Constrained functional additive models for estimating interactions between a treatment and functional covariates

HYUNG G. PARK\textsuperscript{a*}, EVA PETKOVA\textsuperscript{a}, THADDEUS TARPEY\textsuperscript{a}, R. TODD OGDEN\textsuperscript{b}

\textsuperscript{a} Division of Biostatistics, Department of Population Health, New York University
\textsuperscript{b} Department of Biostatistics, Columbia University

Abstract

A novel functional additive model is proposed which is uniquely modified and constrained to model nonlinear interactions between a treatment indicator and a potentially large number of functional/scalar covariates. We generalize functional additive regression models by incorporating treatment-specific components into additive effect components. A structural constraint is imposed on the treatment-specific components, to give a class of orthogonal main and interaction effect additive models. If primary interest is in interactions, we can avoid estimating main effects, obviating the need to specify their form and thereby avoiding the issue of model misspecification. The methods are illustrated with data from a clinical trial with imaging data as predictors.

Keywords: Individualized treatment rules, functional additive regression, sparse additive models, treatment effect-modifiers

1 Introduction

We propose a flexible approach to estimate the interaction effects between a treatment variable and pretreatment covariates on a treatment response, allowing both scalar-valued and functional-valued pretreatment covariates. Recent advances in biomedical imaging, mass spectrometry, and high-throughput gene expression technology produce massive amounts of data on individual patients, and open up the possibility of tailoring treatments to the biosignatures of individual patients from individual-specific data (McKeague and Qian, 2014). Notably, some randomized clinical trials (e.g., Trivedi \textit{et al.}, 2016) are designed to discover biosignatures that characterize patient heterogeneity in treatment responses from vast amounts of patient pretreatment profiles. In such studies, the objective is to make individualized treatment decision rules (ITRs) based on pretreatment patient characteristics such as genetic information (e.g., van’t Veer and Bernards, 2008) and brain structure and function measured from neuroimaging modalities such as magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), among others. Our interest here is in some specific types of high dimensional patient characteristics that are observed in the form of curves or images. Such data can be viewed as functional (e.g., Ramsay and Silverman, 1997), and such data are becoming increasingly prevalent in modern randomized clinical trials as pretreatment covariates.

Much work has been carried out to develop methods for optimizing ITRs using data from randomized clinical trials. Regression-based methodologies are intended to optimize ITRs by estimating treatment-specific mean response functions (e.g., Gunter \textit{et al.}, 2011; Jeng \textit{et al.}, 2018; Lu \textit{et al.}, 2011; Qian and Murphy, 2011; Shi \textit{et al.}, 2016; Tian \textit{et al.}, 2014; Zhang \textit{et al.}, 2012) while attempting to maintain robustness with respect to...
model misspecification. Machine learning approaches for developing ITRs are often framed in the context of classification problems (Zhang et al., 2012; Zhao et al., 2019); for example, outcome weighted learning (e.g., Song et al., 2015; Zhao et al., 2012, 2015) based on support vector machines, tree-based classification (e.g., Laber and Zhao, 2015), and adaptive boosting (Kang et al., 2014), among others. However, to date there has been relatively little research on methods for developing ITRs based on pretreatment functional covariates. McKeage and Qian (2014) propose methods for estimating and evaluating treatment regimes that depend upon a single pretreatment functional covariate. Ciarleglio et al. (2015, 2018) proposed methods that allow ITRs to depend on multiple functional/scalar covariates while performing both estimation and covariate selection for ITRs, when a functional linear model is assumed for the treatment and covariate interactions. Ciarleglio et al. (2016) consider a flexible functional regression approach to estimating ITRs based upon a single pretreatment functional covariate.

In this paper, we propose a flexible functional regression method that allows multiple functional/scalar covariates and performs both estimation and covariate selection for ITRs, without restricting to a linear underlying model. This approach generalizes the method of Ciarleglio et al. (2018) to allow for nonlinear underlying relationships between treatment and pretreatment covariates in their effects on the treatment response and also for more than two treatment conditions. Specifically, we develop a variant of a functional additive regression model of Fan et al. (2014, 2015), uniquely constrained to model possibly nonlinear interaction effects between treatment and pretreatment functional/scalar covariates. A sparse nonlinear combination of 1-dimensional projections of the functional covariates is derived via a sparse additive model formulation (Fan et al., 2014, 2015; Ravikumar et al., 2009). This approach results in a simple yet flexible functional regression model useful for estimating treatment-by-covariates interactions and developing ITRs.

2 Constrained functional additive models

We consider a treatment response variable \( Y \in \mathbb{R} \), a set of \( p \) functional-valued covariates \( X = (X_1, \ldots, X_p) \), and a set of \( q \) scalar-valued covariates \( Z = (Z_1, \ldots, Z_q) \in \mathbb{R}^q \). We will assume that each \( X_j \) is a square integrable random function defined on a compact interval, taken to be \([0, 1]\) without loss of generality. Suppose the treatment \( A \in \{1, \ldots, L\} \) (i.e., there are \( L \) available treatment conditions) is assigned with associated known probabilities \((\pi_1, \ldots, \pi_L)\), such that \( \sum_{i=1}^{L} \pi_i = 1 \) and \( \pi_a > 0 \), independent of all pretreatment covariates. We note that this is readily extended to the case in which treatment assignment is correlated with the covariates, in which case the probabilities \((\pi_1, \ldots, \pi_L)\) should be replaced by the propensity scores \((\pi_1(X, Z), \ldots, \pi_L(X, Z))\), that would typically have to be estimated from observed data.

In this context we focus on the problem of estimating the interactions between the treatment variable \( A \) and a set of functional/scalar covariates \((X, Z)\), on their effects on \( Y \). For a single decision point, an ITR, which we denote as \( D \), maps an individual with pretreatment characteristics \((X, Z)\) to one of the treatment options in \( \{1, \ldots, L\} \) for a treatment recommendation. One natural measure for the effectiveness of an ITR \( D \) is the so-called “value” \((V)\) function (Murphy, 2005), \( V(D) = \mathbb{E}[E(Y|X, Z, A = D(X, Z))] \), which is the expected treatment response under the given treatment assignment regime \( D \). If we assume that a larger value of \( Y \) is better (without loss of generality), then the optimal ITR, which we denote as \( D^{opt} \), can be naturally defined to be the rule \( D \) that maximizes the “value” \( V(D) \). It can be easily shown that such an optimal rule \( D^{opt} \) satisfies: \( D^{opt}(X, Z) = \arg\max_{a \in \{1, \ldots, L\}} \mathbb{E}[Y|X, Z, A = a] \). Qian and Murphy (2011) noted that \( D^{opt}(X, Z) \) depends only on the \((X, Z)\)-by-\( A \) interaction effect, and does not depend on the “main” effect of \((X, Z)\) on \( Y \). Therefore, in terms of developing ITRs, the \((X, Z)\) main effect on \( Y \), i.e., \( \mathbb{E}[Y|X, Z] \), is considered as “nuisance”.

Our approach to developing ITRs is then to express the conditional expectation function \( \mathbb{E}[Y|X, Z, A] \) in terms of a main effect for \((X, Z)\) (represented by a function \( \mu \)) and an \( A \)-by-\((X, Z)\) interaction effect (represented by a function \( \theta \)), and focus on the estimation of the interaction effect, as in advantage learning (A-learning; Robins, 2004). Consider the following decomposition of the conditional expectation:

\[
\mathbb{E}[Y|X, Z, A] = \mu(X, Z) + \theta(X, Z, A),
\]

(1)

\((X, Z)\) “main” effect \( A \)-by-\((X, Z)\) interactions
where the first term $\mu(\mathbf{X}, \mathbf{Z})$ does not depend on $A$ and only the second term $\theta(\mathbf{X}, \mathbf{Z}, A)$ is a function of $A$. Under representation (1), the marginal effect of $(\mathbf{X}, \mathbf{Z})$ on $Y$, i.e., $E[Y|\mathbf{X}, \mathbf{Z}]$, is expressed as:

$$E[Y|\mathbf{X}, \mathbf{Z}] = E[E[Y|\mathbf{X}, \mathbf{Z}, A]|\mathbf{X}, \mathbf{Z}] = E[\mu(\mathbf{X}, \mathbf{Z}) + \theta(\mathbf{X}, \mathbf{Z}, A)|\mathbf{X}, \mathbf{Z}] = \mu(\mathbf{X}, \mathbf{Z}) + E[\theta(\mathbf{X}, \mathbf{Z}, A)|\mathbf{X}, \mathbf{Z}]$$  \hspace{1cm} (2)

In what follows, for the identifiability of decomposition (1), we will set:

$$E[\theta(\mathbf{X}, \mathbf{Z}, A)|\mathbf{X}, \mathbf{Z}] = 0,$$  \hspace{1cm} (3)

which implies, from (2), $E[Y|\mathbf{X}, \mathbf{Z}] = \mu(\mathbf{X}, \mathbf{Z})$, i.e., the term $\mu(\mathbf{X}, \mathbf{Z})$ in (1) represents the $(\mathbf{X}, \mathbf{Z})$ marginal effect, and the second term $\theta(\mathbf{X}, \mathbf{Z}, A)$ in (1) represents the “pure” $A$-by-$(\mathbf{X}, \mathbf{Z})$ interaction effect. We shall focus only on the “pure” interaction effect term $\theta(\mathbf{X}, \mathbf{Z}, A)$ in (1), since the term $\mu(\mathbf{X}, \mathbf{Z})$ in (1) can be viewed a “nuisance” for the purpose of optimizing ITRs. In particular, in this paper, we shall represent $\theta(\mathbf{X}, \mathbf{Z}, A)$ based on a set of functional additive regression models (FAM) of Fan et al. (2015):

$$\theta(\mathbf{X}, \mathbf{Z}, A = a) = \sum_{j=1}^{p} g_{j,a}(<\beta_j, X_j>) + \sum_{k=1}^{q} h_{k,a}(Z_k) \quad (a = 1, \ldots, L),$$  \hspace{1cm} (4)

where the treatment level $a$-specific component functions $\{g_{j,a}(\cdot), j = 1, \ldots, p\}$ and $\{h_{k,a}(\cdot), k = 1, \ldots, q\}$ are general square-integrable 1-dimensional (1-D) functions defined on compact intervals on $\mathbb{R}$. Each functional covariate $X_j(\cdot)$ appears in model (4) as a 1-D projection $<\beta_j, X_j> := \int_{0}^{1} \beta_j(s)X_j(s)ds \in \mathbb{R}$, through an inner product with a square integrable coefficient function $\beta_j(\cdot)$ defined on $[0, 1]$. Due to the unspecified nature of the component function $g_{j,a}(\cdot)$ in (4), for model identifiability, we assume that the coefficient functions $\beta_j(\cdot)$ have a unit $L^2$ norm; let $\Theta$ denote such a space of functions defined over $[0, 1]$, and assume $\beta_j \in \Theta$ (without loss of generality).

Under the general framework (1), model (4) for the $A$-by-$(\mathbf{X}, \mathbf{Z})$ interactions gives the following model:

$$E[Y|\mathbf{X}, \mathbf{Z}, A = a] = \underbrace{\mu(\mathbf{X}, \mathbf{Z})}_{(\mathbf{X}, \mathbf{Z}) \text{ “main” effect}} + \underbrace{\sum_{j=1}^{p} g_{j,a}(<\beta_j, X_j>) + \sum_{k=1}^{q} h_{k,a}(Z_k)}_{A\text{-by-}(\mathbf{X}, \mathbf{Z}) \text{ interactions}} \quad (a = 1, \ldots, L)$$  \hspace{1cm} (5)

where the system of $L$ models (4) $\theta(\mathbf{X}, \mathbf{Z}, A = a)$ $(a = 1, \ldots, L)$ determines the $A$-by-$(\mathbf{X}, \mathbf{Z})$ interaction effect, and the term $\mu(\mathbf{X}, \mathbf{Z})$ represents an unspecified “main” effect of $(\mathbf{X}, \mathbf{Z})$, whose effect does not depend on the treatment variable $A$.

As in the general model (1), we impose the identifiability constraint (3) on model (4). Given the additive model representation of $\theta(\mathbf{X}, \mathbf{Z}, A)$ for each treatment level $A = a$ $(a = 1, \ldots, L)$, it is straightforward to verify that the following condition on the component functions $\{g_{j,a}(\cdot)\}_{a \in \{1, \ldots, L\}}$ and $\{h_{k,a}(\cdot)\}_{a \in \{1, \ldots, L\}}$ of model (4):

$$E[g_{j,a}(<X_j, \beta_j)>|X_j] = 0 \quad (\forall \beta_j \in \Theta) \quad (j = 1, \ldots, p) \quad \text{and} \quad E[h_{k,a}(Z_k)|Z_k] = 0 \quad (\forall k = 1, \ldots, q)$$  \hspace{1cm} (6)

satisfies the identifiability condition (3) of decomposition (1), i.e., imposition of (6) on the proposed model (5) results in: $E[\theta(\mathbf{X}, \mathbf{Z}, A)|\mathbf{X}, \mathbf{Z}] = E[\sum_{j=1}^{p} g_{j,a}(<\beta_j, X_j>) + \sum_{k=1}^{q} h_{k,a}(Z_k)|\mathbf{X}, \mathbf{Z}] = 0$ (almost surely) and satisfies the identifiability condition (3). We call model (5), subject to the constraint (6), a constrained functional additive model (CFAM) for the $A$-by-$(\mathbf{X}, \mathbf{Z})$ interaction effects. For model (5), we assume an additive noise structure $Y = E[Y|\mathbf{X}, \mathbf{Z}, A] + \epsilon$, where $\epsilon \in \mathbb{R}$ is a zero-mean random variable with finite variance, independent of $\mathbf{X}$, $\mathbf{Z}$ and $A$.

Notation. Throughout, we use the notation $<\cdot, \cdot>$ to denote the inner product between two square integrable functions. We consider sets of random variables $(A, <\beta, X>)$ and measurable functions $g(\cdot)$ on $(A, <\beta, X>)$, with the $L^2$ norm of the function $g(\cdot)$ defined as $\|g\| = \sqrt{E[|g|^2(\langle \beta, X \rangle)]]}$ where the expectation is taken with
defined. Sometimes, for the notational simplicity, we also write \( g = \langle \cdot \rangle \) with respect to the joint distribution of \((A, Z)\) (which depends on \( \beta \)). Similarly, we consider sets of random variables \((A, Z)\) and measurable functions \( h_{\cdot, \cdot}(\cdot) \) on \((A, Z)\), with the \( L^2 \) norm of the function \( h_{\cdot, \cdot}(\cdot) \) defined as \( \|h\| = \sqrt{\mathbb{E}[h^2(Z)]} \) where the expectation is taken with respect to the joint distribution of \((A, Z)\). For a given \( \beta \), let us denote the \( L^2 \) space of functions on the variables \((A, \beta, X)\) as \( \mathcal{H}(\beta) = \{ g \mid \|g\|_{A(\beta, X)} = 0, \|g\| < \infty \} \), with inner product on the space defined as \( (g, f) = \mathbb{E}[gA(\beta, X)fA(\beta, X)] \). Similarly, let us denote the \( L^2 \) space of functions on \((A, Z)\) as \( \mathcal{H} = \{ h \mid \|h\|_{A(Z)} = 0, \|h\| < \infty \} \), with inner product on the space similarly defined. Sometimes, for the notational simplicity, we also write \( g := gA(\beta, X) \) (and also \( h := hA(Z) \)).

For CFAM (5), we can specify the “true” underlying functions of interest, which we denote as \( g_j^* := g_j^*(j = 1, \ldots, p) \) and \( h_k^* := h_k^*(k = 1, \ldots, q) \), associated with the \( A \)-by-(\( X, Z) \) interaction effect terms, as the solution to the following constrained optimization:

\[
\begin{align*}
\{g_j^*, \beta_j^*, h_k^*\} &= \argmin_{g_j \in H_j^{(\beta)}, \beta_j \in \Theta, h_k \in H_k} E\left[ \left\{ Y - \mu(X, Z) - \sum_{j=1}^p g_jA(\langle X_j, \beta_j \rangle) - \sum_{k=1}^q h_{k,A}(Z_k) \right\}^2 \right] \\
&\quad \text{subject to} \quad E[g_jA(\langle X_j, \beta_j \rangle)] = 0 \quad \forall \beta_j \in \Theta \quad (j = 1, \ldots, p) \quad \text{and} \quad E[h_{k,A}(Z_k)] = 0 \quad (k = 1, \ldots, q),
\end{align*}
\]

where \( \mu(X, Z) \) is the true “main” effect function specified in model (5) (and is considered as fixed in (7)). Since the minimization in (7) is in terms of \( \{g_j, \beta_j, h_k\} \), the right-hand side of (7) can be reduced to (by expanding the first line, i.e., by expanding the expected squared error term in (7)):

\[
\begin{align*}
\argmin_{g_j \in H_j^{(\beta)}, \beta_j \in \Theta, h_k \in H_k} E\left[ \left\{ Y - \sum_{j=1}^p g_jA(\langle X_j, \beta_j \rangle) - \sum_{k=1}^q h_{k,A}(Z_k) \right\}^2 + 2\mu(X, Z)\left\{ \sum_{j=1}^p g_jA(\langle X_j, \beta_j \rangle) + \sum_{k=1}^q h_{k,A}(Z_k) \right\} \right] \\
&= \argmin_{g_j \in H_j^{(\beta)}, \beta_j \in \Theta, h_k \in H_k} E\left[ \left\{ Y - \sum_{j=1}^p g_jA(\langle X_j, \beta_j \rangle) - \sum_{k=1}^q h_{k,A}(Z_k) \right\}^2 + 2\mu(X, Z)E\left\{ \sum_{j=1}^p g_jA(\langle X_j, \beta_j \rangle) + \sum_{k=1}^q h_{k,A}(Z_k) \mid X, Z \right\} \right] \\
&= \argmin_{g_j \in H_j^{(\beta)}, \beta_j \in \Theta, h_k \in H_k} E\left[ \left\{ Y - \sum_{j=1}^p g_jA(\langle X_j, \beta_j \rangle) - \sum_{k=1}^q h_{k,A}(Z_k) \right\}^2 \right],
\end{align*}
\]

in which the first equality follows from an application of the iterated expectation rule to condition on \((X, Z)\), and the second equality is the result of the constraints imposed in (7): \( E[g_jA(\langle X_j, \beta_j \rangle)] = 0 \quad \forall \beta_j \in \Theta \quad (j = 1, \ldots, p) \) and \( E[h_{k,A}(Z_k)] = 0 \quad (k = 1, \ldots, q) \). Therefore, the optimization-based representation (7) for the “true” underlying functions \( \{g_j^*, \beta_j^*, h_k^*\} \) of CFAM (5) can be simplified to:

\[
\begin{align*}
\{g_j^*, \beta_j^*, h_k^*\} &= \argmin_{g_j \in H_j^{(\beta)}, \beta_j \in \Theta, h_k \in H_k} E\left[ \left\{ Y - \sum_{j=1}^p g_jA(\langle X_j, \beta_j \rangle) - \sum_{k=1}^q h_{k,A}(Z_k) \right\}^2 \right] \\
&\quad \text{subject to} \quad E[g_jA(\langle X_j, \beta_j \rangle)] = 0 \quad \forall \beta_j \in \Theta \quad (j = 1, \ldots, p) \quad \text{and} \quad E[h_{k,A}(Z_k)] = 0 \quad (k = 1, \ldots, q),
\end{align*}
\]

which does not involve the term \( \mu(X, Z) \) of (5). When the “nuisance” functional \( \mu(X, Z) \) in (5) is a complicated functional to specify correctly, utilizing the representation (8) for specifying the component functions \( \{g_j^*, j = 1, \ldots, p\} \cup \{h_k^*, k = 1, \ldots, q\} \) and the single-index coefficient functions \( \{\beta_j^*, j = 1, \ldots, p\} \) of interest is particularly appealing.

In regression models for a treatment outcome, treatment effect-modifiers are covariates that modify the effect of the treatment \( A \) on the outcome variable \( Y \). Under model (5), the potential treatment effect-modifiers among the covariates \( \{X_j, j = 1, \ldots, p\} \cup \{Z_k, k = 1, \ldots, q\} \), that associate the treatment variable \( A \) to the treatment outcome \( Y \), enter the model only through the \( A \)-by-(\( X, Z) \) interaction effect terms \( \sum_{j=1}^p g_{j,a}(\langle X_j, \beta_j \rangle) + \sum_{k=1}^q h_{k,a}(Z_k) \quad (a = 1, \ldots, L) \). Ravikumar et al. (2009) proposed a sparse additive model (SAM) for component selection in a high-dimensional additive regression on scalar-valued covariates.
As in SAM, to deal with a large \( p + q \) and to achieve treatment effect-modifier selection, we impose sparsity on the set of the component functions \( \{g_j, j = 1, \ldots, p; h_k, k = 1, \ldots, q\} \) in CFAM (5), under the often reasonable assumption that most covariates are inconsequential as treatment effect-modifiers. This sparsity structure on the set of component functions can be incorporated into representation (8):

\[
\{g_j^*, \beta_j^*, h_k^*\} = \arg\min_{g_j \in \mathcal{H}_j^{(\beta_j)}, \beta_j \in \Theta, h_k \in \mathcal{H}_k} E\left\{Y - \sum_{j=1}^p g_j,A(\langle X_j, \beta_j \rangle) - \sum_{k=1}^q h_k,A(Z_k)\right\}^2 + \lambda \left\{\sum_{j=1}^p \|g_j\| + \sum_{k=1}^q \|h_k\|\right\}
\]

subject to

\[
E[g_{j,A}(\langle X_j, \beta_j \rangle)|X_j] = 0 \quad \forall \beta_j \in \Theta \quad (j = 1, \ldots, p) \quad \text{and}
\]

\[
E[h_{k,A}(Z_k)|Z_k] = 0 \quad (k = 1, \ldots, q),
\]

for some sparsity-inducing parameter \( \lambda \geq 0 \). The term \( \sum_{j=1}^p \|g_j\| + \sum_{k=1}^q \|h_k\| \) in (9) behaves like an \( L^1 \) ball across different functional components \( \{g_j, j = 1, \ldots, p; h_k, k = 1, \ldots, q\} \) to encourage functional sparsity. For example, a relatively large value of \( \lambda \) in (9) will result in many component functions to be exactly zero, thereby enforcing sparsity on the set of functions \( \{g_j^*, j = 1, \ldots, p; h_k^*, k = 1, \ldots, q\} \).

3 Estimation

We first consider a population characterization of the algorithm for solving (9) in Section 3.1 and then consider a sample counterpart of the population algorithm in Section 3.2.

3.1 Population algorithm

For a set of fixed coefficient functions \( \{\beta_j, j = 1, \ldots, p\} \), the minimizing component function \( g_j \in \mathcal{H}_j^{(\beta_j)} \) (and \( h_k \in \mathcal{H}_k \)) for each \( j \) (and each \( k \)) of the constrained objective function of (9) has a component-wise closed-form expression.

**Theorem 1.** Given \( \lambda \geq 0 \) and a set of fixed single-index coefficient functions \( \{\beta_j, j = 1, \ldots, p\} \), the minimizing component function \( g_j \in \mathcal{H}_j^{(\beta_j)} \) of the constrained objective function of (9) satisfies:

\[
g_{j,A}(\langle X_j, \beta_j \rangle) = \left[1 - \frac{\lambda}{\|f_j\|}\right] f_{j,A}(\langle X_j, \beta_j \rangle) \quad \text{almost surely},
\]

where the function \( f_j \in \mathcal{H}_j^{(\beta_j)} \):

\[
f_{j,A}(\langle X_j, \beta_j \rangle) := \mathbb{E}[R_j|A, \langle X_j, \beta_j \rangle] = \mathbb{E}[R_j|\langle X_j, \beta_j \rangle],
\]

in which

\[
R_j = Y - \sum_{j' \neq j} g_{j',A}(\langle X_j, \beta_j \rangle) - \sum_{k=1}^q h_{k,A}(Z_k)
\]

represents the \( j \)th (functional covariate’s) partial residual. Similarly, the minimizing component function \( h_k \in \mathcal{H}_k \) of the constrained objective function of (9) satisfies:

\[
h_{k,A}(Z_k) = \left[1 - \frac{\lambda}{\|f_k\|}\right] \hat{f}_{k,A}(Z_k) \quad \text{almost surely},
\]

where the function \( \hat{f}_k \in \mathcal{H}_k \):

\[
\hat{f}_{k,A}(Z_k) := \mathbb{E}[\hat{R}_k|A, Z_k] = \mathbb{E}[\hat{R}_k|Z_k],
\]

in which

\[
\hat{R}_k = Y - \sum_{j=1}^p g_{j,A}(\langle X_j, \beta_j \rangle) - \sum_{k' \neq k} h_{k',A}(Z_{k'})
\]
represents the kth (scalar covariate’s) partial residual. (In (10) and (13), \([u]_+ = \max(0,u)\) represents the positive part of \(u\).)

The proof of Theorem 1 is in the Supplementary Material. Given a sparsity tuning parameter \(\lambda \geq 0\), optimization (9) can be split into two iterative steps (Fan et al., 2014, 2015). First (Step 1), for a set of fixed single-indices \(\{X_j, \beta_j\} (j = 1, \ldots, p)\), the component functions \{\(g_j, j = 1, \ldots, p\) \cup \{\(h_k, k = 1, \ldots, q\)\} of the model can be found by a coordinate descent procedure that fixes \{\(g_j; j' \neq j\)\} \cup \{\(h_k, k = 1, \ldots, q\)\} and obtains \(g_j\) by equation (10) (and that fixes \{\(g_j, j = 1, \ldots, p\) \cup \{\(h_k; k' \neq k\)\} and obtains \(h_k\) by equation (13)), and then iterates through all \(j\) and \(k\) until convergence. This step (Step 1) amounts to fitting a sparse additive model (Ravikumar et al., 2009) subject to the constraint (6). Second (Step 2), for a set of fixed component functions \{\(g_j, j = 1, \ldots, p\) \cup \{\(h_k, k = 1, \ldots, q\)\}, the jth single-index coefficient function \(\beta_j \in \Theta\) can be optimized by solving, for each \(j \in \{1, \ldots, p\}\) separately:

\[
\min_{\beta_j \in \Theta} \mathbb{E} \left[ \left( R_j - g_{j,A}(\langle X_j, \beta_j \rangle) \right)^2 \right] \quad (j = 1, \ldots, p). \tag{16}
\]

where the jth partial residual \(R_j\) is defined in (12). These two steps can be iterated until convergence to obtain a population solution \{\(g_j^*, \beta_j^*, h_k^*\)\} on the left-hand side of (9).

To obtain a sample version of the population solution, we can insert sample estimates into the population algorithm, as in standard backfitting in estimating generalized additive models (Hastie and Tibshirani, 1999), which we describe in the next subsection.

### 3.2 Sample version of the population algorithm

Only to simplify the exposition, we describe the optimization of the components \(g_{j,A}(\langle X_j, \beta_j \rangle) (j = 1, \ldots, p)\) associated with the functional covariates \(X_j (j = 1, \ldots, p)\), only. The regression components \(h_{k,A}(Z_k) (k = 1, \ldots, q)\) associated with the scalar covariates \(Z_k (k = 1, \ldots, q)\) in (9) are optimized in the same way, except that we do not need to perform Step 2 of the alternating optimization procedure; that is, when optimizing \(h_{k,A}(Z_k) (k = 1, \ldots, q)\), we only perform Step 1.

#### 3.2.1 Step 1

First, we consider a sample version of Step 1 of the population algorithm. Suppose we are given a set of estimates \{\(\hat{\beta}_j, j = 1, \ldots, p\)\} and the data-version of the jth partial residual \(R_j\) in (12): \(\hat{R}_j = Y_i - \sum_{j' \neq j} \hat{g}_{j',A}(\langle X_{j'}, \beta_{j'} \rangle) - \sum_{k=1}^q \hat{h}_k(Z_{ik}) (i = 1, \ldots, n)\), where \(\hat{g}_{j'}\) represents a current estimate for \(g_{j'}\) and \(\hat{h}_k\) that for \(h_k\). For each \(j\), we update the component function \(\hat{g}_j\) in (10) in two steps: first, estimate the function \(f_j\) in (11); second, plug the estimate of \(f_j\) into \([1 - \frac{\lambda}{\|f_j\|}]_+\) in (10), to obtain the soft-thresholded estimate \(\hat{g}_j\).

Although any linear smoothers for flexible regression methods can be utilized to obtain estimates \{\(\hat{g}_j, j = 1, \ldots, p\)\}, we shall focus on regression spline-type estimators, which are particularly simple and computationally efficient for optimizing the proposed CFAM involving the constraint in (9). For each \(j\), represent the function \(g_j \in \mathcal{H}_j^{(\beta_j)}\) on the right-hand side of (9) as:

\[
g_{j,a}(\langle X_j, \hat{\beta}_j \rangle) = \Psi_j(\langle X_j, \hat{\beta}_j \rangle)^\top \theta_{j,a} \quad (a = 1, \ldots, L) \tag{17}
\]

for some \(d_j\)-dimensional basis function \(\Psi_j(\cdot) \in \mathbb{R}^{d_j}\) (e.g., B-spline basis on a set of evenly spaced knots on a compact interval covering the observed values of \(\langle X_j, \hat{\beta}_j \rangle\)), with a set of \(d_j\)-dimensional unknown basis coefficients \{\(\theta_{j,a}\)\}_{a \in \{1, \ldots, L\}} specific to treatment conditions \(A = a (a = 1, \ldots, L)\). Given representation (17) for the jth component function \(g_j \in \mathcal{H}_j^{(\beta_j)}\), the constraint \(\mathbb{E}[g_{j,A}(\langle X_j, \beta_j \rangle)|X_j] = 0\) in (9) on \(g_j\), for fixed \(\beta_j = \hat{\beta}_j\), can be simplified to: \(\mathbb{E}[\theta_{j,A}] = \sum_{a=1}^L \pi_a \theta_{j,a} = 0\).
Let $\theta_j := (\theta_{j,1}^T, \theta_{j,2}^T, \ldots, \theta_{j,L}^T)^T \in \mathbb{R}^{d_j L}$ be a vectorized version of the treatment $a$-specific basis coefficients \{\theta_{j,a}\}_{a \in \{1, \ldots, L\}}$, and let us introduce the $d_j \times d_j$ matrix $\pi^{(j)} := (\pi_1 I_{d_j}; \pi_2 I_{d_j}; \ldots; \pi_L I_{d_j})$ in which $I_{d_j}$ is the $d_j \times d_j$ identity matrix. If we fix $\beta_j = \hat{\beta}_j$, the constraint in (9) on the function $g_j$ can then be written in matrix form:

$$\pi^{(j)} \theta_j = 0.$$  

(18)

Given $\beta_j = \hat{\beta}_j$, the restriction of $g_j \in \mathcal{H}(\hat{\beta}_j)$ to the form (17) restricts also the minimizer $g_j$ in (10) to have the form (17). In particular, the function $f_j$ in (11), if we fix $\beta_j = \hat{\beta}_j$, is then given by:

$$f_{j,A}(\langle X_j, \hat{\beta}_j \rangle) = \mathbb{E}[R_j | \langle X_j, \hat{\beta}_j \rangle, A] - \sum_{a=1}^L \pi_a \mathbb{E}[R_j | \langle X_j, \hat{\beta}_j \rangle, A = a]$$

$$= \Psi_j(\langle X_j, \hat{\beta}_j \rangle) \theta_{j,A} - \Psi_j(\langle X_j, \hat{\beta}_j \rangle) \{\sum_{a=1}^L \pi_a \theta^{*}_{j,a}\},$$

(19)

where $\{\theta^{*}_{j,a}\}_{a \in \{1, \ldots, L\}} := \arg \min_{\{\theta_{j,a} \in \mathbb{R}^d\}_{a \in \{1, \ldots, L\}}} \mathbb{E} \left[ \{R_j - \Psi_j(\langle X_j, \hat{\beta}_j \rangle)^\top \theta_{j,A}\}^2 \right]$. In (19), the first term $\Psi_j(\langle X_j, \hat{\beta}_j \rangle) \theta_{j,A}$ corresponds to the $L^2$ projection of the $j$th partial residual $R_j$ in (12) onto the class of functions of the form (17) (without the imposition of the constraint $\sum_{a=1}^L \pi_a \theta_{j,a} = 0$), whereas the second term $-\Psi_j(\langle X_j, \hat{\beta}_j \rangle) \{\sum_{a=1}^L \pi_a \theta^{*}_{j,a}\}$ simply centers the first term to satisfy the linear constraint $\sum_{a=1}^L \pi_a \theta_{j,a} = 0$. Note, it can be easily shown that the function $f_j$, as specified in the second line on the right-hand side of (19), corresponds to the $L^2$ projection of $R_j$ onto the subspace of measurable functions of the form (17) subject to the linear constraint (18).

Let the $n \times d_j$ matrices $D_{j,a}$ denote the evaluation matrices of the basis function $\Psi_j(\cdot)$ in (17) on $\langle X_{ij}, \hat{\beta}_j \rangle$ ($i = 1, \ldots, n$) specific to the treatment level $A = a$, whose ith row is the $1 \times d_j$ vector $\Psi_j(\langle X_{ij}, \hat{\beta}_j \rangle)^\top$ if $A_i = a$, and a row of zeros $0^d$ if $A_i \neq a$. Then the column-wise concatenation of the design matrices $\{D_{j,a}\}_{a \in \{1, \ldots, L\}}$, i.e., the $n \times d_j L$ matrix $D_j = (D_{j,1}; D_{j,2}; \ldots; D_{j,L})$, defines the model matrix associated with the vectorized basis coefficient $\theta_j = (\theta_{j,1}^\top, \theta_{j,2}^\top, \ldots, \theta_{j,L}^\top)^\top \in \mathbb{R}^{d_j L}$. Then we can represent the function $g_{j,A}(\langle X_j, \hat{\beta}_j \rangle)$ in (17), based on the sample data, by the length-$n$ vector:

$$g_j = D_j \theta_j \in \mathbb{R}^n$$

(20)

subject to the linear constraint (18) on $\theta_j$. (Similarly, we can represent $h_{k,A}(Z_k)$ by a length-$n$ vector.)

When computing the data version of the function $f_j$ in (11), which corresponds to the projection of $R_j$ onto the class of functions (17) subject to (18) for a given $\beta_j = \hat{\beta}_j$, the linear constraint in (18) on $\theta_j$ can be absorbed into the model matrix $D_j$ in (20) by reparametrization, which we describe next. We can create a $d_j L \times d_j (L-1)$ matrix $Z^{(j)}$, such that, for any arbitrary vector $\tilde{\theta}_j \in \mathbb{R}^{d_j (L-1)}$ if we set $\theta_j = Z^{(j)} \tilde{\theta}_j \in \mathbb{R}^{d_j L}$, then the vector $\theta_j$ automatically satisfies the constraint (18): $\pi^{(j)} \theta_j = 0$. Such a basis matrix $Z^{(j)}$, which spans the null space of the linear constraint (18), can be constructed by a QR decomposition of the matrix $\pi^{(j)\top}$. Then representation (20) can be reparametrized, in terms of the unconstrained vector $\tilde{\theta}_j \in \mathbb{R}^{d_j (L-1)}$ by replacing $D_j$ in (20) with the reparametrized model matrix $\tilde{D}_j = D_j Z^{(j)}$:

$$g_j = \tilde{D}_j \tilde{\theta}_j.$$  

(21)

Theorem 1 indicates that (for fixed $\beta_j = \hat{\beta}_j$) the coordinate-wise minimizing function $g_j$ of the right-hand side of (9) can be estimated based on the sample data by

$$\hat{g}_j = \left[1 - \frac{\lambda}{\sqrt{\pi^2 \| \hat{f}_j \| ^2}}\right] \hat{f}_j$$

(22)

where $\hat{f}_j = \tilde{D}_j (\tilde{D}_j^\top \tilde{D}_j)^{-1} \tilde{D}_j^\top \hat{R}_j$, in which $\hat{R}_j = Y - \sum_{j' \neq j} \hat{g}_{j'} - \sum_{k=1}^q \hat{h}_k$ is the estimated $j$th partial residual vector. (Similarly, we can represent the coordinate-wise minimizing function $h_k$ in (13), based on
the observed data by a length-$n$ vector $\hat{h}_k$.) Based on the sample counterpart (22) of the coordinate-wise solution (10), if we set each $\beta_j$ equal to its corresponding estimate $\hat{\beta}_j$ ($j = 1, \ldots, p$), a highly efficient (unconstrained) coordinate descent algorithm can be conducted to simultaneously optimize all component functions $\{g_j, j = 1, \ldots, p\} \cup \{h_k, k = 1, \ldots, q\}$. In (22), let $\hat{s}_j^{(A)} := \left[1 - \lambda \sqrt{n}/\|\hat{f}_j\|\right]_+$ denote the soft-threshold shrinkage factor associated with the un-shrunken estimate $\hat{f}_j$. Upon convergence of the coordinate descent, we obtain an estimate of the coefficient vector $\hat{\theta}_j$ in (21):

$$\hat{\theta}_j = \hat{s}_j^{(A)}(D_j^T \tilde{D}_j)^{-1} \tilde{D}_j^T \hat{R}_j$$

(23)

which in turn implies an estimate, $\hat{\theta}_j$, of the coefficient vector $\theta_j$ in the original representation (20) for $g_j$:

$$\hat{\theta}_j = (\hat{\theta}_{j,1}, \hat{\theta}_{j,2}, \ldots, \hat{\theta}_{j,L})^T = Z^{(j)} \hat{\eta}_j$$

which gives an estimate of the treatment $a$-specific function $g_{j,a}(\cdot)$ ($a = 1, \ldots, L$):

$$\hat{g}_{j,a}(\cdot) = \Psi_j(\cdot)^\top \hat{\eta}_{j,a} \quad (a = 1, \ldots, L)$$

(24)

estimated within the class of functions (17), given a tuning parameter $\lambda \geq 0$ controlling the soft-threshold shrinkage factor $\hat{s}_j^{(A)}$ in (23). By performing the coordinate-descent with the component-wise update rule (22), we can obtain $\{\hat{g}_j, j = 1, \ldots, p\} \cup \{\hat{h}_k, k = 1, \ldots, q\}$; this completes Step 1 of the alternating optimization procedure.

**Remark 1.** Any scatterplot smoother can be utilized to obtain the sample counterpart (22) of the coordinate-wise solution (10), i.e., estimation of (10) is not restricted to regression splines. To estimate the function $f_j$ in (11), we can estimate the system of treatment $a$-specific functions $\mathbb{E}[R_j(\hat{\beta}_j, X_{ij})]$, $A = a$ ($a = 1, \ldots, L$) (which corresponds to the first term on the right-hand side of (11) if we fix $\beta_j = \hat{\beta}_j$), by performing separate nonparametric regressions of $\hat{R}_j$ on regressor $\langle \hat{\beta}_j, X_{ij} \rangle$ separately for each treatment condition $A = a$ ($a = 1, \ldots, L$). We can also estimate the function $-\mathbb{E}[R_j(\hat{\beta}_j, X_{ij})]$ (which corresponds to the second term $-\mathbb{E}[R_j(\hat{\beta}_j, X_{ij})]$ on the right-hand side of (11) if we fix $\beta_j = \hat{\beta}_j$), by performing a nonparametric regression of $\hat{R}_j$ on regressor $\langle \hat{\beta}_j, X_{ij} \rangle$. Adding these two function estimates provides an estimate for $f_j$ in (11). Evaluating this estimate of $f_j$ at the $n$ points $\langle \langle \hat{\beta}_j, X_{ij} \rangle, A_i \rangle$ ($i = 1, \ldots, n$) gives an estimate $\hat{f}_j \in \mathbb{R}^n$ in (22). Then we can compute the corresponding soft-threshold estimate $\hat{g}_j \in \mathbb{R}^n$ and conduct the coordinate descent procedure described in Algorithm 1.

**3.2.2 Step 2**

We now consider a sample version of Step 2 of the population algorithm that optimizes the coefficient functions $\{\beta_j, j = 1, \ldots, p\}$ on the right-hand side of (9), for a fixed set of the component function estimates $\{\hat{g}_j, j = 1, \ldots, p\} \cup \{\hat{h}_k, k = 1, \ldots, q\}$ provided by Step 1. As an empirical approximation to (16), we consider the optimization problem:

$$\minimize_{\hat{\beta}_j \in \Theta} \sum_{i=1}^n \left( \hat{R}_{ij} - \hat{g}_{j,A}(\langle X_{ij}, \beta_j \rangle) \right)^2 \quad (j = 1, \ldots, p),$$

(25)

where $\hat{R}_{ij}$ is the $i$th element of $\hat{R}_j \in \mathbb{R}^n$ in (22). For each iteration, the minimization (25) can be approximately achieved based on a first-order Taylor series approximation of the term $\hat{g}_{j,A}(\langle X_{ij}, \hat{\beta}_j \rangle)$ around a current estimate of the coefficient function, say, $\hat{\beta}_j^{(c)} \in \Theta$:

$$\sum_{i=1}^n \left( \hat{R}_{ij} - \hat{g}_{j,A}(\langle X_{ij}, \beta_j \rangle) \right)^2 \approx \sum_{i=1}^n \left( \hat{R}_{ij} - \hat{g}_{j,A}(\langle X_{ij}, \hat{\beta}_j^{(c)} \rangle) - \hat{g}_{j,A}(\langle X_{ij}, \hat{\beta}_j^{(c)} \rangle) \langle X_{ij}, \beta_j - \hat{\beta}_j^{(c)} \rangle \right)^2$$

$$= \sum_{i=1}^n \left( \hat{R}_{ij} - \langle X_{ij}, \beta_j \rangle \right)^2,$$

(26)
where the “modified” residuals $\hat{R}_{ij}^*$ and the “modified” regressors $X_{ij}^*$ are defined as:

$$
\begin{align*}
\hat{R}_{ij}^* &= \hat{R}_{ij} - \hat{g}_{j,A_i}((X_{ij}, \hat{\beta}_j^{(c)})) + \hat{g}_{j,A_i}((X_{ij}, \hat{\beta}_j^{(c)}) \langle X_{ij}, \hat{\beta}_j^{(c)} \rangle) \\
X_{ij}^* &= \hat{g}_{j,A_i}((X_{ij}, \hat{\beta}_j^{(c)})) X_{ij} 
\end{align*}
$$

Minimization of the right-hand side of (26) over $\beta_j \in \Theta$ can be performed by employing a standard functional linear regression (e.g., Cardot et al., 2003; Marx and Eilers, 1999) with scalar response $\hat{R}_{ij}^*$ and (functional) regressor $X_{ij}^*$. To elaborate on this briefly, we consider the $P$-splines functional linear regression approach of Marx and Eilers (1999), in which the coefficient function $\beta_j$ in (26) is assumed to be smooth and is represented by a prespecified (and normalized) $m_j$-dimensional $B$-spline basis function $B_j(s) = (b_{j1}(s), \ldots, b_{jm_j}(s))^\top \in \mathbb{R}^{m_j}$ ($s \in [0, 1]$),

$$
\beta_j(s) = \sum_{r=1}^{m_j} b_{jr}(s) \gamma_{jr} \quad s \in [0, 1].
$$

We impose a (2nd order) difference penalty by applying the associated $m_j \times m_j$ penalty matrix $S_j$ to the basis coefficient vector $\gamma_j = (\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jm_j})^\top \in \mathbb{R}^{m_j}$, to ensure appropriate smoothness. The basis $B_j$ in (28) uses equally-spaced knots and is rich enough to allow for the function $\beta_j$ in (26) to be sufficiently well represented by the basis expansion.

Suppose $X_{ij}$ ($i = 1, \ldots, n$) is observed without errors and is discretized at equidistant points $\{s_1: 0 = s_1 < s_2 < \ldots < s_{r_j-1} < s_{r_j} = 1\}$. Based on an approximation of the inner product $\langle X_{ij}, \beta_j^{(c)} \rangle \approx \Delta \sum_{i=1}^{r_j} X_{ij}(s_l) \hat{\beta}_j^{(c)}(s_l)$, in which $\Delta$ denotes the distance between two neighboring discretization points and a given estimate $\hat{g}_{j,a}()$ in (24) available from Step 1 (and the first derivative $\hat{g}_{j,a}'()$), we can easily approximate the quantity $\hat{g}_{j,A_i}((X_{ij}, \hat{\beta}_j^{(c)}))$, and hence compute the “modified” residuals $\hat{R}_{ij}^*$ and the “modified” regressors $X_{ij}^*$ in (27).

Let $X_j^*$ be the $n \times r_j$ matrix in which the $i$th row corresponds to the discretized modified regressor function $X_{ij}^*(s_l)$ ($l = 1, \ldots, r_j$). Let $B_j$ denote the $r_j \times m_j$ matrix, whose $l$th row ($l = 1, \ldots, r_j$) corresponds to the $B$-spline basis $B_j(s) \in \mathbb{R}^{m_j}$, evaluated at the $l$th point $s = s_l$. It follows that, based on the $P$-spline representation (28) for the function $\beta_j(s)$ discretized at the points $s = s_l$ ($l = 1, \ldots, r_j$), we can represent the right-hand side of (26) as:

$$
\|R_j^* - U_j^* \gamma_j\|^2 + \rho_j \gamma_j^\top S_j \gamma_j
$$

where $R_j^* = (R_{1j}^*, \ldots, R_{nj}^*)^\top \in \mathbb{R}^n$ and $U_j^* = \Delta X_j^* B_j$, for some smoothing parameter $\rho_j \geq 0$, which tunes the $P$- spline penalty. We choose the tuning parameter $\rho_j$ based on a generalized cross-validation (GCV) through a grid search. For a chosen $\rho_j$, the penalized least squares minimizer $\hat{\gamma}_j$ of the criterion (29) is scaled to $\|\hat{\gamma}_j\| = 1$, so that the resulting estimate $\hat{\beta}_j(s) = \sum_{r=1}^{m_j} B_{jr}(s) \hat{\gamma}_{jr}$ ($s \in [0, 1]$) satisfies $\hat{\beta}_j \in \Theta$. Minimizing the objective (29), subject to the scale constraint $\|\gamma_j\| = 1$, over $\gamma_j \in \mathbb{R}^d$ for each $j = 1, \ldots, p$ gives a set of estimates $\{\hat{\beta}_j, j = 1, \ldots, p\}$, for a given set $\{\hat{g}_{j}, j = 1, \ldots, p\} \cup \{\hat{h}_k, k = 1, \ldots, q\}$: this completes Step 2 of the alternating optimization procedure.

**Remark 2.** If $X_{ij}$ is either sparsely discretized, discretized at different points across subjects or observed with error, then an initial step for constructing the underlying function is required; the resulting reconstructed function is to be treated as the original function $X_{ij}$ and evaluated on a dense grid of argument values $\{s_i\}$.

### 3.2.3 Initialization and convergence criterion

At the initial iteration, we need some estimates of the single-index coefficient functions $\{\hat{\beta}_j, j = 1, \ldots, p\}$ to initialize the single-indices $\{u_j = \langle \hat{\beta}_j, X_j \rangle, j = 1, \ldots, p\}$, in order to perform Step 1 (i.e., the coordinate-descent procedure) of the estimation procedure described in Section 3.2.1. At the initial iteration, we take $\hat{\beta}_j(s) = 1$ ($s \in [0, 1]$), i.e., we take $u_j = \int_0^1 X_j(s) ds$ ($j = 1, \ldots, p$), which corresponds to the common practice of taking a naïve scalar summary of each functional covariate. The proposed algorithm that alternates between Step 1 and Step 2 terminates when the estimates $\{\hat{\beta}_j, j = 1, \ldots, p\}$ converge. To be specific, the algorithm terminates when $\max_{j \in [1, \ldots, p], r = 1, \ldots, m_j} \| (\hat{\gamma}_{jr} - \hat{\gamma}_{jr}^{(c)}) / \hat{\gamma}_{jr} \|$ is less than a pre-specified convergence
tolerance. Here, \( \hat{\gamma}_{jr}^{(c)} \) represents the current estimate of the coefficient \( \gamma_{jr} \) in (28) at the beginning of Step 1, and \( \hat{\gamma}_{jr} \) denotes the estimate at the end of Step 2. We summarize the computational procedure for solving (9) in Algorithm 1.

**Algorithm 1** Estimation of constrained functional additive models

1: **Input:** Data \( X_j \in \mathbb{R}^n \times \mathbb{R}^{r^j} \) \( (j = 1, \ldots, p) \), \( A \in \mathbb{R}^n \), \( Y \in \mathbb{R}^n \), and tuning parameter \( \lambda \geq 0 \)
2: **Output:** Estimated functions \( \{ \hat{\beta}_j, j = 1, \ldots, p \} \) and \( \{ \hat{g}_j, j = 1, \ldots, p \} \)
3: Initialize \( \hat{\beta}_j(s) = 1 \) \( (s \in [0, 1]) \) \( (j = 1, \ldots, p) \).
4: **while** until convergence of \( \{ \hat{\beta}_j, j = 1, \ldots, p \} \), **do** iteration between Step 1 and Step 2:
5: \( \langle \text{Step 1} \rangle \)
6: Fix \( \{ \hat{\beta}_j, j = 1, \ldots, p \} \), and compute \( \hat{D}_j(\hat{D}_j^\top \hat{D}_j)^{-1}\hat{D}_j^\top \) in (22) \( (j = 1, \ldots, p) \).
7: Initialize \( \hat{g}_j = 0 \) \( (j = 1, \ldots, p) \).
8: **while** until convergence of \( \{ \hat{g}_j, j = 1, \ldots, p \} \), **do** coordinate-descent iteration:
9: \( \langle \text{Step 2} \rangle \)
10: Update \( \hat{g}_j \) by expression (22).
11: Fix \( \{ \hat{g}_j, j = 1, \ldots, p \} \) in (24), and solve (25) based on (29); update \( \hat{\beta}_j \) \( (j = 1, \ldots, p) \).

In Algorithm 1, if the \( j \)-th soft-threshold shrinkage factor \( \hat{s}_j^{(\lambda)} = \left[ 1 - \lambda \sqrt{\bar{n}}/\|\hat{f}_j\| \right]_+ \) in (22) is 0, then the associated \( X_j \) is absent from the model. Therefore, the corresponding coefficient function \( \hat{\beta}_j \) will not be updated, and this greatly reduces the computational cost when most of the shrinkage factors \( \hat{s}_j^{(\lambda)} \) are zeros. In Algorithm 1, the smoother matrix \( \hat{D}_j(\hat{D}_j^\top \hat{D}_j)^{-1}\hat{D}_j^\top \) in (22) \( (j = 1, \ldots, p) \) needs to be computed only once at the beginning of Step 1 given fixed \( \{ \hat{\beta}_j, j = 1, \ldots, p \} \), and therefore the coordinate-descent updates in Step 1 can be performed very efficiently (Fan et al., 2014).

The sparsity tuning parameter \( \lambda \) can be chosen to minimize an estimate of the expected squared error of the estimated models over a dense grid of \( \lambda \)-s, estimated, for example, by a cross-validation. Alternatively, one can utilize the network information criterion (NIC; Murata and Amari, 1994) which is a generalization of the Akaike information criterion (AIC; Akaike, 1974) in approximating the prediction error, for the case in which the true underlying model, i.e., the assumed model (5), is not necessarily in the class of candidate models used to define the squared error criterion in (9).

4 Simulation study

In this section, we assess the optimal ITR estimation performance of the proposed method based on simulations.

4.1 ITR estimation performance

For the illustration in this sub-section, we generate \( n \) independent copies of \( p \) functional covariates \( X_i = (X_{i1}(s), \ldots, X_{ip}(s)) \) \( (i = 1, \ldots, n) \) based on \( X_{ij}(s_l) = \Phi(s_l)^\top \hat{x}_{ij} + u_{ij} \) \( (l = 1, \ldots, 50) \) \( (j = 1, \ldots, p) \), where each function \( X_{ij}(s) \) is observed at 50 equally spaced points \( \{ s_l \}_{l=1}^{50} \) between 0 and 1. To generate such functions \( X_{ij}(s) \), we use a 4-dimensional Fourier basis \( \Phi(s) = (\sqrt{2} \sin(2\pi s), \sqrt{2} \cos(2\pi s), \sqrt{2} \sin(4\pi s), \sqrt{2} \cos(4\pi s))^\top \in \mathbb{R}^4 \) \( (s \in [0, 1]) \), along with randomly generated Fourier basis coefficients \( \hat{x}_{ij} \in \mathbb{R}^4 \) that follow \( \mathcal{N}(0, I_4) \), and the measurement noise \( u_{ij} \in \mathbb{R} \) following \( \mathcal{N}(0, 0.25^2) \), all sampled independently from each other. In addition, we generate \( n \) independent copies of \( q \) scalar covariates \( Z_i = (Z_{i1}, \ldots, Z_{iq})^\top \in \mathbb{R}^q \) \( (i = 1, \ldots, n) \) based on the multivariate Gaussian distribution with each component having mean 0 and variance 1; correlations between
the components are given by \( \text{corr}(Z_{ij}, Z_{ik}) = 0.5 |j-k| \). We generate the responses \( Y_i \) \( (i = 1, \ldots, n) \) from:

\[
Y_i = \delta \left\{ 9 \sum_{j=2} \sin(\langle \eta_j, X_{ij} \rangle) + 9 \sum_{k=2} \sin(\langle Z_{ik} \rangle) \right\}
\]

\[
+ 2(-1)^d \left\{ \sin(\langle \beta_1, X_{i1} \rangle) + \sin(\langle \beta_2, X_{i2} \rangle) + \xi \sin(\langle X_{i1}, X_{i2} \rangle) + Z_{i1}/2 + \sin(Z_{i2}) + \xi \sin(Z_{i1}Z_{i2}) \right\} + \epsilon_i
\]

where we sample \( A_i \in \{1, 2\} \) with the distribution \( \text{pr}(A_i = 1) = \text{pr}(A_i = 2) = 0.5 \) independently of the covariates \( (X_{i1}, Z_{i}) \) and \( \epsilon_i \sim \mathcal{N}(0, 0.5^2) \). In model (30), there are only four “true” treatment effect-modifiers \( (X_{i1}, X_{i2}, Z_{i1} \text{and} Z_{i2}) \) which associate the treatment variable \( A_i \) with the response \( Y_i \). The other \( p + q - 4 \) covariates are “noise” covariates which are not useful for optimizing ITRs. In this example, we set \( p = q = 20 \), therefore we consider a total of 40 pretreatment covariates. In model (30), we set the coefficient functions, \( \beta_1 \) and \( \beta_2 \), associated with the \( A \)-by-\( X_j \) \( (j = 1, 2) \) interaction effect terms to be: \( \beta_1(s) = \Phi(s)^\top \{0.5, 0.5, 0.5, 0.5\} \) and \( \beta_2(s) = \Phi(s)^\top \{0.5, -0.5, 0.5, -0.5\} \), respectively. Further, we set the coefficient functions \( \eta_j \) \( (j = 2, \ldots, 9) \) associated with the \( X_j \) “main” effect terms to be: \( \eta_j(s) = \Phi(s)^\top \eta_j \), where, for each simulation replication, the vector \( \eta_j \in \mathbb{R}^3 \) \( (j = 2, \ldots, 9) \) is randomly generated from a multivariate Gaussian distribution and is then rescaled to a unit \( L^2 \) norm \( ||\eta_j|| = 1 \). Model (30) is indexed by a pair \( (\delta, \xi) \). The parameter \( \delta \in \{1, 2\} \) controls the the contribution of the \( (X, Z) \) “main” effect term \( \delta \left\{ \sum_{j=2} \cos(\langle \eta_j, X_{ij} \rangle) + \sum_{k=2} \cos(\langle Z_{ik} \rangle) \right\} \) to the variance of \( Y \), in which \( \delta = 1 \) represents a relatively moderate \( (X, Z) \) main effect (contributing about the same variance as the interaction effect does) and \( \delta = 2 \) represents a relatively large \( (X, Z) \) main effect (about 4 times greater than the interaction effect) when \( \xi = 0 \). The parameter \( \xi \in \{0, 1\} \) determines whether the \( A \)-by-\( (X, Z) \) interaction effect component has an additive regression structure \( (\xi = 0) \) of the form (5) or whether it deviates from an additive regression structure \( (\xi = 1) \), in the case of \( \xi = 0 \), the proposed CFAM (5) is correctly specified, whereas, for the case of \( \xi = 1 \), it is misspecified. For each simulation replication, we consider the following four approaches to estimating the optimal ITR \( D_{\text{opt}} \):

1. The functional additive regression approach (9), estimated via Algorithm 1, with the dimension of the \( B \)-spline basis for the functions \( \{g_j, h_k\} \) and \( \{\beta_j\} \) set at \( d_j = d_k = 6 \) and \( m_j = 8 \). The tuning parameter \( \lambda > 0 \) is chosen to minimize 5-fold cross-validated prediction error of the fitted models.

2. The functional linear regression approach of Ciarleglio et al. (2018),

\[
\min_{\beta_j \in L^2[0,1], \alpha_k \in \mathbb{R}} \mathbb{E} \left[ \left\{ Y - \sum_{j=1}^p (A-1.5)\langle \beta_j, X_j \rangle - \sum_{k=1}^q (A-1.5)\alpha_k Z_k \right\}^2 \right] + \lambda \left\{ \sum_{j=1}^p (||\beta_j|| + \rho_j \gamma_j^\top S_j \gamma_j) + \sum_{k=1}^q ||\alpha_k|| \right\},
\]

which tends to result in a sparse set \( \{\beta_j\} \cup \{\alpha_k\} \), which performs estimation based on the \( P \)-spline representation (28) for the coefficient function \( \beta_j \) with dimension \( m_j = 8 \). For each simulated dataset, the tuning parameters \( \lambda > 0 \) and \( \rho_j > 0 \) \( (j = 1, \ldots, p) \) are chosen to minimize a 5-fold cross-validated prediction error (Ciarleglio et al., 2018), and the ITR is given by:

\[
D_{\text{opt}}^p \left( X, Z \right) = \arg \max_{\alpha \in \{1, \ldots, L\}} \left\{ \sum_{j=1}^p (a-1.5)\langle \beta_j, X_j \rangle + \sum_{k=1}^q \alpha_k Z_k \right\}.
\]

Since the component functions \( \{g_j, h_k\} \) of the regression model of Ciarleglio et al. (2018) are constrained to be linear, i.e., \( g_{j,a}(\langle \beta_j, X_j \rangle) = (a-1.5)\langle \beta_j, X_j \rangle \) and \( h_{k,a}(Z_k) = (a-1.5)\alpha_k Z_k \), which correspond to a special case of CFAM, we call the model of Ciarleglio et al. (2018), a constrained functional additive model with linear component functions (CFAM-lin) for the notational simplicity.

3. The outcome weighted learning (OWL; Zhao et al., 2012) method based on a linear kernel (OWL-lin), implemented in the R-package \( \text{DTRlearn} \). To improve its efficiency, we employ the augmented outcome weighted learning approach of Liu et al. (2018). The tuning parameter \( \kappa \) in Zhao et al. (2012) is chosen from the grid of \( \{0.25, 0.5, 1, 2, 4\} \) (the default setting of \( \text{DTRlearn} \) based on a 5-fold cross-validation. Since there is no currently available OWL method that directly deals with functional covariates, we simply compute a scalar summary of each functional covariate, i.e., \( X_j = \int_0^1 X_j(s) ds \in \mathbb{R} \), and use \( X_j \) along with the other scalar covariates \( Z_k \) as inputs to the augmented OWL procedure.

11
4. The same approach as in 3 but based on a Gaussian radial basis function kernel (OWL-Gauss) in place of a linear kernel. The inverse bandwidth parameter $\sigma^2$ in Zhao et al. (2012) is chosen from the grid of (0.01, 0.02, 0.04, ..., 0.64, 1.28) and $\kappa$ is chosen from the grid of (0.25, 0.5, 1, 2, 4), based on a 5-fold cross-validation.

For each simulation run, we estimate $D^{opt}$ from each of the four methods based on a training set (of size $n \in \{250, 500\}$), and for evaluation of these methods, we compute the value $V(\hat{D}^{opt}) = E[|E[Y|X, Z, A = D^{opt}(X, Z)]|^2]$, given each estimate $\hat{D}^{opt}$, using a Monte Carlo approximation based on a random sample of size $10^3$. Since we know the true data generating model in simulation studies, the optimal $D^{opt}$ can be determined for each simulation run. Given each estimate $\hat{D}^{opt}$ of $D^{opt}$, we report $V(\hat{D}^{opt}) - V(D^{opt})$, as the performance measure of $\hat{D}^{opt}$. A larger value of the measure indicates better performance.

![Figure 1: Boxplots obtained from 200 Monte Carlo simulations comparing 4 approaches to estimating $D^{opt}$, given each scenario indexed by $\xi \in \{0, 1\}$, $\delta \in \{1, 2\}$ and $n \in \{250, 500\}$. The dotted horizontal line represents the optimal value corresponding to $D^{opt}$.](image)

In Figure 1, we present the boxplots, obtained from 200 simulation runs, of the normalized values $V(\hat{D}^{opt})$ (normalized by the optimal values $V(D^{opt})$) of the decision rules $\hat{D}^{opt}$ estimated from the four approaches, for each combination of $n \in \{250, 500\}$, $\xi \in \{0, 1\}$ (corresponding to correctly-specified or mis-specified CFAM interaction models, respectively) and $\delta \in \{1, 2\}$ (corresponding to moderate or large main effects, respectively). The results in Figure 1 indicate that the proposed method (CFAM) outperforms all other approaches. In particular, if the sample size is relatively large ($n = 500$), for a correctly-specified CFAM ($\xi = 0$) interaction underlying model, the proposed method gives a close-to-optimal performance in comparison to $D^{opt}$. With nonlinearities present in the data model (30), CFAM-lin, which assumes a stringent linear structure on the interaction effect term, is outperformed by CFAM that utilizes flexible component functions $g_{j,a}(\cdot)$ and $h_{k,a}(\cdot)$ for the approximation of the $A$-by-$Z$ interaction effects. In the absence of prior knowledge about the interaction effects, this suggests that employing CFAM for optimizing ITRs is more suitable than the linear interaction effect regression approach. The estimated values of the OWL methods using linear and Gaussian kernels, respectively, are similar to each other, but both are outperformed by CFAM, even when the true interaction effect structure deviates from CFAM (i.e., when $\xi = 1$), as the current CFAM does not directly deal with the functional covariates. If the $(X, Z)$ “main” effect dominates the $A$-by-$Z$ interaction effect (i.e., when $\delta = 2$), although the increased magnitude of this nuisance effect dampens the performance of all approaches to estimating $D^{opt}$, the proposed approach outperforms all other methods. The superior performance of CFAM is a result of its targeted estimation of the interaction effects while allowing for an unspecified $(X, Z)$ main effect.

### 4.2 Treatment effect-modifier variable selection performance

In this subsection, we will report simulation results illustrating the performance of the treatment effect-modifier covariate selection among the covariates $\{X_j, j = 1, \ldots, p\} \cup \{Z_k, k = 1, \ldots, q\}$. The complexity of CFAM (5) for the $A$-by-$Z$ interaction effect can be summarized in terms of the size of the index set for the component functions $\{g_j, j = 1, \ldots, p\} \cup \{h_k, k = 1, \ldots, q\}$ that are not identically zero, each of which
can be either correctly or incorrectly estimated to be equal to zero. We generate 200 datasets using the same simulation settings as in Section 4.1, i.e., using the data generating model (30) indexed by \( \xi \in \{0, 1\} \) and \( \delta \in \{1, 2\} \), for each sample size \( n \in \{100, 150, \ldots, 450, 500\} \). As in Section 4.1, we set \( p = q = 20 \), i.e., we consider a total of \( p + q = 40 \) potential treatment effect-modifiers, among which there are only 4 true treatment effect-modifiers.

Figure 2: The proportion of the relevant covariates (i.e., the treatment effect-modifiers) correctly selected (the “true positives”; the top gray panels), and the “noise” covariates incorrectly selected (the “false positives”; the bottom white panels), respectively (and \( \pm 1 \) standard deviation), with a varying sample size \( n \in \{100, 150, \ldots, 450, 500\} \), for each combination of \( \xi \in \{0, 1\} \) and \( \delta \in \{1, 2\} \).

Figure 2 summarizes the results of the treatment effect-modifier covariate selection performance with respect to the true/false positive rates (the top/bottom panels, respectively), comparing the proposed CFAM approach to the linear interaction modeling approach, CFAM-lin, of Ciarleglio et al. (2018). The results are reported as the averages (and \( \pm 1 \) standard deviations) across the 200 simulated datasets, for each simulation scenario. Figure 2 illustrates that the proportion of the correctly selecting treatment effect-modifiers out of the 4 true treatment effect-modifiers (i.e., the true positive rate; the top gray panels) tends to 1 as \( n \) increases from \( n = 100 \) to \( n = 500 \), while the proportion of incorrectly selecting treatment effect-modifiers (i.e., the false positive rate; the bottom white panels) out of the 36 irrelevant “noise” covariates tends to 0; the proportions tend to either 1 or 0 quickly for moderate main effect (\( \delta = 1 \)) scenarios. Although not presented in Figure 2, when the sample size is large, for example, when \( n = 1000 \), the false positive rate for CFAM is very close to 0 (the false positive rate = 0.04, with standard deviation 0.03) even for a large main effect (\( \delta = 2 \)) and misspecified CFAM (\( \xi = 1 \)) case, while the true positive rate stays close to 1 (the true positive rate = 0.98, with standard deviation 0.08). On the other hand, the proportion of correctly selecting treatment effect-modifiers for CFAM-lin (the blue dotted curves), even with a large \( n \) (= 1000), tends to be only around 0.70 for all combinations of \( \xi \) and \( \delta \), due to the stringent linear regression restriction on the \( A \)-by-(\( X, Z \)) interaction effect.

5 Application

In this section, we apply the proposed functional additive regression approach to a dataset from a study comparing an antidepressant and placebo for treating major depressive disorder. The main objective of our
investigation in this study was to use baseline functional covariates to guide treatment decisions when a patient presents for treatment. The study collected baseline scalar and functional data, including electroencephalogram (EEG) amplitude spectra curves, prior to treatment assignment. Following these baseline assessments, study participants were randomized to either placebo \( (a = 1) \) or an antidepressant \( (a = 2) \). Subjects were monitored for 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, hence a larger value of the outcome \( Y \) is considered desirable.

There were \( n = 179 \) subjects in the study. We considered \( p = 19 \) baseline functional covariates, a subset of EEG channels from a total of 72 EEG electrodes which gives a fairly good spatial coverage of the scalp. The locations for these 19 electrodes are indicated in Figure 3. Specifically, the functional data of interest consist of the curves giving the current source density (CSD) amplitude spectrum values over a frequency range of 3 to 16 Hz, observed while the participants’ eyes were closed. This frequency range was scaled to \([0, 1]\), hence each of the functional covariates \( X = (X_1(s), \ldots, X_{19}(s)) \) was defined on the interval \([0, 1]\). In addition, we considered a set of 4 baseline scalar covariates, consisting of the baseline (i.e., week 0) HRSD score \( (Z_1) \), sex \( (Z_2) \) \((1 \text{ for female, } 0 \text{ for male})\), age \( (Z_3) \), and the baseline HRSD-by-age interaction \( (Z_4 = Z_1 \times Z_3) \). In this dataset, 46\% of the subjects were randomized to the sertraline. The average outcomes \( Y \) for the sertraline and placebo groups were 7.75 and 6.29, respectively. The mean age was 38.3 years, the mean baseline HSRD score was 18.78, and 64\% of the subjects were female.

The proposed CFAM estimation approach (9) selected two functional covariates: “F7” \( (X_2) \) and “O2” \( (X_{12}) \) (the selected electrodes are indicated by the red dashed circles in Figure 3), and one scalar covariate: “baseline HRSD” \( (Z_1) \). In the left column of Figure 4, we display CSD curves corresponding to the selected functional covariates, \( X_2(s) \) and \( X_{12}(s) \), observed from the 179 subjects. In the middle column of Figure 4, we display the estimated single-index coefficient functions \( \hat{\beta}_j(s) \) \((j = 2, 12)\) (along with the 95\% confidence bands), associated with the selected two functional covariates.

![Figure 3: The locations for the 19 electrode channels (“A1” and “A2” were not used). Those marked in red circles are the selected electrodes from the modeling method (9): “F7” and “O2”.

Figure 3: The locations for the 19 electrode channels (“A1” and “A2” were not used). Those marked in red circles are the selected electrodes from the modeling method (9): “F7” and “O2”.

14
Figure 4: Left column: observed current source density (CSD) curves from the selected channels “F7” ($X_2$) and “O2” ($X_{12}$), for the sertraline group (red dashed curves) and for the placebo group (blue dotted curves), over a frequency range of 3 to 16 Hz (this frequency range is scaled to $[0,1]$), when the participants’ eyes are closed. Middle column: the estimated single-index coefficient functions ($\beta_2$ and $\beta_{12}$) associated with the selected channels $X_2$ and $X_{12}$ (and the associated 95% confidence bands). Right column: the scatter plots of the ($j$th: $j=2,12$) partial residuals vs. the estimated single-indices $\langle X_2, \beta_2 \rangle$ and $\langle X_{12}, \beta_{12} \rangle$. Overlaid are the estimated treatment-specific component functions $\hat{g}_{j,a}(\cdot)$ ($a=1,2; j=2,12$) for the placebo group in the dotted blue curves, and the sertraline group in the solid red curves.

The estimated single-index coefficient functions $\hat{\beta}_j$ lead to data-driven scalar variables $\langle \hat{\beta}_j, X_j \rangle \in \mathbb{R}$. Hence, each of the EEG amplitude spectra was reduced to an weighted average, the weights determined by the function $\hat{\beta}_j$. These averages were selected as potential (scalar) modifiers of treatment effect, and they are linked to differential treatment response by two corresponding nonzero component functions $\hat{g}_j$ ($j=2,12$) in this example. In the right column of Figure 4 (and in Figure 5), the estimated treatment $a$-specific component functions $\hat{g}_{j,a}(\cdot)$ (and $\hat{h}_{k,a}(\cdot)$) ($a=1,2$) of the selected functional (and scalar) covariates, respectively, are displayed on the corresponding partial residual plots. These plots illustrate some nonlinear $A$-by-$X_j$ ($j=2,12$) (and $A$-by-$Z_1$) interaction effects captured by the estimated component functions $\{\hat{g}_j, \hat{h}_k\}$ of the model.

Figure 5: The scatter plots of the $k$th partial residual vs. the $k$th scalar covariates $Z_k$ ($k=1$): the baseline HRSD ($Z_1$). Overlaid are the estimated associated treatment-specific component functions $\hat{h}_{1,a}(\cdot)$ ($a=1,2$) for the placebo group in the dotted blue, and the active drug group in the solid red curves.
To evaluate the performance of ITRs ($\hat{D}_{opt}$) estimated from the four different approaches described in Section 4, we randomly split the data into a training set and a testing set (of size $\tilde{n}$) using a ratio of 5 : 1, replicated 500 times, each time estimating an ITR $\hat{D}_{opt}$ based on the training set, and its “value” $V(\hat{D}_{opt}) = E[E[Y|X, Z, A = \hat{D}_{opt}(X, Z)]]$, by an inverse probability weighted estimator (Murphy, 2005)

\[
\hat{V}(\hat{D}_{opt}) = \sum_{i=1}^{\tilde{n}} Y_i I(A_i = \hat{D}_{opt}(X_i, Z_i)) / \sum_{i=1}^{\tilde{n}} I(A_i = \hat{D}_{opt}(X_i, Z_i)),
\]

computed based on the testing set (of size $\tilde{n}$). For comparison, we also include two naïve rules: treating all patients with placebo (“All PBO”) and treating all patients with the active drug (“All DRUG”), each regardless of the individual patient’s characteristics $(X, Z)$. The resulting boxplots obtained from the 500 random splits are illustrated in Figure 6. A larger value of the measure indicates better performance.

![Figure 6: Boxplots of the estimated values of the treatment rules $\hat{D}_{opt}$ estimated from 6 approaches, obtained from 500 randomly split testing sets. Higher values are preferred.](image.png)

The results in Figure 6 demonstrates that CFAM for optimizing ITRs tends to outperform other approaches in terms of the averaged estimated values, showing some modest superiority over the naïve rule of assigning everyone to the active drug. CFAM-lin basically assigns all subjects to the active drug (see “CFAM-lin” and “All DRUG”, which give the identical boxplots). Both the OWL-lin and OWL-Gauss approaches are outperformed by the naïve rule that assigns every subject to the active drug. Although the difference between CFAM and the active drug (sertraline) is quite small, sertraline has been one of the most successful antidepressant treatments available for several decades, and in comparison to the difference between the active drug and placebo, we argue that this small difference between CFAM and the active drug still provides a meaningful step towards “personalizing” treatment decision rules based on potentially complex patients’ pretreatment characteristics. The proposed CFAM provides a means of simultaneously identifying and selecting treatment effect-modifiers in addition to providing a visualization for heterogeneous effects explained by each estimated treatment effect-modifier as in Figure 5, which is an appealing feature in practice.

### 6 Discussion

We have developed a functional additive regression model estimation approach specifically focused on extracting pertinent interaction effects between treatment and multiple functional/scalar covariates which is of paramount importance in developing effective ITR’s for precision medicine. This is accomplished by imposing an appropriate structural constraints and performing variable selection. The estimation approach utilizes an efficient coordinate-descent algorithm for the additive component functions, coupled with a standard functional linear regression model fitting procedure for the coefficient functions. The proposed functional regression for ITRs extends existing (functional) linear model-based regression methods by incorporating possibly nonlinear treatment-by-functional covariates interactions. Encouraged by our simulation results and the application, future work will investigate the asymptotic properties of the method related to variable selection and estimation consistency, and development of a hypothesis testing procedures for testing nonzero interaction effects between treatment and functional covariates based on the proposed models.
SUPPLEMENTARY MATERIAL

Supplementary Material available online includes the proof of Theorem 1.

**R-package**: R-package *famTEMsel* (Functional Additive Models for Treatment Effect-Modifier Selection) contains R-codes to perform the methods proposed in the article, and is publicly available on GitHub (syhyunpark/famTEMsel).

**Acknowledgments**

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

**Conflict of interest**

None declared.

**References**

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* 13:571–592.

Ciarleglio, A., Petkova, E., Ogden, R. T., and Tarpey, T. (2015). Treatment decisions based on scalar and functional baseline covariates based on scalar and functional baseline covariates. *Biometrics* 71:884–894.

Ciarleglio, A., Petkova, E., Ogden, R. T., and Tarpey, T. (2018). Constructing treatment decision rules based on scalar and functional predictors when moderators of treatment effect are unknown. *Journal of Royal Statistical Society: Series C* 67:1331–1356.

Ciarleglio, A., Petkova, E., Tarpey, T., and Ogden, R. T. (2016). Flexible functional regression methods for estimating individualized treatment rules. *Stat* 5:185–199.

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

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

Gunter, L., Zhu, J., and Murphy, S. (2011). Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. *Journal of Biopharmaceutical Statistics* 21:1063–1078.

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

Jeng, X., Lu, W., and Peng, H. (2018). High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics* 12:2074–2089.

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

Laber, E. B. and Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* 102:501–514.
Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine* **37**:3776–3788.

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

Marx, B. and Eilers, P. (1999). Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics* **41**:1–13.

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

Murata, N. and Amari, S. (1994). Network Information Criterion- Determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks* **5**:865–872.

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

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

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer, New York.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of Royal Statistical Society: Series B* **71**:1009–1030.

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

Shi, C., Song, R., and Lu, W. (2016). Robust learning for optimal treatment decision with np-dimensionality. *Electronic Journal of Statistics* **10**:2894–2921.

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

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

Trivedi, M., McGrath, P., Fava, M., Parsey, R., Kurian, B., Phillips, M., Quenodo, M., Bruder, G., Pizzagalli, D., Toups, M., Cooper, C., Adams, P., Weyandt, S., Morris, D., Grannemann, B., Ogden, R., Buckner, R., McInnis, M., Kraemer, H., Petkova, E., Carmody, T., and Weissman, M. (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. *Journal of Psychiatric Research* **78**:11–23.

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

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

Zhao, Y., Laber, E., Ning, Y., Saha, S., and Sands, B. (2019). Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research* **20**:1–23.

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

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

Proof of Theorem 1

Proof. In order to simplify the exposition, we focus on the derivation of the minimizing component functions $g_j \in \mathcal{H}^{(\beta_j)}_j$ ($j = 1, \ldots, p$) associated with the functional covariates $X_j$ ($j = 1, \ldots, p$), only. The minimizing component functions $h_k \in \mathcal{H}_k$ ($k = 1, \ldots, q$) associated with the scalar covariates $Z_k$ ($k = 1, \ldots, q$) are derived in a similar fashion. Further, for fixed $\beta_j \in \Theta$ ($j = 1, \ldots, p$), we write $X_{\beta_j} = (X_j, \beta_j) \in \mathbb{R}$ ($j = 1, \ldots, p$), for notational simplicity.

The squared error criterion on the right-hand side of (8) is

$$
\mathbb{E} \left[ \{ Y - \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \}^2 \right] = \mathbb{E} \left[ Y \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) - \left( \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \right)^2 \right] / 2
$$

(with respect to $\{g_j\}$)

$$
= \mathbb{E} \left[ \{ \mu(X) + \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \} \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) - \left( \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \right)^2 \right] / 2
$$

$$
= \mathbb{E} \left[ \{ \mu(X) \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \} + \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \{ \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \} - \left( \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \right)^2 \right] / 2
$$

$$
= \mathbb{E} \left[ \{ \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \} \{ \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \} - \left( \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \right)^2 \right] / 2,
$$

where the last equality follows from the constraint $\mathbb{E}[g_{j,A}(X_{\beta_j})|X_j] = 0$ ($j = 1, \ldots, p$) in (8) imposed on $\{g_j\}$, which implies $\mathbb{E}[\mu(X) \sum_{j=1}^{p} g_{j,A}(X_{\beta_j})] = \mathbb{E}[\mu(X) \sum_{j=1}^{p} g_{j,A}(X_{\beta_j})|X_j] = 0$. From (31), for fixed $\{\beta_j, j = 1, \ldots, p\}$, we can rewrite the squared error criterion in (8) by (omitting the components associated with the scalar covariates):

$$
\arg\min_{\{g_j \in \mathcal{H}^{(\beta_j)}_j\}} \mathbb{E} \left[ \{ Y - \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \}^2 \right] = \arg\min_{\{g_j \in \mathcal{H}^{(\beta_j)}_j\}} \mathbb{E} \left[ \{ \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) - \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \}^2 \right].
$$

(32)

In the following, we closely follow the proof of Theorem 1 in Ravikumar et al. (2009). The Lagrangian in (9) for fixed $\{\beta_j, j = 1, \ldots, p\}$ can be rewritten as:

$$
Q(\{g_j\}; \lambda) := \mathbb{E} \left[ \{ \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) - \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) \}^2 \right] + \lambda \sum_{j=1}^{p} \|g_j\|
$$

(33)

Fixing $\{\beta_j, j = 1, \ldots, p\}$, for each $j$, let us consider the minimization of (33) with respect to the component function $g_j \in \mathcal{H}^{(\beta_j)}_j$, holding the other component functions $\{g_{j'}, j' \neq j\}$ fixed. The stationary condition is obtained by setting its Fréchet derivative to 0. Denote by $\partial_j Q(\{g_j\}; \lambda; \eta_j)$ the directional derivative with respect to $g_j \in \mathcal{H}^{(\beta_j)}_j$ ($j = 1, \ldots, p$) in the direction, say, $\eta_j \in \mathcal{H}^{(\beta_j)}_j$. Then, for fixed $\{\beta_j, j = 1, \ldots, p\}$, the stationary point of the Lagrangian (33) can be formulated as:

$$
\partial_j Q(\{g_j\}; \lambda; \eta_j) = 2 \mathbb{E} \left[ (g_j - \hat{R}_j + \lambda \nu_j) \eta_j \right] = 0,
$$

(34)

where

$$
\hat{R}_j := \sum_{j=1}^{p} g_{j,A}(X_{\beta_j}) - \sum_{j' \neq j} g_{j',A}(X_{\beta_{j'}})
$$

(35)

is the partial residual for the $j$th component function $g_j$, and the function $\nu_j$ is an element of the subgradient $\partial\|g_j\|$, which satisfies $\nu_j = g_j/\|g_j\|$ if $\|g_j\| \neq 0$, and $\nu_j \in \{s \in \mathcal{H}^{(\beta_j)}_j | \|s\| \leq 1\}$, otherwise. Applying the iterated expectations to condition on $(X_{\beta_j}, A)$, the stationary condition (34) can be rewritten as:

$$
2 \mathbb{E} \left[ (g_j - \mathbb{E} \left[ \hat{R}_j|X_{\beta_j}, A \right] + \lambda \nu_j ) \eta_j \right] = 0.
$$

(36)
Since the function $g_j - \mathbb{E} \left[ \hat{R}_j | X_{\beta_j}, A \right] + \lambda \nu_j \in \mathcal{H}^{(\beta)}_j$, we can evaluate (34) (i.e., expression (36)) in the particular direction: $\eta_j = g_j - \mathbb{E} \left[ \hat{R}_j | X_{\beta_j}, A \right] + \lambda \nu_j$, which gives $\mathbb{E} \left[ (g_j - \mathbb{E} \left[ \hat{R}_j | X_{\beta_j}, A \right] + \lambda \nu_j)^2 \right] = 0$. This equation implies:

$$g_j + \lambda \nu_j = \mathbb{E} \left[ \hat{R}_j | X_{\beta_j}, A \right] \quad \text{(almost surely).} \quad (37)$$

Let $f_j$ denote the right-hand side of (37), i.e., $f_j = f_j, A(X_{\beta_j}) := \mathbb{E} \left[ \hat{R}_j | X_{\beta_j}, A \right]$. If $\|g_j\| \neq 0$, then $\nu_j = g_j / \|g_j\|$. Therefore, by (37), we have $\|f_j\| = \|g_j + \lambda g_j / \|g_j\|\| = \|g_j\| + \lambda \geq \lambda$. On the other hand, if $\|g_j\| = 0$, then $g_j = 0$ (almost surely), and $\|\nu_j\| \leq 1$. Then, condition (37) implies that $\|f_j\| \leq \lambda$. This gives us the equivalence between $\|f_j\| \leq \lambda$ and the statement $g_j = 0$ (almost surely). Therefore, condition (37) leads to the following expression:

$$(1 + \lambda / \|g_j\|) g_j = f_j \quad \text{(almost surely)}$$

if $\|f_j\| > \lambda$, and $g_j = 0$ (almost surely), otherwise. This gives the soft thresholding update rule for $g_j$.

Note, the underlying model (5) (if we omit the components associated with the scalar covariates) implies that $\sum_{j'=1}^p g_{j', A}(X_{\beta_{j'}}) = \mathbb{E}[Y | X, A] - \mu(X)$. Thus, (35) can be equivalently written as: $\hat{R}_j = \mathbb{E}[Y | X, A] - \mu(X) - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}})$. Therefore, the function $f_{j, A}(X_{\beta_j}) = \mathbb{E} \left[ \hat{R}_j | X_{\beta_j}, A \right]$ can be written by:

$$f_{j, A}(X_{\beta_j}) = \mathbb{E} \left[ Y | X, A \right] - \mu(X) - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}}) | X_{\beta_j}, A$$

$$= \mathbb{E} \left[ Y - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}}) | X_{\beta_j}, A \right] - \mathbb{E} \left[ \mu(X) | X_{\beta_j}, A \right]$$

$$= \mathbb{E} \left[ Y - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}}) | X_{\beta_j}, A \right] - \mathbb{E} \left[ \mu(X) | X_{\beta_j} \right]$$

$$= \mathbb{E} \left[ Y - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}}) | X_{\beta_j}, A \right] - \mathbb{E} \left[ Y | X_{\beta_j} \right]$$

$$= \mathbb{E} \left[ Y - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}}) | X_{\beta_j}, A \right] - \mathbb{E} \left[ Y | X_{\beta_j} \right]$$

$$= \mathbb{E} \left[ Y - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}}) | X_{\beta_j}, A \right] - \mathbb{E} \left[ Y - \sum_{j' \neq j} g_{j', A}(X_{\beta_{j'}}) | X_{\beta_j} \right]$$

$$= \mathbb{E} \left[ R_j | X_{\beta_j}, A \right] - \mathbb{E} \left[ R_j | X_{\beta_j} \right],$$

where the fourth equality follows from the identifiability constraint (6) of the underlying model (5), and the sixth equality follows from the optimization constraint $\mathbb{E}[g_{j', A}(X_{\beta_{j'}})|X_j] = 0 \ (j' \neq j)$ implied by (8) imposed on $\{g_{j', j' \neq j} \}$. This gives the desired expression (11).