Supporting Information for Functional additive models for optimizing individualized treatment rules

Hyung Park[a0], Eva Petkova[a], Thaddeus Tarpey[a], R. Todd Ogden[b]

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

Abstract

In this document, we provide additional technical details supplementing the main manuscript of the paper, including the derivations of mathematical expressions presented in the manuscript and proof of Theorem 1 of the main manuscript, as well as the details of the estimation procedure referenced in the main manuscript. We also provide additional simulation examples, including a set of simulation experiments with a "linear" A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect scenario.

Web Appendix A: Technical details and additional simulations

A.1. Description of the constrained least squares criterion in Section 2.1

In Section 2 of the main manuscript, we introduce the constrained functional additive model (CFAM) for the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect:

$$Y \ = \mu(\boldsymbol{X}, \boldsymbol{Z}) + \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, A) + \sum_{k=1}^{q} h_k(Z_k, A) \ + \ \epsilon, \tag{S.1}$$

with $\beta_j \in \Theta$, subject to the constraint on the component functions $g_j \in \mathcal{H}_j^{(\beta_j)}$ $(j = 1, \dots, p)$ and $h_k \in \mathcal{H}_k$ $(k = 1, \dots, q)$:

$$\begin{aligned} E[g_j(\langle X_j, \beta_j \rangle, A)|X_j] = 0 \quad &(\text{almost surely}) \quad (\forall \beta_j \in \Theta) \quad (j = 1, \dots, p) \quad \text{and} \\ E[h_k(Z_k, A)|Z_k] = 0 \quad &(\text{almost surely}) \quad (k = 1, \dots, q), \end{aligned} \tag{S.2}$$

in which the expectation is taken with respect to the distribution of A given X_j (or Z_k), and $\epsilon \in \mathbb{R}$ is a mean zero noise with finite variance, and the form of the squared integrable functional μ in (S.1) is left unspecified.

Under model (S.1) subject to (S.2), the "true" (i.e., optimal) functional components, which we denote by $\{g_j^*, j = 1, \dots, p\} \cup \{\beta_j^*, j = 1, \dots, p\} \cup \{h_k^*, k = 1, \dots, q\}$ that constitute the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect, can be specified and viewed as the solution to the following constrained least squares problem:

$$\begin{aligned} \{g_j^*, \beta_j^*, h_k^*\} \quad = \quad &\underset{g_j \in \mathcal{H}_j^{(\beta_j)}, \beta_j \in \Theta, h_k \in \mathcal{H}_k}{\text{argmin}} \quad E\left[\left\{Y - \mu(\boldsymbol{X}, \boldsymbol{Z}) - \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, A) - \sum_{k=1}^{q} h_k(Z_k, A)\right\}^2\right] \\ &\text{subject to} \quad E[g_j(\langle X_j, \beta_j \rangle, A)|X_j] = 0 \quad \forall \beta_j \in \Theta \quad (j = 1, \dots, p) \quad \text{and} \\ &\qquad\qquad\quad E[h_k(Z_k, A)|Z_k] = 0 \quad (k = 1, \dots, q), \end{aligned} \tag{S.3}$$

[0] To whom correspondence should be addressed; parkh15@nyu.edu

in which $\mu(\boldsymbol{X},\boldsymbol{Z})$ is the "main" effect component that is assumed in model (S.1) (and is considered as fixed in (S.3)). In particular, on the right-hand side of (S.3), the expected squared error criterion term can be expanded as:

$$\begin{aligned}
&\underset{g_j\in\mathcal{H}_j^{(\beta_j)},\beta_j\in\Theta,h_k\in\mathcal{H}_k}{\arg\min} \quad E\left[\left\{Y-\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle,A)-\sum_{k=1}^{q} h_k(Z_k,A)\right\}^2+2\mu(\boldsymbol{X},\boldsymbol{Z})\left\{\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle,A)+\sum_{k=1}^{q} h_k(Z_k,A)\right\}\right]\\
=&\underset{g_j\in\mathcal{H}_j^{(\beta_j)},\beta_j\in\Theta,h_k\in\mathcal{H}_k}{\arg\min} \quad E\left[\left\{Y-\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle,A)-\sum_{k=1}^{q} h_k(Z_k,A)\right\}^2+2\mu(\boldsymbol{X},\boldsymbol{Z})E\left[\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle,A)+\sum_{k=1}^{q} h_k(Z_k,A)|\boldsymbol{X},\boldsymbol{Z}\right]\right]\\
=&\underset{g_j\in\mathcal{H}_j^{(\beta_j)},\beta_j\in\Theta,h_k\in\mathcal{H}_k}{\arg\min} \quad E\left[\left\{Y-\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle,A)-\sum_{k=1}^{q} h_k(Z_k,A)\right\}^2\right],
\end{aligned}$$

in which the first equality follows from an application of the iterated expectation rule to condition on $(\boldsymbol{X},\boldsymbol{Z})$, and the second equality follows from the constraint imposed in (S.3), that is, $E[g_{j,A}(\langle X_j,\beta_j\rangle)|X_j]=0,\ \forall\beta_j\in\Theta\ (j=1,\dots,p)$ and $E[h_{k,A}(Z_k)|Z_k]=0\ \ (k=1,\dots,q)$, which makes the second term on the second line of the above expression vanish to zero.

Since the minimization in (S.3) is in terms of the components $\{g_j,\beta_j,h_k\}$, the right-hand side of (S.3) can then be reduced to:

$$\begin{array}{rcll}
\{g_j^*,\beta_j^*,h_k^*\} & = & \underset{g_j\in\mathcal{H}_j^{(\beta_j)},\beta_j\in\Theta,h_k\in\mathcal{H}_k}{\operatorname{argmin}} & E\left[\left\{Y-\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle,A)-\sum_{k=1}^{q} h_k(Z_k,A)\right\}^2\right]\\
 & & \text{subject to} & E\left[g_j(\langle X_j,\beta_j\rangle,A)|X_j\right]=0 \quad \forall\beta_j\in\Theta \quad (j=1,\dots,p) \quad \text{and}\\
 & & & E\left[h_k(Z_k,A)|Z_k\right]=0 \quad (k=1,\dots,q),
\end{array} \tag{S.4}$$

which is as appeared in the representation (3) of the main manuscript.

A.2. Proof of Theorem 1

In this subsection, we provide the proof of Theorem 1 in Section 3.1 of the main manuscript. In order to simplify the exposition, we focus on the derivation of the minimizing functions $g_j \in \mathcal{H}_j^{(\beta_j)}$ $(j = 1, \ldots, p)$ associated with the functional covariates X_j $(j = 1, \ldots, p)$, only. The minimizing 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 the similar way. For fixed $\beta_j \in \Theta$ $(j = 1, \ldots, p)$, let us write $X_{\beta_j} = \langle X_j, \beta_j \rangle \in \mathbb{R}$ $(j = 1, \ldots, p)$, for notational simplicity.

The squared error criterion on the right-hand side of (S.4) is

$$\begin{aligned}
E\Big[\big\{Y - \sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\}^2\Big] &\propto E\Big[Y \sum_{j=1}^{p} g_j(X_{\beta_j}, A) - \big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\}^2/2\Big] \quad \text{(with respect to } \{g_j\}) \\
&= E\Big[\big\{\mu(\boldsymbol{X}) + \sum_{j=1}^{p} g_j^*(X_{\beta_j^*}, A)\big\} \sum_{j=1}^{p} g_j(X_{\beta_j}, A) - \big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\}^2/2\Big] \\
&= E\Big[\mu(\boldsymbol{X}) \sum_{j=1}^{p} g_j(X_{\beta_j}, A)\Big] + E\Big[\big\{\sum_{j=1}^{p} g_j^*(X_{\beta_j^*}, A)\big\}\big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\} - \big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\}^2/2\Big] \\
&= E\Big[\big\{\sum_{j=1}^{p} g_j^*(X_{\beta_j^*}, A)\big\}\big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\} - \big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\}^2/2\Big],
\end{aligned} \tag{S.5}$$

where the last equality follows from the constraints $E[g_j(X_{\beta_j}, A)|X_j] = 0$ $(j = 1, \ldots, p)$ in (S.4) that we imposed on $\{g_j\}$, that imply $E\big[\mu(\boldsymbol{X})\big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\}\big] = E\big[E\big[\mu(\boldsymbol{X})\big\{\sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big\} \mid \boldsymbol{X}\big]\big] = E\big[\mu(\boldsymbol{X}) \sum_{j=1}^{p} E\big[g_j(X_{\beta_j}, A) \mid X_j\big]\big] = 0$. From (S.5), for fixed $\{\beta_j, j = 1, \ldots, p\}$, we can rewrite the squared error criterion in (S.4) as:

$$\underset{\{g_j \in \mathcal{H}_j^{(\beta_j)}\}}{\text{argmin}} \; E\Big[\big(Y - \sum_{j=1}^{p} g_j\big(X_{\beta_j}, A\big)\big)^2\Big] = \underset{\{g_j \in \mathcal{H}_j^{(\beta_j)}\}}{\text{argmin}} \; E\Big[\big(\sum_{j=1}^{p} g_j^*\big(X_{\beta_j^*}, A\big) - \sum_{j=1}^{p} g_j\big(X_{\beta_j}, A\big)\big)^2\Big], \tag{S.6}$$

where we omitted the components associated with the scalar covariates to simplify the exposition. In the following, we closely follow the proof of Theorem 1 in Ravikumar et al. (2009). The Lagrangian in (4) of the main manuscript, for fixed $\{\beta_j, j = 1, \ldots, p\}$ can be rewritten as:

$$Q(\{g_j\}; \lambda) := E\Big[\big(\sum_{j=1}^{p} g_j^*(X_{\beta_j^*}, A) - \sum_{j=1}^{p} g_j(X_{\beta_j}, A)\big)^2\Big] + \lambda \sum_{j=1}^{p} \|g_j\|. \tag{S.7}$$

Fixing $\{\beta_j, j = 1, \ldots, p\}$, for each j, let us consider the minimization of (S.7) with respect to the jth component function $g_j \in \mathcal{H}_j^{(\beta_j)}$, holding the other component functions $\{g_{j'}, j' \neq j\}$ fixed. The stationary condition is obtained by setting its Fréchet derivative to 0. Let $\partial_j Q(\{g_j\}; \lambda; \eta_j)$ denote the directional derivative with respect to $g_j \in \mathcal{H}_j^{(\beta_j)}$ $(j = 1, \ldots, p)$ in some arbitrary direction $\eta_j \in \mathcal{H}_j^{(\beta_j)}$. Then, for fixed $\{\beta_j, j = 1, \ldots, p\}$, the stationary point of the Lagrangian (S.7) can be formulated as:

$$\partial_j Q(\{g_j\}; \lambda; \eta_j) = 2E\left[(g_j - \widetilde{R}_j + \lambda \nu_j)\eta_j\right] = 0, \tag{S.8}$$

where

$$\widetilde{R}_j := \sum_{j=1}^{p} g_j^*(X_{\beta_j^*}, A) - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A), \tag{S.9}$$

which represents the partial residual for the jth component function g_j, and the function ν_j in (S.8) is an element of the subgradient $\partial\|g_j\|$, which satisfies $\nu_j = g_j/\|g_j\|$ if $\|g_j\| \neq 0$, and $\nu_j \in \{s \in \mathcal{H}_j^{(\beta_j)} \mid \|s\| \leq 1\}$, otherwise. Applying the iterated expectations to condition on (X_{β_j}, A), the stationary condition (S.8) can be rewritten as:

$$2E\left[\left(g_j - E\left[\widetilde{R}_j | X_{\beta_j}, A\right] + \lambda \nu_j\right) \eta_j\right] = 0. \tag{S.10}$$

Since the function $g_j - E\left[\widetilde{R}_j|X_{\beta_j}, A\right] + \lambda\nu_j$ is in $\mathcal{H}_j^{(\beta_j)}$, we can evaluate (S.8) (i.e., expression (S.10)) in this particular direction: $\eta_j := g_j - E\left[\widetilde{R}_j|X_{\beta_j}, A\right] + \lambda\nu_j$, which gives $E\left[\left(g_j - E\left[\widetilde{R}_j|X_{\beta_j}, A\right] + \lambda\nu_j\right)^2\right] = 0$. This equation implies:

$$g_j + \lambda\nu_j = E\left[\widetilde{R}_j|X_{\beta_j}, A\right] \quad \text{(almost surely).} \tag{S.11}$$

Now, let P_j denote the right-hand side of (S.11), i.e., $P_j(= P_j(X_{\beta_j}, A)) := E\left[\widetilde{R}_j|X_{\beta_j}, A\right]$. We note that, if $\|g_j\| \neq 0$, then $\nu_j = g_j/\|g_j\|$. Therefore, by (S.11), we have $\|P_j\| = \|g_j + \lambda g_j/\|g_j\|\| = \|g_j\| + \lambda \geq \lambda$. On the other hand, if $\|g_j\| = 0$, then $g_j = 0$ (almost surely) and $\|\nu_j\| \leq 1$. Then, condition (S.11) implies that $\|P_j\| \leq \lambda$. This gives us the equivalence between $\|P_j\| \leq \lambda$ and the statement $g_j = 0$ (almost surely). Therefore, condition (S.11) leads to the following expression:

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

if $\|P_j\| > \lambda$, and $g_j = 0$ (almost surely), otherwise; this implies the soft thresholding update rule for g_j appeared in (5) of the main manuscript.

Now we will derive the expression (6) of the main manuscript for the function P_j. Note, the underlying model (S.1) implies that $\sum_{j=1}^{p} g_j^*(X_{\beta_j^*}, A) = E[Y|\boldsymbol{X}, A] - \mu(\boldsymbol{X})$ (if we omit the components associated with the scalar covariates). Thus, (S.9) can be equivalently written as: $\widetilde{R}_j = E[Y|\boldsymbol{X}, A] - \mu(\boldsymbol{X}) - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A)$. Therefore, the function $P_j(X_{\beta_j}, A) = E\left[\widetilde{R}_j|X_{\beta_j}, A\right]$ can be written as:

$$\begin{aligned}
P_j(X_{\beta_j}, A) &= E\big[E[Y|\boldsymbol{X}, A] - \mu(\boldsymbol{X}) - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A) \mid X_{\beta_j}, A\big] \\
&= E\big[E[Y|\boldsymbol{X}, A] - \sum_{j' \neq j} g_{j'}(X_{j'}, A) \mid X_{\beta_j}, A\big] - E\big[\mu(\boldsymbol{X}) \mid X_{\beta_j}, A\big] \\
&= E\big[Y - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A) \mid X_{\beta_j}, A\big] - E\big[\mu(\boldsymbol{X}) \mid X_{\beta_j}\big] \\
&= E\big[Y - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A) \mid X_{\beta_j}, A\big] - E\big[\mu(\boldsymbol{X}) + \sum_{j=1}^{p} g_j^*(X_{\beta_j^*}, A) \mid X_{\beta_j}\big] \\
&= E\big[Y - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A) \mid X_{\beta_j}, A\big] - E\big[Y \mid X_{\beta_j}\big] \\
&= E\big[Y - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A) \mid X_{\beta_j}, A\big] - E\big[Y - \sum_{j' \neq j} g_{j'}(X_{\beta_{j'}}, A) \mid X_{\beta_j}\big] \\
&= E\big[R_j \mid X_{\beta_j}, A\big] - E\big[R_j \mid X_{\beta_j}\big],
\end{aligned}$$

where the fourth equality follows from the identifiability constraint (S.2) that we imposed on the underlying model (S.1), and the sixth equality follows from the optimization constraint $E[g_{j'}(X_{\beta_{j'}}, A)|X_j] = 0$ $(j' \neq j)$ implied by (S.4) that we imposed on $\{g_{j'}, j' \neq j\}$. This gives the expression (6) of the main manuscript for P_j.

A.3. Description of general linear smoothers for the component functions

As referenced in Section 3.2.1 of the main manuscript, in general, estimation of the component functions g_j is not restricted to regression splines and any scatterplot smoother can be utilized. In this paper, the estimate $\widehat{g}_j$ corresponds to a soft-thresholded estimate of the function P_j specified in (6) of the main manuscript. To estimate P_j, we can first estimate the system of treatment a-specific functions $E[R_{ij}|\langle\widehat{\beta}_j, X_{ij}\rangle, A=a]$ $(a=1,\ldots,L)$ (which corresponds to the first term on the right-hand side of (6) if we fix $\beta_j = \widehat{\beta}_j$), by performing separate nonparametric regressions of $\widehat{R}_{ij}$ on regressor $\langle\widehat{\beta}_j, X_{ij}\rangle$, separately for each treatment condition $A=a$ $(a=1,\ldots,L)$. We can then estimate the function $-E[R_{ij}|\langle\beta_j, X_{ij}\rangle]$ (which corresponds to the second term on the right-hand side of (6) if we fix $\beta_j = \widehat{\beta}_j$), by performing a nonparametric regression of $\widehat{R}_{ij}$ on regressor $\langle\widehat{\beta}_j, X_{ij}\rangle$. Adding these two function estimates evaluated on the observed values provides an estimate for P_j in (6) evaluated at the n points $(\langle\widehat{\beta}_j, X_{ij}\rangle, A_i)$ $(i=1,\ldots,n)$ (analogous to the vector $\widehat{\boldsymbol{P}}_j \in \mathbb{R}^n$ given in (S.17) in Section A.4 below). Given this estimate of P_j evaluated at the n points, we can compute the corresponding soft-thresholded estimate $\widehat{\boldsymbol{g}}_j \in \mathbb{R}^n$ and conduct the coordinate descent procedure described in Algorithm 1 in Section A.6 of this document.

A.4. Estimation details for Step 1

As indicated in Section A.3 above, although any linear smoothers can be utilized to obtain estimators $\{\widehat{g}_j, j=1,\ldots,p\}$, we shall focus on regression spline-type estimators, which are simple and computationally efficient to implement. For each j and $\beta_j = \widehat{\beta}_j$, we will represent the component function $g_j \in \mathcal{H}_j^{(\widehat{\beta}_j)}$ on the right-hand side of (4) of the main manuscript as:

$$g_j(\langle X_j, \widehat{\beta}_j\rangle, a) = \boldsymbol{\Psi}_j(\langle X_j, \widehat{\beta}_j\rangle)^\top \boldsymbol{\theta}_{j,a} \quad (a=1,\ldots,L) \tag{S.12}$$

for some prespecified d_j-dimensional basis $\boldsymbol{\Psi}_j(\cdot)$ (e.g., cubic B-spline basis with $d_j - 4$ interior knots, evenly placed over the range (scaled to, say, $[0,1]$) of the observed values of $\langle X_j, \widehat{\beta}_j\rangle$) and a set of unknown treatment a-specific basis coefficients $\{\boldsymbol{\theta}_{j,a} \in \mathbb{R}^{d_j}\}_{a\in\{1,\ldots,L\}}$. Based on representation (S.12) of $g_j \in \mathcal{H}_j^{(\widehat{\beta}_j)}$ for fixed $\widehat{\beta}_j$, the constraint $E[g_j(\langle X_j, \beta_j\rangle, A)|X_j] = 0$ in (4) of the main manuscript on g_j, for fixed $\beta_j = \widehat{\beta}_j$, can be simplified to: $E[\boldsymbol{\theta}_{j,A}] = \sum_{a=1}^{L} \pi_a \boldsymbol{\theta}_{j,a} = \boldsymbol{0}$ (if π_a depends on the covariates, then we can reparametrize model (1) of the main manuscript and accommodate $\pi_a(\boldsymbol{X}, \boldsymbol{Z})$ in the estimation; see Section A.17). If we fix $\beta_j = \widehat{\beta}_j$, the constraint in (4) of the main manuscript on the function g_j can be succinctly written in matrix form:

$$\boldsymbol{\pi}^{(j)} \boldsymbol{\theta}_j = \boldsymbol{0}, \tag{S.13}$$

where $\boldsymbol{\theta}_j := (\boldsymbol{\theta}_{j,1}^\top, \boldsymbol{\theta}_{j,2}^\top, \ldots, \boldsymbol{\theta}_{j,L}^\top)^\top \in \mathbb{R}^{d_j L}$ is the vectorized version of the basis coefficients $\{\boldsymbol{\theta}_{j,a}\}_{a\in\{1,\ldots,L\}}$, and the $d_j \times d_j L$ matrix $\boldsymbol{\pi}^{(j)} := (\pi_1 \boldsymbol{I}_{d_j}; \pi_2 \boldsymbol{I}_{d_j}; \ldots; \pi_L \boldsymbol{I}_{d_j})$ where $\boldsymbol{I}_{d_j}$ is the $d_j \times d_j$ identity matrix.

Let the $n \times d_j$ matrices $\boldsymbol{D}_{j,a}$ $(a=1,\ldots,L)$ denote the evaluation matrices of the basis $\boldsymbol{\Psi}_j(\cdot)$ on $\langle X_{ij}, \widehat{\beta}_j\rangle$ $(i=1,\ldots,n)$ specific to the treatment $A=a$ $(a=1,\ldots,L)$, whose ith row is the $1 \times d_j$ vector $\boldsymbol{\Psi}_j(\langle X_{ij}, \widehat{\beta}_j\rangle)^\top$ if $A_i = a$, and a row of zeros $\boldsymbol{0}^\top$ if $A_i \neq a$. Then the column-wise concatenation of the design matrices $\{\boldsymbol{D}_{j,a}\}_{a\in\{1,\ldots,L\}}$, i.e., the $n \times d_j L$ matrix $\boldsymbol{D}_j = (\boldsymbol{D}_{j,1}; \boldsymbol{D}_{j,2}; \ldots; \boldsymbol{D}_{j,L})$, defines the model matrix associated with the vectorized basis coefficient $\boldsymbol{\theta}_j \in \mathbb{R}^{d_j L}$, vectorized across $\{\boldsymbol{\theta}_{j,a}\}_{a\in\{1,\ldots,L\}}$ in representation (S.12). We can then represent $g_j(\langle X_j, \widehat{\beta}_j\rangle, A)$ of (S.12), based on the sample data, by the length-n vector:

$$\boldsymbol{g}_j = \boldsymbol{D}_j \boldsymbol{\theta}_j \in \mathbb{R}^n \tag{S.14}$$

subject to the linear constraint (S.13) on the parameters $\boldsymbol{\theta}_j$. (Similarly, we can represent $h_k(Z_k, A)$ by a length-n vector.)

The linear constraint in (S.13) on $\boldsymbol{\theta}_j$ can be conveniently absorbed into the model matrix $\boldsymbol{D}_j$ in (S.14) by reparametrization, which we describe next. We can find a $d_j L \times d_j(L-1)$ basis matrix $\boldsymbol{n}^{(j)}$ (that spans the *null* space of the linear constraint (S.13)), such that, if we set $\boldsymbol{\theta}_j = \boldsymbol{n}^{(j)}\widetilde{\boldsymbol{\theta}}_j$ for any arbitrary vector $\widetilde{\boldsymbol{\theta}}_j \in \mathbb{R}^{d_j(L-1)}$, then the vector $\boldsymbol{\theta}_j \in \mathbb{R}^{d_jL}$ automatically satisfies the constraint (S.13): $\boldsymbol{\pi}^{(j)}\boldsymbol{\theta}_j = \mathbf{0}$. Such a basis matrix $\boldsymbol{n}^{(j)}$ can be constructed by a QR decomposition of the matrix $\boldsymbol{\pi}^{(j)\top}$. Then representation (S.14) can be reparametrized, in terms of the unconstrained $\widetilde{\boldsymbol{\theta}}_j \in \mathbb{R}^{d_j(L-1)}$ by replacing $\boldsymbol{D}_j$ in (S.14) with a reparametrized model matrix $\widetilde{\boldsymbol{D}}_j = \boldsymbol{D}_j\boldsymbol{n}^{(j)}$:

$$\boldsymbol{g}_j = \widetilde{\boldsymbol{D}}_j\widetilde{\boldsymbol{\theta}}_j. \tag{S.15}$$

Theorem 1 of the main manuscript, together with the results in Section A.5 below, indicates that (for fixed $\beta_j = \widehat{\beta}_j$) the coordinate-wise minimizing function g_j of the right-hand side of (4) of the main manuscript can be estimated based on the sample by:

$$\widehat{\boldsymbol{g}}_j = \left[1 - \frac{\lambda}{\sqrt{\frac{1}{n}\|\widehat{\boldsymbol{P}}_j\|^2}}\right]_+ \widehat{\boldsymbol{P}}_j \tag{S.16}$$

where

$$\widehat{\boldsymbol{P}}_j = \widetilde{\boldsymbol{D}}_j(\widetilde{\boldsymbol{D}}_j^\top\widetilde{\boldsymbol{D}}_j)^{-1}\widetilde{\boldsymbol{D}}_j^\top\widehat{\boldsymbol{R}}_j, \tag{S.17}$$

in which $\widehat{\boldsymbol{R}}_j = \boldsymbol{Y} - \sum_{j'\neq j}\widehat{\boldsymbol{g}}_{j'} - \sum_{k=1}^q \widehat{\boldsymbol{h}}_k$ corresponds to the estimated jth partial residual vector. (Similarly, we can represent the coordinate-wise minimizing function h_k in (8) of the main manuscript, based on the observed data by a length-n vector $\widehat{\boldsymbol{h}}_k$.) If we set each β_j equal to its corresponding estimate $\widehat{\beta}_j$ $(j = 1, \ldots, p)$, then based on the sample counterpart (S.16) of the coordinate-wise solution (5) of the main manuscript, a highly efficient coordinate descent algorithm can be conducted to optimize $\{g_j, j = 1, \ldots, p\} \cup \{h_k, k = 1, \ldots, q\}$ simultaneously. Let $\widehat{s}_j^{(\lambda)} := \left[1 - \lambda\sqrt{n}/\|\widehat{\boldsymbol{P}}_j\|\right]_+$ in (S.16) denote the soft-threshold shrinkage factor associated with the un-shrunk estimate $\widehat{\boldsymbol{P}}_j$ in (S.17). At convergence of the coordinate descent, we obtain a basis coefficient estimate of $\widetilde{\boldsymbol{\theta}}_j$ associated with representation (S.15):

$$\widehat{\widetilde{\boldsymbol{\theta}}}_j = \widehat{s}_j^{(\lambda)}(\widetilde{\boldsymbol{D}}_j^\top\widetilde{\boldsymbol{D}}_j)^{-1}\widetilde{\boldsymbol{D}}_j^\top\widehat{\boldsymbol{R}}_j, \tag{S.18}$$

which in turn implies an estimate of $\boldsymbol{\theta}_j$ under representation (S.14): $\widehat{\boldsymbol{\theta}}_j = (\widehat{\boldsymbol{\theta}}_{j,1}^\top, \widehat{\boldsymbol{\theta}}_{j,2}^\top, \ldots, \widehat{\boldsymbol{\theta}}_{j,L}^\top)^\top = \boldsymbol{n}^{(j)}\widehat{\widetilde{\boldsymbol{\theta}}}_j$. Specifically, this gives an estimate of the treatment a-specific function $g_j(\cdot, a)$ $(a = 1, \ldots, L)$ in model (1) of the main manuscript:

$$\widehat{g}_j(\cdot, a) = \boldsymbol{\Psi}_j(\cdot)^\top\widehat{\boldsymbol{\theta}}_{j,a} \quad (a = 1, \ldots, L) \quad (j = 1, \ldots, p) \tag{S.19}$$

estimated within the class of functions (S.12), for a given tuning parameter $\lambda \geq 0$ that controls the soft-threshold shrinkage factor $\widehat{s}_j^{(\lambda)}$ in (S.18), resulting in the functions $\{\widehat{g}_j, j = 1, \ldots, p\} \cup \{\widehat{h}_k, k = 1, \ldots, q\}$; this completes Step 1 of the alternating optimization procedure. (We note that Step 2 of the optimization procedure is provided in Section 3.2.2 of the main manuscript.)

A.5. Supplementary information for Section A.4

The restriction of the function g_j to the form (12) of the main manuscript (i.e., (S.12)) restricts also the minimizing function g_j in (5) to the form (S.12) $\big($note, $g_j(\langle X_j, \widehat{\beta}_j\rangle, A) = s_j^{(\lambda)}P_j(\langle X_j, \widehat{\beta}_j\rangle, A)$ in (5) of the main manuscript, where $s_j^{(\lambda)} = [1 - \lambda/\|P_j\|]_+$ is a scaling factor$\big)$. In particular, we can express the function P_j in (6) of the main manuscript as:

$$\begin{aligned} P_j(\langle X_j, \widehat{\beta}_j\rangle, A) &= E[R_j|\langle X_j, \widehat{\beta}_j\rangle, A] - \sum_{a=1}^L \pi_a E[R_j|\langle X_j, \widehat{\beta}_j\rangle, A = a] \\ &= \boldsymbol{\Psi}_j(\langle X_j, \widehat{\beta}_j\rangle)\boldsymbol{\theta}_{j,A}^* - \boldsymbol{\Psi}_j(\langle X_j, \widehat{\beta}_j\rangle)\{\sum_{a=1}^L \pi_a\boldsymbol{\theta}_{j,a}^*\}, \end{aligned} \tag{S.20}$$

where $\{\boldsymbol{\theta}^*_{j,a}\}_{a\in\{1,\ldots,L\}} := \underset{\{\boldsymbol{\theta}_{j,a}\in\mathbb{R}^{d_j}\}_{a\in\{1,\ldots,L\}}}{\operatorname{argmin}} E\left[\left\{R_j - \boldsymbol{\Psi}_j(\langle X_j,\widehat{\beta}_j\rangle)^\top\boldsymbol{\theta}_{j,A}\right\}^2\right]$. In (S.20), the first term, $\boldsymbol{\Psi}_j(\langle X_j,\widehat{\beta}_j\rangle)\boldsymbol{\theta}^*_{j,A}$, corresponds to the L^2 projection of the jth partial residual R_j in (7) (of the main manuscript) onto the class of functions of the form (S.12) (without the imposition of the constraint (S.13), that is, without the constraint $\sum_{a=1}^{L}\pi_a\boldsymbol{\theta}_{j,a} = \mathbf{0}$), whereas the second term, $-\boldsymbol{\Psi}_j(\langle X_j,\widehat{\beta}_j\rangle)\{\sum_{a=1}^{L}\pi_a\boldsymbol{\theta}^*_{j,a}\}$, simply centers the first term to satisfy the linear constraint, $\sum_{a=1}^{L}\pi_a\boldsymbol{\theta}_{j,a} = \mathbf{0}$. Then it follows that P_j, as given in (S.20), corresponds to the L^2 projection of R_j onto the subspace of measurable functions of the form (S.12) subject to the linear constraint (S.13).

A.6. Estimation algorithm

Algorithm 1 Estimation of constrained functional additive models, given each $\lambda \geq 0$

```
**Input**: Data $\boldsymbol{Y}\in\mathbb{R}^n$, $\boldsymbol{A}\in\mathbb{R}^n$, $\boldsymbol{X}_j\in\mathbb{R}^n\times\mathbb{R}^{r_j}$ $(j=1,\ldots,p)$, and $\lambda\geq 0$
**Output**: Estimated functions $\{\widehat{\beta}_j, j=1,\ldots,p\}$ and $\{\widehat{g}_j, j=1,\ldots,p\}$
Initialize $\widehat{\beta}_j(s)=1$ $(s\in[0,1])$ $(j=1,\ldots,p)$.
**while** until convergence of $\{\widehat{\beta}_j, j=1,\ldots,p\}$, **do** iteratate between Step 1 and Step 2:
$\langle$Step 1$\rangle$
    Fix $\{\widehat{\beta}_j, j=1,\ldots,p\}$, and compute $\widetilde{\boldsymbol{D}}_j(\widetilde{\boldsymbol{D}}_j^\top\widetilde{\boldsymbol{D}}_j)^{-1}\widetilde{\boldsymbol{D}}_j^\top$ in (S.17) $(j=1,\ldots,p)$.
    Initialize $\widehat{\boldsymbol{g}}_j=\mathbf{0}\in\mathbb{R}^n$ $(j=1,\ldots,p)$.
    **while** until convergence of $\{\widehat{\boldsymbol{g}}_j, j=1,\ldots,p\}$, **do** iterate through $j=1,\ldots,p$ :
        Compute the partial residual $\widehat{\boldsymbol{R}}_j=\boldsymbol{Y}-\sum_{j'\neq j}\widehat{\boldsymbol{g}}_{j'}$ (in (S.17)).
        Compute $\widehat{\boldsymbol{P}}_j$ in (S.17); then compute the thresholded estimate $\widehat{\boldsymbol{g}}_j$ in (S.16).
$\langle$Step 2$\rangle$
    Fix $\{\widehat{g}_j, j=1,\ldots,p\}$ in (S.19), and solve (15) (of the main text) based on (19); update $\widehat{\beta}_j$ $(j=1,\ldots,p)$.
```

A.7. Computational note

In Algorithm 1, if the jth soft-threshold shrinkage factor $\widehat{s}_j^{(\lambda)}$ in (S.18) is 0, then the associated X_j is absent from the model. Therefore, the corresponding projection function $\widehat{\beta}_j$ will not be updated, and this greatly reduces the computational cost when most of the shrinkage factors $\widehat{s}_j^{(\lambda)}$ are zeros. Furthermore, in Algorithm 1, the smoother matrix $\widetilde{\boldsymbol{D}}_j(\widetilde{\boldsymbol{D}}_j^\top\widetilde{\boldsymbol{D}}_j)^{-1}\widetilde{\boldsymbol{D}}_j^\top$ $(j=1,\ldots,p)$ (defined in (S.17)) needs to be computed only once at the beginning of Step 1 given fixed $\{\widehat{\beta}_j, j=1,\ldots,p\}$, and therefore the coordinate-descent updates in Step 1 can be performed very efficiently (Fan et al., 2014).

Specifically, in Step 1, holding all other components and after computing $\widehat{\boldsymbol{P}}_j$ in (S.17), the update rule (S.16) for each (the jth) block-update just amounts to soft-thresholding. For a fixed set of $\widehat{\beta}_j$ $(j=1,\ldots,p)$ (given from the previous iterate for Step 2), in Step 1, we can estimate g_j $(j=1,\ldots,p)$ on a grid of λ values, from $\lambda_{\max}$ to $\lambda_{\min}$, using warm starts: we can fit a sequence of models from $\lambda_{\max}$ down to $\lambda_{\min}$, where $\lambda_{\max}$ is the smallest value of λ for which all coefficients are zero. Solutions do not change much from one λ to the next. By decreasing λ slowly over a dense grid, construction of a solution path over λ does not require much cycling, and thus Step 1 can be performed very efficiently for all λ values. In Step 2 of the estimation algorithm, the projection functions $\widehat{\beta}_j$ $(j=1,\ldots,p)$ are updated for only the covariate indices j's associated with nonzero component functions, i.e., $\{j|\widehat{g}_j\neq 0\}$ for each fixed λ.

To investigate computation time of the above estimation algorithm, we consider different combinations of number of subjects $n\in\{250,500,750,1000\}$ and number of functional covariates $p\in\{20,40,80\}$. We do not consider the scalar covariates (i.e., $q=0$) to focus on the case of the functional regression. We use the same data generation model as in the simulation section (Section 4.1) of the main manuscript, with $\xi=0$ and $\delta=1$.

Figure S.1 provides the averaged computation time (in seconds), averaged over 200 simulation runs, for the four ITR estimation methods considered in Section 4.1 of the main manuscript. For each method, we optimize the associated tuning parameters via 10-fold cross validation given each scenario and for these optimized tuning parameters, we measure the computation time. All method are implemented in R (R Development Core Team, 2020). Computation times were measured on a MacBook computer running 64-bit, 2.3 GHz Intel Core i7, with 32 GB random access memory. Again, CFAM and CFAM-lin are implemented through the R package `famTEMsel` (Park et al., 2020b) and the outcome weighted learning (OWL) approaches (OWL-lin and OWL-Gauss) are implemented through the R-package `DTRlearn` (Chen et al., 2020).

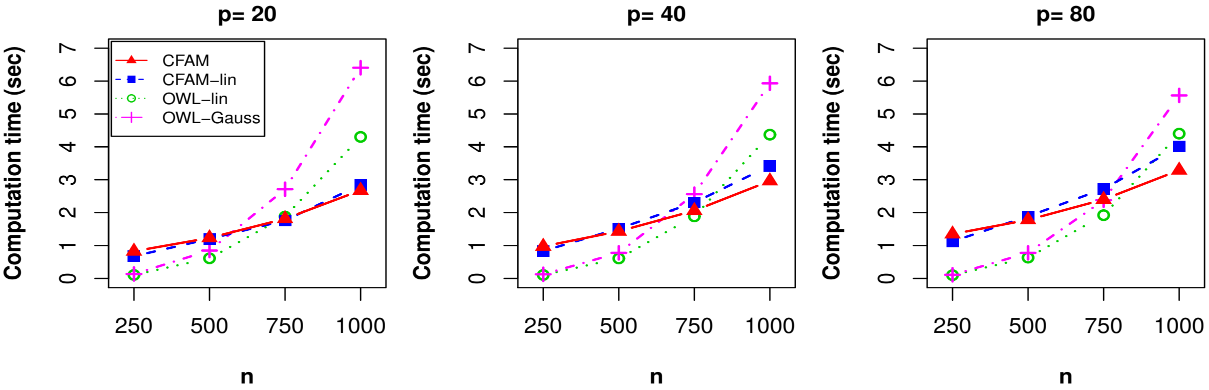

Figure S.1: The averaged computation time (in seconds) with varying $n \in \{250, 500, 750, 1000\}$ for the four estimation methods considered in Section 4.1 of the main manuscript, for each $p \in \{20, 40, 80\}$.

Under the considered simulation settings, the Algorithm 1 converges often in 5 to 8 iterations for CFAM (i.e., the maximum difference in the p estimates, $\widehat{\beta}_j$ $(j = 1, \ldots, p)$, over any two consecutive iterations is less than a prespecified small convergence tolerance). In Figure S.1, for the methods CFAM and CFAM-lin, the computation time tends to increase linearly with n, whereas for the support vector machine (SVM)-based approach (i.e., OWL-lin and OWL-Gauss), the computation time tends to scale exponentially with n, although, when n was small ($n = 250, 500$), training was faster for the SVM-based approaches than CFAM. For CFAM, the computation time depends on the number of iterations alternating between Step 1 and Step 2 in Algorithm 1. We note that Algorithm 1 was also used to fit CFAM-lin by restricting the component function g_j to be linear. In Figure S.1, the computation times for CFAM and CFAM-lin were comparable to each other. This is because CFMA-lin, which is a misspecified linear model, often takes more iterations to converge to its solution than CFAM which is a correctly specified model. However, within each iteration, the computation time for CFAM is typically (slightly) larger than CFAM-lin, since CFAM is required to represent each observation in terms of a nonlinear spline basis. We also note that the computation time reported in Figure S.1 is the time required for completing the training. When making predictions given new data, the OWL approaches can take a substantial amount of additional time to match the new data to the training data in order to make predictions.

Another observation from Figure S.1 is that the model training time of all the methods considered do not necessarily increase with the number of functional covariates (p), given sparsity-inducing (or regularization) tuning parameters, as the computation times are comparable across different $p \in \{20, 40, 80\}$. However, we note that the OWL approaches (OWL-lin and OWL-Gauss) do not directly handle the functional covariates, and prespecified naïve averages of the functional covariates were used as input to the OWL approaches. This is in contrast to the CFAM and CFAM-lin methods that use a supervised fit for projecting the functional predictors into the best one-dimensional space.

A.8. Discussion on the confidence bands associated with the model parameters

In the minimization of (19) in the main text, the discretized coefficient function $\beta_j(s)$ in (18), discretized at $\{s_1, s_2, \ldots, s_{r_j}\}$, is represented by the length-r_j vector $\boldsymbol{B}_j\widehat{\boldsymbol{\gamma}}_j$, where $\widehat{\boldsymbol{\gamma}}_j \in \mathbb{R}^{m_j}$ is the minimizer of (19) (scaled to unit norm for model identifiability). Given this optimization-based representation of the basis coefficient $\boldsymbol{\gamma}_j$ associated with $\beta_j(s)$, the typical inferential machinery (e.g., Ruppert et al., 2003; Wood, 2017) can be used to obtain variance–covariance estimate associated with the discretized function estimate $\widehat{\beta}_j(s)$. Specifically, we can use $\boldsymbol{B}_j\widehat{\boldsymbol{V}}\boldsymbol{B}_j^\top$, where $\widehat{\boldsymbol{V}}$ represents the variance–covariance estimate of $\widehat{\boldsymbol{\gamma}}_j$, based on which a 95% normal-approximated point-wise confidence band for the function $\beta_j(s)$ is constructed, in Figure 4 of the main manuscript. At the time of convergence of the algorithm, the minimizer $\widehat{\boldsymbol{\gamma}}_j$ of (19) essentially satisfies the condition $\|\widehat{\boldsymbol{\gamma}}_j\| = 1$, thus we do not need to re-scale the associated variance–covariance matrix at convergence of the algorithm, when we compute $\widehat{\boldsymbol{V}}$.

Similarly, for the treatment-specific component function $g_j(s,a)$ (and $h_k(s,a)$) ($a = 1, \ldots, L$), appearing in Figure 5 of the main manuscript, we utilize representation (S.12) (estimated by (S.19) given the partial residual $\widehat{\boldsymbol{R}}_j$ and the shrinkage scale $\widehat{s}_j^{(\lambda)}$) to construct a 95% normal-approximated, point-wise confidence band of the function $\widehat{g}_j(s,a)$ (and $\widehat{h}_k(s,a)$) evaluated over the observed values of each function's argument s.

However, an important limitation of this approach is that, for example, the confidence band associated with $\widehat{g}_j(s,a)$ is computed conditional on the estimated projection $s = \langle X_j, \widehat{\beta}_j \rangle$, as well as the other components that constitute the jth partial residual. The computed standard error returned for components does not include the uncertainty about the other conditioning components. However, correctly accounting for the uncertainty in the other components in a projection-pursuit regression is inherently challenging. Specifically, the fact that the domain of $g_j(\cdot, a)$ varies depending on the projection direction estimate $\widehat{\beta}_j(s)$ complicates the confidence band construction for the component functions $g_j(\cdot, a)$ ($a = 1, \ldots, L$). One could potentially consider the regression surface functional $k_j(X_j, a) := g_j(\langle X_j, \beta_j \rangle, a)$, however, due to the infinite dimensionality of the domain of the functional, construction of a confidence band is generally challenging.

One potential alternative approach is to condition on the observed data and perform a posterior inference on the model parameters using a Bayesian framework. We can consider a posterior distribution of $k_j(X_j, A)$ and make probabilistic statements about the prediction of the component $k_j(X_j, A)$ given each $(\boldsymbol{X}, \boldsymbol{Z}, A)$. Our future work will investigate the development of a Bayesian framework for the model accounting for the posterior uncertainty in the parameters β_j g_j, h_k and the unmodeled noise variance, to allow a posterior inference on the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effects and predictions, allowing one to conduct inference for the components of the model.

A.9. Discussion on convergence of the estimation algorithm

Although our approach uses a supervised fit to project the functional predictors into one-dimensional space, the proposed estimation approach (described in Algorithm 1) to fitting the model can be generally suboptimal with respect to the empirical version of (4) of the main manuscript, i.e., the criterion $\|\boldsymbol{Y} - \sum_{j=1}^{p} \boldsymbol{g}_j - \sum_{k=1}^{q} \boldsymbol{h}_k\|^2 + \sqrt{n}\lambda\big\{\sum_{j=1}^{p}\|\boldsymbol{g}_j\| + \sum_{k=1}^{q}\|\boldsymbol{h}_k\|\big\}$, to be optimized over $\tilde{\boldsymbol{\theta}}_j \in \mathbb{R}^{d_j(L-1)}$ given in (S.15) and over $\boldsymbol{\gamma}_j \in \mathbb{R}^{m_j}$ (subject to a unit norm constraint) associated with β_j (see (19) of the main manuscript) (and over the corresponding parameter associated with $\boldsymbol{h}_k$).

The proposed flexible method for modeling high dimensional treatment effect modification of functional covariates adopts the methodology of single index regression, a particular instance of projection pursuit regression. In projection pursuit regression, with or without the estimation of the component functions g_j ($j = 1, \ldots, p$), the optimization for the projection directions β_j entails non-convex estimation problems. For example, for general function g_j, the estimation problem (11) in the main manuscript is generally a non-convex estimation task. This means that the proposed iterative estimation approach in Algorithm 1, in which the iteratively-defined loss function (i.e., (19) of the main manuscript) applied to β_j is adapted as a function of the current estimate of g_j and β_j (based on a local approximation of the objective function around the current estimate of β_j, as described in (16) of the main text) may not converge to a global optimum.

The algorithm requires local convexity of the objective function around the initial estimate, which appears to be the case in our simulation studies as well as in the application example, where the estimation algorithm converges relatively quickly and stably, typically within 5 to 8 iterates. For a fixed $\lambda \geq 0$, Step 1 of the Algorithm 1 converges (Tseng, 2001) and under local convexity of the objective function (11) around the initial estimate, alternating between Step 1 and Step 2 can converge stably.

However, given the high dimensionality of functional covariates and the inherent non-convexity of the least squares loss function in projection pursuit regression, finding a global optimum is considered to be practically challenging. For the case of monotonic component functions g_j, there were some studies to deal with this non-convex optimization. For example, Ravikumar et al. (2008) investigated a procedure involving only tractable convex optimization steps by using appropriate classes of Bregman divergences. However, imposition of a monotonic g_j may not be appropriate, because the treatment effect moderation function can commonly be, for example, quadratically shaped. Thus, our strategy and motivation in this paper was to instead start off with the most common practice of taking a naïve scalar summary of each functional covariate and to improve over these most common but naïve summaries. Alternatively, we can utilize scientifically plausible/informed estimates of β_j or estimates obtained from a functional linear regression as a starting point. In our simulation experiments, β_j optimized to incorporate possibly nonlinear interactions, provides a significant improvement over the initial naïve flat functions, as well as the estimates of β_j obtained based on the linear component functions g_j (i.e., CFAM-lin), as illustrated by the simulations in Section 4.

A.10. Model parameter estimation performance in terms of root squared error

In this subsection, we report the results that supplement the results reported in Section 4.1 of the main manuscript. Specifically, we provide the performance of the proposed estimation method for the model parameters $\{g_j, h_k, \beta_j\}$ when $\xi = 0$ (i.e., when CFAM is correctly specified) with varying $\delta \in \{1, 2\}$ and $n \in \{250, 500, 1000\}$. In particular, we focus on the estimation performance for the model parameters β_1, β_2, g_1, g_2, h_1 and h_2, that are associated with the "signal" covariates X_1, X_2, Z_1 and Z_2 that have nonzero interactions with the treatment variable A. We note that the treatment effect-modifiers selection performance is reported in Section 4.2 of the main manuscript.

As a measure of the estimation performance, for β_j, we report the root squared error $\text{RSE}(\beta_j) = \sqrt{\int_0^1 (\widehat{\beta}_j(s) - \beta_j(s))^2 ds}$ $(j = 1, 2)$, where the parameters β_1 and β_2 are specified in model (20) of the main manuscript, and $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the corresponding CFAM estimates as described in Section 4.1 of the main manuscript. Since the domain of $\beta_j(s)$ is $[0, 1]$ (i.e., bounded), the above RSE can be easily evaluated. However, the domain of the component functions g_j and h_k are unbounded in our setting. In particular, the functions $g_j(s, a)$ and $h_k(Z_k, a)$ $(a = 1, 2)$ can be estimated only over the observed range of $s = \langle X_j, \beta_j \rangle$ (and Z_k). Since we know the observed values of the "true" $s_i = \langle X_{ij}, \beta_j \rangle$ (and Z_{ki}) $(i = 1, \ldots, n)$ (for each simulation run) where β_j is the true value, we can truncate the domain of the functions g_j and h_k based on the observed "true" s_i, and evaluate them on a truncated range of the "true" index $\langle X_{ij}, \beta_j \rangle$ (and Z_k) (to be indicated below), for each simulation run, to compute the RSE.

Under the data generating model (20) of the main manuscript, the true component functions: $g_1(s, a) = 4(a - 1.5)\sin(s)$ and $g_2(s, a) = -4(a - 1.5)\sin(s)$. If we set $a = 2$, then $g_1(s, 2) = 2\sin(s)$ and $g_2(s, 2) = -2\sin(s)$. Since the component function g_j for $a = 1$ is completely determined by that for $a = 2$ due to the identifiability condition (2) that we impose on g_j, it is sufficient to focus only on the function $g_j(s, 2)$ when evaluating the performance of the component function estimate $\widehat{g}_j$. Similarly, we can write $h_1(s, 2) = 2\cos(s)$ and $h_2(s, 2) = -2\cos(s)$ under the given simulation setting. Then we report the root squared error $\text{RSE}(g_j) = \sqrt{\int_{c_1}^{c_2} (\widehat{g}_j(s, 2) - g_j(s, 2))^2 ds}$ $(j = 1, 2)$ and $\text{RSE}(h_k) = \sqrt{\int_{c_1}^{c_2} (\widehat{h}_k(s, 2) - h_k(s, 2))^2 ds}$ $(k = 1, 2)$ for g_j and h_k, respectively, where c_1 and c_2 correspond to the 5% and 95% quantiles of the "true" observed values of s (we excluded the tails of each variable because they often have very few observed values). In addition to RSE, we also report the optimal ITR estimation performance of CFAM, in terms of the (normalized) value, $V^*(\widehat{\mathcal{D}}^{opt}) = V(\widehat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$ (where a larger value of $V^*(\widehat{\mathcal{D}}^{opt})$ is desired). In Table S.1, we report the mean (and standard deviation) of these performance measures obtained from 200 simulation replications for each scenario.

Table S.1: The CFAM parameter estimation performance assessed by the root squared error RSE(β_j), RSE(g_j), and RSE(h_k) (a smaller value of RSE is desired) and the optimal ITR estimation performance assessed by $V^*(\widehat{\mathcal{D}}^{opt}) = V(\widehat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$ (a larger value of $V^*(\widehat{\mathcal{D}}^{opt})$ is desired), for varying $\delta \in \{1, 2\}$ and $n \in \{250, 500, 1000\}$. The entries report the mean (and standard deviation) obtained from 200 simulation replications for each scenario.

	$\delta = 1$ (*Moderate* "main" effect)			$\delta = 2$ (*Large* "main" effect)		
	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
RSE(β_1)	0.53(0.08)	0.34(0.02)	0.26(0.02)	0.60(0.14)	0.38(0.05)	0.29(0.03)
RSE(β_2)	0.53(0.06)	0.34(0.02)	0.27(0.01)	0.59(0.13)	0.39(0.07)	0.29(0.03)
RSE(g_1)	0.20(0.07)	0.14(0.05)	0.11(0.05)	0.34(0.33)	0.23(0.11)	0.20(0.09)
RSE(g_2)	0.22(0.08)	0.17(0.09)	0.12(0.05)	0.32(0.25)	0.25(0.15)	0.20(0.10)
RSE(h_1)	0.31(0.09)	0.15(0.04)	0.12(0.03)	0.36(0.10)	0.21(0.09)	0.16(0.06)
RSE(h_2)	0.31(0.10)	0.15(0.05)	0.12(0.04)	0.36(0.10)	0.21(0.09)	0.16(0.06)
$V^*(\widehat{\mathcal{D}}^{opt})$	-0.07(0.04)	-0.03(0.01)	-0.01(0.01)	-0.16(0.07)	-0.07(0.02)	-0.04(0.01)

In Figure 2 of the main manuscript, an illustration of typical 10 CFAM sample estimates $\widehat{\beta}_j(s)$ for the parameters $\beta_j(s)$ for $j = 1$ and 2 when $\xi = 1$ is provided. The results in Table S.1 indicate that the estimation performance for all the nonzero model parameters $\{\beta_j, g_j, h_k\}$, as measured by the corresponding RSE, improves with an increasing n, in both cases of $\delta = 1$ (moderate "main" effect) and $\delta = 2$ (large "main" effect). In particular, when $n = 1000$ and $\delta = 1$, the ITR estimation performance is very close to the optimal one ($V^*(\widehat{\mathcal{D}}^{opt}) = -0.01$), indicating that the proposed estimation approach performs very well in this setting.

A.11. Separate modeling of the $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect component

Under model (S.1), constraint (S.2) (i..e, constraint (2) of the main manuscript) ensures that

$$E\left[\mu(\boldsymbol{X},\boldsymbol{Z})\left\{\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle, A)+\sum_{k=1}^{q} h_k(Z_k,A)\right\}\right] = E\left[\mu(\boldsymbol{X},\boldsymbol{Z})E\left\{\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle, A)+\sum_{k=1}^{q} h_k(Z_k,A) \mid \boldsymbol{X},\boldsymbol{Z}\right\}\right] = 0,$$

where, on the right-hand side, we apply the iterated expectation rule to condition on $(\boldsymbol{X}, \boldsymbol{Z})$, which implies:

$$\mu(\boldsymbol{X},\boldsymbol{Z}) \quad \perp \quad \sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle, A)+\sum_{k=1}^{q} h_k(Z_k,A) \tag{S.21}$$

in L^2. The orthogonality (S.21) implies that, under the squared error minimization criterion, the optimization for μ and the components $\{g_j, \beta_j, h_k\}$ in model (S.1) (subject to (S.2)) can be performed separately, without iterating between the two optimization procedures. To be specific, we can solve for the $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect:

$$\mu^* \quad = \quad \underset{\mu\in\mathcal{H}}{\text{argmin}} \quad E\left[\{Y-\mu(\boldsymbol{X},\boldsymbol{Z})\}^2\right], \tag{S.22}$$

and can separately solve for the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect via optimization (S.4). In optimization (S.22), $\mathcal{H}$ represents a (possibly misspecified) L^2 space of functionals over $(\boldsymbol{X}, \boldsymbol{Z})$. Even if the true μ in (S.1) is not in the class $\mathcal{H}$, the representation (S.4) that specifies the optimal $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect components $\{g_j^*, \beta_j^*, h_k^*\}$ is not affected by the possible misspecification for μ, due to the orthogonality (S.21). Thus, the property (S.21) is both conceptually and practically appealing for the estimation of the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect components.

For the case of a continuous outcome Y, Tian et al. (2014) (in the linear regression context with scalar-valued covariates and Lasso regularization) and Park et al. (2020a) (in the single-index regression context with scalar-valued covariates) proposed to separately model and fit the main effect component μ^* using the approach (S.22), by exploiting the orthogonality property analogous to (S.21), and then use the residualized outcome $Y - \widehat{\mu}^*(\boldsymbol{X}, \boldsymbol{Z})$ (instead of using the original outcome Y) for the estimation of the interaction effect components. This residualization procedure that uses a separately fitted main effect model was termed *efficiency augmentation* by Tian et al. (2014), and can improve the efficiency of the estimator for the interaction effect component (while maintaining the consistency of the estimator). In what follows, we illustrate an additional set of simulations supplementing the results of Section 4.1 of the main manuscript, to demonstrate some performance improvement of the CFAM method by an efficiency augmentation procedure.

Under the simulation model (20) of the main manuscript (with the corresponding estimation performance results reported in Table S.1 in Section A.10) for generating the data, we report additional simulation results from the CFAM method with efficiency augmentation, where the $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect component of the data generating model (20) is separately modeled by a functional additive regression, i.e., by the model: $\mu(\boldsymbol{X}, \boldsymbol{Z}) = \sum_{j=1}^{p} \widetilde{g}_j(\langle X_j, \widetilde{\beta}_j\rangle) + \sum_{k=1}^{q} \widetilde{h}_k(Z_k)$, estimated based on an L^1 regularization that is similar to (4) of the main manuscript, with the associated tuning parameters selected as in the CFAM method, using a 10-fold cross validation) and the corresponding residualized outcome is used to implement the CFAM method.

In Table S.2 below, we report the estimation performance of this approach, assessed by RSE(β_j), RSE(g_j) and RSE(h_k) associated with the model parameters $\{\beta_j, g_k, h_k\}$ and the overall ITR estimation performance is assessed by $V^*(\widehat{\mathcal{D}}^{opt})$, as in Section A.10.

Table S.2: The CFAM parameter estimation performance, for the case where the efficiency augmentation is implemented via the functional additive regression model for the $(\boldsymbol{X}, \boldsymbol{Z})$ main effect, assessed by the root squared error RSE(β_j), RSE(g_j), and RSE(h_k) (a smaller value of RSE is desired) and the optimal ITR estimation performance assessed by $V^*(\widehat{\mathcal{D}}^{opt}) = V(\widehat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$ (a larger value of $V^*(\widehat{\mathcal{D}}^{opt})$ is desired), for varying $\delta \in \{1, 2\}$ and $n \in \{250, 500, 1000\}$. The entries report the mean (and standard deviation) obtained from 200 simulation replications for each scenario.

	$\delta = 1$ (*Moderate* "main" effect)			$\delta = 2$ (*Large* "main" effect)		
	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
RSE(β_1)	0.52(0.06)	0.33(0.02)	0.26(0.01)	0.57(0.13)	0.36(0.03)	0.28(0.02)
RSE(β_2)	0.52(0.06)	0.33(0.01)	0.26(0.01)	0.56(0.11)	0.36(0.04)	0.28(0.02)
RSE(g_1)	0.19(0.07)	0.13(0.05)	0.09(0.04)	0.27(0.20)	0.19(0.09)	0.15(0.07)
RSE(g_2)	0.22(0.08)	0.17(0.08)	0.11(0.05)	0.27(0.19)	0.20(0.11)	0.16(0.08)
RSE(h_1)	0.30(0.09)	0.14(0.04)	0.11(0.03)	0.35(0.10)	0.19(0.08)	0.14(0.05)
RSE(h_2)	0.29(0.09)	0.14(0.04)	0.11(0.03)	0.34(0.10)	0.18(0.08)	0.14(0.05)
$V^*(\widehat{\mathcal{D}}^{opt})$	-0.06(0.04)	-0.02(0.01)	-0.01(0.00)	-0.12(0.05)	-0.05(0.01)	-0.02(0.01)

By comparing the entries of Table S.2 with those of Table S.1, one can observe that the CFAM estimation with efficiency augmentation using the functional additive regression model for the term $\mu(\boldsymbol{X}, \boldsymbol{Z})$ (whose performance is reported in Table S.2 above) improves the estimation without the functional additive regression model for the term $\mu(\boldsymbol{X}, \boldsymbol{Z})$ (whose performance is reported in Table S.1). The superiority of this efficiency augmentation appears to be particularly prominent when $\delta = 2$ (i.e, for the large "main" effect cases), in terms of both the model parameter estimation performance, i.e., RSE(β_j), RSE(g_j), and RSE(h_k), and the optimal ITR estimation performance, i.e., $V^*(\widehat{\mathcal{D}}^{opt})$.

A.12. Simulation results under a “linear” A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect scenario

In this subsection, as an extension of the simulation example in Section 4.1 of the main manuscript, we consider a case where the treatment effect varies *linearly* in the covariates $(\boldsymbol{X}, \boldsymbol{Z})$, i.e., a “*linear*” A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect scenario and assess the ITR estimation performance of the methods. Specifically, we consider the data generation model:

$$\begin{aligned} Y_i = \delta\Big\{\sum_{j=1}^{8}\sin(\langle\eta_j, X_{ij}\rangle) + \sum_{k=1}^{8}\sin(Z_{ik})\Big\} &+ \\ 4(A_i - 1.5)\Big[\langle\beta_1, X_{i1}\rangle/1.5 - \langle\beta_2, X_{i2}\rangle/1.5 + Z_{i1}/1.5 - Z_{i2}/1.5 + \xi\big\{\langle X_{i1}, X_{i2}\rangle/1.5 + Z_{i1}Z_{i2}/1.5\big\}\Big] &+ \epsilon_i, \end{aligned} \tag{S.23}$$

in which, when $\xi = 0$, a functional *linear* model specifies the A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect term (i.e., the second term on the right-hand side of (S.23)). However, when $\xi = 1$, the underlying model (S.23) deviates from the exact linear A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect structure, and in such a case, the model CFAM-lin (as well as CFAM) is misspecified. The contribution to the variance of Y from the $(\boldsymbol{X}, \boldsymbol{Z})$ main and the A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect terms in (S.23) was made similar to that of the data generating model (25) of Section 4.1 of the main manuscript.

Figure S.2 below illustrates the boxplots, obtained from 200 simulation runs, of the normalized values $V(\widehat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$ (normalized by the optimal values $V(\mathcal{D}^{opt})$) of the decision rules $\widehat{\mathcal{D}}^{opt}$ estimated from the four ITR estimation approaches described in Section 4.1 of the main manuscript, for each combination of $n \in \{250, 500\}$, $\xi \in \{0, 1\}$ (corresponding to *correctly-specified* or *mis-specified* CFAM scenarios, respectively) and $\delta \in \{1, 2\}$ (corresponding to *moderate* or *large* main effects, respectively).

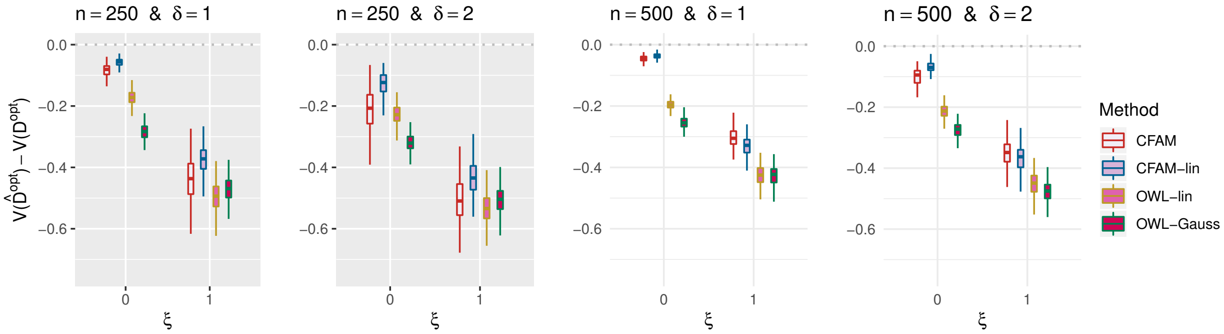

Figure S.2: Boxplots obtained from 200 Monte Carlo simulations comparing 4 approaches to estimating $\mathcal{D}^{opt}$, given each scenario indexed by $\xi \in \{0, 1\}$, $\delta \in \{1, 2\}$ and $n \in \{250, 500\}$. The dotted horizontal line represents the optimal value corresponding to $\mathcal{D}^{opt}$.

The results in Figure S.2 indicate that, in all scenarios with $\xi = 0$ (i.e., when the linear interaction model is correctly specified), CFAM-lin outperforms CFAM, but by a relatively small amount in comparison to the difference in performance appearing in Figure 1 of the main manuscript, in which CFAM outperforms CFAM-lin. Moreover, if the underlying model deviates from the exact linear structure (i.e., $\xi = 1$ in model (S.23)) and $n = 500$, the more flexible CFAM tends to outperform CFAM-lin. Given the outstanding performance of CFAM compared to CFAM-lin in the *nonlinear* A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect scenarios considered in the main manuscript, this result suggests that, in the absence of prior knowledge about the form of the interaction effect, flexible modeling of the interaction effect using the proposed CFAM can lead to at least comparable or better results in comparison to CFAM-lin.

A.13. The estimated model parameters of CFAM-lin

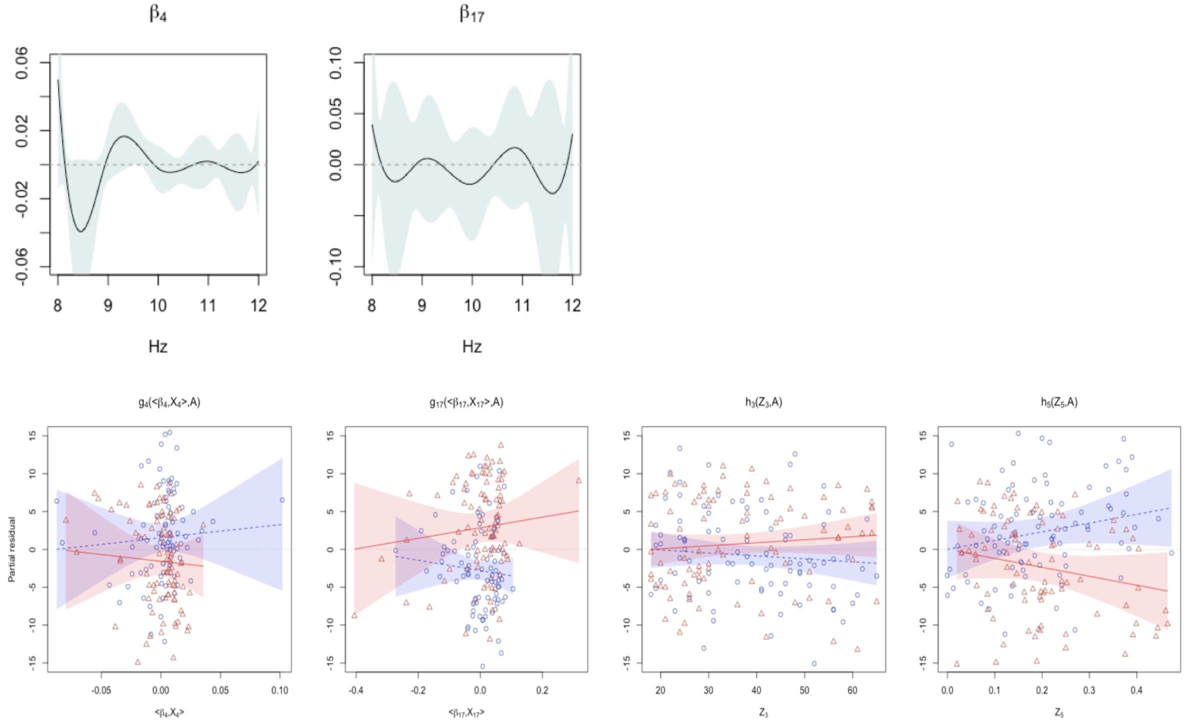

Figure S.3: **Top**: The estimated single-index coefficient functions (β_4 and β_{17}) for the selected channels X_4 (electrode "C3") and X_{17} (electrode "T5") from CFAM-lin (and the associated 95% confidence bands, given the jth partial residual and $\widehat{g}_j$). **Bottom**: The scatter plots of the CFAM-lin's (4th and 17th) partial residual vs. the estimated single-indices $\langle X_4, \beta_4 \rangle$ and $\langle X_{17}, \beta_{17} \rangle$, respectively (in the left two panels), and those of the (3rd and 5th) partial residual vs. the scalar covariates Z_3 (age) and Z_5 (Flanker accuracy test), respectively (in the right two panels), for the placebo $A = 1$ (blue circles) and sertraline $A = 2$ (red triangles) treated individuals. The estimated treatment A-specific component functions $g_j(\langle X_j, \beta_j \rangle, A)$ $(A = 1, 2)$ $(j = 4, 17)$ and $h_k(Z_k, A)$ $(A = 1, 2)$ $(k = 3, 5)$ are overlaid (with the associated 95% confidence bands, given the jth (kth) partial residual and $\widehat{\beta}_j$), for the placebo condition $(A = 1)$ in blue dotted lines and for the sertraline condition $(A = 2)$ in red solid lines.

As referenced in Section 5 of the main manuscript, Figure S.3 reports the estimated model parameters (β_j, g_j and h_k) from CFAM-lin. Both CFAM and CFAM-lin selected the functional covariate X_4 (electrode "C3") and the scalar covariate Z_5 ("Flanker accuracy test score") as treatment effect-modifiers. The estimated coefficient function β_4 (in the top left panel in Figure S.3) for X_4 from CFAM-lin appears to have a similar shape as that of CFAM (see Figure 4 of the main manuscript for comparison). However, as the component functions g_j are restricted to be linear in CFAM-lin, the shape of the component function g_4 from CFAM-lin (see the bottom left panel in Figure S.3) and that from CFAM (see Figure 5 of the main manuscript) for the selected variable X_4 appears to be quite different. In contrast to CFAM, CFAM-lin unselected the covariate X_5 (electrode "P3"), and instead selected X_{17} (electrode "T5") and Z_3 ("Word fluency test score"). Although the ITR performance (displayed in Figure 6 of the main manuscript) is very similar between CFAM-lin and CFAM, looking at their component functions, the proposed CFAM appears to provides a more natural fit to the data compared to CFAM-lin, as it allows more flexibility to model the component functions. The data-driven projection functions β_j and component functions g_j, that provide better fidelity to the data, while retaining simplicity and interpretability, can be useful for understanding scientific basis behind treatment effect moderation.

In Table S.3 below, we report the cross-stratification tables comparing the four ITR approaches (CFAM, CFAM-lin, OWL-lin, OWL-Gauss) considered in Section 5 of the main manuscript, in terms of recommended treatments from the rules, on the depression study data ($n = 180$).

Table S.3: The cross-tables of the recommended treatments (Placebo or Sertraline) comparing the four ITRs (CFAM, CFAM-lin, OWL-lin, OWL-Gauss), for the $n = 180$ subjects considered in Section 5 of the main manuscript. Each entry reports the number (and proportion in %) of subjects classified into the respective cross-stratification category.

		CFAM-lin		
		Placebo	Sertraline	
CFAM	Placebo	40 (22%)	7 (4%)	47 (26%)
	Sertraline	19 (11%)	114 (63%)	133 (74%)
		59 (33%)	121 (67%)	

		OWL-lin		
		Placebo	Sertraline	
CFAM	Placebo	30 (17%)	17 (9%)	47 (26%)
	Sertraline	37 (21%)	96 (53%)	133 (74%)
		67 (38%)	113 (62%)	

		OWL-Gauss		
		Placebo	Sertraline	
CFAM	Placebo	32 (18%)	15 (8%)	47 (26%)
	Sertraline	42 (23%)	91 (51%)	133 (74%)
		74 (41%)	106 (59%)	

		OWL-lin		
		Placebo	Sertraline	
CFAM-lin	Placebo	36 (20%)	23 (13%)	59 (33%)
	Sertraline	31 (17%)	90 (50%)	121 (67%)
		67 (37%)	113 (63%)	

		OWL-Gauss		
		Placebo	Sertraline	
CFAM-lin	Placebo	35 (19%)	24 (13%)	59 (32%)
	Sertraline	39 (22%)	82 (46%)	121 (68%)
		74 (41%)	106 (59%)	

		OWL-Gauss		
		Placebo	Sertraline	
OWL-lin	Placebo	48 (27%)	19 (11%)	67 (38%)
	Sertraline	26 (14%)	87 (48%)	113 (62%)
		74 (41%)	106 (59%)	

In Table S.3, the proportion of agreement of the recommended treatments between CFAM and CFAM-lin is 85%(= 22% + 63%), between CFAM and OWL-lin is 70%(= 17% + 53%), and that of CFAM and CFAM-Gauss is 69%(= 18% + 51%). Expectedly, CFAM behaved more similarly to CFAM-lin than to the SVM-based approaches (OWL-lin and OWL-Gauss) in making treatment recommendations. Although the two methods chose different sets of predictors for optimizing the corresponding ITRs, both CFAM and CFAM-lin assumed an additivity of the selected covariates' effects on the heterogeneous treatment responses, with CFAM giving a more flexible additivity than CFAM-lin, while sharing two common predictors, under a functional regression model specifically designed for the heterogeneous treatment effects. The Pearson correlation between the associated treatment benefit indexes (defined as $\widehat{f}(\boldsymbol{X}, \boldsymbol{Z}) = \sum_{j=1}^{p} \widehat{g}_j(\langle X_j, \widehat{\beta}_j \rangle, 2) + \sum_{k=1}^{q} \widehat{h}_k(Z_k, 2)$, see Section 5 of the main manuscript) for CFAM and CFAM-lin was 0.73.

A.14. Noisy and sparse functional covariates case

Several "preprocessing" steps are typically taken before modeling the data. Aside from smoothing the functional data, in some cases it is appropriate to apply registration or feature alignment, or if the grid points differ across observations, to interpolate to a dense common grid (Reiss et al., 2017). Measurement error is expected to be low in some (e.g., chemometric and EEG power spectral analysis considered in the main text) applications but when it is not it can have important effects on the regression relation. In particular, some methods (e.g., James (2002)) account explicitly for such error. In the paper, we assumed the functional data observed on a common dense grid with negligible error. When it is not the case, an initial step to de-noise and re-construct the underlying curves is required, which was the general approach taken in Goldsmith et al. (2011) in their functional linear regression model estimation.

In this subsection, we consider a set of simulations for the case where the functional predictors are not directly observed but observed with measurement errors. As in Section 4 of the main manuscript, we generate n independent copies of p functional-valued covariates $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ $(i = 1, \ldots, n)$, where we use a 4-dimensional Fourier basis, $\boldsymbol{\Phi}(s) = (\sqrt{2}\sin(2\pi s), \sqrt{2}\cos(2\pi s), \sqrt{2}\sin(4\pi s), \sqrt{2}\cos(4\pi s))^\top \in \mathbb{R}^4$ $(s \in [0,1])$, and random coefficients $\widetilde{\boldsymbol{x}}_{ij} \in \mathbb{R}^4$, each independently following $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_4)$, to form the functional covariates $X_{ij}(s) = \boldsymbol{\Phi}(s)^\top \widetilde{\boldsymbol{x}}_{ij}$ $(s \in [0,1])$ $(i = 1, \ldots, n;\ j = 1, \ldots, p)$. We consider the situation where we measure, instead of $X_{ij}(s)$ directly, a proxy functional covariate $W_{ij}(s)$,

$$W_{ij}(s) = X_{ij}(s) + u_{ij}(s), \tag{S.24}$$

where $u_{ij}(s) \sim \mathcal{N}(0,1)$. Specifically, we observe the functions $W_{ij}(s)$ $(i = 1, \ldots, n; j = 1, \ldots, p)$ at points $s_{ijl} \in [0,1]$ $(l = 1, \ldots, r_{ijl})$, for which we consider two cases: 1) the functions are measured on a regular grid of $r_{ijl} = 50$ equally-spaced points, $\{s_{ij1}, \ldots, s_{ij50}\}$ (for each i and j); 2) the functions are measured on a sparse grid of points, where the number of points $r_{ijl} \in \{5, 6, 7, 8\}$ randomly takes a value with equal probabilities and then the points $s_{ijl} \sim \text{Unif}[0,1]$ $(l = 1, \ldots, r_{ijl})$, for each i and j.

When the functions $X_{ij}(s)$ are observed on a set of sparse grid points possibly subject to measurement errors $u_{ij}(s)$, we recommend to use the principal component (PC) decomposition in the first step of the analysis. We estimate $X_{ij}(s)$ using a finite series expansion into the PC basis obtained from its covariance operator. We use the PC decomposition to represent the functions with a small number of bases. Indeed, using decompositions of the functional covariates in terms of other bases (e.g., P-splines (Marx and Eilers, 1999), Fourier basis) is straightforward.

We note that the covariance operator for $W_{ij}(s)$ in (S.24) corresponds to $\text{cov}\{W_{ij}(s), W_{ij}(v)\} = \text{cov}\{X_{ij}(s), X_{ij}(v)\} + \delta_{sv}$, where $\delta_{sv} = 1$ if $s = v$ and is 0, otherwise. Employing the work of Yao et al. (2005), we first use a fine grid of points on $[0,1]$ to obtain an undersmooth of the observed covariance matrix (using a very small bandwidth smoother to obtain a rough estimate of the covariance operator for sparsely observed subject-specific functional regressors, see Di et al. (2009)). We then smooth the off-diagonal elements of this rough covariance matrix to estimate the covariance operator, $\text{cov}\{X_{ij}(s), X_{ij}(v)\}$, of the functional regressors $X_{ij}(s)$ in (S.24), which is then evaluated on our regular grid $s_1, \ldots, s_{r_j}$ defined in Section 3.2.2 of the main text, to estimate the proposed model.

Specifically, the spectral decomposition of the estimated covariance function yields a finite series expansion of the subject-specific functional regressors $X_{ij}(s)$ in (S.24), using conditional expectation (Yao et al., 2005) in the basis of eigenfunctions, evaluated on these grid points. We employed the simple criterion of using percentage (90%) of explained variance to select the finite truncation of the eigenfunction expansion in this simulation study. For the both cases (of the sparsely and the densely measured functions), we use `fpca.sc()` (Di et al., 2009) from the R package `refund` to carry out the functional principal component analysis.

In this simulation, the scalar-valued covariates and the outcomes are generated in the same way as in the simulation illustrated in Section 4 of the main manuscript, using the same set of parameters with $p = q = 20$ and varying $n \in \{250, 500\}$, $\xi \in \{0, 1\}$ and $\delta \in \{1, 2\}$. The results are reported in Figures S.4 and S.5 below, for the case where the functions are measured on a regular grid of 50 equally-spaced points (in Figures S.4), and for the case where the functions are measured on a sparse grid of points (in Figures S.5), respectively.

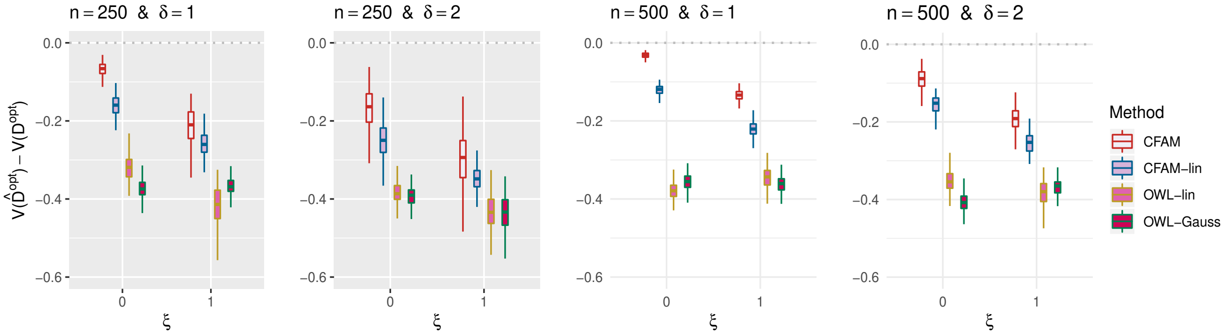

Figure S.4: For the case of densely measured functional covariates subject to measurement errors, we report the boxplots of the normalized values $V(\widehat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$, obtained from 200 Monte Carlo simulations, comparing the 4 estimation approaches, for each simulation setting indexed by $n \in \{250, 500\}$, $\delta \in \{1, 2\}$ and $\xi \in \{0, 1\}$. The dotted horizontal line represents the optimal value corresponding to $\mathcal{D}^{opt}$.

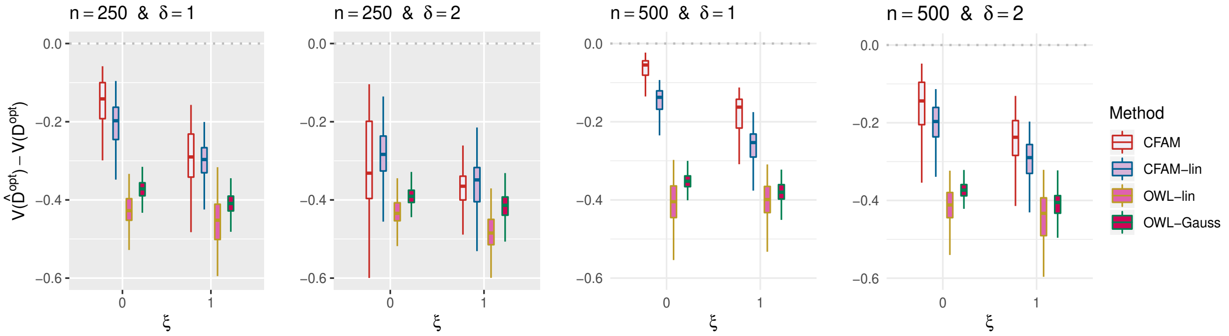

Figure S.5: For the case of sparsely and irregularly measured functional covariates subject to measurement errors, we report the boxplots of the normalized values $V(\widehat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$, obtained from 200 Monte Carlo simulations, comparing the 4 estimation approaches, for each simulation setting indexed by $n \in \{250, 500\}$, $\delta \in \{1, 2\}$ and $\xi \in \{0, 1\}$. The dotted horizontal line represents the optimal value corresponding to $\mathcal{D}^{opt}$.

In Figure S.4 that reports the results for the regularly observed functions with measurement errors, the proposed approach (CFAM) outperforms the other approaches. In particular, its performance is comparable to that presented in Figure 1 of the main manuscript, where the functional covariates are measured without error.

Given sparsely observed functional covariates, the de-noising and re-construction (preprocessing) performance associated with the initial step of the analysis generally degrades. Due to the sparse sampling nature, an accurate prediction of the underlying functions observed with errors would require more subjects than what would be required for the case of densely measured functional covariates subject to measurement errors. In Figure S.5 that reports the results for the irregularly and sparsely observed (5 to 8 irregularly spaced points) functions with measurement errors, when $n = 250$ and a large nuisance effect ($\delta = 2$) is present, the relative performance of CFAM has degraded. Given a set of inaccurately recovered functional covariates, the simpler method (CFAM-lin) tends to work relatively well. However, with the increased sample size ($n = 500$), especially for the case of a correctly specified CFAM (i.e., when $\xi = 0$), the relative performance of CFAM significantly improves compared to the case of $n = 250$ and is close to the optimal one.

A.15. Treatment benefit index parametrization

In model (1) of the main manuscript (i.e., model (S.1) of this document), without loss of generality, we assumed that the treatment's main effect is centered at 0, i.e., $E[Y|A = a] = 0$ $(a = 1, \dots, L)$. This is only to suppress the treatment a-specific intercepts in the regression model in order to simplify the exposition, and can be achieved by removing the treatment level a-specific means from Y. In model (1), $E[Y|A = a] = 0$ also indicates that $E[\mu(\boldsymbol{X}, \boldsymbol{Z})] = 0$, and since the model is subject to constraint (2) (i.e., constraint (S.2)), all the additive components $\{\mu, g_j, h_k\}$ of the model are centered at 0.

In the more general case where the treatment's main effect is not centered at 0, we introduce treatment-specific intercepts, $\tau_a \in \mathbb{R}$ $(a = 1, \dots, L)$, in the model. Then model (S.1) can be written as:

$$Y \ = \mu(\boldsymbol{X}, \boldsymbol{Z}) + \tau_a + \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, a) + \sum_{k=1}^{q} h_k(Z_k, a) \ + \ \epsilon, \tag{S.25}$$

subject to the identifiability conditions (S.2) and $E[\mu(\boldsymbol{X}, \boldsymbol{Z})] = 0$ (the latter constraint can be also easily absorbed into the estimation by appropriately constraining the basis elements of μ in the estimation). For the most common situation of binary treatments (i.e., when $L = 2$), in model (S.25), $\tau_2 - \tau_1$ represents the marginal treatment effect (comparing $a = 2$ with $a = 1$).

Given the underlying model (S.25) (with $L = 2$), let us define a 1-dimensional index,

$$f(\boldsymbol{X}, \boldsymbol{Z}) := \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, a = 2) + \sum_{k=1}^{q} h_k(Z_k, a = 2),$$

which parameterizes the treatment response contrast (evaluating the efficacy of $a = 2$ vs. $a = 1$) as follows:

$$\begin{aligned}
&E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = 2] - E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = 1] \\
&= \tau_2 - \tau_1 + \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, a = 2) + \sum_{k=1}^{q} h_k(Z_k, a = 2) - \Big\{ \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, a = 1) + \sum_{k=1}^{q} h_k(Z_k, a = 1) \Big\} \\
&= \tau_2 - \tau_1 + f(\boldsymbol{X}, \boldsymbol{Z}) + \frac{\pi_2}{\pi_1} f(\boldsymbol{X}, \boldsymbol{Z}) \\
&= \tau_2 - \tau_1 + f(\boldsymbol{X}, \boldsymbol{Z}) \frac{1}{\pi_1},
\end{aligned} \tag{S.26}$$

as a linear function, with $\tau_2 - \tau_1$ acting as the intercept and $\frac{1}{\pi_1}$ acting as the slope. This function $f(\boldsymbol{X}, \boldsymbol{Z})$ provides a continuous gradient of the benefit from one treatment $(a = 2)$ to another $(a = 1)$. In (S.26), the second equality follows from the following condition implied by constraint (S.2) (i.e., constraint (2) of the main manuscript): $E\big[\sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, A) + \sum_{k=1}^{q} h_k(Z_k, A) | \boldsymbol{X}, \boldsymbol{Z}\big] = \sum_{a=1}^{2} \pi_a \big\{ \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, a) + \sum_{k=1}^{q} h_k(Z_k, a) \big\} = \pi_1 \big\{ \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, 1) + \sum_{k=1}^{q} h_k(Z_k, 1) \big\} + \pi_2 \big\{ f(\boldsymbol{X}, \boldsymbol{Z}) \big\} = 0$. In (S.26), the third equality follows from the property that probabilities $\pi_a = \mathrm{pr}(A = a)$ sum to 1, i.e., $\pi_1 + \pi_2 = 1$.

Expression (S.26) indicates that the differential treatment effect is conveniently and continuously indexed by the patient-specific parsimonious index $f(\boldsymbol{X}, \boldsymbol{Z})$, a one-dimensional summary of the patient's pretreatment characteristics $(\boldsymbol{X}, \boldsymbol{Z})$. This is parsimonious because the $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect term $\mu(\boldsymbol{X}, \boldsymbol{Z})$ is separately specified in (S.25). We can assess each individual's benefit from one treatment versus the other using the index $f(\boldsymbol{X}, \boldsymbol{Z})$ and may consider providing a more refined decision using three or more group (see the bottom right panel of Figure 5 of the main manuscript) than a simple binary recommendation.

A.16. Suboptimality of the proposed approach to optimizing ITRs when A depends on $(\boldsymbol{X}, \boldsymbol{Z})$

Let $Y^{(a)} \in \mathbb{R}$ be the potential outcome under treatment $A = a$ (as defined in Section 2 of the main manuscript). As referenced in Section 2 of the main manuscript, in this paper we assume the following standard causal inference assumptions (Rubin, 1974): Assumption 1) consistency, i.e., $A = a$ implies $Y = Y^{(a)}$; Assumption 2) no unmeasured confoundedness, i.e., conditional independence $Y^{(a)} \perp A$ given $(\boldsymbol{X}, \boldsymbol{Z})$; Assumption 3) positivity, i.e., for every covariate $(\boldsymbol{X}, \boldsymbol{Z})$, the probability of receiving every level of treatment is positive.

In this subsection, we discuss the potential suboptimality of the proposed approach to optimizing ITRs when A depends on $(\boldsymbol{X}, \boldsymbol{Z})$. Under the aforementioned causal inference assumptions, the functions P_j $(j = 1, \ldots, p)$ in (6) in Section 3.1 of the main manuscript:

$$P_j(\langle X_j, \beta_j \rangle, A = a) \;=\; E[R_j | \langle X_j, \beta_j \rangle, A = a] \;-\; E[R_j | \langle X_j, \beta_j \rangle] \qquad (j = 1, \ldots, p), \tag{S.27}$$

in which

$$R_j = Y - \sum_{j' \neq j} g_{j'}(\langle X_{j'}, \beta_{j'} \rangle, A = a) - \sum_{k=1}^{q} h_k(Z_k, A = a), \tag{S.28}$$

can be ideally defined in terms of the potential outcome framework, as:

$$P_j(\langle X_j, \beta_j \rangle, A = a) \;:=\; E[R_j^{(a)} | \langle X_j, \beta_j \rangle] \;-\; E[R_j | \langle X_j, \beta_j \rangle] \qquad (j = 1, \ldots, p), \tag{S.29}$$

in which

$$R_j^{(a)} = Y^{(a)} - \sum_{j' \neq j} g_{j'}(\langle X_{j'}, \beta_{j'} \rangle, A = a) - \sum_{k=1}^{q} h_k(Z_k, A = a). \tag{S.30}$$

If the treatment A is independent of $(\boldsymbol{X}, \boldsymbol{Z})$ (as can happen in randomized studies), (S.29) reduces to (S.27), i.e., (6) of the main manuscript. The right hand side of (S.27) can be estimated from observed data, for example, using the procedure described in Section 3.2, for each fixed set of β_j $(j = 1, \ldots, p)$. However, if A depends on $(\boldsymbol{X}, \boldsymbol{Z})$ (as can happen in observational studies), the expression on the right-hand side of (S.27) for the function $P_j(\langle X_j, \beta_j \rangle, A = a)$ defined in (S.29) is generally not valid. To elaborate on this, under the consistency assumption (Assumption 1), the right-hand side of (S.27) can be written as:

$$E[R_j^{(a)} | \langle X_j, \beta_j \rangle, A = a] \;-\; E[R_j | \langle X_j, \beta_j \rangle],$$

where $R_j^{(a)}$ is defined in (S.30). The no unmeasured confoundedness assumption (Assumption 2) implies that, given $(\boldsymbol{X}, \boldsymbol{Z})$, we have $R_j^{(a)} \perp A$ (since $Y^{(a)} \perp A$, given $(\boldsymbol{X}, \boldsymbol{Z})$). However, given only $(\langle X_j, \beta_j \rangle, j = 1, \ldots, p, \boldsymbol{Z})$ (as in the case of (S.27)), $R_j^{(a)}$ and A need not be independent each other. Therefore, expression (S.27) is generally not equal to the right-hand side of (S.29), i.e., $E[R_j^{(a)} | \langle X_j, \beta_j \rangle, A = a] \neq E[R_j | \langle X_j, \beta_j \rangle, A = a]$. It follows that, in observational studies, even if one could consistently estimate the right-hand side of (S.27), the estimators would not be generally consistent for the functions (S.29). Thus, the associated individualized treatment rules are potentially suboptimal in the context of observational studies.

However, if we relax the no unmeasured confoundedness assumption (i.e., Assumption 2) to: conditional independence $Y^{(a)} \perp A$ given additive measurable functions of $(\langle X_j, \beta_j \rangle, j = 1, \ldots, p, \boldsymbol{Z})$, then the proposed approach that utilizes the right-hand side of (S.27) to update each P_j (given $\langle X_j, \beta_j \rangle$ and A) can lead to an optimal treatment decision rule in the context of observational studies.

Although this relaxed no unmeasured confoundedness condition may not strictly hold in practical applications, if the distribution of A given additive measurable functions of $(\langle X_j, \beta_j \rangle, j = 1, \ldots, p, \boldsymbol{Z})$ can reasonably approximate the distribution of A given $(\boldsymbol{X}, \boldsymbol{Z})$, the proposed approach can provide a reasonable approximation to the optimal treatment decision regime, even when A depends on $(\boldsymbol{X}, \boldsymbol{Z})$.

A.17. For more general case when A depends on $(\boldsymbol{X}, \boldsymbol{Z})$ and treatment propensity information is available

Suppose that the treatment assignment propensities, $\pi_1(\boldsymbol{X}, \boldsymbol{Z}), \ldots, \pi_L(\boldsymbol{X}, \boldsymbol{Z})$, associated with the L treatment conditions, satisfying $\sum_{a=1}^{L} \pi_a(\boldsymbol{X}, \boldsymbol{Z}) = 1$ and $\pi_a(\boldsymbol{X}, \boldsymbol{Z}) > 0$, are available. For notational convenience, let us write the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect term in model (1) of the main text, as

$$Q_a(\boldsymbol{X}, \boldsymbol{Z}) := \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, A = a) + \sum_{k=1}^{q} h_k(Z_k, A = a) \quad (a = 1, \ldots, L). \tag{S.31}$$

The identifiability condition (2) for CFAM in the main text implies:

$$E[Q_A(\boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X}, \boldsymbol{Z}] = \sum_{a=1}^{L} Q_a(\boldsymbol{X}, \boldsymbol{Z}) \pi_a(\boldsymbol{X}, \boldsymbol{Z}) = 0,$$

or equivalently,

$$Q_1(\boldsymbol{X}, \boldsymbol{Z}) = -\frac{1}{\pi_1(\boldsymbol{X}, \boldsymbol{Z})} \sum_{a=2}^{L} Q_a(\boldsymbol{X}, \boldsymbol{Z}) \pi_a(\boldsymbol{X}, \boldsymbol{Z}). \tag{S.32}$$

Thus, the proposed CFAM (1), subject to (2), of the main manuscript can be reparametrized to:

$$\begin{aligned} E[Y | \boldsymbol{X}, \boldsymbol{Z}, A] &= \mu(\boldsymbol{X}, \boldsymbol{Z}) + \sum_{a=1}^{L} I(A = a) Q_a(\boldsymbol{X}, \boldsymbol{Z}) \\ &= \mu(\boldsymbol{X}, \boldsymbol{Z}) - I(A = 1) \sum_{a=2}^{L} Q_a(\boldsymbol{X}, \boldsymbol{Z}) \frac{\pi_a(\boldsymbol{X}, \boldsymbol{Z})}{\pi_1(\boldsymbol{X}, \boldsymbol{Z})} + \sum_{a=2}^{L} I(A = a) Q_a(\boldsymbol{X}, \boldsymbol{Z}) \\ &= \mu(\boldsymbol{X}, \boldsymbol{Z}) + \sum_{a=2}^{L} Q_a(\boldsymbol{X}, \boldsymbol{Z}) \left\{ I(A = a) - \frac{\pi_a(\boldsymbol{X}, \boldsymbol{Z})}{\pi_1(\boldsymbol{X}, \boldsymbol{Z})} I(A = 1) \right\}, \end{aligned} \tag{S.33}$$

which is an unconstrained formulation of CFAM without constraint (2) of the main manuscript.

In (S.33) (as in the CFAM formulation), the second term, i.e., the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect term, satisfies:

$$\begin{aligned} &E\left[\sum_{a=2}^{L} Q_a(\boldsymbol{X}, \boldsymbol{Z}) \left\{ I(A = a) - \frac{\pi_a(\boldsymbol{X}, \boldsymbol{Z})}{\pi_1(\boldsymbol{X}, \boldsymbol{Z})} I(A = 1) \right\} | \boldsymbol{X}, \boldsymbol{Z} \right] \\ =&\sum_{a=2}^{L} Q_a(\boldsymbol{X}, \boldsymbol{Z}) E\left[\left\{ I(A = a) - \frac{\pi_a(\boldsymbol{X}, \boldsymbol{Z})}{\pi_1(\boldsymbol{X}, \boldsymbol{Z})} I(A = 1) \right\} | \boldsymbol{X}, \boldsymbol{Z} \right] = 0, \end{aligned} \tag{S.34}$$

which implies the orthogonality (in L^2) between the "main" and the interaction effect components, i.e.,

$$\mu(\boldsymbol{X}, \boldsymbol{Z}) \quad \perp \quad \sum_{a=2}^{L} Q_a(\boldsymbol{X}, \boldsymbol{Z}) \left\{ I(A = a) - \frac{\pi_a(\boldsymbol{X}, \boldsymbol{Z})}{\pi_1(\boldsymbol{X}, \boldsymbol{Z})} I(A = 1) \right\}. \tag{S.35}$$

When we estimate the "signal" component (S.31) ($a = 2, \ldots, L$) (and the $a = 1$ case is determined by the condition (S.32)) of the proposed model (1), the orthogonality (S.35) makes the estimation robust to misspecification of the $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect term $\mu(\boldsymbol{X}, \boldsymbol{Z})$ and allows us to use the procedure described in Section 3.2 of the main manuscript (which is detailed in Section A.4 of this document), with some modifications which we describe next.

In the estimation, we define the design matrix $\widetilde{\boldsymbol{D}}_j$, that is used for the block-wise coordinate descent update expression in (S.16) and (S.17), to be the $n \times d_j(L-1)$ matrix $\widetilde{\boldsymbol{D}}_j = (\widetilde{\boldsymbol{D}}_{j,2}; \ldots; \widetilde{\boldsymbol{D}}_{j,L})$, where each element (the $n \times d_j$ matrix) $\widetilde{\boldsymbol{D}}_{j,a}$ $(a = 2, \ldots, L)$ specific to each treatment condition $A = a$ $(a = 2, \ldots, L)$, denotes the evaluation matrix of the basis $\boldsymbol{\Psi}_j(\cdot)$ on $\langle X_{ij}, \widehat{\beta}_j \rangle$ $(i = 1, \ldots, n)$, multiplied by subject-specific constants $c_i = I(A_i = a) - \frac{\pi_a(\boldsymbol{X}_i, \boldsymbol{Z}_i)}{\pi_1(\boldsymbol{X}_i, \boldsymbol{Z}_i)} I(A_i = 1)$ $(i = 1, \ldots, n)$, whose ith row is the $1 \times d_j$ vector $c_i \boldsymbol{\Psi}_j(\langle X_{ij}, \widehat{\beta}_j \rangle)^\top$. Upon construction of these design matrices $\widetilde{\boldsymbol{D}}_j$ $(j = 1, \ldots, p)$, we can obtain the vector $\widehat{\widetilde{\boldsymbol{\theta}}}_j = (\widehat{\boldsymbol{\theta}}_{j,2}^\top, \ldots, \widehat{\boldsymbol{\theta}}_{j,L}^\top)^\top$, corresponding to the vector in (S.18), at convergence of the coordinate descent in Step 1. From this, we can obtain:

$$\widehat{g}_j(\cdot, a) = \boldsymbol{\Psi}_j(\cdot)^\top \widehat{\boldsymbol{\theta}}_{j,a} \quad (a = 2, \ldots, L) \ (j = 1, \ldots, p).$$

Since the functions $\widehat{h}_k(\cdot, a)$ $(a = 2, \ldots, L)$ $(k = 1, \ldots, q)$ associated with the scalar covariates can be also obtained similarly, for fixed $\widehat{\beta}_j$ $(j = 1, \ldots, p)$ (available from the previous iterate for Step 2), we have all the components needed to specify the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect term in (S.34), i.e.,

$$\sum_{a=2}^{L} \left(\sum_{j=1}^{p} \widehat{g}_j(\langle X_j, \widehat{\beta}_j \rangle, A = a) + \sum_{k=1}^{q} \widehat{h}_k(Z_k, A = a) \right) \left\{ I(A = a) - \frac{\pi_a(\boldsymbol{X}, \boldsymbol{Z})}{\pi_1(\boldsymbol{X}, \boldsymbol{Z})} I(A = 1) \right\},$$

completing Step 1 of the estimation algorithm. For Step 2 of the estimation algorithm, to update β_j around the current estimate $\widehat{\beta}_j^{(c)}$, as in (16) of the main text we can utilize the partial residual $\widehat{\boldsymbol{R}}_j$ in (S.17) (available from Step 1) and the first-order Taylor series approximation around the current estimate $\widehat{\beta}_j^{(c)}$. However, in (16) of the main text, one needs to multiply the subject-specific components $\widehat{g}_j(\langle X_{ij}, \widehat{\beta}_j^{(c)} \rangle, A_i)$ and $\dot{\widehat{g}}_j(\langle X_{ij}, \widehat{\beta}_j^{(c)} \rangle, A_i)$ by $I(A_i = a) - \frac{\pi_a(\boldsymbol{X}_i, \boldsymbol{Z}_i)}{\pi_1(\boldsymbol{X}_i, \boldsymbol{Z}_i)} I(A_i = 1)$, to account for the treatment propensities in the estimation.

A.18. Discussion on possible theoretical development

Fan et al. (2015), in their functional additive regression model estimation, use the empirical process approach to the asymptotics of nonlinear least-squares estimation and obtain the stochastic bound for the regression function estimate. However, direct extension of the theoretical results of Fan et al. (2015) using their method of proof to our setting is not feasible, since their method of proof treats all the predictors (i.e., baseline covariates and treatment indicators in our context) as deterministic (which is not the case in our setting, where the treatment indicators are considered as random, under model specification (2) of the main manuscript) and their model is assumed to be correctly specified.

The main challenge to the theoretical development of the proposed estimation approach is in that the working model associated with the proposed estimation criterion is misspecified. Specifically, the working model, $Y \approx \sum_{j=1}^{p} g_j(\langle X_j, \beta_j \rangle, A) + \sum_{k=1}^{q} h_k(Z_k, A) + \epsilon$, implied by the population characterization of the model parameters in (3) of the main text (see also, Section A.1 of Supporting Information), is an approximation model, due to the omission of the unspecified term $\mu(\boldsymbol{X}, \boldsymbol{Z})$ that is present in model (1), and thus it is generally misspecified. In such a case, one would need to conduct the asymptotic analysis using the ideas from semiparametric M-estimation, following, for example, the approaches in Ichimura and Lee (2010); Wang and Yang (2009), that deal with semiparametric least squares estimation under model misspecification.

To be more specific, to establish the consistency of the estimators of the model components, one would need to establish the estimation consistency for the function (see, (6) of the main manuscript)

$$P_j(\langle X_j, \beta_j \rangle, A) \ := \ E[R_j | \langle X_j, \beta_j \rangle, A] \ - \ E[R_j | \langle X_j, \beta_j \rangle], \tag{S.36}$$

in which

$$R_j = Y - \sum_{j' \neq j} g_{j'}(\langle X_{j'}, \beta_{j'} \rangle, A) - \sum_{k=1}^{q} h_k(Z_k, A) \tag{S.37}$$

represents the jth functional covariate's partial residual, given β_j's, h_k's, and $g_{j'}$'s ($j' \neq j$). Here, the partial residual R_j, that is used to define the function P_j in (S.36), is not a function only of $(\langle X_j, \beta_j \rangle, A)$, because of the unspecified term $\mu(\boldsymbol{X}, \boldsymbol{Z})$ that is assumed to be present in the variable Y in (S.37), under the true model (1). In particular, in (S.36), the function $E[R_j | \langle X_j, \beta_j \rangle, A]$ is defined as the best L^2 *approximation* based on a measurable function of $(\langle X_j, \beta_j \rangle, A)$ to the response R_j, rather than as an exact model given $(\langle X_j, \beta_j \rangle, A)$.

Semiparametric least squares estimation under general model misspecification is a challenging theoretical problem and typically requires a two-step semiparametric M-estimation formulation (Ichimura and Lee, 2010) for theoretical development. The investigation could entail uniformly consistent (spline) estimators of the conditional expectations $E[R_j | \langle X_j, \beta_j \rangle, A]$ and $E[R_j | \langle X_j, \beta_j \rangle]$ (uniformly over $\beta_j \in \Theta$), using spline basis expansion as employed in the proposed method, based on the ideas from Wang and Yang (2009). Based on these uniformly consistent estimators of the functions in (S.36) (uniformly over $\beta_j \in \Theta$), one could establish the consistency of the estimators of the projection directions β_j, given all the other model components. Then the investigation could entail the large number of estimated components in the additive regression model for the response, for which we could use ideas from high dimensional additive model (e.g., Bühlmann and van de Geer (2011); Meier et al. (2009)). We note that the functional aspect of the data (i.e., infinite dimensional predictors X_j) and of the coefficients β_j adds further complexity to the already challenging problem of developing a semiparametric estimation theory under model misspecification. This will entail a significant amount of additional work and we leave this theoretical investigation on the estimation consistency for the model components as future work.

References

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Methods, Theory and Applications.* Springer, Heidelberg.

Chen, Y., Liu, Y., Zeng, D., and Wang, Y. (2020). DTRlerarn: Statistical learning methods for optimizing dynamic treatment regimes. *R package version 1.1* .

Di, C.-Z., Crainiceanu, C., Caffo, B., and Punjabi, N. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics* **3,** 458–488.

Fan, Y., Foutz, N., James, G., and Jank, W. (2014). Functional response additive model 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.

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

Ichimura, H. and Lee, S. (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics* **159,** 252–266.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society Series B* **64,** 411–432.

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

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37,** 3779–3821.

Park, H., Petkova, E., Tarpey, T., and Ogden, R. T. (2020a). A constrained single-index regression for estimating interactions between a treatment and covariates. *Biometrics. https://doi.org/10.1111/biom.13320* .

Park, H., Petkova, E., Tarpey, T., and Ogden, R. T. (2020b). famTEMsel: Functional additive models with treatement effect modifier selection. *R package version 0.1.0. https://github.com/syhyunpark/famTEMsel* .

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

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

Ravikumar, P., Wainwright, M., and Yu, B. (2008). Single index convex experts: Efficient estimation via adapted bregman losses. *Snowbird Workshop* .

Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). Methods for scalar-on-function regression. *International statistical review* **85,** 228–249.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*, volume 66. Cambridge University Press.

Tian, L., Alizadeh, A., Gentles, A., and Tibshrani, 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.

Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable maximation. *Journal of optimization theory and applications* **109,** 475–494.

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

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, 2nd edition edition.

Yao, F., Müller, H., and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of American Statistical Association* **100,** 577–590.