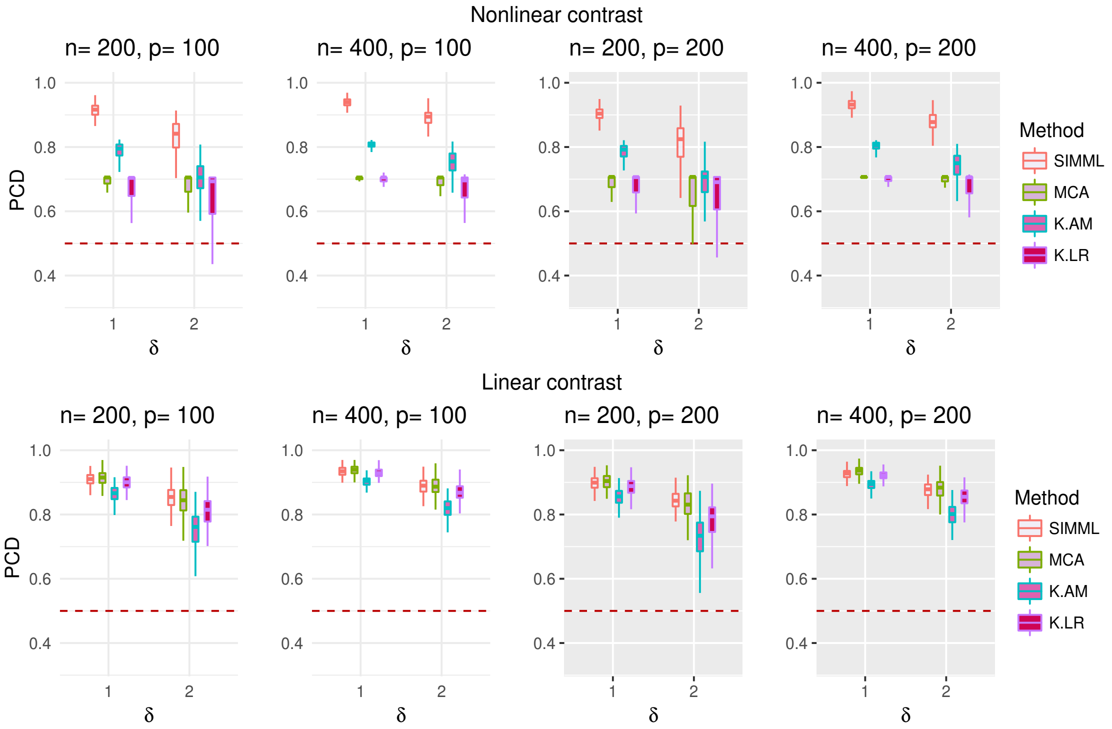

Figure 4.1: Top panels: boxplots of the PCDs of the ITRs estimated from the four methods (SIMML, MCA, K.AM, and K.LR) for the nonlinear contrast case. Lower panels: boxplots of the PCDs of the ITRs for the linear contrast case. For each case, $n \in \{200, 400\}$, $p \in \{100, 200\}$, and $\delta \in \{1, 2\}$ were considered.

since the MCA is correctly specified in the linear contrast case, and is a special case of the SIMML but under a more parsimonious model for estimating the T-by-X interactions on the outcome. However, when the contrast functions became nonlinear, the performance of the MCA sharply deteriorated, which was in contrast to the performance of the SIMML. For the nonlinear contrast, the performance of the SIMML was outstanding in all cases. On the other hand, the benefit of fitting the MCA and the K separate linear regression (K.LR) was small in all cases. The K separate (sparse) additive models (K.AM) performed reasonably well in comparison to the MCA and K.LR for the nonlinear contrast, however, was significantly outperformed by the SIMML, particularly when the main effect intensity increased to $\delta = 2$.

4.3.2 ITR performance for $K = 3$ case

In this section, we considered $K = 3$. To generate data, we used the same settings as in the settings of Section 3.6.3, except that we increased the number of predicdtors to $p = 100$ and $p = 200$. $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ were set at the same as in the settings of Section 4.3.1, thus there were only 4 "signal" predictors associated with the nonzero coefficients of $\boldsymbol{\alpha}$, interacting with the treatment indicator $T \in \{1, 2, 3\}$ in their effects on the outcome. In Figure 4.2, we displayed the boxplots of the PCDs (computed from testing datasets) obtained from 200 simulation runs, for the estimated ITRs from the 4 different methods (SIMML, L.GEM, K.AM, and K.LR) (described in the beginning of Section 4.3.1), for each $n \in \{200, 400\}$, $p \in \{100, 200\}$ and $\delta \in \{1, 3\}$.

When $K \geq 3$, separately estimating $\mathbb{E}(Y \mid X, T = t)$ for each group $t \in \{1, \ldots, K\}$ is a typical approach of modeling the T-by-X interactions. However, when the main effect intensity increases from $\delta = 1$ to $\delta = 3$, due to the large main effect variance, these K separate regression approaches tend to focus more on modeling the main effect of X missing important T-by-X interaction effect-related variabilities. On the other hand, the SIMML models the interaction effect only. As a result, in Figure 4.2, although the increased magnitude of the main effect affected the performance of all methods, it had least effect for the SIMML, and bigger effect for the K separate regression approaches. Moreover, the SIMML is a more parsimonious model than the K separate regression models, that provides

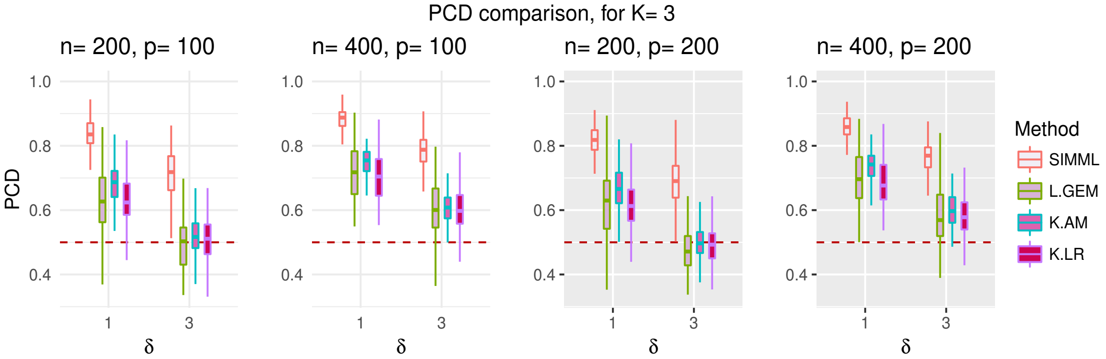

Figure 4.2: Boxplots of the PCDs of the ITRs estimated from the four methods (SIMML, L.GEM, K.AM, and K.LR), for each combination of $n \in \{200, 400\}$, $p \in \{100, 200\}$, and $\delta \in \{1, 3\}$, obtained from 200 simulated runs.

a superior interpretability.

4.3.3 Treatment effect modifier selection performance

In this section, we compare the performance of the SIMML and the MCA, in terms of the treatment effect modifier selection performance. Here, we report the results from the simulation example of Section 4.3.1, in which we computed the average number of correctly (C.) selected treatment effect modifiers (i.e., the true positives), and the average number of incorrectly (I.C.) selected treatment effect modifiers (i.e., the false positives), averaged out of the 200 simulation runs. The maximum possible number for the true positives was 4. The maximum possible number for the false positives was 96 for the $p = 100$ case and 196 for the $p = 200$ case. The selection performance results are presented in Table 4.1.

Not surprisingly, the cases with the nonlinear contrast was much more favorable to the SIMML in comparison to the MCA. In Table 4.1, the SIMML tended to correctly select the true 4 treatment effect modifiers, while the MCA selected almost 0 true treatment effect modifiers. The cases with the linear contrast, on the other hand, was slightly more favorable to the MCA than to the SIMML, however, their performance levels were quite comparable. In fact, although the averaged numbers of correctly selected treatment effect modifiers (C.) were slightly larger for the MCA, the average numbers of incorrectly selected treatment effect modifiers (I.C.) were actually small for the SIMML. Generally, due to the increasing

Contrast	Main Eff.	n	Avg. number							
shape	intensity		$p = 100$				$p = 200$			
			C. (max.4)		I.C. (max.96)		C. (max. 4)		I.C. (max.196)	
			SIMML	MCA	SIMML	MCA	SIMML	MCA	SIMML	MCA
Nonlinear	Small	200	**3.31**	0.17	6.27	**3.01**	**3.20**	0.15	7.53	**5.39**
		400	**3.78**	0.18	7.61	**3.35**	**3.60**	0.16	8.79	**4.24**
	Large	200	**1.98**	0.18	5.04	**3.22**	**1.70**	0.09	5.65	**4.89**
		400	**3.00**	0.17	6.12	**3.17**	**2.71**	0.09	6.90	**3.32**
Linear	Small	200	3.15	**3.32**	**6.35**	11.34	**2.93**	3.15	**7.01**	14.81
		400	3.66	**3.72**	**7.43**	12.83	**3.44**	3.56	**7.93**	14.39
	Large	200	2.07	**2.32**	**6.10**	10.39	**2.07**	2.32	**6.10**	10.39
		400	2.75	**2.96**	**5.78**	10.95	**2.61**	2.82	**5.97**	11.80

Table 4.1: Comparison of the treatment effect modifier selection performance of the SIMML and the MCA. The averaged number of correctly (C.) selected treatment effect modifiers and incorrectly (I.C.) selected treatment effect modifiers, averaged out of 200 simulation runs, are reported. Superior performances are indicated in bold-faced.

variabilities that were not related to the treatment (i.e., the "noise" variabilities), the main effect intensity affected the treatment effect modifier selection performance. However, as the sample size increased, the SIMML tended to recover the true treatment effect modifiers, while the MCA did not recover for the nonlinear contrast. Overall, there was a clear advantage of utilizing the flexible link functions for discovery of treatment effect modifiers when there was a nonlinear association between the treatment and a set of covariates, while the performance level of the MCA and the SIMML was similar when the interaction effect was linear.

4.4 Application: Depression RCT

In this section, we illustrate the utility of the constrained SIMML for estimating interactions on a real dataset. We considered a dataset from a RCT comparing an antidepressant ($t = 2$) and placebo ($t = 1$) (i.e., the number of treatments $K = 2$) for treating major depressive disorder (MDD), with a primary focus on the discovery of baseline clinical characteristics that potentially modify the effect of treatment for the MDD patients. In order to develop an ITR, the treatment-by-baseline predictors interactions need to be estimated. In many psychiatric diseases, simple moderators have already been discovered if there were any, and most patient characteristics at best have a weak moderating treatment effect. Therefore, the proposed approach of creating a composite baseline predictor (i.e., a single-index) that could exhibit stronger (nonlinear) interaction effects with the treatment could be particularly appealing.

Of the 166 subjects, 88 were randomized to placebo and 78 to drug. Some pretreatment clinical characteristics were collected, including: (i) current patient age; (ii) sex; (iii) severity of depressive symptoms measured by the Hamilton Rating Scale for Depression (HRSD); (iv) duration of the current major depressive episode (MDE); and (v) age of onset of first MDE. In addition to those more standard clinical assessments, patients underwent neuropsychiatric testing prior to treatments. Patients were tested on the following tasks: Flanker [Flanker and Eriksen, 1974], Choice reaction time (CRT) [Deary *et al.*, 2011], Word Fluency (WF) [Loonstra *et al.*, 2001], A not B working memory (AnotB) [Herrera-Guzman *et al.*, 2009] and several others. The purpose of these tests was to assess psychomotor slowing, working memory, reaction time (RT) and cognitive control (e.g., post-error recovery), as these behavioral characteristics were believed to correspond to biological phenotypes related to response to antidepressants. Table 4.2 displays the means and the standard deviations of the $p = 22$ baseline patient characteristics that were considered. The outcome Y was the improvement in symptoms severity (assessed by the HRSD scores) from baseline to week 8 taken as the difference (week 0 - week 8), and thus larger values of the outcome were better. The estimated coefficients $\boldsymbol{\alpha}$ obtained from the MCA under the efficiency augmentation with a linear model $\boldsymbol{\mu}^\top X$, estimated via the Lasso with a 10 fold cross validation, and the $\boldsymbol{\alpha}$ estimated from the proposed SIMML approach are presented in Table 4.2.

In the upper panel of Figure 4.3, we plotted the outcomes against the projection $\hat{\boldsymbol{\alpha}}^\top X$, in which $\hat{\boldsymbol{\alpha}}$ was estimated from the SIMML under L^1 regularization, and an estimated pair of the treatment-specific B-spline approximated link functions $\hat{g}_t$, $t \in \{1, 2\}$ were superimposed. As indicated in Table 4.2, the SIMML estimate $\hat{\boldsymbol{\alpha}}$ had two nonzero components, associated with the predictor x_{18} ("Flanker Accuracy") and the predictor x_{22} ("Flanker

(Label) Baseline	Mean	SIMML coef.		MCA coef.	
predictors	(SD)	L^1-regul.	Unregul.	L^1-regul.	Unregul.
(x_1) Age at evaluation	38.00 (13.84)	0	0.20	0	-0.11
(x_2) Sex	0.64 (0.48)	0	-0.06	0	0.05
(x_3) HRSD(baseline)	18.8 (4.29)	0	0.37	0	0.18
(x_4) Dur. MDE	38.19 (53.17)	0	0.21	0	-0.04
(x_5) Age at MDE	16.4 (6.09)	0	-0.04	0	-0.13
(x_6) Family history	0.87 (0.68)	0	-0.09	0	-0.02
(x_7) Fatigue	2.87 (0.39)	0	-0.01	0	-0.00
(x_8) Hypersomnia	0.20 (0.40)	0	0.11	0	0.03
(x_9) Axis II	3.92 (1.43)	0	-0.09	0	-0.05
(x_{10}) Anger attack	3.00 (2.11)	0	0.02	0	-0.04
(x_{11}) Anxiety	5.33 (1.87)	0	-0.21	0	-0.09
(x_{12}) AnotB, RT(negative)	0.30 (2.13)	0	-0.24	0	-0.10
(x_{13}) AnotB, RT(non-neg.)	0.32 (1.63)	0	-0.13	0	0.10
(x_{14}) AnotB, RT(all)	0.37 (1.77)	0	0.21	0	0.06
(x_{15}) AnotB, total correct	0.16 (0.77)	0	-0.06	0	0.06
(x_{16}) Median choice RT	0.23 (1.45)	0	-0.10	0	-0.03
(x_{17}) Word Fluency	37.42 (11.68)	0	-0.10	0	-0.04
(x_{18}) Flanker Accuracy	0.22 (0.15)	0.84	-0.64	-0.78	-0.72
(x_{19}) Flanker RT	59.51 (26.63)	0	-0.17	0	0.01
(x_{20}) Post-conflict adjus.	0.07 (0.12)	0	0.07	0	0.03
(x_{21}) $x_1 \times x_3$ interaction	722.99 (336.70)	0	-0.12	0.59	-0.10
(x_{22}) $x_1 \times x_{18}$ interaction	7.93 (6.27)	-0.53	0.25	0	0.59

Table 4.2: Description of $p = 22$ baseline predictors and the estimated L^1 regularized/unregularized index coefficients $\boldsymbol{\alpha}$ from the SIMML and the MCA, respectively. All coefficient estimates were scaled to have unit L^2 norm.

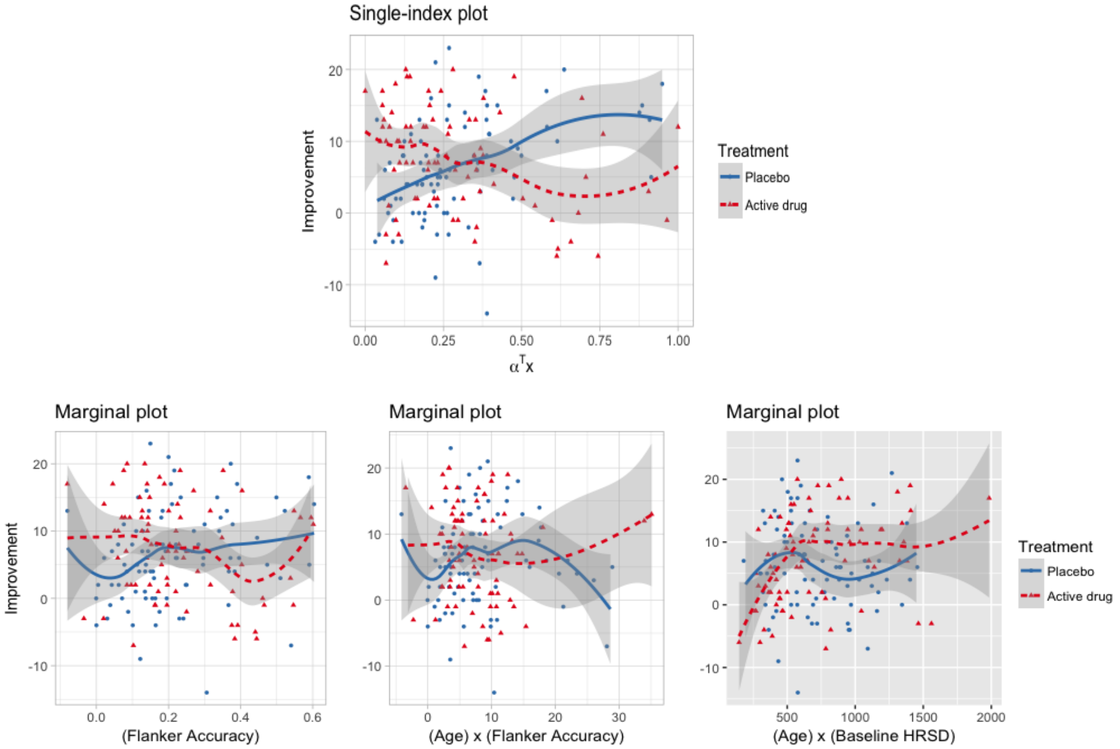

Figure 4.3: Observed values of the response variable against the estimated single-index ($= \boldsymbol{\alpha}^\top X$) from the SIMML in the top panel, and against individual predictors, "Flanker Accuracy", "(Age) x (Flanker Accuracy)", and "(Age) x (Baseline HRSD)", from left to right, respectively, in the bottom panels. Pairs of estimated treatment-specific B-spline approximated link functions with the associated 95% confidence bands were overlaid; in the top panel, the confidence bands were constructed conditioning on $\boldsymbol{\alpha}$.

Accuracy"-by-"Age at evaluation" interaction). In the first two panels in the bottom row of Figure 4.3, we display the marginal plots of the outcomes against each of these two predictors. In each of the plots, an estimated pair of the treatment-specific B-spline approximated curves were overlaid to describe the relationship with the outcomes. In the plots, although the interaction effect between the treatment and the predictor "Flanker Accuracy" seemed almost linear, the interaction effect between the treatment and the ("Flanker Accuracy"-by-"Age at evaluation" interaction) exhibited substantial nonlinearities, which, if one used only the linear regression lines to describe the relationship, then it would have been difficult to be detected. The estimated composite treatment effect modifier $\hat{\boldsymbol{\alpha}}^{\top}X$ was the linear combination (with the weights 0.84 and -0.53) of the predictor "Flanker Accuracy" and the predictor ("Flanker Accuracy"-by-"Age at evaluation" interaction). Note, the composite treatment effect modifier $\hat{\boldsymbol{\alpha}}^{\top}X$ exhibited a stronger interaction effect compared to each individual predictor marginally, as can be observed in Figure 4.3. The MCA identified the predictor x_{18} ("Flanker Accuracy") and the predictor x_{21} ("baseline HRSD"-by-"Age at evaluation" interaction) as important treatment effect modifiers. The marginal plot for the predictor("baseline HRSD"-by-"Age at evaluation" interaction) is displayed in the bottom right panel of Figure 4.3.

To evaluate the performance of the ITRs estimated from the SIMML and the MCA, we randomly split the data into a training set and a testing set using a ratio of 10 to 1, replicated 500 times, each time fitting the methods on the training set and computing the estimated value of ITR (2.22) based on the test set. In addition to the SIMML and the MCA, we included the K separate additive models (K.AM) and the K separate linear regression (K.LR), described in Section 4.3.1, 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).

In Figure 4.4, the proposed SIMML approach, in terms of the average estimated values, outperformed all other alternatives. In particular, the SIMML outperformed the MCA and the K separate linear regressions, which indicated that the SIMML that employs unspecified link functions to approximate the heterogeneous treatment effect was better suited for developing ITRs. We also note that the SIMML has a significant interpretational advantage

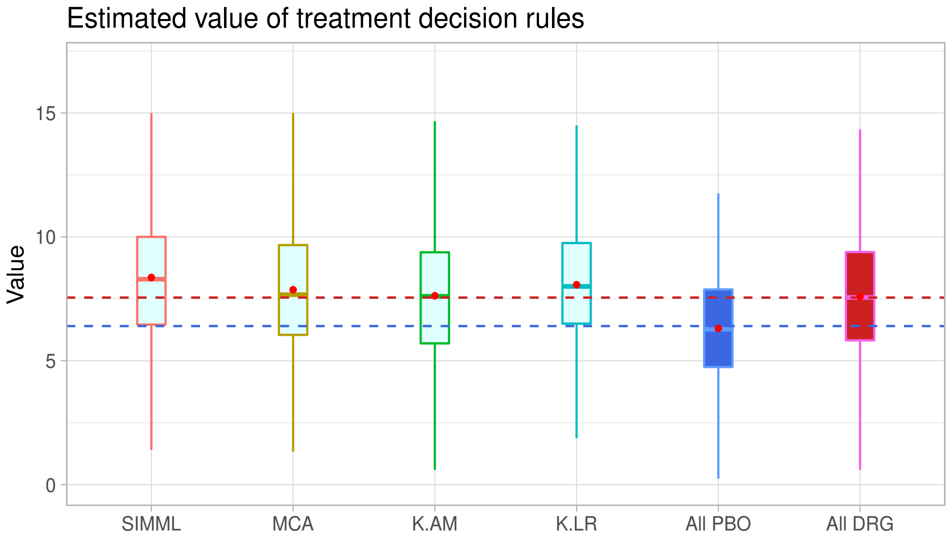

Figure 4.4: Boxplots of the values (2.22) of the ITRs estimated from the six different approaches, obtained from the 500 randomly split testing sets. Higher values are preferred.

over the K separate regression approaches (K.AM and K.LR). The estimated single-index $\boldsymbol{\alpha}^\top X$, as displayed in the top panel of Figure 4.3, allows visualization, and the single-index coefficient $\boldsymbol{\alpha}$ describes the relative importance of each predictor in characterizing the heterogeneous treatment effect.

4.5 Discussion/Extension to a partially linear single-index model (PLSIM) with multiple-links

We focused on the class of a dimension reduction using a single-index of the form $\boldsymbol{w}_{\boldsymbol{\alpha}}(X) = \boldsymbol{\alpha}^\top X \in \mathbb{R}$, however, more general nonlinear multiple-indices, for example, a vector-valued function, $\boldsymbol{w}_{\boldsymbol{\alpha}}(X) \in \mathbb{R}^q$, can also be considered, where q (i.e., multiple) indices need to be estimated and can be utilized to model the (high dimensional) interaction effects. This approach extends the SIMML to a multiple-index model with multiple-links (MIMML) for estimating interactions. In the following, we consider one instance of the MIMML.

For $K = 2$, the SIMML model (3.2) can be extended to the context of the partially linear single-index models (PLSIM), by adding a modified covariate (MC) linear component in the model. Assuming the equal probability $\pi_1 = \pi_2 = 1/2$ for the treatment assignment, we consider the following extended SIMML model that embeds the MCA model (3.3) inside

$$\mathbb{E}\left(Y \mid X, T = t\right) = \mu(X) + \boldsymbol{\beta}^\top X(-1)^t/2 + g_t(\boldsymbol{\alpha}^\top X), \quad t = 1, 2, \tag{4.5}$$

where the unspecified link functions g_t, $t \in \{1, 2\}$, are again subject to the constraint (3.5), i.e., $\sum_{t=1}^{2} \pi_t g_t(\boldsymbol{\alpha}^\top X) = 0$, and the main effect $\mu(X)$ is completely unspecified. The interaction component $\boldsymbol{\beta}^\top X(-1)^t/2 + g_t(\boldsymbol{\alpha}^\top X)$ under the constraint (3.5) satisfies the "orthogonality" condition, i.e., $\mathbb{E}_T\left(\boldsymbol{\beta}^\top X(-1)^T/2 + g_T(\boldsymbol{\alpha}^\top X) \mid X\right) = 0$, that gives the orthogonality with the unspecified main effects, making the approach robust to a potential misspecification of any working model for $\mu(X)$. Model (4.5) is in the class of extended PLSIM ([Lian and Liang, 2016], [Xia *et al.*, 1999]), in which the linear term $\boldsymbol{\beta}^\top X(-1)^t/2$ and the nonparametric term $g_t(\boldsymbol{\alpha}^\top X)$ of the interaction component share the same set of predictors X. For model identifiability of the extended PLSIM, an additional constraint, for example, $\boldsymbol{\alpha} \perp \boldsymbol{\beta}$ ([Yuan, 2011]) is required. In the Appendix, the details of estimating the extended SIMML (4.5) under such constraint are described.

Future work in refining and developing the proposed approach will investigate the incorporation of baseline functional predictors (considered in Chapter 5), and an extension to incorporate longitudinal outcomes.

Chapter 5

A Sparse functional additive model with multiple-links

5.1 Introduction

In this chapter, we will develop an additive regression approach for estimating interactions between a treatment variable and a large number of functional/scalar predictors. If the main effect of baseline predictors is misspecified or high-dimensional (or, infinite dimensional), any standard nonparametric or semiparametric approach for estimating the treatment-by-predictors interactions tends to be unsatisfactory because it is prone to (possibly severe) inconsistency and poor approximation to the true treatment-by-predictors interaction effect. This is particularly problematic for modeling the treatment-by-functional predictors interactions, due to the infinite dimensional nature of functional predictors. The infinite dimensionality requires some sort of dimension reduction when we formulate a model for the main effect, which essentially involves a main effect model misspecification in a finite sample. Thus, estimating treatment-by-predictors interactions in the context of a functional regression is particularly challenging.

To address this issue, we will apply the methodology developed in the previous chapters that gives the orthogonality between the main and the interaction effect components in a regression model, and estimate the main and the interaction effect components separately. If our interest is in interactions, we can estimate the interaction component only and no main effects. Again, this approach obviates the need to specify the form of the main effect component, thus side-stepping the issue of misspecification of the main effects. Further, we will impose a concave penalty in estimation, and the proposed estimation method will simultaneously select functional/scalar treatment effect modifiers that exhibit possibly nonlinear interaction effects with the treatment indicator.

Earlier attempts to model interactions in a regression setting include modified covariate approach ([Tian *et al.*, 2014]). The method simply codes the treatment variable as ±0.5 and then includes the products of this variable with each covariate in an appropriate working mode. The [Tian *et al.*, 2014]'s approach obviates the need to directly model the main effect of covariates. MCA was extended to incorporate multiple functional and scalar predictors ([Ciarleglio *et al.*, 2015b]). See also [McKeague and Qian, 2014] for a functional data-analytic approach for developing ITRs with a functional predictor. However, these methods are limited to parametric regression models for estimating the interactions.

In this chapter, we remove the parametric model restriction by developing an additive regression ([Hastie and Tibshirani, 1999]) model for estimating the interactions. Specifically, we will base our method on a functional additive model (FAM) ([Fan *et al.*, 2015]). However, FAM will be extended to have treatment-specific multiple link functions for charactering the treatment-by-functional predictors interactions. The proposed functional additive model will be optimized for the treatment-by-predictors interaction effects. A sparse nonlinear combination of the baseline functional/scalar predictors is derived via a sparse additive model formulation ([Ravikumar *et al.*, 2009], [Fan *et al.*, 2014]), and the method achieves a simultaneous treatment effect modifier selection. The approach provides a natural semiparametric framework for estimating interactions between the treatment and multiple functional/scalar predictors.

5.2 Method

5.2.1 Functional additive model with multiple-links (FAMML)

We consider a RCT that consists of K treatment arms, with a scalar treatment outcome Y. The treatment assignment variable, T, which takes a value t from the set $\{1, \ldots, K\}$, is assumed to be randomized to follow a discrete probability distribution, say, $\{\pi_1, \ldots, \pi_K \mid \sum_{t=1}^{K} \pi_t = 1, \pi_t > 0\}$. We consider p different baseline functional predictors, denoted by $\boldsymbol{x}_j(s)$, $j = 1, \ldots, p$, each of them is squared integrable, defined on a compact interval, say, $[0, 1]$. For convenience, we will collectively write $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$. We also consider a set of q baseline scalar predictors, written by $Z = (z_1, \ldots, z_q)^\top$. For notation, $\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle = \int_0^1 \boldsymbol{\alpha}_j(s)\boldsymbol{x}_j(s)ds$, where $\boldsymbol{\alpha}_j(s)$ is a square integrable function defined on $[0, 1]$, with $\|\boldsymbol{\alpha}_j\| = 1$, corresponding to a direction that we project $\boldsymbol{x}_j(s)$ into.

We will consider a FAM ([Fan *et al.*, 2014], [Fan *et al.*, 2015]), equipped with treatment T-specific unspecified link functions ([Park *et al.*, 2017]) for modeling interactions, a functional additive model with multiple-links (FAMML)

$$E(Y \mid X, Z, T) = \sum_{j=1}^{p} \Big(\underbrace{\mu_j(\boldsymbol{x}_j)}_{\text{main}} + \underbrace{g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big)}_{\text{interaction}} \Big) + \sum_{k=1}^{q} \Big(\underbrace{h_k(z_k)}_{\text{main}} + \underbrace{h_{k,T}(z_k)}_{\text{interaction}} \Big), \qquad (5.1)$$

with $T \in \{1, \ldots, K\}$, where $g_{j,T}(\cdot)$ is designed to capture the nonlinear treatment T-by-$\boldsymbol{x}_j$ interaction effects, as a general smooth function of the jth functional index $\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle$, for each $j = 1, \ldots, p$. Here, the projection functions $\boldsymbol{\alpha}_j$ need to be determined. For scalar predictors, $h_{k,T}(\cdot)$ is designed to capture the nonlinear T-by-z_k interaction effects, as a general smooth function of the scalar predictor z_k, for each $k = 1, \ldots, q$. Importantly, the potentially complicated main effect function $\mu_j(\boldsymbol{x}_j)$ is left unspecified in FAMML (5.1). Model (5.1) provides a useful semiparametric framework to estimate the nonlinear interaction effects, while treating the main effect component as a "nuisance" component. Without loss of generality, we assume that the outcome Y is centered at 0 per each treatment group $t \in \{1, \ldots, K\}$.

With nontrivial main effects, a necessary and sufficient condition for the orthogonality

$$\begin{aligned} \mu_j(\boldsymbol{x}_j) \ &\perp \ g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big), \quad j = 1, \ldots, p, \quad \text{a.s.,} \quad \text{and} \\ h_k(z_k) \ &\perp \ h_{k,T}(z_k), \quad k = 1, \ldots, q, \quad \text{a.s.,} \end{aligned} \qquad (5.2)$$

between the main effect component and the interaction effect component in model (5.1) is given by

$$\begin{aligned} \mathbb{E}_T \left(g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big) \mid \boldsymbol{x}_j \right) = 0, \quad j = 1, \ldots, p, \quad \text{a.s.,} \quad \text{and} \\ \mathbb{E}_T \left(h_{k,T}(z_k) \mid z_k \right) = 0, \quad k = 1, \ldots, q, \quad \text{a.s.,} \end{aligned} \qquad (5.3)$$

i.e., the nonparametrically-defined link functions, $g_{j,T}$'s and $h_{k,T}$'s, have mean zero with respect to the treatment indicator T.

For model (5.1), the orthogonality (5.2) is attractive, since the main effect $\sum_{j=1}^{p} \mu_j(\boldsymbol{x}_j) + \sum_{k=1}^{q} h_k(z_k)$, and the interaction effect $\sum_{j=1}^{p} g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big) + \sum_{k=1}^{q} h_{k,T}(z_k)$, can be estimated separately. This suggests a simpler working model than (5.1), using the interaction components only and no main effects, if our interest is in estimating interactions

$$\mathbb{E}(Y \mid X, Z, T) = \sum_{j=1}^{p} g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big) + \sum_{k=1}^{q} h_{k,T}(z_k), \tag{5.4}$$

where $g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big)$'s and $h_{k,T}(z_k)$'s are constrained by (5.3) to give the orthogonality (5.2). Working with model (5.4) under (5.3) is appealing, since we do not have to specify the form of the main effects, side-stepping the issues with misspecification of $\mu_j(\boldsymbol{x}_j)$, a potentially complicated function.

To deal with a large number of p and q, we can seek a further structure on model (5.4), with the following reparametrization

$$\mathbb{E}(Y \mid X, Z, T) = \sum_{j=1}^{p} \beta_j g^*_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big) - \sum_{k=1}^{q} \gamma_k h^*_{k,T}(z_k), \tag{5.5}$$

under the orthogonality constraint

$$\mathbb{E}_T\left(g^*_{j,T}\right) = \mathbb{E}_T\left(h^*_{k,T}\right) = 0, \tag{5.6}$$

where $\|g^*_{j,T}\| = \|h^*_{k,T}\| = 1$, i.e., the scales are taken out, and impose sparsity on the scales, $(\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)^\top \in \mathbb{R}^{p+q}$, i.e., we assume that most of the predictors are unrelated to the response as treatment effect modifiers. A geometric intuition of model (5.5) under the constraint (5.6) is described in Figure 5.1.

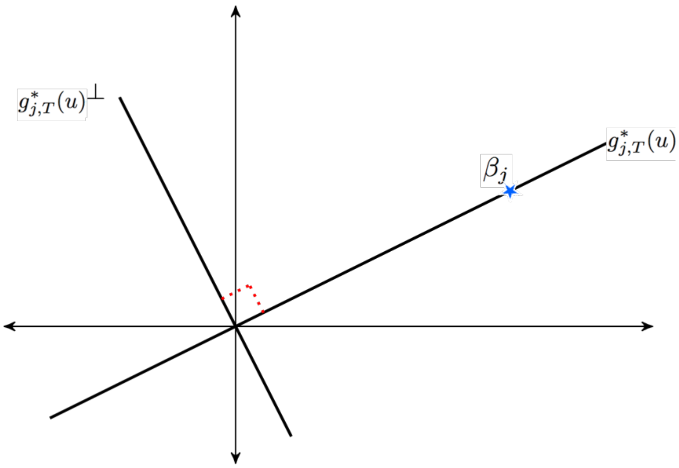

Figure 5.1: Figure describes a set of orthogonal coordinate axes. $g^*_{j,T}(u)$ corresponds to the axis for the T-by-u interaction effect. Its orthogonal complement, $g^*_{j,T}(u)^\perp$, corresponds to the axis for the main effect of u. The interaction effect is quantified by β_j. Here, $u = \langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle$, to be optimized over $\boldsymbol{\alpha}_j$. The regression plane is represented by the two orthogonal axes. The nonlinearity is captured by $g^*_{j,T}(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle)$, and the scale is captured by β_j.

5.2.2 Criterion

In this section, we propose the criterion for optimizing the interaction effect component of model (5.1). First, we formulate an optimization problem in the population setting. We suggest to minimize the following constrained mean squared error of the working model (5.5), over the functions $g^*_{j,t}$'s, $h^*_{k,t}$'s, and $\boldsymbol{\alpha}_j$'s

$$\begin{aligned}
\underset{\beta_j,\gamma_k,g^*_{j,T},h^*_{k,T},\boldsymbol{\alpha}_j}{\text{minimize}} \quad & \mathbb{E}\left(Y - \sum_{j=1}^{p} \beta_j g^*_{j,T}\left(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\right) - \sum_{k=1}^{q} \gamma_k h^*_{k,T}(z_k)\right)^2 \\
\text{subject to} \quad & \sum_{j=1}^{p} |\beta_j| + \sum_{k=1}^{q} |\gamma_k| \quad \leq \quad L \\
& \|g^*_{j,T}\| = \|h^*_{k,T}\| = 1,\ j = 1,\dots,p,\ k = 1,\dots,q \\
& \mathbb{E}_T\left(g^*_{j,T}\right) = \mathbb{E}_T\left(h^*_{k,T}\right) = 0,\ j = 1,\dots,p,\ k = 1,\dots,q.
\end{aligned} \tag{5.7}$$

The constraint that $(\beta_1, \dots, \beta_p, \gamma_1, \dots, \gamma_q)^\top \in \mathbb{R}^{p+q}$ lies in the L^1 ball (the 2nd line in

(5.7)) encourages sparsity on the index set for the link functions that are nonzero. The orthogonality constraint (the 4th line in (5.7)) gives the orthogonality with the unspecified main effect. Additionally, we need identifiability constraints $\|\boldsymbol{\alpha}_j\| = 1$, $j = 1, \dots, p$.

Consider the following convenient equivalent form ([Ravikumar *et al.*, 2009]) for (5.7):

$$\begin{aligned}
\underset{g_{j,T}, h_{k,T}, \boldsymbol{\alpha}_j}{\text{minimize}} \quad & \mathbb{E}\Big(Y - \sum_{j=1}^{p} g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle\big) - \sum_{k=1}^{q} h_{k,T}(z_k)\Big)^2 \\
\text{subject to} \quad & \sum_{j=1}^{p} \|g_{j,T}\| + \sum_{k=1}^{q} \|h_{k,T}\| \quad \leq \quad L \\
& \mathbb{E}_T\left(g_{j,T}\right) = \mathbb{E}_T\left(h_{k,T}\right) = 0, \ j = 1, \dots, p, \ k = 1, \dots, q
\end{aligned} \tag{5.8}$$

with $\|\boldsymbol{\alpha}_j\| = 1, \ j = 1, \dots, p$.

Given $\boldsymbol{\alpha}_j$'s, (5.8) can be written in the following equivalent penalized Lagrangian form

$$\begin{aligned}
&\mathbb{E}\Big(Y - \sum_{j=1}^{p} g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle\big) - \sum_{k=1}^{q} h_{k,T}(z_k)\Big)^2 \\
&\qquad + \lambda\Big(\sum_{j=1}^{p} \|g_{j,T}\| + \sum_{k=1}^{q} \|h_{k,T}\|\Big) + \sum_{j=1}^{p} \tau_j \mathbb{E}_T\left(g_{j,T}\right) + \sum_{k=1}^{q} \kappa_k \mathbb{E}_T\left(h_{k,T}\right),
\end{aligned} \tag{5.9}$$

where $\lambda \geq 0$ is the sparsity parameter for prediction selection, in a similar fashion to the group Lasso ([Yuan and Lin, 2006]). Given $\boldsymbol{\alpha}_j$'s, the minimizing functions in (5.9) have a closed-form expression in the population setting, as follows.

Theorem 4. *Given $\boldsymbol{\alpha}_j$'s, the minimizers $g_{j,T}$ of (5.9) satisfy*

$$g_{j,T}(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle) \ = \ \left[1 - \frac{\lambda}{\|P_{j,T}(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle)\|}\right]_+ P_{j,T}(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle), \quad a.s., \tag{5.10}$$

where

$$P_{j,T}(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle) \quad = \quad \mathbb{E}(R_j \mid \langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle, T) \quad - \quad \mathbb{E}(R_j \mid \langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j\rangle), \tag{5.11}$$

in which the residual $R_j = Y - \sum_{j' \neq j} g_{j',T}(\langle \boldsymbol{\alpha}_{j'}, \boldsymbol{x}_{j'}\rangle) - \sum_{k=1}^{q} h_{k,T}(z_k)$. The minimizers $h_{k,T}$ of (5.9) satisfy

$$h_{k,T}(z_k) \ = \ \left[1 - \frac{\lambda}{\|Q_{k,T}(z_k)\|}\right]_+ Q_{k,T}(z_k), \quad a.s., \tag{5.12}$$

where

$$Q_{k,T}(z_k) \quad = \quad \mathbb{E}(R_k \mid z_k, T) \quad - \quad \mathbb{E}(R_k \mid z_k), \tag{5.13}$$

in which the residual $R_k = Y - \sum_{j=1}^{p} g_{j,T}(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle) - \sum_{k' \neq k}^{q} h_{k',T}(z_{k'})$. *Here,* $Z_+ = \max(0, Z)$ *represents the positive part of* Z.

The proof of Theorem 4 is in the Appendix. In the population setting, for each $L \geq 0$ (or equivalently, for some $\lambda \geq 0$), the problem of optimizing (5.7) can be split into two iterative steps ([Fan *et al.*, 2014], [Fan *et al.*, 2015]). First (Step 1), given $\boldsymbol{\alpha}_j$'s, the links $g_{j,T}$'s and $h_{k,T}$'s can be found by a coordinate descent procedure that fixes $g_{j',T}$'s (or $h_{k',T}$'s) at all $j' \neq j$ (or $k' \neq k$), and obtain $g_{j,T}$ (or $h_{k,T}$) by equation (5.10) (or (5.12)), then iterate over j and k. This step corresponds to estimating under the constraint (5.3) a sparse additive model (SAM) ([Ravikumar *et al.*, 2009]). Second (Step 2), given the links $g_{j,T}$'s and $h_{k,T}$'s, $\boldsymbol{\alpha}_j$'s can be obtained by minimizing

$$\mathbb{E}\left(Y - \sum_{j=1}^{p} g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big) - \sum_{k=1}^{q} h_{k,T}(z_k)\right)^2, \tag{5.14}$$

under the constraints $\|\boldsymbol{\alpha}_j\| = 1$, $j = 1, \ldots, p$. These two steps can be iterated until convergence to obtain a population solution. To obtain a sample version of the population solution, we can insert sample estimates into the population algorithm, as in standard backfitting in fitting generalized additive models ([Hastie and Tibshirani, 1999]).

5.2.3 Estimation

We assume that we are given data $\{(Y_i, T_i, X_i, Z_i)_{i=1}^{n}\}$, with the set of p functional predictors $X_i = \{\boldsymbol{x}_{i1}(s), \ldots, \boldsymbol{x}_{ip}(s)\}$ (assumed to be observed without errors), the vector of q scalar predictors $Z_i = (z_{i1}, \ldots, z_{iq})^\top$, and the treatment indicator variable T_i that takes a value $t \in \{1, \ldots, K\}$, corresponding to the ith subject, $i = 1, \ldots, n$, where $n = \sum_{t=1}^{K} n_t$ is the total sample size, with n_t denoting the sample size for the tth treatment group, i.e., $\{i \mid T_i = t, i = 1, \ldots, n\}$.

5.2.3.1 Representation

Let us represent the projection functions $\boldsymbol{\alpha}_j(s) = \tilde{\boldsymbol{\alpha}}_j^\top \boldsymbol{\Phi}_j(s)$ and the predictors $\boldsymbol{x}_{ij}(s) = \tilde{\boldsymbol{x}}_{ij}^\top \boldsymbol{\Psi}_j(s)$, with appropriately chosen (orthogonal) basis functions, $\boldsymbol{\Phi}_j(s)$ and $\boldsymbol{\Psi}_j(s)$, respectively. To ensure that $\boldsymbol{\alpha}_j$ and the links $g_{j,T}$ are identifiable, we constrain $\|\tilde{\boldsymbol{\alpha}}_j\| = 1$ for all

j. Then we write

$$\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_{ij} \rangle = \int_0^1 \boldsymbol{\alpha}_j(s)\boldsymbol{x}_{ij}(s)ds = \tilde{\boldsymbol{\alpha}}_j^\top \left[\int_0^1 \boldsymbol{\Phi}_j(s)\boldsymbol{\Psi}_j(s)^\top ds \right] \tilde{\boldsymbol{x}}_{ij} = \tilde{\boldsymbol{\alpha}}_j^\top \boldsymbol{\xi}_{ij},$$

where $\boldsymbol{\xi}_{ij} = \left[\int_0^1 \boldsymbol{\Phi}_j(s)\boldsymbol{\Psi}_j(s)^\top ds \right] \tilde{\boldsymbol{x}}_{ij}$. Here, the coefficients $\tilde{\boldsymbol{\alpha}}_j$ need to be estimated, whereas the subject i- and the predictor j-specific scores $\tilde{\boldsymbol{x}}_{ij}$ can be easily computed, since $\boldsymbol{x}_{ij}(s)$ are directly observed.

Let us write the $n \times 1$ vector $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_K^\top\right)^\top$, in which the $n_t \times 1$ vector $\boldsymbol{Y}_t = (Y_{1_t}, \ldots, Y_{n_t})^\top$ is the observed response vector corresponding to the tth treatment group. For the regression function $g_{j,T}$ of the jth functional predictor, let us write the $n \times 1$ vector of evaluations, $\boldsymbol{g}_{\tilde{\boldsymbol{\alpha}}_j} = \left(\boldsymbol{g}_{\tilde{\boldsymbol{\alpha}}_j,1}^\top, \ldots, \boldsymbol{g}_{\tilde{\boldsymbol{\alpha}}_j,K}^\top\right)^\top$, the $n_t \times 1$ vector $\boldsymbol{g}_{\tilde{\boldsymbol{\alpha}}_j,t} = \left(g_{j,t}(\tilde{\boldsymbol{\alpha}}_j^\top \boldsymbol{\xi}_{j1}), \ldots, g_{j,t}(\tilde{\boldsymbol{\alpha}}_j^\top \boldsymbol{\xi}_{jn_t})\right)^\top$ corresponds to the vector of evaluations from the tth treatment group, $t \in \{1, \ldots, K\}$. Similarly, for the regression function $h_{k,T}$ of the kth scalar predictor, let us write the $n \times 1$ vector of evaluations, $\boldsymbol{h}_k = \left(\boldsymbol{h}_{k,1}^\top, \ldots, \boldsymbol{h}_{k,K}^\top\right)^\top$, in which the $n_t \times 1$ vector $\boldsymbol{h}_{k,t} = (h_{k,t}(z_{k1}), \ldots, h_{k,t}(z_{kn_t}))^\top$ corresponds to the observations from the tth treatment group. Both the link functions, $g_{j,t}(\cdot)$'s and $h_{k,t}(\cdot)$'s, and the projection directions, $\tilde{\boldsymbol{\alpha}}_j$'s, need to be estimated.

For approximating the conditional expectations in (5.11) and (5.13), any nonparametric smoothers can be used, for example, B-splines [de Boor, 2001] and local kernel regression ([Ruppert and Wand, 1994], [Hardle and Muller, 2012]). For each candidate $\tilde{\boldsymbol{\alpha}}_j$, let us denote a suitable nonparametric smoother for approximating the bivariate conditional expectation $\mathbb{E}(R_j \mid \langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle, T)$ in (5.11) by $\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}}^{(**)}$. Also, let us denote a suitable nonparametric smoother for approximating the univariate conditional expectation $\mathbb{E}(R_j \mid \langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle)$ in (5.11) by $\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}}^{(*)}$. Then, a smoother for (5.11) can be given by

$$\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} = \boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j}^{(**)} - \boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j}^{(*)}, \tag{5.15}$$

for the jth functional predictor. Similarly, for (5.13), with a smoother for $\mathbb{E}(R_k \mid z_k, T)$ denoted by $\boldsymbol{H}_k^{(**)}$ and a smoother for $\mathbb{E}(R_k \mid z_k)$ denoted by $\boldsymbol{H}_k^{(*)}$, a smoother for (5.13) can be given by

$$\boldsymbol{H}_k = \boldsymbol{H}_k^{(**)} - \boldsymbol{H}_k^{(*)}, \tag{5.16}$$

for the kth scalar predictor.

5.2.3.2 Algorithm

Given $\tilde{\boldsymbol{\alpha}}_j$ and the smoother matrices (5.15) and (5.16), the corresponding plug-in estimates for (5.10) and (5.12) are

$$\hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j} = c_j \boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j, \quad j = 1, \ldots, p, \tag{5.17}$$

where $\boldsymbol{R}_j = \boldsymbol{Y} - \sum_{j' \neq j} \hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_{j'}} - \sum_{k=1}^{q} \hat{\boldsymbol{h}}_k$, with the shrinkage factor $c_j = \left(1 - \lambda / \|\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|\right)_+$, and

$$\hat{\boldsymbol{h}}_k = c_k \boldsymbol{H}_k \boldsymbol{R}_k, \quad k = 1, \ldots, q, \tag{5.18}$$

where $\boldsymbol{R}_k = \boldsymbol{Y} - \sum_{j=1}^{p} \hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j} - \sum_{k' \neq k} \hat{\boldsymbol{h}}_{k'}$, with the shrinkage factor $c_k = (1 - \lambda / \|\boldsymbol{H}_k \boldsymbol{R}_k\|)_+$, respectively.

Theorem 4 and the expressions (5.17) and (5.18) for each fixed $\tilde{\boldsymbol{\alpha}}_j$ suggest an iterative algorithm to approximately solve (5.8) for the links $g_{j,T}$ and $h_{k,T}$, and the projection weights $\tilde{\boldsymbol{\alpha}}_j$. Following [Fan *et al.*, 2014], we summarize below an algorithm that alternates between a coordinate descent (CD) for updating the link functions $g_{j,T}$'s and $h_{k,T}$'s (i.e., Step 1), and a gradient descent (GD) for updating the projection weights $\tilde{\boldsymbol{\alpha}}_j$'s (i.e., Step 2).

Algorithm 4 Estimation of a sparse additive model with treatment specific links

```
Compute $\boldsymbol{H}_k$ defined in (5.16) for $k = 1, \ldots, q$.
Initialize $\tilde{\boldsymbol{\alpha}}_j$ for $j = 1, \ldots, p$.
for outer iteration until convergence do
    (Step 1: the link updates via CD)
    Fix all $\tilde{\boldsymbol{\alpha}}_j$, and compute $\boldsymbol{S}_{\boldsymbol{\alpha}_j}$ defined in (5.15) for $j = 1, \ldots, p$.
    for inner iteration until convergence do
        for each $j \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, q\}$ do
            Fix $\hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_{j'}}$ for all $j' \neq j$ and all $\hat{\boldsymbol{h}}_k$, and compute $\hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j}$ defined in (5.17).
            Fix $\hat{\boldsymbol{h}}_{k'}$ for all $k' \neq k$ and all $\hat{\boldsymbol{g}}_j$, and compute $\hat{\boldsymbol{h}}_k$ defined in (5.18).
        end for
    end for
    (Step 2: $\tilde{\boldsymbol{\alpha}}_j$ updates via GD)
    Fix all $\hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j}$ and all $\hat{\boldsymbol{h}}_k$, and minimize (5.19) over $\tilde{\boldsymbol{\alpha}}_j$, for $j = 1, \ldots, p$.
end for
```

In Algorithm 4, if the shrinkage factor c_j defined in (5.17) is 0, then the jth functional predictor is absent from the model, and the corresponding value of $\tilde{\boldsymbol{\alpha}}_j$ will not be updated. For Step 1, the smoothing matrix $\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j}$ only needs to be computed once, therefore the link function updates can be performed efficiently ([Fan *et al.*, 2014]). For Step 2, the following sample counterpart of (5.14), the residual sum of squares

$$\|\boldsymbol{Y} - \sum_{j=1}^{p} \hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j} - \sum_{k=1}^{q} \hat{\boldsymbol{h}}_k\|^2 = \|\sum_{j=1}^{p} \left(\boldsymbol{R}_j - \hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j}\right)\|^2, \tag{5.19}$$

where $\boldsymbol{R}_j$ is defined in (5.17) and the link estimates $\hat{g}_{j,T}$ in $\hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j}$ fixed, can be minimized over $\tilde{\boldsymbol{\alpha}}_j$, subject to $\|\tilde{\boldsymbol{\alpha}}_j\| = 1$ (with its first element constrained to be positive, i.e., $\tilde{\alpha}_{j1} > 0$, for model identifiability) for each $j = 1, \ldots, p$. In implementation, we can approximate $\hat{g}_{j,T}$ by a cubic polynomial, and perform an IWLS to approximately minimize (5.19), and rescale $\tilde{\boldsymbol{\alpha}}_j$ to have unit L^2 norm with $\tilde{\alpha}_{j1} > 0$.

5.2.3.3 Some intuition

In this subsection, we briefly describe some intuition of optimizing (5.19) over $\tilde{\boldsymbol{\alpha}}_j$. In optimizing (5.19), we can optimize the jth component separately, for each $j = 1, \ldots, p$. Hence, we can minimize over $\tilde{\boldsymbol{\alpha}}_j$, $\|\boldsymbol{R}_j - \hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j}\|^2 = \|\boldsymbol{R}_j - c_j \boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2 = \|\boldsymbol{R}_j\|^2 - \|\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2 \left(2 - c_j^2\right)$, where c_j and $\boldsymbol{R}_j$ are defined in (5.17). Note, minimizing $\|\boldsymbol{R}_j - \hat{\boldsymbol{g}}_{\tilde{\boldsymbol{\alpha}}_j}\|^2$ over $\tilde{\boldsymbol{\alpha}}_j$ is equivalent to maximizing $\|\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2 = \|\boldsymbol{S}^{(**)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j - \boldsymbol{S}^{(*)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2$ over $\tilde{\boldsymbol{\alpha}}_j$. Next, we consider a geometric intuition of maximizing $\|\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2$ over $\tilde{\boldsymbol{\alpha}}_j$. Similar to optimizing the SIMML (3.6) of Chapter 3, this is analogous to maximizing $\|\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}} \boldsymbol{1}_n\|^2$ in Figure 3.1.

Analogous to the orthogonal sum of squares decomposition in (3.34) which was obtained from centering by the intercept $\bar{\boldsymbol{Y}} \boldsymbol{1}_n$, in our case, the observed ($\boldsymbol{R}_j$) and the fitted ($\boldsymbol{S}^{(**)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j$) are centered by the vector $\boldsymbol{S}^{(*)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j$, that gives

$$\|\boldsymbol{R}_j - \boldsymbol{S}^{(*)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2 = \|\boldsymbol{R}_j - \boldsymbol{S}^{(**)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2 + \|\boldsymbol{S}^{(**)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j - \boldsymbol{S}^{(*)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2. \tag{5.20}$$

Here, the second component $\|\boldsymbol{S}^{(**)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j - \boldsymbol{S}^{(*)}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2 = \|\boldsymbol{S}_{\tilde{\boldsymbol{\alpha}}_j} \boldsymbol{R}_j\|^2$ quantifies the intensity of how much the effect of $\tilde{\boldsymbol{\alpha}}_j^\top \boldsymbol{\xi}_{ij}$ (which approximates $\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_{ij} \rangle$) gets modified by T, i.e., the magnitude of treatment effect modification, which is to be maximized over $\tilde{\boldsymbol{\alpha}}_j$. Equivalently, (5.19) is minimized over $\tilde{\boldsymbol{\alpha}}_j$, $j = 1, \ldots, p$. Note, for the simple linear regression on T with no regularization, the counterpart of (5.11) is (3.33).

5.2.3.4 Choosing the regularization parameters

The sparsity parameter λ in Section 5.2.3.2 can be chosen to minimize an estimate of the prediction error. We can define the total effective degrees of freedom ([Ravikumar *et al.*, 2009]) of FAMML as

$$\text{d.f.}(\lambda) = \sum_{j=1}^{p} s_j I(\|\hat{\boldsymbol{g}}^{(\lambda)}_{\boldsymbol{\alpha}_j}\| \neq 0) + \sum_{k=1}^{q} r_k I(\|\hat{\boldsymbol{h}}^{(\lambda)}_{k}\| \neq 0)$$

where $s_j = \text{trace}(\boldsymbol{S}_{\boldsymbol{\alpha}_j})$ is the effective degrees of freedom for the smoother on the jth functional predictor; $r_k = \text{trace}(\boldsymbol{H}_k)$ is that for the smoother on the kth scalar predictor; $\hat{\boldsymbol{g}}^{(\lambda)}_{\boldsymbol{\alpha}_j}$ and $\hat{\boldsymbol{h}}^{(\lambda)}_{k}$ are given in (5.17) and (5.18), respectively, evaluated at a particular given $\lambda \geq 0$. Then, λ can be chosen to minimize the following C_p statistic

$$C_p(\lambda) = \|\boldsymbol{Y} - \sum_{j=1}^{p} \hat{\boldsymbol{g}}^{(\lambda)}_{\tilde{\boldsymbol{\alpha}}_j} - \sum_{k=1}^{q} \hat{\boldsymbol{h}}^{(\lambda)}_{k}\|^2 + 2\hat{\sigma}^2 \text{d.f.}(\lambda),$$

where $\hat{\sigma}^2$ is an estimate of the error variance, based on regression on the whole set of predictors. In our empirical studies, we took $\hat{\sigma}^2 = \|\boldsymbol{Y} - \sum_{j=1}^{p} \hat{\boldsymbol{g}}^{(\lambda_{\min})}_{\tilde{\boldsymbol{\alpha}}_j} - \sum_{k=1}^{q} \hat{\boldsymbol{h}}^{(\lambda_{\min})}_{k}\|^2/n$, in which the smallest grid value $\lambda_{\min} \geq 0$ was taken at some small fraction of the largest grid value $\lambda_{\max}$, say, at $\lambda_{\min} = \lambda_{\max}/100$. $\lambda_{\max}$ was derived from data as the smallest regularization parameter giving the sparsest model (i.e., all the functions are identically zeros). Other methods for choosing the tuning parameter, for example, minimizing the AIC, the BIC, or cross-validated predictor errors, can also be used.

5.3 Simulation illustration

5.3.1 Treatment effect modifier selection performance

In this section, we will report the treatment effect modifier selection performance. The complexity of the model can be summarized in terms of the size of the index set for the link functions ($g_{j,T}$'s) that are not identically zero, which can be either correctly or incorrectly estimated as nonzeros. The following model was used for generating the data

$$Y_T = \sum_{j=1}^{p} \Big(\underbrace{\mu_j(\boldsymbol{x}_j)}_{\text{main}} + \underbrace{g_{j,T}\big(\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle\big)}_{\text{interaction}} \Big) + \epsilon, \quad T \in \{-1, 1\}, \tag{5.21}$$

with $p \in \{50, 100\}$.

We used a 4-dimensional Fourier basis $\boldsymbol{b}(s) = (1, \sqrt{2}\sin(\pi s), \sqrt{2}\sin(2\pi s), \sqrt{2}\sin(3\pi s))^\top$ to generate p functional predictors, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip}$

$$\boldsymbol{x}_{ij}(s_k) = \boldsymbol{b}(s_k)^\top \boldsymbol{\theta}_{ij} + w_{ijk}, \quad w_{ijk} \sim \mathcal{N}(0, \sigma_x^2), \quad \boldsymbol{\theta}_{ij} \sim \mathcal{N}(0, \boldsymbol{I}_4), \quad i = 1, \ldots, n,$$

where the basis coefficients, $\boldsymbol{\theta}_{ij}$, and the random noise term, w_{ijk}, were all sampled independently from each other, and the predictors $\boldsymbol{x}_{ij}$ were observed at 200 equally spaced points, $0 = s_1, s_2, \ldots, s_{200} = 1$. For the first 2 predictors, the link functions $g_{j,T}$ were set at

$$\begin{aligned} g_{1,T}(u) &= 0.5uT, \quad T \in \{-1, 1\} \\ g_{2,T}(u) &= \sqrt{2}\cos(u)T, \quad T \in \{-1, 1\} \end{aligned} \tag{5.22}$$

and $g_{j,T}(u) = 0$, for all remaining $j = 3, \ldots, p$. Therefore, only the first 2 predictors were relevant treatment effect modifiers. For simplicity, we set the projection functions $\boldsymbol{\alpha}_j(s) = \frac{1}{2}\left[1, 1, 1, 1\right]^\top \boldsymbol{b}(s)$, for all $j = 1, \ldots, p$. Note $\|\boldsymbol{\alpha}_j\| = 1$. For the main effect, we took $\mu_j(\boldsymbol{x}_j) = \langle \boldsymbol{\mu}_j, \boldsymbol{x}_j \rangle$, in which we set

$$\boldsymbol{\mu}_j(s) = \boldsymbol{b}(s)^\top \boldsymbol{\eta}_j, \quad j = 1, \ldots, 4, \tag{5.23}$$

for the first 4 predictors, and $\boldsymbol{\mu}_j(s) = 0$, for all remaining $j = 5, \ldots, p$. Therefore, only the first 4 predictors had nonzero main effects. In (5.23), the vectors $\boldsymbol{\eta}_j \in \mathbb{R}^4$ were first generated independently from a multivariate normal distribution, and then rescaled to have unit L^2 norm, $\|\boldsymbol{\eta}_j\| = 1$. In (5.21), the error term ϵ was generated from $\mathcal{N}(0, 0.5^2)$.

Under the setting (5.22) and (5.23), the contribution to the variance of the outcome from the main effect component was about 3 times larger than that from the interaction effect component. There were 2 true treatment effect modifiers, and the other $p - 2$ predictors were "noise" predictors.

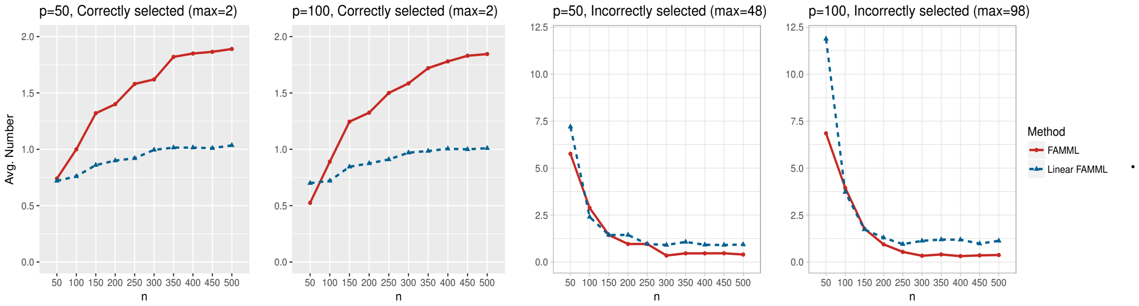

Figure 5.2: The average number of treatment effect modifiers "correctly selected" (in the panels with gray background), and "incorrectly selected" (in the panels with white background), respectively, as the training sample size n varies from 50 to 500, for each $p \in \{50, 100\}$. Two methods were compared: 1) the proposed semiparametric FAMML, and 2) a FAMML with the links $g_{j,T}$ restricted to be linear (Linear-FAMML).

Figure 5.2 summarizes the results of the treatment effect modifier selection performance, comparing the proposed semiparametric FAMML and a FAMML with the links $g_{j,T}$ restricted to be linear (Linear-FAMML). In Figure 5.2, as n increased from $n = 50$ to $n = 500$, the proposed FAMML (the red solid curves) correctly recovered the index set of the true 2 treatment effect modifiers that had nonzero link functions, with probabilities tending to 1. In both cases of $p = 50$ and $p = 100$, the average number of correctly selected treatment effect modifier converged to 2, while that of incorrectly selected treatment effect modifier converged to 0, as n increased. On the other hand, the Linear-FAMML (the blue dotted curves) failed to recover the true 2 treatment effect modifiers. For the Linear-FAMML, the average number of correctly selected treatment effect modifier tended to 1 instead of 2, selecting only the 1st predictor with the link $g_{1,T}$ that had a linear interaction effect with the treatment on the outcome, while failing to select the 2nd predictor that had a nonlinear interaction effect.

5.3.2 ITR performance

In this section, we will assess the performance of the proposed method with respect to selection of optimal ITRs on simulated data in various settings.

5.3.2.1 Setting

The vector of baseline scalar covariates, $Z = (z_1, \ldots, z_q)^\top$, was generated from a multivariate normal distribution with each component having mean 0 and variance 1. Correlation between the components was given by $\mathrm{corr}(z_j, z_k) = 0.5^{|j-k|}$. We wrote $\boldsymbol{z} = (1, Z)^\top \in \mathbb{R}^{q+1}$, including the intercept. A set of p functional baseline predictors, $X = \{\boldsymbol{x}_1(s), \ldots, \boldsymbol{x}_p(s)\}$, was generated to be similar to the EEG curves observed in the motivating data set discussed in Section 5.4, in which the domain for each function was set at $[0, 1]$. In this illustration, we considered relatively low dimensional scenarios with $(q = 5, p = 3)$, and made comparisons with other functional regression approaches of constructing ITRs.

Responses were generated, for 1) "*nonlinear*" treatment-by-predictors interactions:

$$\begin{aligned} Y = &\Big\{0.5\cos(\boldsymbol{\beta}_0^\top \boldsymbol{z}) + 0.5\cos(\langle \boldsymbol{\beta}_1, \boldsymbol{x}_1\rangle) + 0.5\cos(\langle \boldsymbol{\beta}_2, \boldsymbol{x}_2\rangle)\Big\}\,\delta \\ &+ \Big\{-0.25\cos(\boldsymbol{\gamma}^\top \boldsymbol{z}) - 0.25\cos(\langle \boldsymbol{\alpha}_1, \boldsymbol{x}_1\rangle) + 0.5(\langle \boldsymbol{\alpha}_3, \boldsymbol{x}_3\rangle)^2\Big\}\,T + \epsilon, \end{aligned} \tag{5.24}$$

and for 2) "*linear*" treatment-by-predictors interactions:

$$\begin{aligned} Y = &\Big\{0.5\cos(\boldsymbol{\beta}_0^\top \boldsymbol{z}) + 0.5\cos(\langle \boldsymbol{\beta}_1, \boldsymbol{x}_1\rangle) + 0.5\cos(\langle \boldsymbol{\beta}_2, \boldsymbol{x}_2\rangle)\Big\}\,\delta \\ &+ \Big\{-0.25\boldsymbol{\gamma}^\top \boldsymbol{z} - 0.25\langle \boldsymbol{\alpha}_1, \boldsymbol{x}_1\rangle + 0.5\langle \boldsymbol{\alpha}_3, \boldsymbol{x}_3\rangle\Big\}\,T + \epsilon, \end{aligned} \tag{5.25}$$

where the treatment variable $T \in \{-1, 1\}$ was generated independently of the other covariates, such that $P(T = 1) = P(T = -1) = 1/2$. In (5.24) and (5.25), the first term on the right hand side corresponded to the main effect of (Z, X), and the second term corresponded to the T-by-(Z, X) interaction. The parameter $\delta \in \{0, 1, 2\}$ controlled the proportion of the variance of the response Y attributable to the main effect and the interaction: $\delta = 0$ corresponded to the zero contribution of the main effect; $\delta = 1$ to a moderate main effect contribution (the variance of Y attributable to the main effect and the interaction was about the same); $\delta = 2$ to a large main effect contribution (the variance of Y attributable to the main effect was approximately 4 times larger than that from the interaction). If δ is large, the model has a large main effect, and it is more difficult to estimate the interaction effect. The error term, ϵ, followed $\mathcal{N}(0, \sigma_\epsilon^2)$, where σ_ϵ^2 was chosen such that R^2 of the model when $\delta = 0$ was about 0.85.

Two simulation settings were considered, denoted by set "A" and "B", respectively. For the simulation set "A", the functional coefficients for the interaction effect component were set at $\boldsymbol{\alpha}_1(s) = \frac{1}{6\sqrt{2\pi}}\left\{-e^{-\frac{8}{9}(12s-5)^2} + e^{-\frac{8}{9}(12s-7)^2}\right\}$, $\boldsymbol{\alpha}_2(s) = 0$, and $\boldsymbol{\alpha}_3(s) = \frac{1}{6\sqrt{2\pi}}\left\{e^{-\frac{8}{9}(12s-5)^2} - e^{-\frac{8}{9}(12s-7)^2}\right\}$, as described in the left two panels of Figure 5.3. This was the setting where the true functional contrast coefficients, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$, were such that, reducing the 1-D predictors to scalars via simple averaging resulted in the loss of important features that were needed to correctly estimate the ITRs. We set the main effect functional coefficients to be $\boldsymbol{\beta}_1(s) = 0.055\sin(2\pi s)$, $\boldsymbol{\beta}_2(s) = -0.055\sin(2\pi s)$, and $\boldsymbol{\beta}_3(s) = 0$. For the coefficients corresponding to the scalar predictors, we set $\boldsymbol{\beta}_0 = [5, 0.5, 0, -0.5, 0, 0.5]^\top$ for the main effect, and $\boldsymbol{\gamma} = [-0.25, 0.5, 0, 0, 0, -0.5]^\top$ for the interaction.

For the simulation set "B", we took the contrast functional coefficients to be $\boldsymbol{\alpha}_1(s) = \frac{25}{3}s^2e^{-10s}$, $\boldsymbol{\alpha}_2(s) = 0$, and $\boldsymbol{\alpha}_3(s) = -\frac{25}{3}(1-s)^2e^{-10(1-s)}$, as described in the right two panels of Figure 5.3. This was the setting where the true functional contrast coefficients were such that, reducing the 1-D predictors to scalars via simple averaging did not affect much in making correct ITRs. For the main effect functional coefficients, we set $\boldsymbol{\beta}_1(s) = 0.05\sin(2\pi s)$, $\boldsymbol{\beta}_2(s) = -0.05\sin(2\pi s)$, and $\boldsymbol{\beta}_3(s) = 0$. For the coefficients corresponding to the scalar predictors, we set $\boldsymbol{\beta}_0 = [5, 1, 0, -1, 0, 1]^\top$ for the main effect, and $\boldsymbol{\gamma} = [-0.5, 1, 0, 0, 0, -1]^\top$ for the contrast.

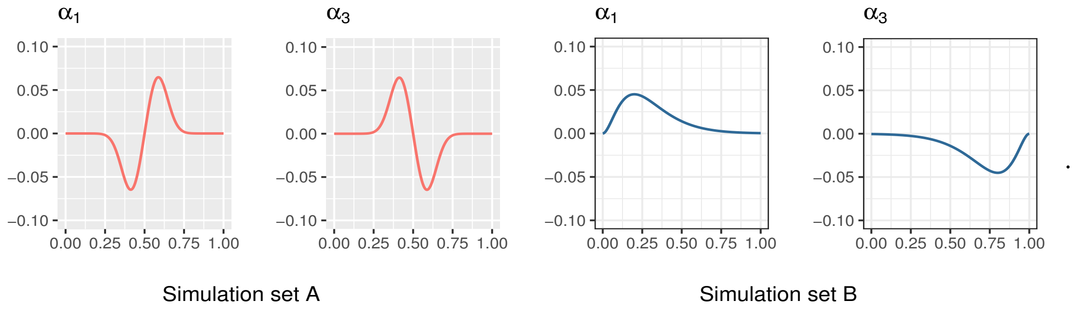

Figure 5.3: The true functional contrast coefficients (α_1, α_3) for the simulation set "A" on the left two panels, and for the simulation set "B" on the right two panels, respectively.

We considered a class of ITRs of the form based on X and Z

$$\mathcal{D}(X, Z) = \underset{t\in\{-1,1\}}{\arg\max}\ \mathbb{E}[Y \mid X, Z, T = t], \tag{5.26}$$

where $\mathbb{E}[Y \mid X, Z, T]$ was the conditional expectation given (X, Z, T). In this example, we considered the following three approaches for estimating $\mathbb{E}[Y \mid X, Z, T]$ from the training set.

FAMML Fit the proposed FAMML (5.5) under the constraint (5.6), by minimizing the criterion (5.7).

F-MCA Estimate the functional modified covariate model with efficiency augmentation (F-MCA) ([Ciarleglio *et al.*, 2015a]). Here, the efficiency augmentation was performed by fitting a functional linear regression. Note, the MCA ([Tian *et al.*, 2014]) was extended by [Ciarleglio *et al.*, 2015b] to incorporate both functional and scalar baseline predictors under efficiency augmentation, which we employed here for comparison.

Separate FLR Fit a functional linear regression (FLR) model via a penalized FLR ([Goldsmith *et al.*, 2011]) for each treatment group separately. Then the corresponding ITR was obtained from the $K(=2)$ number of separately fitted FLR models by (5.26).

The evaluation measure was the proportion of correct decisions (PCD) made from each method. PCD can be calculated for each of the ITRs estimated from the methods considered, since we know the true models (5.24) and (5.25). This measure was calculated from an independent testing set of size $n = 10000$.

5.3.2.2 Simulation set "A" results

This was the setting where misspecification of the nonlinear contrast component in model (5.24) as a linear model (i.e., the cases for F-MCA and Separate FLR models) has a detrimental effect in developing ITRs. The true functional contrast coefficients were set in a way that reducing the functional predictors to scalars via a naïve averaging resulted in the loss of salient features that were needed to correctly estimate the ITRs. In this setting, the proposed FAMML performed significantly better than F-MCA and Separate FLR. As the sample size increased from $n = 100$ to $n = 200$ and to $n = 400$, the FAMML performed contrastingly better. F-MCA performed generally better than Separate FLR, since F-MCA was robust to misspecification of the main effect component. The proposed method, as an extension of the F-MCA, was robust to the main effect as well as was flexible to model the nonlinear contrast, giving a superior performance level.

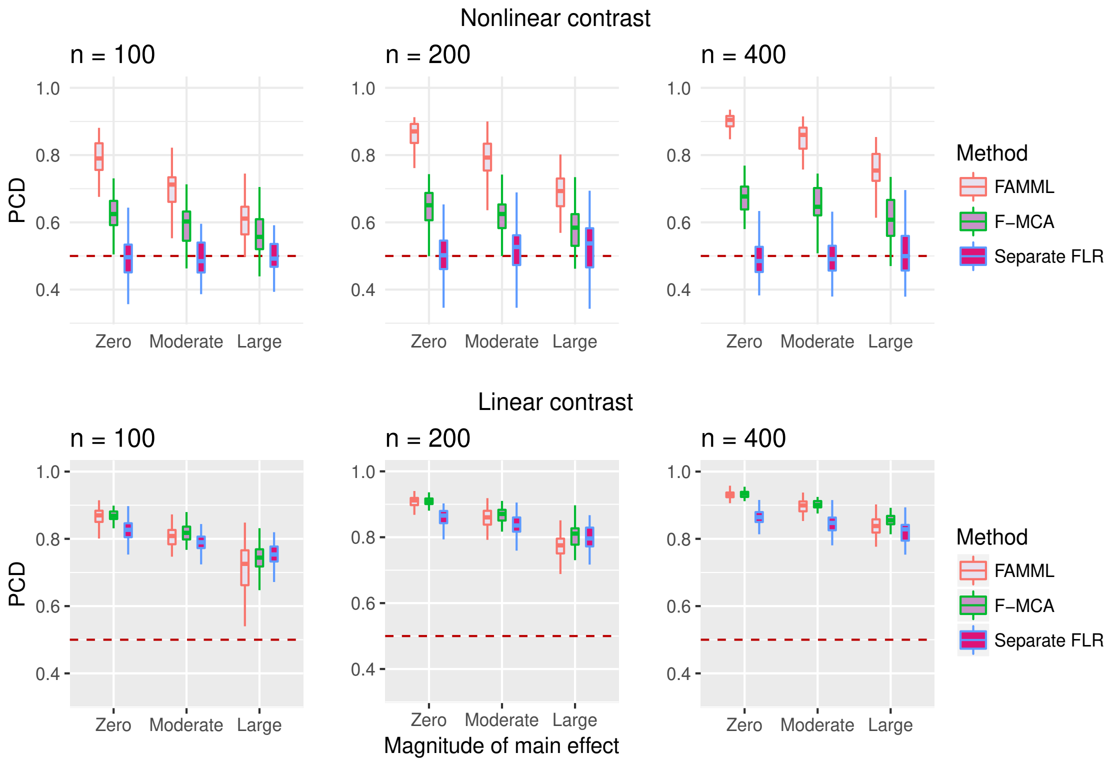

Figure 5.4: Boxplots of the PCDs of the ITRs for the simulation set "A", obtained from 100 replications, estimated from the following three methods: (1) FAMML: the proposed semi-parametric method satisfying the orthogonality constraint (5.6); (2) F-MCA: the modified covariate approach with efficiency augmentation; (3) Separate FLR: two separate FLR models estimated separately for each of the two treatment groups. For each case, $n \in \{100, 200, 400\}$ and $\delta \in \{0, 1, 2\}$ were considered.

5.3.2.3 Simulation set "B" results

This was the setting where a simple naïve averaging of the functional predictors provided similar ITRs as with those of a functional regression. Hence, this was the setting where misspecification of the nonlinear contrast component in (5.24) as a linear model (i.e., the cases for F-MCA and Separate FLR models) is not so detrimental in terms of developing ITRs, compared to the simulation set "A". Even in this case, the proposed approach give some advantage over the other two methods especially as the sample size became larger, for the nonlinear contrast. Additionally, Figure 5.5 shows that the performance of the parametric linear models (i.e., F-MCA and Separate FLR) did not improve considerably when the sample size increased, since these two methods are limited by a misspecified linear model for the true interaction effect.

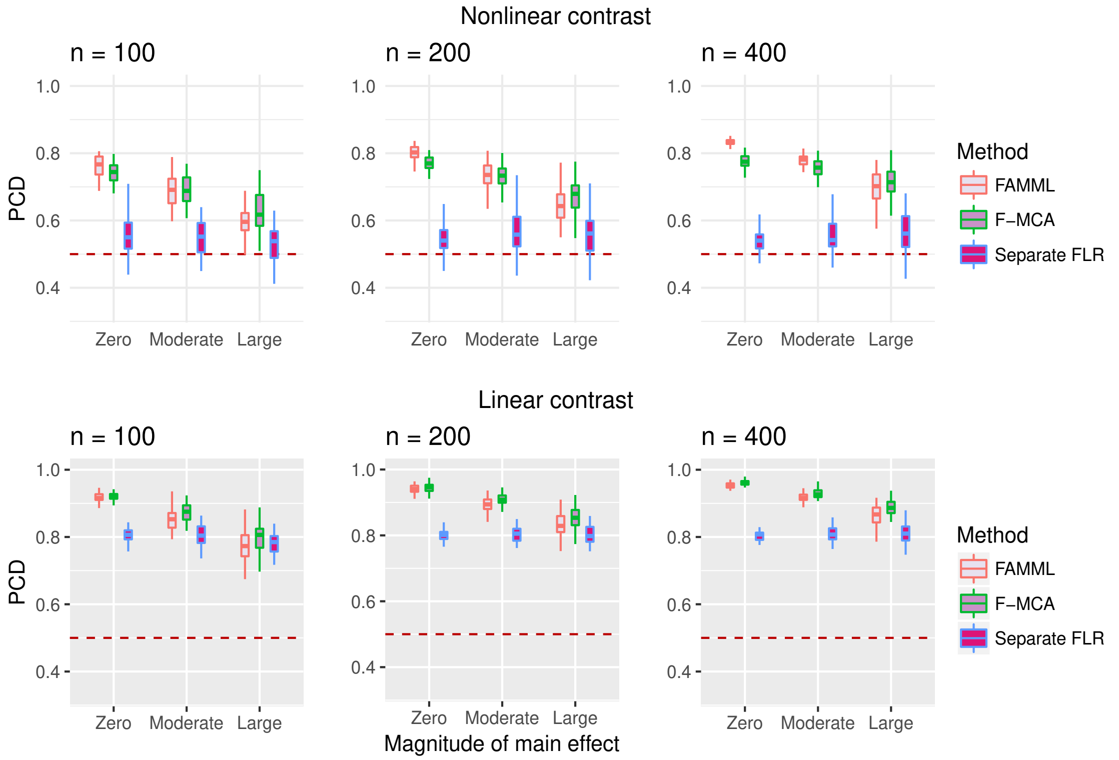

Figure 5.5: Boxplots of the PCDs of the ITRs for the simulation set "B", obtained from 100 replications, estimated from the following three methods: (1) FAMML: the proposed semi-parametric method satisfying the orthogonality constraint (5.6); (2) F-MCA: the modified covariate approach with efficiency augmentation; (3) Separate FLR: two separate FLR models estimated separately for each of the two treatment groups. For each case, $n \in \{100, 200, 400\}$ and $\delta \in \{0, 1, 2\}$ were considered.

5.4 Application

In this section, we apply the proposed FAMML to a dataset from a study comparing an antidepressant and placebo for treating major depressive disorder (MDD). The main objective of our investigation in this study was to use baseline functional/scalar predictors to guide treatment decisions when a patient presents for treatment. The study collected baseline scalar and functional data, including EEG amplitude spectra curves, prior to treatment assignment. Then the study participant was randomized to either placebo ($t = 1$) or an antidepressant (sertraline) ($t = 2$). Subjects were monitored via depression assessments at 1,2,3,4,6, and 8 weeks after initiation of treatment. The primary endpoint of interest was the Hamilton Rating Scale for Depression (HRSD) score at week 8. The outcome Y was taken to be the improvement in symptoms severity from baseline to week 8 taken as the difference (week 0 HRSD score - week 8 HRSD score). (Lower scores on the HRSD correspond to lower depression severity.)

There were $n = 156$ subjects in the study. We considered $p = 19$ baseline functional predictors, a subset of EEG channels from a 72-EEG montage. Specifically, the functional data of interest consisted of the curves giving the current source density (CSD) amplitude spectrum values over a frequency range of 3 to 16 Hz, when the participants' eyes were closed. This frequency range was scaled to $[0, 1]$, hence each of the functional predictors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{15}\}$ was defined on the interval $[0, 1]$. The locations for these 15 electrodes are described in Figure 5.6. In addition, we considered a set of $q = 4$ baseline scalar predictors, consisting of the HRSD scores at baseline (i.e., the week 0)(z_1), sex (1 for female, 0 otherwise)(z_2), age (z_3), and the baseline HRSD-by-age interaction (z_4). In this dataset, 46% of the subjects were randomized to the sertraline. The average outcomes Y for the sertraline group and placebo groups were 7.75, and 6.29 respectively. The mean age was 38.3, the mean baseline HSRD score 18.78, and 64% of the subjects were female.

The FAMML (5.5) was estimated and the method simultaneously selected the 5 functional predictors with non-zero link functions associated with the electrodes indicated by the red dashed circles in Figure 5.6: the electrodes "FP1" ($\boldsymbol{x}_1$), "C3" ($\boldsymbol{x}_4$), "O1" ($\boldsymbol{x}_6$), "O2" ($\boldsymbol{x}_{12}$), and "PZ" ($\boldsymbol{x}_{15}$). In the top panels of Figure 5.7, we display the CSD curves corresponding to these electrodes observed from the 156 subjects. In the bottom panels of Figure 5.7, we display the estimated projection functions $\boldsymbol{\alpha}_j$, $j \in \{1, 4, 6, 12, 15\}$, for the selected 5 functional predictors.

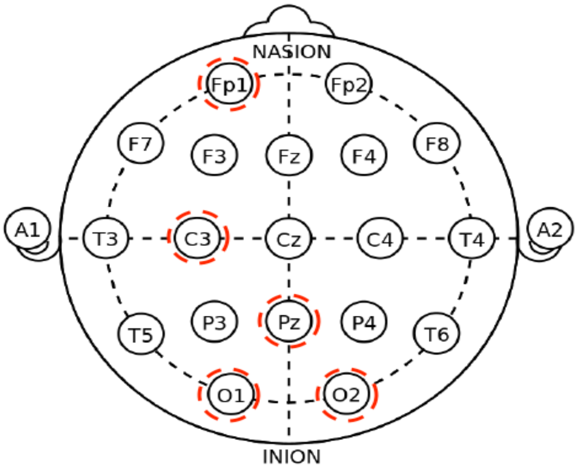

Figure 5.6: The locations for the 19 electrode channels. “A1” and “A2” were not used. Those marked in red circles are the selected electrodes from the fitted FAMML (5.5): “FP1”, “C3”, “O1”, “O2”, and “PZ”.

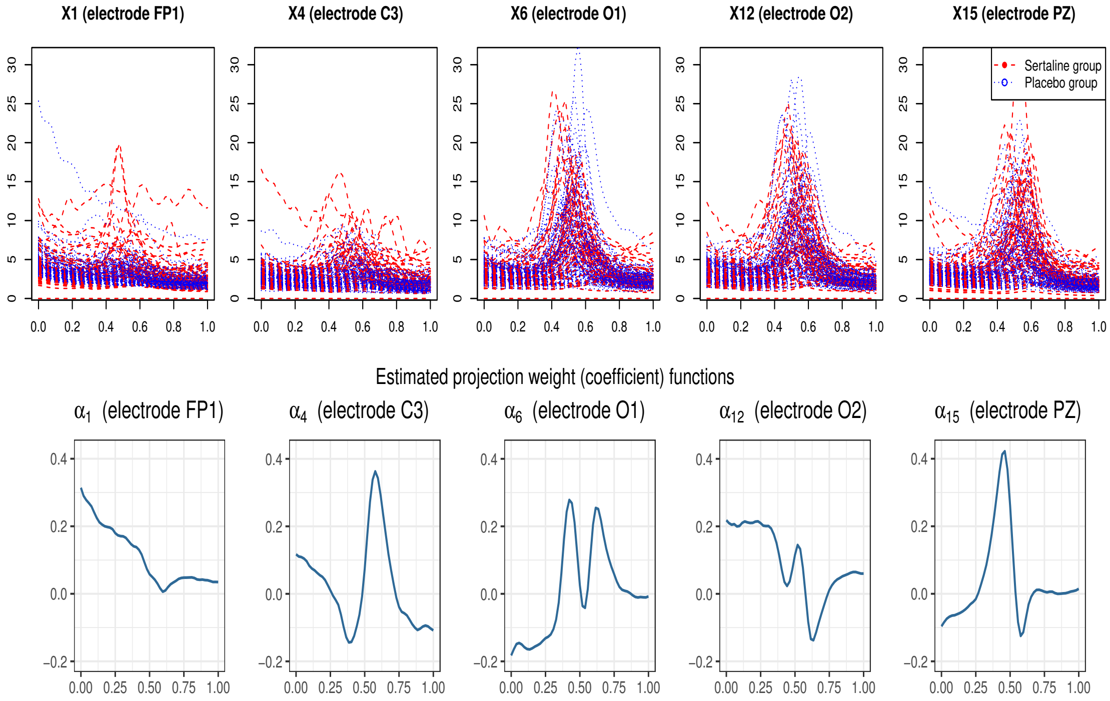

Figure 5.7: Top panels: observed CSD curves from the 5 channels, FP1, C3, O1, O2, and PZ, for the active drug group (red dashed curves) and for the placebo group (blue dotted curves), over a frequency range of 3 to 16 Hz, when the participants' eyes are closed. Bottom panels: estimated projection functions ($\boldsymbol{\alpha}_j$'s) for the selected 5 functional predictors (from left to right: FP1, C3, O1, O2, and PZ).

The estimated projection directions $\boldsymbol{\alpha}_j$, $\|\boldsymbol{\alpha}_j\| = 1$, produce data-driven scalar variables $\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle$. Hence, each of the EEG amplitude spectra was reduced to an weighted average, weighted by the function $\boldsymbol{\alpha}_j$. These averages were investigated as potential scalar effect modifiers of treatment effect, that gave 5 sets of treatment-specific links, $g_{j,T}(\cdot)$, $j \in \{1, 4, 6, 12, 15\}$, $T \in \{1, 2\}$, that were not identically zero. These processes were performed simultaneously by fitting the FAMML (5.5). In Figure 5.8 and 5.9, the estimated treatment-specific link functions of the selected functional and the selected scalar predictors, respectively, are displayed, on the corresponding partial residual plots. These plots display some nonlinear contrast patterns captured by the treatment-specific link functions, which would have been difficult to capture if the link functions were restricted to be linear.

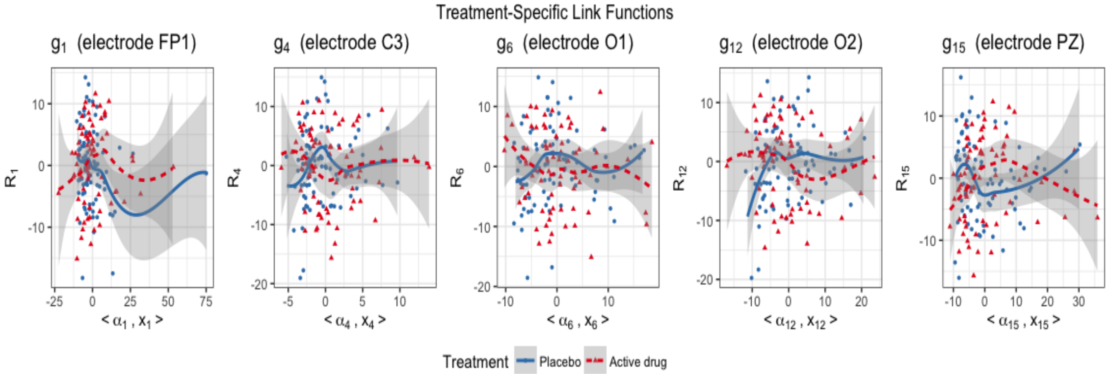

Figure 5.8: The scatter plots of the jth partial residual vs. the jth projection $\langle \boldsymbol{\alpha}_j, \boldsymbol{x}_j \rangle$, $j \in \{1, 4, 6, 12, 15\}$, corresponding to the channels, FP1, C3, O1, O2, and PZ. Overlaid are the estimated treatment-specific link functions $(g_{j,1}, g_{j,2})$ for the placebo group in the dotted blue, and the active drug group in the dotted red curves; the associated 95% confidence bands were constructed conditioning on the projection functions, $\boldsymbol{\alpha}_j$'s.

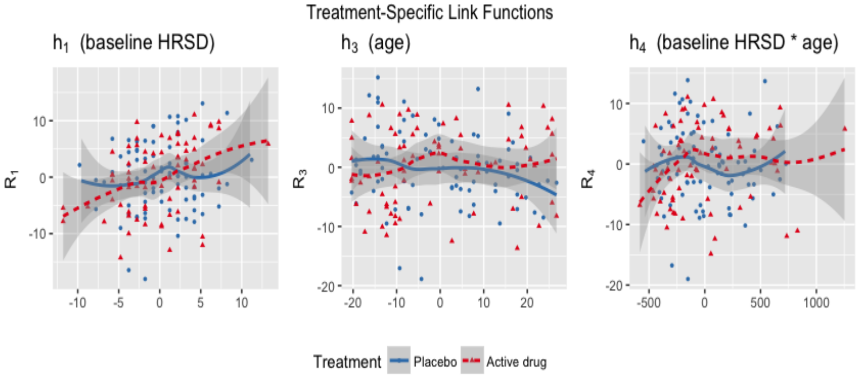

Figure 5.9: The scatter plots of the kth partial residual vs. the kth scalar predictors, $k \in \{1, 3, 4\}$, the baseline HRSD, age, and the baseline HRSD-by-age interaction, respectively, where all variables are centered at 0. The variable "sex" (z_2) was not selected by the model. Overlaid are the estimated treatment-specific functions $(h_{k,1}, h_{k,2})$ for the placebo group in the solid blue, and the active drug group in the dotted red curves; the associated 95% confidence bands of the regression curves were plotted.

To evaluate the performance of the ITRs obtained from the proposed method, we randomly split the RCT data into a training set and a testing set using a ratio of 10 to 1, replicated 300 times, each time estimating an ITR $\mathcal{D}$ based on the training set, then estimating its value ([Murphy, 2005]) $\mathrm{V}(\mathcal{D}) = \mathbb{E}\left[\mathbb{E}[Y \mid X, Z, T = \mathcal{D}(X, Z)]\right]$, using an inverse probability weighted estimator ([Murphy, 2005])

$$\hat{V}(\mathcal{D}) = \sum_{i=1}^{\tilde{n}} Y_i I_{T_i = \mathcal{D}(X_i, Z_i)} / \sum_{i=1}^{\tilde{n}} I_{T_i = \mathcal{D}(X_i, Z_i)}, \tag{5.27}$$

based on the testing set $\{(Y_i, T_i, X_i, Z_i)_{i=1}^{\tilde{n}}\}$, where $\tilde{n}$ denoted the sample size of the testing set. We compared four approaches for constructing an ITR (5.26): 1) using the estimated FAMML, 2) using the estimated Linear-FAMML, a FAMML under the linear link restriction, 3) giving everyone the placebo ("All PLACEBO"), and 4) giving everyone the active drug ("All DRUG"). The resulting boxplots are illustrated in Figure 5.10.

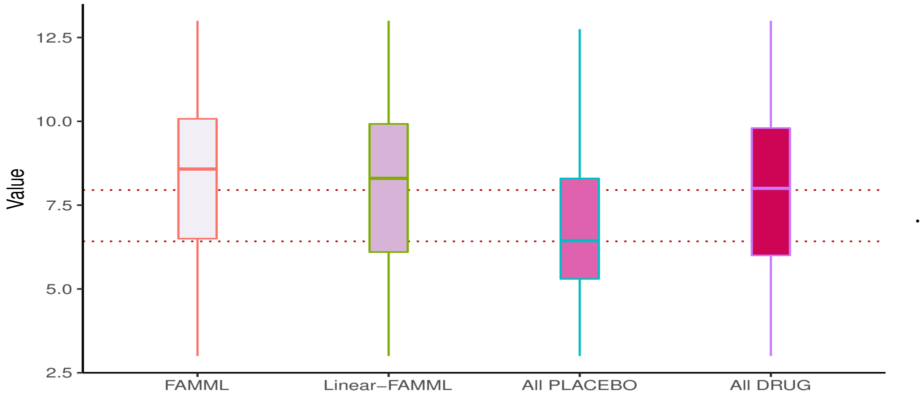

Figure 5.10: Boxplots of the values of the ITRs, obtained from 300 randomly split testing sets, estimated from the four approaches: FAMML, Linear-FAMML, giving everyone the placebo, and giving everyone the active drug. Higher values are preferred.

The results in Figure 5.10 demonstrate that the proposed FAMML approach for constructing ITRs may be beneficial to overall patient population. The FAMML, in terms of the averaged estimated value, outperformed the simple naïve approaches of giving everyone either the active drug or the placebo. The FAMML also shows a modest improvement over the Linear-FAMML which could only model the linear T-by-(X, Z) interaction effects.

5.5 Discussion

In this chapter, we used a L^1 penalized and constrained least squares optimization criterion to optimize the proposed FAMML with respect to the interaction effect, and to simultaneously select treatment effect modifiers in situations involving multiple functional and scalar predictors. We used an efficient coordinate descent algorithm to fit general nonlinear additive relationships between the treatment effect modifiers and the response. The proposed sparse FAMML provides a natural semi-parametric framework for modeling treatment-by-multiple predictors interactions, particularly when we use a large number of functional covariates as predictors for estimating the treatment effect modification.

Chapter 6

Conclusion

The theme of this dissertation was on the problem of effectively estimating interactions between a treatment variable and a set of baseline predictors in their effect on the outcome, without restriction to a linear relationship. Single-index models generalize linear regression by replacing the linear predictor with a semiparametric component that has an unspecified link function. Due to their flexibility and interpretability of the coefficients, single-index models are becoming popular in many scientific field. In addition, additive models generalize linear regression by replacing each individual predictor with a nonparametric component defined by an unspecified univariate transformation of each individual predictor. Because of their flexibility, interpretability, and efficient computation with coordinate descent, additive models are becoming increasingly popular for analyzing data. In this dissertation, we developed novel approaches for modeling the treatment effect modifications by tailoring the single-index models and the additive models to have treatment-specific links, and estimate the interaction effect without the need to specify the main effect. The proposed approaches provide flexible extensions of the linear model for estimating the interactions between a discrete treatment variable and baseline covariates.

Future work in refining and developing the proposed approach will investigate an extension to a continuous treatment variable, an extension to longitudinal outcomes, and incorporation of hypothesis testing procedures for testing possibly nonlinear interactions based on the proposed models.

Bibliography

[Adragni and Cook, 2009] K. P. Adragni and D. R. Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society*, 367:4385–4405, 2009.

[Antoniadis *et al.*, 2004] A. Antoniadis, G. Gregoire, and I. McKeague. Bayesian estimation in single-index models. *Statistica Sinica*, 14:1147–1164, 2004.

[Brillinger, 1982] R. D. Brillinger. *A generalized linear model with "Gaussian" regressor variables In A Festschrift for Erich L. Lehman (Edited by P. J. Bickel, K. A. Doksum and J. L. Hodges).* Wadsworth, New York, 1982.

[Cai *et al.*, 2011] T. Cai, L. Tian, P. H. Wong, and L. J. Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, 2011.

[Ciarleglio *et al.*, 2015a] A. Ciarleglio, E. Petkova, R. T. Ogden, and T. Tarpey. Treatment decisions based on scalar and functional baseline covariates. *Biometrics*, 71:884–894, JUNE 2015.

[Ciarleglio *et al.*, 2015b] A. Ciarleglio, E. Petkova, R. T. Ogden, and Thaddeus Tarpey. Treatment decisions based on scalar and functional baseline covariatesecisions based on scalar and functional baseline covariates. *Biometrics*, 71(4):884–894, 2015.

[Cook, 2007] R. D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

[de Boor, 2001] C. de Boor. *A Practical Guide to Splines.* Springer-Verlag, New York, 2001.

[Deary *et al.*, 2011] I. J. Deary, D. Liewald, and J. Nissan. A free, easy-to-use, computer-based simple and four-choice reaction time programme: The deary-liewald reaction time task. *Behavioral Research Methods*, 43:258–268, 2011.

[DiCiccio and Efron, 1996] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11:189–228, 1996.

[Fan and Li, 2001] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, December 2001.

[Fan *et al.*, 2014] Y. Fan, N. Foutz, G. M. James, and W. Jank. Functional response additive model estimation with online virtual stock markets. *The Annals of Applied Statistics*, 8:2435–2460, 2014.

[Fan *et al.*, 2015] Y. Fan, G. M. James, and P. Radchanko. Functional additive regression. *The Annals of Statistics*, 43:2296–2325, 2015.

[Flanker and Eriksen, 1974] B. A. Flanker and C. W. Eriksen. Effects of noise letters upon identification of a target letter in a non-search task. *Perception and Psychophysics*, 16:143–149, 1974.

[Friedman and Stuetzle, 1981] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[Friedman *et al.*, 2007] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[Goldsmith *et al.*, 2011] J. Goldsmith, J. Bobb, C. Crainiceanu, B. Caffo, and D. Reich. Penalized functional regression. *Journal of computational and graphical statististics*, 20(4):830–851, 2011.

[Gunter *et al.*, 2011] L. Gunter, J. Zhu, and S. A. Murphy. Variable selection for qualitative interactions in presonalized medicine while controlling the family-wise error rate. *Journal of Biopharmaceutical Statistics*, 21:1063–1078, 2011.

[Hardle and Muller, 2012] W. Hardle and M. Muller. *Multivariate and Semiparametric Kernel Regression.* Wiley Series in Probability and Statistics, 2012.

[Hardle *et al.*, 1993] W. Hardle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *Annals of Statistics*, 21:157–178, 1993.

[Hastie and Tibshirani, 1999] T. Hastie and R. Tibshirani. *Generalized Additive Models.* Chapman & Hall Ltd., 1999.

[Herrera-Guzman *et al.*, 2009] I. Herrera-Guzman, E. Guidayol-Ferre, D. Herrera-Guzman, J. Guardia-Olmos, E. Hinojosa-Calvo, and J. E. Herrera-Abarca. Effects of selective serotonin reuptake and dual serotonergic–noradrenergic reuptake treatments on memory and mental processing speed in patients with major depressive disorder. *Psychiatric Research*, 43:855–863, 2009.

[Horowitz, 2009] J. L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics.* Springer, 2009.

[Kang *et al.*, 2014] C. Kang, H. Janes, and Y. Huang. Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70:696–707, 2014.

[Laber and Zhao, 2015] E. B. Laber and Y. Zhao. Tree-based methods for individualized treatment regimes. *Biometrika*, 102:501–514, 2015.

[Lian and Liang, 2016] H. Lian and H. Liang. Separation of linear and index covariates in partially linear single-index models. *Journal of Multivariate Analysis*, 143:56–70, January 2016.

[Lin and Kulasekera, 2007] W. Lin and K. B. Kulasekera. Uniqueness of a single index model. *Biometrika*, 94:496–501, 2007.

[Lin *et al.*, 2014] W. Lin, P. Shi, R. Feng, and H. Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

[Loonstra *et al.*, 2001] A. Loonstra, A. R. Tarlow, and A. H. Sellers. Cowat metanorms across age, education, and gender. *Applied Neuropsychology*, 8:161–166, 2001.

[Lu *et al.*, 2011] W. Lu, H. Zhang, and D. Zeng. Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22:493–504, 2011.

[Mackay, 2003] D. J. Mackay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[McKeague and Qian, 2014] I. McKeague and M. Qian. Estimation of treatment policies based on functional predictors. *Statistica Sinica*, 24:1461–1485, 2014.

[Murphy, 2003] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, May 2003.

[Murphy, 2005] S. A. Murphy. A generalization error for q-learning. *Journal of Machine Learning*, 6:1073–1097, 2005.

[Nelder and Wedderburn, 1972] J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135:370–384, 1972.

[Ong, 2014] P. Ong. Adaptive cuckoo search algorithm for unconstrained optimization. *The Scientific World Journal*, 2014:8, 2014.

[Park *et al.*, 2017] H. Park, E. Petkova, T. Tarpey, and R. T. Ogden. Single-index with multiple-links models. *Preprint*, 2017.

[Peng and Huang, 2011] H. Peng and T. Huang. Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141:1362–1379, 2011.

[Petkova *et al.*, 2016] E. Petkova, T. Tarpey, Z. Su, and R. T. Ogden. Generated effect modifiers in randomized clinical trials. *Biostatistics*, 18(1):105–118, July 2016.

[Powell *et al.*, 1989] J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57:1403–1430, 1989.

[Qian and Murphy, 2011] M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011.

[Radchanko, 2015] P. Radchanko. High dimensional single index model. *Journal of Multivariate Analysis*, 139:266–282, 2015.

[Ravikumar *et al.*, 2009] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of Royal Statistical Society: Series B*, 2009.

[Robins, 2004] J. Robins. *Optimal Structural Nested Models for Optimal Sequential Decisions.* Springer, New York, 2004.

[Royston and Sauerbrei, 2008] P. Royston and W. Sauerbrei. Interactions between treatment and continuous covariates: A step toward individualizing therapy. *Journal of Clinical Oncology*, 26(9):1397–99, 2008.

[Ruppert and Wand, 1994] D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370, 1994.

[Seber and Lee, 2012] G. Seber and A. Lee. *Linear Regression Analysis.* John Wiley & Sons, 2012.

[Song *et al.*, 2015] R. Song, M. Kosorok, D. Zeng, Y. Zhao, E. B. Laber, and M. Yuan. On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat*, 4:59–68, 2015.

[Stoker, 1986] T. M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54:1461–1481, 1986.

[Tian *et al.*, 2014] L. Tian, A. Alizadeh, A. Gentles, and R. Tibshrani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

[Tibshirani, 1996] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288, 1996.

[van't Veer and Bernards, 2008] L.J. van't Veer and R. Bernards. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452(7187):564–570, 2008.

[Wang and Wang, 2015] G. Wang and L. Wang. Spline estimation and variable selection for single-index prediction models with diverging number of index parameters. *Journal of Statistical Planning and Inference*, 162:1–19, 2015.

[Wang and Yang, 2007] L. Wang and L. Yang. Spline single-index prediction model. *Technical Report. https://arxiv.org/abs/0704.0302*, 2007.

[Wang and Yang, 2009] L. Wang and L. Yang. Spline estimation of single-index models. *Statistica Sinica*, 19:765–783, 2009.

[Wang and Yin, 2008] Q. Wang and X. Yin. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave,. *Computational Statistics and Data Analysis*, 52:4512–4512, 2008.

[Wu and Rolling, 2016] W. Wu and C. A. Rolling. Sufficient dimension reduction for treatment effect estimation. In *Joint Statistical Meeting (JSM)*, 2016.

[Xia and Li, 1999] Y. Xia and W. Li. On single index coefficient regression models. *Journal of the American Statistical Association*, 94(448):1275–1285, 1999.

[Xia *et al.*, 1999] Y. Xia, H. Tong, and W. K. Li. On extended partially linear single-index models. *Biometrika*, 86:831–842, 1999.

[Xia *et al.*, 2002] Y. Xia, H. Tong, W. Li, and L. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 64:363–410, 2002.

[Xia, 2008] Y. Xia. A multiple-index model and dimension reduction. *Journal of American Statistical Association*, 103(484):1631–1640, December 2008.

[Yang and Deb, 2009] X. S. Yang and S. Deb. Cuckoo search via levy flights. *Proc. of World Congress on Nature and Biologically Inspired Computing (NaBIC 2009)*, pages 210–214, 2009.

[Yang and Deb, 2014] X. S. Yang and S. Deb. Cuckoo search: recent advances and applications. *Neural Computing and Applications*, 24(1):169–174, 2014.

[Yuan and Lin, 2006] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[Yuan, 2011] M. Yuan. On the identifiabliity of additive index models. *Statistica Sinica*, 21:1901–1911, 2011.

[Zhang *et al.*, 2012] B. Zhang, A. A. Tsiatis, M. Davidian, M. Zhang, and E. Laber. Estimating optimal treatment regimes from classification perspective. *Stat*, 1:103–114, 2012.

[Zhao *et al.*, 2012] Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107:1106–1118, 2012.

[Zhao *et al.*, 2015] Y. Zhao, D. Zheng, E. B. Laber, and M. R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110:583–598, 2015.

[Zhu *et al.*, 2011] L. Zhu, L. Qian, and J. Lin. Variable selection in a class of single-index models. *Annals of the Institute of Statistical Mathematics*, 63:1277–1293, 2011.

Appendix A

Appendix for Chapter 2

A.1 Assumptions for Theorem 1 and Theorem 2

Assumption 1. *The objective function (2.17), $R(\boldsymbol{\alpha}_{-p})$, 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. *(Regularity on the underlying mean functions)*
The underlying mean functions $m_t(X)$ in (2.4) are in $C^{(4)}(B_a^p)$, $t \in \{1, \dots, K\}$ for some finite $a > 0$, where B_a^p is the p-dimensional ball with center 0 and radius a and $C^{(q)}(B_a^p) = \big\{f \mid \text{ the qth order partial derivatives of } f \text{ are continuous in } B_a^p\big\}$.

Assumption 3. *(Regularity on the probability density function of X)*
The probability density function of $X, f_X(x) \in C^{(4)}(B_a^p)$, and there are constants $0 < c_f < C_f$ such that

$$\begin{cases} c_f / \mathit{Vol}_p(B_a^p) \leq f_X(x) \leq C_f / \mathit{Vol}_p(B_a^p), & \text{if} \quad x \in B_a^p \\ f_X(x) = 0, & \text{if} \quad x \notin B_a^p \end{cases}$$

Assumption 4. *(Regularity on the underlying noise distribution)*
The underlying noise ϵ in (2.4) 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 $t \in \{1, \dots, K\}$, 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 for the B-spline satisfies: $n_{\max}^{1/6} \ll N = d - 4 \ll n_{\min}^{1/5}(\log(n_{\min}))^{-(2/5)}$, where $n_{\max} = \max\{n_1, \ldots, n_t\}$ and $n_{\min} = \min\{n_1, \ldots, n_t\}$.*

A.2 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

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

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

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

where $u_{\boldsymbol{\alpha}} = F_p(\boldsymbol{\alpha}^\top X)$ and F_p is the rescaled centered Beta$\{(p+1)/2, (p+1)/2\}$ cumulative distribution function defined in Section 2.4.

A.3 Proof

A.3.1 Proof of Theorem 1

Proof. Under Assumptions 2–4 from the Appendix, by the results from $A.14$ of [Wang and Yang, 2007], we have

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

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

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

almost surely.

By the continuous mapping theorem,

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

almost surely, therefore, we have

$$\sup_{\boldsymbol{\alpha}\in\Theta_c} |Q(\boldsymbol{\alpha}) - R(\boldsymbol{\alpha})| \to 0, \tag{A.3}$$

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 (A.3), 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 (A.3), 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. □

A.3.2 Proof of Theorem 2

Proof. We first derive the expression (A1) 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 $R_t(\boldsymbol{\alpha}_{-p}) = R_t\left(\alpha_1, \dots, \alpha_{p-1}, \sqrt{1-(\alpha_1^2 + \cdots + \alpha_{p-1}^2)}\right)$. Applying the chain rule for taking the derivative of $R_t(\boldsymbol{\alpha}_{-p})$ with respect to α_j, we obtain

$$\frac{\partial}{\partial \alpha_j} R_t(\boldsymbol{\alpha}_{-p}) = \frac{\partial}{\partial \alpha_j} R_t(\boldsymbol{\alpha}) - \frac{\alpha_j}{\alpha_p} \frac{\partial}{\partial \alpha_p} R_t(\boldsymbol{\alpha}) \tag{A.4}$$

for each $j \in \{1, \dots, p-1\}$. Taking another derivative of (A.4) with respect to α_q, for each $q \in \{1, \dots, p-1\}$, again by applications of the chain rule,

$$\begin{aligned}\frac{\partial^2}{\partial\alpha_q\partial\alpha_j}R_t(\boldsymbol{\alpha}_{-p}) = \frac{\partial^2}{\partial\alpha_q\partial\alpha_j}R_t(\boldsymbol{\alpha}) - \frac{\alpha_q}{\alpha_p}\frac{\partial^2}{\partial\alpha_p\partial\alpha_j}R_t(\boldsymbol{\alpha}) - \frac{\alpha_j}{\alpha_p}\frac{\partial^2}{\partial\alpha_q\partial\alpha_p}R_t(\boldsymbol{\alpha}) \\ - \frac{\partial}{\partial\alpha_q}\left(\frac{\alpha_j}{\alpha_p}\right)\frac{\partial}{\partial\alpha_p}R_t(\boldsymbol{\alpha}) + \frac{\alpha_q\alpha_j}{\alpha_p^2}\frac{\partial^2}{\partial\alpha_p\partial\alpha_p}R_t(\boldsymbol{\alpha}). \quad \text{(A.5)}\end{aligned}$$

After summing (A.5) over the groups $t \in \{1, \dots, K\}$, weighted by the group probabilities $\pi_1, \dots, \pi_K$, evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, we obtain (A1).

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

$$\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\frac{\partial}{\partial\alpha_j}\{Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})\} - \frac{1}{n_t}\sum_{i=1}^{n_t}\xi_{\boldsymbol{\alpha},i,j,t}\right| = o\big(n_t^{-1/2}\big) \quad \text{(A.6)}$$

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

$$\begin{aligned}\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\frac{\partial}{\partial\alpha_j}\{Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})\}\right| = o(1), \\ \sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le q,j\le p}\left|\frac{\partial^2}{\partial\alpha_q\partial\alpha_j}\{Q_t(\boldsymbol{\alpha}) - R_t(\boldsymbol{\alpha})\}\right| = o(1),\end{aligned} \quad \text{(A.7)}$$

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

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 $\boldsymbol{\eta}_i \in \mathbb{R}^{p-1}$, $i \in \{1, \dots, n\}$, 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

$$\sup_{1\le j\le p-1}\left|\hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p}) - \frac{1}{n}\sum_{i=1}^n \eta_{i,j}\right| = o(n^{-1/2}), \quad \text{(A.8)}$$

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 $\boldsymbol{\eta}_i$. Now, in order to employ the result (A.6), we first consider the score function defined on the set Θ_c, i.e., $\hat{\Psi}_j(\boldsymbol{\alpha})$, instead of the "$p$th component removed" score function defined on $\mathbb{R}^{p-1}$, $\hat{\Psi}_j(\boldsymbol{\alpha}_{-p})$. We will show that, for some mean-zero independent random variables, which we will denote by $\xi^*_{\boldsymbol{\alpha},i,j}, i \in \{1,\ldots,n\}, j \in \{1,\ldots,p\}$,

$$\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\frac{\partial}{\partial\alpha_j}\{Q(\boldsymbol{\alpha})-R(\boldsymbol{\alpha})\}-\frac{1}{n}\sum_{i=1}^{n}\xi^*_{\boldsymbol{\alpha},i,j}\right| = o(n^{-1/2}) \tag{A.9}$$

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

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

which must satisfy the following:

$$\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p}\left|\sum_{t=1}^{K}\pi_t\left[\frac{\partial}{\partial\alpha_j}Q_t(\boldsymbol{\alpha})-\frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha})\right]-\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{K}\xi^*_{\boldsymbol{\alpha},i,j,t}\right| = o(n^{-1/2}). \tag{A.10}$$

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 (A.6). Therefore, applying the continuous mapping theorem and Slutsky's theorem to (A.6) leads to the desired result (A.10).

Next, we will show (A.8), 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 (A.8) to be

$$\begin{aligned}\eta_{i,j} &= \xi^*_{\boldsymbol{\alpha},i,j}-\frac{\alpha_j}{\alpha_p}\xi^*_{\boldsymbol{\alpha},i,p}\\ &=\sum_{t=1}^{K}\left[2\{g_t(u_{\boldsymbol{\alpha},i})-Y_i\}\left\{\frac{\partial}{\partial\alpha_j}g_t(u_{\boldsymbol{\alpha},i})-\frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}g_t(u_{\boldsymbol{\alpha},i})\right\}\right. \qquad \text{(A.11)}\\ &\qquad\left.+\frac{\partial}{\partial\alpha_j}R_t(\boldsymbol{\alpha})-\frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}R_t(\boldsymbol{\alpha})\right]I(T_i=t),\end{aligned}$$

then we can show

$$\begin{aligned}
&\sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p-1}\left|\left(\frac{\partial}{\partial\alpha_j}-\frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}\right)\{Q(\boldsymbol{\alpha})-R(\boldsymbol{\alpha})\}-\frac{1}{n}\sum_{i=1}^{n}\eta_{i,j}\right| \\
&\le \sup_{\boldsymbol{\alpha}\in\Theta_c}\sup_{1\le j\le p-1}\left|\frac{\partial}{\partial\alpha_j}\big(Q(\boldsymbol{\alpha})-R(\boldsymbol{\alpha})\big)-\frac{1}{n}\sum_{i=1}^{n}\xi^*_{\boldsymbol{\alpha},i,j}\right| \qquad (\text{A.12})\\
&+\sup_{\boldsymbol{\alpha}\in\Theta_c}\frac{\alpha_j}{\alpha_p}\left|\frac{\partial}{\partial\alpha_p}\big(Q(\boldsymbol{\alpha})-R(\boldsymbol{\alpha})\big)-\frac{1}{n}\sum_{i=1}^{n}\xi^*_{\boldsymbol{\alpha},i,p}\right|=o(n^{-1/2}),
\end{aligned}$$

by the triangle inequality and the result of (A.9). Since $\Psi_j(\boldsymbol{\alpha}_{-p})$ is evaluated at the minimum $\boldsymbol{\alpha}_{0,-p}$, we have

$$\Psi_j(\boldsymbol{\alpha}_{0,-p})=\left(\frac{\partial}{\partial\alpha_j}-\frac{\alpha_j}{\alpha_p}\frac{\partial}{\partial\alpha_p}\right)\{Q(\boldsymbol{\alpha})\}\bigg|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}=0, \qquad (\text{A.13})$$

by the local convexity under Assumption 1. Then we obtain the desired result of (A.8), by (A.12) and (A.13).

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 (A.7) 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 (A.8), and the following parametrization: for each component $j\in\{1,\dots,p-1\}$

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

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 $j=1,\dots,p-1$, there exists $s_j^*\in[0,1]$ by the mean value theorem, such that

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

which is just

$$\Bigg[\frac{\partial^2}{\partial\alpha_1\partial\alpha_j}\hat{Q}\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p}+(1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big),\\ \ldots,\frac{\partial^2}{\partial\alpha_{p-1}\partial\alpha_j}\hat{Q}\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p}+(1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big)\Bigg]\Bigg[\hat{\boldsymbol{\alpha}}_{-p}-\boldsymbol{\alpha}_{0,-p}\Bigg], \quad \text{(A.14)}$$

where $\Bigg[\hat{\boldsymbol{\alpha}}_{-p}-\boldsymbol{\alpha}_{0,-p}\Bigg]$ is a $p-1$ dimensional random vector. Writing (A.14) in matrix notation, we have

$$-\hat{\Psi}(\boldsymbol{\alpha}_{0,-p})=\Bigg[\frac{\partial^2}{\partial\alpha_q\partial\alpha_j}\hat{Q}\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p}+(1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big)\Bigg]_{j,q=1}^{p-1}\Bigg[\hat{\boldsymbol{\alpha}}_{-p}-\boldsymbol{\alpha}_{0,-p}\Bigg]. \quad \text{(A.15)}$$

Then, by (A.15) one can write

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_{-p}-\boldsymbol{\alpha}_{0,-p})=-\Bigg\{\Bigg[\frac{\partial^2}{\partial\alpha_q\partial\alpha_j}\hat{Q}\big(s_j^*\hat{\boldsymbol{\alpha}}_{-p}+(1-s_j^*)\boldsymbol{\alpha}_{0,-p}\big)\Bigg]_{j,q=1}^{p-1}\Bigg\}^{-1}\sqrt{n}\hat{\Psi}(\boldsymbol{\alpha}_{0,-p}). \quad \text{(A.16)}$$

Meanwhile, by (A.8), for each component $j\in\{1,\ldots,p-1\}$ of $\hat{\Psi}(\boldsymbol{\alpha}_{0,-p})$, we can write

$$\hat{\Psi}_j(\boldsymbol{\alpha}_{0,-p})=\frac{1}{n}\sum_{i=1}^{n}\eta_{i,j}+o(n^{-1/2}), \quad \text{(A.17)}$$

almost surely with $\mathbb{E}(\eta_{i,j})=0$. The variance-covariance matrix of the random vector $\boldsymbol{\eta}_i=\big[\eta_{i,1},\ldots,\eta_{i,p-1}\big]^\top\in\mathbb{R}^{p-1}$ evaluated at $\boldsymbol{\alpha}_{-p}=\boldsymbol{\alpha}_{0,-p}$, where $\eta_{i,j}$ are specified in (A.11), is given in ($A2$), where it is denoted by $\boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}}$. From (A.17), the central limit theorem ensures that $\sqrt{n}\hat{\Psi}(\boldsymbol{\alpha}_{0,-p})\to\mathcal{N}(\boldsymbol{0},\boldsymbol{W}_{\boldsymbol{\alpha}_{0,-p}})$ in distribution. Now, by the representation of (A.16) together with an application of Slutsky's theorem on the observed Hessian, we obtain $\sqrt{n}(\hat{\boldsymbol{\alpha}}_{0,-p}-\boldsymbol{\alpha}_{0,-p})\to\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{0,-p}})$ 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. □

A.4 Table for Section 2.6.3 Coverage probability of asymptotic 95% confidence intervals

		$\omega = 0$ (single-crossing)					$\omega = 1$ (multiple-crossing)				
n		α_1	α_2	α_3	α_4	α_5	α_1	α_2	α_3	α_4	α_5
100	Coverage	0.61	0.56	0.56	0.57	0.60	0.57	0.57	0.49	0.50	0.49
200	Coverage	0.69	0.71	0.70	0.66	0.67	0.77	0.70	0.70	0.67	0.66
400	Coverage	0.80	0.81	0.81	0.79	0.81	0.84	0.82	0.78	0.74	0.73
800	Coverage	0.89	0.89	0.87	0.89	0.90	0.90	0.90	0.86	0.81	0.81
1600	Coverage	0.93	0.93	0.94	0.93	0.94	0.92	0.89	0.87	0.85	0.83
3200	Coverage	0.94	0.93	0.95	0.96	0.94	0.92	0.91	0.89	0.81	0.81
6400	Coverage	0.95	0.94	0.94	0.92	0.92	0.93	0.95	0.87	0.86	0.84

Table A.1: The proportion of time ("Coverage") that the asymptotic 95% confidence interval contains the true value of α_j, $j \in \{1, \ldots, 5\}$, for contrast functions with a single crossing ($\omega = 0$), and contrasts functions with multiple crossings ($\omega = 1$), with varying $n(= n_1 + n_2$, where $2n_1 = 3n_2)$

Appendix B

Appendix for Chapter 3

B.1 A justification for the equivalence between MCA and SIMML under a linear link restriction

With the $K = 2$ treatment groups, for sample data and each given $\boldsymbol{\alpha}$, in order to represent the treatment t-specific linear link function that consists of an intercept and a slope (note, we consider a special case of the constrained SIMML in which the link function is restricted to be linear), we can define the $n_t \times 2$ design matrix $\mathbb{Z}_{\boldsymbol{\alpha},t}$ for the tth treatment group, $t = 1, 2$, whose ith row is $\left[1, \boldsymbol{\alpha}^\top X_{ti}\right]$, $i = 1, \dots, n_t$. Then, the "smoother" $\mathbb{S}^*$ is given by the projection matrix

$$\mathbb{S}^* = \begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} \\ \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix} \left[\begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} \\ \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix}^\top \begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} \\ \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix} \right]^{-1} \begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} \\ \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix}^\top,$$

and similarly, the "smoother" $\mathbb{S}^{**}$ is given by

$$\mathbb{S}^{**} = \begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} & \mathbf{0} \\ \mathbf{0} & \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix} \left[\begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} & \mathbf{0} \\ \mathbf{0} & \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix}^\top \begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} & \mathbf{0} \\ \mathbf{0} & \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix} \right]^{-1} \begin{bmatrix} \mathbb{Z}_{\boldsymbol{\alpha},1} & \mathbf{0} \\ \mathbf{0} & \mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix}^\top,$$

and therefore, $\mathbb{S}_{\boldsymbol{\alpha}} = \mathbb{S}_{\boldsymbol{\alpha}}^{(**)} - \mathbb{S}_{\boldsymbol{\alpha}}^{(*)}$. The constrained SIMML minimizes $\|\boldsymbol{Y} - \mathbb{S}_{\boldsymbol{\alpha}}\boldsymbol{Y}\|^2$ over $\boldsymbol{\alpha}$.

On the other hand, the design matrix under the working model of the MCA (3.3) is given by $\begin{bmatrix} -0.5\mathbb{Z}_{\boldsymbol{\alpha},1} \\ +0.5\mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix}_{n\times 2}$. The corresponding "smoother" is given by

$$\mathbb{S}_{\boldsymbol{\alpha}}^{(MCA)} = \begin{bmatrix} -0.5\mathbb{Z}_{\boldsymbol{\alpha},1} \\ +0.5\mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix} \left[\begin{bmatrix} -0.5\mathbb{Z}_{\boldsymbol{\alpha},1} \\ +0.5\mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix}^\top \begin{bmatrix} -0.5\mathbb{Z}_{\boldsymbol{\alpha},1} \\ +0.5\mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix} \right]^{-1} \begin{bmatrix} -0.5\mathbb{Z}_{\boldsymbol{\alpha},1} \\ +0.5\mathbb{Z}_{\boldsymbol{\alpha},2} \end{bmatrix}^\top$$

Then the MCA minimizes the squared error criterion $\|\boldsymbol{Y} - \mathbb{S}_{\boldsymbol{\alpha}}^{(MCA)}\boldsymbol{Y}\|^2$ over $\boldsymbol{\alpha}$. It can be verified that $\mathbb{S}_{\boldsymbol{\alpha}}^{(MCA)} = \mathbb{S}_{\boldsymbol{\alpha}}$ by simple algebra, implying that the two approaches minimize the same criterion (up to a scale, that depends on the identifiability conditions on $\boldsymbol{\alpha}$), for the case where $K = 2$ and the link function in model (3.6) is linear.

B.2 Proof

B.2.1 Proof of Theorem 3

Proof. For each $\boldsymbol{\alpha}$ the criterion (3.7) as a function of $\boldsymbol{g} = (g_1, \dots, g_K)$ can be given as $\sum_{t=1}^{K} \pi_t \mathbb{E}\left(Y - g_t\left(\boldsymbol{\alpha}^\top X\right) \mid T = t\right)^2$. For each $u = \boldsymbol{\alpha}^\top X$, consider the minimization of the Lagrangian:

$$H(\boldsymbol{g}; \lambda) = \sum_{t=1}^{K} \pi_t \mathbb{E}\left(Y - g_t(u) \mid u, T = t\right)^2 + \lambda \sum_{t=1}^{K} \pi_t g_t(u)$$

with respect to $g_t \in L^2$, where L^2 denotes the L^2 space of functions, holding the other components $\{g_{t'}, t' \neq t\}$ fixed, where $\lambda > 0$ is the Lagrange multiplier. The stationary condition is obtained by setting the Fréchet derivative to 0. Denote by $\partial_t H(\boldsymbol{g}; \lambda; \eta_t)$ the directional derivative with respect to g_t in the direction $\eta_t \in L^2$. The stationary point can be formulated as

$$\partial_t H(\boldsymbol{g}; \lambda; \eta_t) = 2\mathbb{E}\left((g_t - R_t + \lambda)\eta_t\right) = 0,$$

where $R_t = Y - \sum_{t' \neq t} g_{t'}$ is the partial residual for g_t. Using iterated expectations, the condition above can be rewritten as

$$\mathbb{E}\left((g_t + \lambda - \mathbb{E}\left(R_t \mid u\right))\eta_t\right) = 0.$$

Since $g_t + \lambda - \mathbb{E}\left(R_t \mid u\right) \in L^2$, hence, we can compute the derivative in the direction $\eta_t = g_t + \lambda - \mathbb{E}\left(R_t \mid u\right)$, giving $\mathbb{E}\left(g_t + \lambda - \mathbb{E}\left(R_t \mid u\right)\right)^2 = 0$, hence,

$$g_t = \mathbb{E}\left(R_t \mid u\right) - \lambda, \quad t = 1, \dots, K, \quad \text{a.s..} \tag{B.1}$$

Similarly, taking derivative with respect to λ and set it to 0, we obtain

$$\lambda = \sum_{t=1}^{K} \pi_t \mathbb{E}\left(Y \mid u, T=t\right) = \mathbb{E}\left(Y \mid u\right).$$

Solving the system of equation (B.1), we have $g_t(u) = \mathbb{E}\left(Y \mid u, T=t\right) - \mathbb{E}\left(Y \mid u\right)$, $t = 1, \ldots, K$, almost surely. $\nabla H(\boldsymbol{g};\lambda)(u) = (\partial H/\partial g_1(u), \ldots, \partial H/\partial g_I(u), \partial H/\partial \lambda)$, where $\partial H/\partial g_t$'s are the Fréchet derivatives of H at g_t, $t = 1, \ldots, K$. Then, setting $\nabla H(\boldsymbol{g};\lambda)(u)$ to $\mathbf{0}$ for all u to obtain the unique critical point yields the following system of solutions: $g_t(u) = \mathbb{E}\left(Y \mid u, T=t\right) - \lambda$, $t = 1, \ldots, K$, and $\lambda = \sum_{t=1}^{K} \pi_t \mathbb{E}\left(Y \mid u, T=t\right) = \mathbb{E}\left(Y \mid u\right)$, almost surely. Therefore, we have $g_t(u) = \mathbb{E}\left(Y \mid u, T=t\right) - \mathbb{E}\left(Y \mid u\right)$, $t = 1, \ldots, K$, almost surely. □

B.2.2 Proof of Lemma 1

Proof. Since $\bar{\boldsymbol{\beta}} \sum_{t=1}^{K} c_t = 0$, one can write $C(X) = \sum_{t=1}^{K} c_t(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})^\top X$, where the main effect $\mu(X)$ was canceled out, under model (3.19), $\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}$, $t \in \{1, \ldots, K\}$, can be uniquely expressed in terms of coordinates, say $\boldsymbol{\gamma}_t \in \mathbb{R}^{K-1}$, with respect to the basis $\boldsymbol{\Phi}$; i.e., $\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}} = \boldsymbol{\Phi}\boldsymbol{\gamma}_t$, $t \in \{1, \ldots, K\}$. Then $C(X) = \sum_{t=1}^{K} c_t \left(\boldsymbol{\Phi}\boldsymbol{\gamma}_t\right)^\top X = \sum_{t=1}^{K} c_t(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})^\top \boldsymbol{\Phi}\boldsymbol{\Phi}^\top X = \sum_{t=1}^{K} c_t \boldsymbol{\beta}_t^\top \boldsymbol{\Phi}\boldsymbol{\Phi}^\top X$.

On the other hand, under model (3.19), $C(\boldsymbol{\Phi}^\top X) = \sum_{t=1}^{K} c_t \mathbb{E}\left(Y \mid \boldsymbol{\Phi}^\top X, T=t\right) = \sum_{t=1}^{K} c_t(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}_x \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Psi}_x \boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top X$, where $\boldsymbol{\Psi}_x = \mathrm{cov}(X)$, and note that $\mu(X)$ was canceled out, due to $\sum_{t=1}^{K} c_t = 0$. Since X is "scaled", we have $\mathrm{cov}(X) = \boldsymbol{I}_p$. Therefore, $C(\boldsymbol{\Phi}^\top X) = \sum_{t=1}^{K} c_t \boldsymbol{\beta}_t^\top \boldsymbol{\Phi}\boldsymbol{\Phi}^\top X$. Hence, the two expressions agree with each other. □

B.2.3 Proof of Lemma 2

Proof. Conditioning on X, model (3.26) is written as

$$\mathbb{E}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{bmatrix}_{K\times 1} = \begin{bmatrix} X^\top \boldsymbol{\xi}_1 & 0 & \ldots & 0 \\ 0 & X^\top \boldsymbol{\xi}_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & X^\top \boldsymbol{\xi}_1 \end{bmatrix}_{K\times K} \begin{bmatrix} \gamma_1^* \\ \gamma_2^* \\ \vdots \\ \gamma_K^* \end{bmatrix}_{K\times 1}, \tag{B.2}$$

subject to the identifiability constraint (3.24) $\boldsymbol{\pi}^\top \boldsymbol{\gamma}^* = 0$ in matrix-vector notation, where $\boldsymbol{\pi} := [\pi_1, \ldots, \pi_K]^\top$ and $\boldsymbol{\gamma}^* := [\gamma_1^*, \ldots \gamma_K^*]^\top$. In (B.2), Y_t denotes the treatment t-specific outcome, $t \in \{1, \ldots, K\}$. Let $\boldsymbol{\gamma} \in \mathbb{R}^K$ denote the unconstrained least squares solution for equation (B.2). In terms of the population parameters of model (3.19), we have

$$\boldsymbol{\gamma} = \begin{bmatrix} \text{cov}(X^\top \boldsymbol{\xi}_1, Y_1) \\ \vdots \\ \text{cov}(X^\top \boldsymbol{\xi}_1, Y_K) \end{bmatrix}_{K \times 1} = \begin{bmatrix} \boldsymbol{\xi}_1^\top \boldsymbol{\beta}_1 + \text{cov}(X^\top \boldsymbol{\xi}_1, \mu(X)) \\ \vdots \\ \boldsymbol{\xi}_1^\top \boldsymbol{\beta}_K + \text{cov}(X^\top \boldsymbol{\xi}_1, \mu(X)) \end{bmatrix}_{K \times 1} . \tag{B.3}$$

The constrained solution $\boldsymbol{\gamma}^*$ in (B.2) can be obtained by a linear projection approach available in linear model and regression books (e.g., [Seber and Lee, 2012]), however, we present the derivation for completeness. For convenience, let us write the $K \times K$ design matrix of (B.2) by $\boldsymbol{D}$. Suppose $\boldsymbol{\gamma}_0^*$ is any vector satisfying $\boldsymbol{\pi}^\top \boldsymbol{\gamma}^* = 0$. Let us consider the "shifted" response vector adjusted by $\boldsymbol{D}\boldsymbol{\gamma}_0^*$, i.e., $\tilde{\boldsymbol{Y}} = [Y_1, \ldots, Y_K]^\top - \boldsymbol{D}\boldsymbol{\gamma}_0^*$, and the corresponding "shifted" parameter $\tilde{\boldsymbol{\gamma}}^* = \boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0^*$, to write

$$\tilde{\boldsymbol{Y}} = \boldsymbol{D}\tilde{\boldsymbol{\gamma}}^* + \boldsymbol{\epsilon}, \tag{B.4}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^K$ is a mean-zero noise. In (B.4), the constraint $\boldsymbol{\pi}^\top \tilde{\boldsymbol{\gamma}}^* = \boldsymbol{\pi}^\top \boldsymbol{\gamma}^* - \boldsymbol{\pi}^\top \boldsymbol{\gamma}_0^* = 0$ is still satisfied. Let us define $\boldsymbol{\Omega} = \mathcal{C}(\boldsymbol{D})$, the span of $\boldsymbol{D}$. We can write model $\tilde{\boldsymbol{Y}} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. Since $\boldsymbol{D}$ is full-rank, it follows that $\boldsymbol{\pi}^\top (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{\theta} = \boldsymbol{\pi}^\top \tilde{\boldsymbol{\gamma}}^* = 0$. Let us write $\boldsymbol{\pi}_1^\top = \boldsymbol{\pi}^\top (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top$. The subspace of interest, which we denote by $\boldsymbol{\omega}$, corresponds to the space of mean responses specified by solutions $\boldsymbol{\gamma}^*$ satisfying the constraint (3.24), i.e., $\boldsymbol{\omega} = \mathcal{N}(\boldsymbol{\pi}_1^\top) \cap \boldsymbol{\Omega}$. Furthermore, we can write $\boldsymbol{\omega}^\perp \cap \boldsymbol{\Omega} = \mathcal{C}(\boldsymbol{H}_{\boldsymbol{\Omega}} \boldsymbol{\pi}_1)$, i.e., by the span of the matrix $\boldsymbol{H}_{\boldsymbol{\Omega}} \boldsymbol{\pi}_1$, in which $\boldsymbol{H}_{\boldsymbol{\Omega}} \boldsymbol{\pi}_1 = \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{D} (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi} = \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi}$. Then, we can write the projection matrix onto the space of interest $\boldsymbol{\omega}$, i.e., $\boldsymbol{H}_{\boldsymbol{\omega}} = \boldsymbol{H}_{\boldsymbol{\Omega}} - \boldsymbol{H}_{\boldsymbol{\omega}^\perp \cap \boldsymbol{\Omega}}$ by

$$\boldsymbol{H}_{\boldsymbol{\omega}} = \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top - \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi}^\top (\boldsymbol{\pi}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi}^\top)^{-1} \boldsymbol{\pi} (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top, \tag{B.5}$$

which leads to the following identity

$$\begin{aligned} \boldsymbol{D}\boldsymbol{\gamma}^* - \boldsymbol{D}\boldsymbol{\gamma}_0^* = \boldsymbol{H}_{\boldsymbol{\omega}} \boldsymbol{\theta} \\ &= \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{\gamma} - \boldsymbol{D}\boldsymbol{\gamma}_0^*) \\ &\quad - \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi}^\top (\boldsymbol{\pi}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi}^\top)^{-1} \boldsymbol{\pi} (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{\gamma} - \boldsymbol{D}\boldsymbol{\gamma}_0^*) \\ &= \boldsymbol{D}\boldsymbol{\gamma} - \boldsymbol{D}\boldsymbol{\gamma}_0^* - \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi}^\top (\boldsymbol{\pi}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\pi}^\top)^{-1} \boldsymbol{\pi} \boldsymbol{\gamma}, \end{aligned} \tag{B.6}$$

where the left hand side is from (B.4). Canceling $\boldsymbol{D}\boldsymbol{\gamma}_0^*$ and multiplying by $(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top$ on the both sides of (B.6), we can obtain the form of the constrained least squares solution $\boldsymbol{\gamma}^*$

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma} - (\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{\pi}^\top(\boldsymbol{\pi}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{\pi}^\top)^{-1}\boldsymbol{\pi}\boldsymbol{\gamma}, \tag{B.7}$$

where the second term on the right-hand side is a length-K vector

$$\left[\mathbf{1}_K \otimes \left(\boldsymbol{\xi}_1^\top\bar{\boldsymbol{\beta}} + \mathrm{cov}(X^\top\boldsymbol{\xi}_1, \mu(X))\right)\right],$$

$\mathbf{1}_K$ denotes the vector of ones of length K, and $\boldsymbol{\gamma}$ is given in (B.3). The term $\mathrm{cov}(X^\top\boldsymbol{\xi}_1, \mu(X))$ cancels out in (B.7), and the constrained solution in (B.2) is $\gamma_t^* = \boldsymbol{\xi}_1^\top(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})$, $t \in \{1, \ldots, K\}$. □

Appendix C

Appendix for Chapter 4

C.1 Derivation of the updating rule

To obtain the expression Δ_j in (4.3), we will minimize the objective function in a small neighborhood of the current $\boldsymbol{\alpha}$, say at $\tilde{\boldsymbol{\alpha}}$, for the (j, L)th block, with $\Delta_j := \alpha_j - \tilde{\alpha}_j$ and $\Delta_L := \alpha_L - \tilde{\alpha}_L$, by the following local linear approximation of the regression model

$$Q(\boldsymbol{\alpha}) \approx \left\| \boldsymbol{R} - \hat{\boldsymbol{g}}_{\boldsymbol{\alpha}}^{'} * (X_j \Delta_j + X_L \Delta_L) \right\|^2 . \tag{C.1}$$

Then the update rule for the (j, L)th block is $\left\{ \alpha_j^{\text{new}} \leftarrow \tilde{\alpha}_j + \hat{\Delta}_j, \quad \alpha_L^{\text{new}} \leftarrow \tilde{\alpha}_L + \hat{\Delta}_L \right\}$, where $\left[\hat{\Delta}_j, \hat{\Delta}_L \right]^\top$ is the constrained minimizer of the approximated objective (C.1) over $[\Delta_j, \Delta_L]^\top \in \mathbb{R}^2$, under the constraint (4.4). After substituting Δ_L in (C.1) by $-\Delta_j S_{jL}$ in (4.4), we take derivative of (C.1) with respect to Δ_j, set it to 0, and solve the estimating equation for Δ_j

$$\left(\boldsymbol{R} - \hat{\boldsymbol{g}}_{\boldsymbol{\alpha}}^{'} * (X_j - S_{jL} X_L) \Delta_j \right)^\top \left(-\hat{\boldsymbol{g}}_{\boldsymbol{\alpha}}^{'} * (X_j - S_{jL} X_L) \right) = 0,$$

and solving for Δ_j gives the expression (4.3).

C.2 Some computational notes 1

In this subsection, we focus on the constrained estimation of $\boldsymbol{\beta}$ in model (4.5), under the orthogonality (identifiability) constraint $\boldsymbol{\alpha} \perp \boldsymbol{\beta}$, given each candidate $\boldsymbol{\alpha}$, that we discussed in Section 5.4. Let us denote the "$\hat{a}$th component removed" $\boldsymbol{\alpha}$ by $\boldsymbol{\alpha}^{(-\hat{a})} \in \mathbb{R}^{p-1}$. Similarly, let us also denote the "$\hat{a}$th component removed" $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(-\hat{a})} \in \mathbb{R}^{p-1}$. Given $\boldsymbol{\alpha}^{(-\hat{a})}$, the orthogonality constraints will be satisfied, if we set the $\hat{a}$th component of $\boldsymbol{\beta}$ by

$$\beta_{\hat{a}} = \sum_{j \neq \hat{a}, j=1}^{p} \alpha_j \beta_j \tag{C.2}$$

The estimation is performed on $\left\{\boldsymbol{\alpha}^{(-\hat{a})}, \boldsymbol{\beta}^{(-\hat{a})}\right\}$, and the remaining components are determined by the identifiability constraint $\alpha_{\hat{a}} = 1$ in (4.1) and by equation (C.2). Let us denote the $p \times (p-1)$ Jacobian transformation matrix from $\boldsymbol{\beta}^{(-\hat{a})} \in \mathbb{R}^{p-1}$ to the full vector $\boldsymbol{\beta} \in \mathbb{R}^{p}$ by $\boldsymbol{J} = \begin{bmatrix} \boldsymbol{\beta}^{(-1)} & \boldsymbol{I}_{p-1} \end{bmatrix}^{\top}$, where we set $\hat{a}$ to be 1 (i.e., the first component of $\boldsymbol{\alpha}$) for the ease of illustration without loss of generality. Depending on what actual $\hat{a}$ is, the structure of $\boldsymbol{J}$ will be changed accordingly (i.e., the $\hat{a}$th row of $\boldsymbol{J}$ should be the vector $\boldsymbol{\beta}^{(-\hat{a})\top}$). Taking the approach of [Lian and Liang, 2016], the asymptotic first-order condition for optimizing $\boldsymbol{\beta}$ under the constraint (C.2) for each $\boldsymbol{\alpha}$ is given in the form of a penalized estimating function

$$\boldsymbol{J}^{\top}\left[\left(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}}\right)\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)\boldsymbol{X}\right] + \boldsymbol{J}^{\top}\left[\boldsymbol{p}'_{\lambda_2}(|\boldsymbol{\beta}|)\mathbf{sgn}(\boldsymbol{\beta})\right] = \mathbf{0}, \tag{C.3}$$

where $\boldsymbol{p}'_{\lambda_2}(|\boldsymbol{\beta}|) := (p'_{\lambda_2}(|\beta_1|), \ldots, p'_{\lambda_2}(|\beta_p|))^{\top}$, and $\mathbf{sgn}(\boldsymbol{\beta}) := (\mathrm{sgn}(\beta_1), \ldots, \mathrm{sgn}(\beta_p))^{\top}$, and $\mathbf{0}$ is a vector of zeros of length $(p-1)$. Many penalty choices are available for regularizing $\boldsymbol{\beta}$ (e.g., Lasso, SCAD). Given a current estimate of $\boldsymbol{\beta}$, say, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^{\top}$, one can perform the local quadratic approximation (LQA) ([Fan and Li, 2001]), i.e., linearly approximate the first derivative of the penalty at $\tilde{\boldsymbol{\beta}}$, $p'_{\lambda_2}(|\beta_j|) \approx \left\{p'_{\lambda_2}(|\tilde{\beta}_j|)/|\tilde{\beta}_j|\right\}\beta_j$, $j = 1, \ldots, p$, in which $\tilde{\boldsymbol{\beta}}$ is to be updated over iterations until convergence (the "outer" loop).

If we define the $p \times p$ penalty matrix $\tilde{\boldsymbol{\Omega}}_{\lambda_2} = \mathrm{diag}\left(p'_{\lambda_2}(|\tilde{\beta}_1|)/|\tilde{\beta}_1|, \ldots, p'_{\lambda_2}(|\tilde{\beta}_p|)/|\tilde{\beta}_p|\right)$, where the subscript, λ_2, highlights its dependency. then we can perform a gradient descent on $\boldsymbol{\beta}^{(-\hat{a})}$ until convergence (the "inner" loop), as suggested by [Lian and Liang, 2016]

$$\tilde{\boldsymbol{\beta}}^{(-\hat{a})} \leftarrow \tilde{\boldsymbol{\beta}}^{(-\hat{a})} + \kappa\left\{\boldsymbol{J}^{\top}\left[\left(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}}\right)\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)\boldsymbol{X}\right] - \boldsymbol{J}^{\top}\tilde{\boldsymbol{\Omega}}_{\lambda_2}\tilde{\boldsymbol{\beta}}\right\}, \tag{C.4}$$

where $\kappa > 0$ is a small number corresponding to the size of the descent. The remaining component $\tilde{\beta}_{\hat{a}}$ is determined via (C.2), for each iterative step. To select the tuning parameters, we can choose, for example, the minimizer of $\mathrm{BIC}(\lambda_2) = \log(\mathrm{MSE}(\lambda_2)) + \log(n)\mathrm{d.f.}_{\lambda_2}/n$, over the sequence of candidate values of λ_2's. Here, $\mathrm{d.f.}_{\lambda_2}$ is the number of nonzero coefficients in the estimate $\hat{\boldsymbol{\beta}}^{(\lambda_2)}$ obtained under the regularization parameter λ_2, and $\mathrm{MSE}(\lambda_2) = \|\big(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}}\big)\big(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_2)}\big)\|^2/n$.

C.3 Some computational notes 2

Given $\boldsymbol{\alpha}$, the estimation of model (4.5) can be viewed as the penalized estimation with a zero-sum constraint (i.e., the orthogonality constraint $\boldsymbol{\alpha}^\top\boldsymbol{\beta} = 0$), once the link functions are represented by a set of B-spline basis functions. [Lin *et al.*, 2014] developed an algorithm to solve such constrained optimization problem with a L^1 regularization. However, we notice that their numerical results of the estimator are not significantly different to that of the regular Lasso that simply ignores the zero-sum constraint, in terms of support recovery (i.e., the predictor selection), although the resulting regular Lasso estimator would violate the zero-sum constraint in finite samples. In practice, we can use the regular Lasso approach for the covariate selection, followed by refitting the unpenalized least squares problem with the zero-sum (i.e., the orthogonality condition) with the selected variables, using the method of Lagrange multipliers. Since the coordinate descent algorithms have been shown to be very efficient for solving large-scale regular Lasso problems ([Friedman *et al.*, 2007]), this "ad-hoc" approach described in the following can save a significant amount of computing time.

For each assumed $\boldsymbol{\alpha}$, regress the vector $(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}})\,\boldsymbol{Y}$ on $(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}})\,\boldsymbol{X}$ using the regular Lasso, and obtain an estimate $\hat{\boldsymbol{\beta}}$, where we choose the regularization parameters by a cross-validation. We can select the model via the Lasso (i.e., $\hat{S} = \mathrm{support}(\hat{\boldsymbol{\beta}})$), which consists of the (modified) predictors associated with nonzero coefficients of $\hat{\boldsymbol{\beta}}$. Then, one can solve for $\boldsymbol{\beta}$, using the method of Lagrange multipliers, that satisfies the linear equation

$$\begin{bmatrix} 2\check{\boldsymbol{X}}^\top\check{\boldsymbol{X}} & \check{\boldsymbol{\alpha}} \\ \check{\boldsymbol{\alpha}}^\top & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \check{\boldsymbol{\beta}} \\ \nu \end{bmatrix} = \begin{bmatrix} 2\check{\boldsymbol{X}}^\top\,(\boldsymbol{I}_n - \mathbb{S}_{\boldsymbol{\alpha}})\,\boldsymbol{Y} \\ \boldsymbol{0} \end{bmatrix}, \tag{C.5}$$

in which ν is the Lagrange multiplier, and $\check{\boldsymbol{\alpha}}$ stands for the original $\boldsymbol{\alpha}$ with the components selected by $\hat{S}$ only; similarly we define $\check{\boldsymbol{X}}$ and $\check{\boldsymbol{\beta}}$. Provided that $\begin{bmatrix} 2\check{\boldsymbol{X}}^\top\check{\boldsymbol{X}} & \check{\boldsymbol{\alpha}} \\ \check{\boldsymbol{\alpha}}^\top & \boldsymbol{0} \end{bmatrix}^{-1}$ exists, the equation (C.5) can be solved immediately. We can obtain $\hat{\boldsymbol{\beta}}$ satisfying $\boldsymbol{\alpha}^\top\hat{\boldsymbol{\beta}} = 0$ by setting the components of $\hat{\boldsymbol{\beta}}$ that correspond to $\hat{S}$ by the solution $\boldsymbol{\beta}^*$ in (C.5), and those in the complement of $\hat{S}$ by zeros.

Appendix D

Appendix for Chapter 5

D.1 Proof

D.1.1 Proof of Theorem 4

Proof. Given α_j, $j = 1, \ldots, p$, the mean squared error criterion can be given as $\sum_{t=1}^{K} \pi_t \mathbb{E}\left(Y - \sum_{j=1}^{p} g_{j,t}\left(\langle \alpha_j, x_j \rangle\right) \mid T = t\right)^2$, where the expectation inside is taken with respect to the conditional distribution of (Y, T, X) given that $T = t$.

Given $u_j = \langle \alpha_j, x_j \rangle$, $j = 1, \ldots, p$, let us write $\tilde{\boldsymbol{g}} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_p)$, in which $\boldsymbol{g}_j = (g_{j,1}, \ldots, g_{j,K})$ for each $j = 1, \ldots, p$. In the following, we closely follow the proof of Theorem 1 in [Ravikumar *et al.*, 2009]. We consider the minimization of the Lagrangian:

$$H(\tilde{\boldsymbol{g}}; \lambda) = \sum_{t=1}^{K} \pi_t \left[\mathbb{E}\bigg(Y - \sum_{j=1}^{p} g_{j,t}(u_j) \mid T = t\bigg)^2 + \lambda\bigg(\sum_{j=1}^{p} \|g_{j,t}\|\bigg) + \sum_{j=1}^{p} \tau_j \mathbb{E}\left[g_{j,t}(u_j) \mid T = t\right] \right],$$

with respect to $g_{j,t}$, holding the other components $\{g_{j',t'}, j' \neq j, t' \neq t\}$ fixed, for each j and t. The stationary condition is obtained by setting its Fréchet derivative to 0. Denote by $\partial_{j,t} H(\tilde{\boldsymbol{g}}; \lambda; \eta_{j,t})$ the directional derivative with respect to $g_{j,t}$ in the direction, say, $\eta_{j,t} \in L^2$, where L^2 denotes the L^2 space of functions. Then, the stationary point can be formulated as

$$\partial_{j,t} H(\tilde{\boldsymbol{g}}; \lambda; \eta_{j,t}) = 2 \sum_{t=1}^{K} \pi_t \mathbb{E}\left((g_{j,t}(u_j) - R_{j,t} + \lambda \nu_{j,t} + \tau_j)\eta_{j,t} \mid T = t\right) = 0,$$

where $R_{j,t} = Y - \sum_{j' \neq j} g_{j',t}(u_{j'})$ is the residual for $g_{j,t}(u_j)$; $\nu_{j,t}$ is an element of the subgradient $\partial \|g_{j,t}\|$, which satisfies $\nu_{j,t} = g_{j,t}(u_j) / \|g_{j,t}\|$ if $\|g_{j,t}\| \neq 0$, and $\nu_{j,t} \in \{s \in L^2 \mid \|s\| \leq 1\}$, otherwise.

Using iterated expectations, the condition above can be rewritten as

$$\mathbb{E}\left(\left(g_{j,t}(u_j) + \lambda\nu_{j,t} + \tau_j - \mathbb{E}\left[R_{j,t} \mid u_j, T = t\right]\right)\eta_{j,t}\right) = 0, \quad t = 1, \ldots, K.$$

Since $g_{j,t}(u_j) + \lambda\nu_{j,t} + \tau_j - \mathbb{E}\left(R_j \mid u_j, T = t\right) \in L^2$, we can compute the derivative in the direction $g_{j,t}(u_j) + \lambda\nu_{j,t} + \tau_j - \mathbb{E}\left(R_j \mid u_j, T = t\right)$, implying

$$\mathbb{E}\left(g_{j,t}(u_j) + \lambda\nu_{j,t} + \tau_j - \mathbb{E}\left(R_j \mid u_j, T = t\right)\right)^2 = 0.$$

Therefore, we have

$$g_{j,t}(u_j) + \lambda\nu_{j,t} = \mathbb{E}\left(R_j \mid u_j, T = t\right) - \tau_j, \quad t = 1, \ldots, K, \quad \text{a.s..} \tag{D.1}$$

Let $P_{j,t}$ denote the right-hand side of (D.1), $\mathbb{E}\left(R_j \mid u_j, T = t\right) - \tau_j$. If $\|g_{j,t}\| \neq 0$, then $\nu_{j,t} = g_{j,t}(u_j)/\|g_{j,t}\|$. Therefore, by (D.1), we have $\|P_{j,t}\| = \|g_{j,t}(u_j) + \lambda g_{j,t}(u_j)/\|g_{j,t}\|\| = \|g_{j,t}\| + \lambda \geq \lambda$. On the other hand, if $\|g_{j,t}\| = 0$, then $g_{j,t}(u_j) = 0$ almost surely, and $\|\nu_{j,t}\| \leq 1$. Then, condition (D.1) implies that $\|P_{j,t}\| \leq \lambda$. This gives us the equivalence between $\|P_{j,t}\| \leq \lambda$ and the statement $g_{j,t}(u_j) = 0$ almost surely.

Taking derivative of $H(\tilde{\boldsymbol{g}}; \lambda)$ with respect to τ_j and setting it to 0, we obtain $\tau_j = \sum_{t=1}^{K} \pi_t \mathbb{E}\left(R_j \mid u_j, T = t\right) = \mathbb{E}\left(R_j \mid u_j\right)$, therefore $P_{j,t} = \mathbb{E}\left(R_j \mid u_j, T = t\right) - \mathbb{E}\left(R_j \mid u_j\right)$.

Condition (D.1) leads to the following expression:

$$\left(1 + \lambda/\|g_{j,t}\|\right) g_{j,t}(u_j) = P_{j,t}, \quad t = 1, \ldots, K, \quad \text{a.s.,}$$

if $\|P_{j,t}\| > \lambda$, and $g_{j,t}(u_j) = 0$, almost surely, otherwise. This gives the soft thresholding update rule for $g_{j,t}$.

□