Statistical Papers
https://doi.org/10.1007/s00362-024-01546-0

REGULAR ARTICLE

A high-dimensional single-index regression for interactions between treatment and covariates

Hyung Park[1] · Thaddeus Tarpey[1] · Eva Petkova[1] · R. Todd Ogden[2]

Received: 17 July 2023 / Revised: 24 January 2024

Abstract

This paper explores a methodology for dimension reduction in regression models for a treatment outcome, specifically to capture covariates' moderating impact on the treatment-outcome association. The motivation behind this stems from the field of precision medicine, where a comprehensive understanding of the interactions between a treatment variable and pretreatment covariates is essential for developing individualized treatment regimes (ITRs). We provide a review of sufficient dimension reduction methods suitable for capturing treatment-covariate interactions and establish connections with linear model-based approaches for the proposed model. Within the framework of single-index regression models, we introduce a sparse estimation method for a dimension reduction vector to tackle the challenges posed by high-dimensional covariate data. Our methods offer insights into dimension reduction techniques specifically for interaction analysis, by providing a semiparametric framework for approximating the minimally sufficient subspace for interactions.

Keywords Precision medicine · Modified covariate method · Single-index model · Sufficient reduction · Central mean subspace

1 Introduction

Cook (2007, Sect. 8.2) considered the notion of sufficiency to the realm of regression as a dimension reduction concept (see also, (Li 1991, 1992; Cook 1994, 1996; Bura and Cook 2001; Adragni and Cook 2009; Ma and Zhu 2012, 2013) for discussions on sufficient dimension reduction). Given a set of p covariates $\boldsymbol{X} \in \mathbb{R}^p$ and an outcome variable $Y \in \mathbb{R}$, Cook's notion of a sufficient subspace in regression can be summarized

✉ Hyung Park
parkh15@nyu.edu

[1] Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY 10016, USA

[2] Department of Biostatistics, Columbia University, New York, NY 10032, USA

Published online: 13 April 2024

as $Y|\boldsymbol{X} \overset{d}{=} Y|R(\boldsymbol{X})$ where $R : \mathbb{R}^p \mapsto \mathbb{R}^q$, $q < p$, i.e., the covariate-outcome association is captured by a lower-dimensional association based on $R(\boldsymbol{X}) \in \mathbb{R}^q$.

The central subspace, which we denote by $S_{Y|\boldsymbol{X}}$, is the subspace with the smallest possible dimension q in $\mathbb{R}^p$, such that Y is independent of $\boldsymbol{X}$ given $R(\boldsymbol{X}) = \boldsymbol{B}'\boldsymbol{X}$ for some $p \times q$ matrix $\boldsymbol{B}$, $q < p$, in which the columns of $\boldsymbol{B}$ form a basis for the subspace (Cook and Li 2002). For comprehensive discussion on the central subspace, see Cook (1998). Dimension reduction is often aimed at reducing dimensionality for modeling the conditional mean function $\mathbb{E}[Y|\boldsymbol{X}]$ alone, while leaving the rest of the distribution $Y|\boldsymbol{X}$ as the "nuisance parameter." For this case, Cook and Li (2002) introduced the central mean subspace, denoted as $S_{\mathbb{E}[Y|\boldsymbol{X}]}$, defined to be the smallest subspace, span$(\boldsymbol{B})$ for some basis matrix $\boldsymbol{B}$, sufficient to model the conditional mean $\mathbb{E}[Y|\boldsymbol{X}]$.

In this paper, we focus on the notion of a sufficient subspace in regression with an outcome variable Y when our interest is in the interaction effect between the vector of covariates $\boldsymbol{X} \in \mathbb{R}^p$ and another variable $T \in \mathcal{T}$. This paper considers the case when T is a discrete random variable on a space $\mathcal{T} = \{1, \ldots, K\}$, i.e., there are K possible levels for the random variable T, and the dimensionality of $\boldsymbol{X}$ is potentially large. The primary focus is on reducing the dimension of $\boldsymbol{X}$, using a (regularized) single index regression (Stoker 1986; Carroll et al. 1997; Xia et al. 1999; Wang and Yang 2009; Poon and Wang 2013; Radchenko 2015; Liu et l. 2023) to parsimoniously model the effects of interactions between $\boldsymbol{X}$ and T on Y. The motivation for this work is in the context of precision medicine, where we seek to optimize an individualized treatment rule (ITR) that assigns a treatment to each patient according to the patient's specific characteristics. Typically, individual-specific clinical characteristics are represented by a vector of covariates $\boldsymbol{X}$ measured before treatment assignment, and treatment option can be represented by the variable $T \in \mathcal{T}$. An optimal ITR relies on the $\boldsymbol{X}$-by-T interaction effects on Y (see, e.g., Qian and Murphy 2011), rather than the main effects of $\boldsymbol{X}$ that are unrelated with treatment T. Therefore, a sufficient reduction subspace for $\boldsymbol{X}$ in this setting will typically be defined in terms of a subspace sufficient to model the $\boldsymbol{X}$-by-T interaction effect, whereas the pure main effect of $\boldsymbol{X}$ on Y can be viewed as a "nuisance" effect.

In this paper, we define a sufficient dimension reduction subspace for $\boldsymbol{X}$ in terms of a parsimonious characterization of the $\boldsymbol{X}$-by-T interaction effect in the conditional mean $\mathbb{E}[Y|\boldsymbol{X}, T]$. Specifically, we introduce a semiparametric framework to produce a leading dimension reduction vector within this sufficient subspace, extending the approach developed in Park et al. (2021) to a higher dimensional covariate setting. The proposed framework to estimate the $\boldsymbol{X}$-by-T interactions takes the linear model based approaches as its special cases (e.g., Lu et al. 2011; Tian et al. 2014a; Shi et al. 2016, 2018; Jeng et al. 2018). Luo et al. (2018) considered sufficient dimension reduction to estimate a lower dimensional linear combination of $\boldsymbol{X}$ that is sufficient to model the regression causal effect, defined as the mean difference in the potential outcomes (Rubin 1974) conditional on $\boldsymbol{X}$ (see also, Luo et al. 2017), when the treatment variable T is binary-valued. Our framework, instead, focuses on the interaction between covariates and treatment that allows general K treatment levels, with emphasis on establishing connections with existing linear model-based methods.

The paper is organized as follows. We begin by providing an overview of sufficient dimension reduction methods for capturing interaction effects between treatment and covariates in Sect. 2. We then establish connections between these methods and existing linear model-based approaches in Sect. 3, followed by development of a semiparametric modeling framework that focuses on achieving a single-dimensional reduction, with an L^1 regularization to address the challenges posed by high-dimensional data in Sect. 4. We present simulation studies to assess the performance of the proposed method in Sect. 5, and an application to a depression randomized clinical trial (RCT) data in Sect. 6. The paper concludes with discussion in Sect. 7.

2 Sufficient reduction for interactions

2.1 Preliminaries

We express the conditional mean $\mathbb{E}[Y|\boldsymbol{X}, T]$ in terms of a main effect of $\boldsymbol{X}$ (i.e., the term $\mu(\boldsymbol{X})$) and a $\boldsymbol{X}$-by-T interaction effect (i.e., the term $g(\boldsymbol{X}, T)$),

$$\mathbb{E}[Y|\boldsymbol{X}, T] \;=\; \mu(\boldsymbol{X}) + g(\boldsymbol{X}, T), \tag{2.1}$$

where we impose a constraint on $g(\boldsymbol{X}, T)$,

$$\mathbb{E}[g(\boldsymbol{X}, T)|\boldsymbol{X}] = 0, \tag{2.2}$$

for the identifiability of the decomposition in Eq. (2.1). Of note, the main effect of T is included in the term $g(\boldsymbol{X}, T)$ in Eq. (2.1). The constraint Eq. (2.2) ensures that the first component $\mu(\boldsymbol{X})$ in Eq. (2.1) does not involve T. Specifically, in representation Eq. (2.1), the term $\mu(\boldsymbol{X})$ captures the $\boldsymbol{X}$-related effect that is consistent across levels of T. On the other hand, the term $g(\boldsymbol{X}, T)$ captures the $\boldsymbol{X}$-related effect that *interacts* with the specific value of $T = t$ ($t \in \mathcal{T}$), hence it is called the $\boldsymbol{X}$-by-T interaction term. Throughout the paper, we write $\boldsymbol{\Sigma}_X = \text{var}(\boldsymbol{X})$, and assume an additive mean zero noise with finite outcome variance.

2.2 Central mean subspace

For a discrete treatment space $\mathcal{T} = \{1, \ldots, K\}$ with K available treatments, an ITR, $\mathcal{D}(\boldsymbol{X}) : \mathbb{R}^p \mapsto \mathcal{T}$, is defined to map each individual's pretreatment covariates $\boldsymbol{X} \in \mathbb{R}^p$ to one of the K available treatment options (Murphy 2003; Robins 2004; Cai et al. 2011; Qian and Murphy 2011; Zhang et al. 2012). The average outcome when all individuals are treated according to such an ITR is referred to as the "value" (V) of the ITR (Qian and Murphy 2011); we can express the "value" of an ITR $\mathcal{D}$ as: $V(\mathcal{D}) = \mathbb{E}[\mathbb{E}[Y|\boldsymbol{X}, T = \mathcal{D}(\boldsymbol{X})]]$. Without loss of generality, let us assume that a larger value of Y is desirable, so that we wish to maximize the value $V(\mathcal{D})$. Then it is straightforward to verify that the optimal ITR, denoted as $\mathcal{D}_{\text{opt}}$, which results in the largest value $V(\mathcal{D}_{\text{opt}})$, is of the form:

$$\mathcal{D}_{\text{opt}}(\boldsymbol{X}) \;=\; \arg\max_{t\in\mathcal{T}}\; \mathbb{E}\big[Y|\boldsymbol{X}, T=t\big], \tag{2.3}$$

that is, the optimal ITR $\mathcal{D}_{\text{opt}}(\boldsymbol{X})$ assigns a treatment $t \in \mathcal{T}$ to an individual with pretreatment characteristics $\boldsymbol{X}$ that yields the highest expected quality treatment given $\boldsymbol{X}$.

We will cast the notion of sufficient reduction specifically for the $\boldsymbol{X}$-by-T interaction effects, $g(\boldsymbol{X}, T)$ in Eq. (2.1). We define a *contrast vector* $\boldsymbol{c} = (c_1, \ldots, c_K)' \in \mathbb{R}^K$ as a vector such that $\sum_{t=1}^{K} c_t = 0$ (zero-sum constraint) with the elements c_t ($t = 1, \ldots, K$), where c_t's are not all zeros to avoid the trivial case.

Definition 1 For an arbitrary contrast vector $\boldsymbol{c} = (c_1, \ldots, c_K)'$, we define the mean outcome contrast function given $\boldsymbol{X}$, as the following linear contrast:

$$\mathcal{C}(\boldsymbol{X}|\boldsymbol{c}) = \sum_{t=1}^{K} c_t \mathbb{E}\left[Y|\boldsymbol{X}, T=t\right]. \tag{2.4}$$

For example, if $K = 2$ with $c_1 = 1$ and $c_2 = -1$, the contrast function Eq. (2.4), i.e., the function $\mathcal{C}(\boldsymbol{X}|\boldsymbol{c}) = \mathbb{E}\left[Y|\boldsymbol{X}, T=1\right] - \mathbb{E}\left[Y|\boldsymbol{X}, T=2\right]$, is reduced to the case studied by Luo et al. (2018), where its sign would determine the optimal ITR defined in Eq. (2.3).

In this paper, we consider a lower dimensional linear transformation of $\boldsymbol{X}$ that is sufficient, for any contrast vector $\boldsymbol{c} \in \mathbb{R}^K$, to recover the mean contrast function $\mathcal{C}(\boldsymbol{X}|\boldsymbol{c})$ in Eq. (2.4).

Definition 2 Let $\boldsymbol{B}$ denote a $p \times q$ matrix with full column rank. The transformation $R(\boldsymbol{X}) = \boldsymbol{B}'\boldsymbol{X}$ is said to be a sufficient dimension reduction for $\boldsymbol{X}$-by-T interactions, if

$$\mathcal{C}(\boldsymbol{X}|\boldsymbol{c}) = \mathcal{C}(\boldsymbol{B}'\boldsymbol{X}|\boldsymbol{c}) = \sum_{t=1}^{K} c_t g_t(\boldsymbol{B}'\boldsymbol{X}) \tag{2.5}$$

for any contrast vector $\boldsymbol{c}$, where the treatment t-specific functions $\{g_t(\boldsymbol{B}'\boldsymbol{X})\}_{t\in\mathcal{T}}$ are unspecified functions associated with each level of $T \in \mathcal{T}$, defined over $\boldsymbol{B}'\boldsymbol{X} \in \mathbb{R}^q$. Correspondingly, the column space of $\boldsymbol{B}$ will be called a sufficient reduction subspace for $\boldsymbol{X}$-by-T interactions.

For any $p \times q$ matrix $\boldsymbol{B} = [\boldsymbol{\beta}; \ldots; \boldsymbol{\beta}_q]$ satisfying Eq. (2.5) and any $p \times p$ nonsingular matrix $\boldsymbol{\eta}$, the matrix $\boldsymbol{\eta}\boldsymbol{B}$ still satisfies Eq. (2.5) if the $\{g_t\}_{t\in\mathcal{T}}$ are adjusted accordingly, and a further constraint on $\boldsymbol{B}$ is needed for an identifiable parametrization. To remove trivial ambiguity, let us define the space of $p \times q$ matrices, denoted as Θ_q, that have a positive first nonzero entry and consists of q distinct orthonormal vectors. In Eq. (2.5), without loss of generality, we assume $\boldsymbol{B} \in \Theta_q$.

As the notion of sufficiency Eq. (2.5) relies on the outcome contrast function $\mathcal{C}(\boldsymbol{X}|\boldsymbol{c})$ not involving the term $\mu(\boldsymbol{X})$ in Eq. (2.1), we can formalize a minimally sufficient dimension reduction in $\boldsymbol{X}$ by solely focusing on the term $g(\boldsymbol{X}, T)$ in Eq. (2.1).

Definition 3 A sufficient reduction subspace for $\boldsymbol{X}$ for $\boldsymbol{X}$-by-T interactions is said to be minimal, if the dimension of its span is less than or equal to that of any other sufficient reduction subspace for such interactions. We denote the minimally sufficient reduction subspace (also called the central mean subspace) for $\boldsymbol{X}$-by-T interactions as $S_{\mathcal{C}|\boldsymbol{X}}$, and $\dim(S_{\mathcal{C}|\boldsymbol{X}})$ will denote its dimension.

The central mean subspace of Cook and Li (2002) refers to the minimally sufficient subspace in $\mathbb{R}^p$ associated with the mean response function. The subspace $S_{\mathcal{C}|\boldsymbol{X}}$ defined above is a special case of the central mean subspace for the mean function Eq. (2.1) in which only the interaction term $g(\boldsymbol{X}, T)$ is considered for dimension reduction. We assume that the central mean subspace for interactions, $S_{\mathcal{C}|\boldsymbol{X}}$, uniquely exists throughout this article. The uniqueness of the central mean subspace is guaranteed under fairly general conditions (Cook and Li 2002; Luo et al. 2018; Yin et al. 2008); for example, it is guaranteed when the domain of $\boldsymbol{X}$ is open and convex.

Dimension reduction using a minimal number of directions, $\dim(S_{\mathcal{C}|\boldsymbol{X}})$, is important for interpretability and parsimonious parametrization. By solely concentrating on $g(\boldsymbol{X}, T)$, the dimension reduction process can be streamlined to capture the essential aspects of $\boldsymbol{X}$ that are crucial for estimating treatment effects or studying interactions, allowing for a more concise and targeted representation of the data while maintaining the necessary information for decision-making. In practice, 1-dimensional reductions often provide reasonable approximations to capture pertinent interaction effects. Examples of 1-dimensional reductions for $\boldsymbol{X}$-by-T interactions include performing a regression with a linear model that focuses on a single vector of coefficients (e.g., Tian et al. 2014a; Petkova et al. 2016) and its semiparametric generalization with a set of flexible link functions, a single-index model with treatment level-specific link functions (Park et al. 2021). In the remainder of this section, we introduce a semiparametric regression framework for approximating the minimally sufficient subspace $S_{\mathcal{C}|\boldsymbol{X}}$ for $\boldsymbol{X}$-by-T interactions, and in Sect. 3, we build connections to other linear model-based approaches as its special cases.

2.3 Models

Given the notion of sufficiency Eq. (2.5), we posit that the $\boldsymbol{X}$-by-T interaction effect within $\mathbb{E}[Y|\boldsymbol{X}, T]$ in Eq. (2.1) has an intrinsic q-dimensional structure with some dimension reduction matrix $\boldsymbol{B}^* \in \Theta_q$ of rank-q:

$$\mathbb{E}[Y \mid \boldsymbol{X}, T = t] = \mu^*(\boldsymbol{X}) + g_t^*\left(\boldsymbol{B}^{*\prime}\boldsymbol{X}\right) \quad (t \in \mathcal{T}), \tag{2.6}$$

where the term $\mu^*(\boldsymbol{X})$ is a square integrable unspecified function of $\boldsymbol{X}$ only, and the expected value of the second term $g_T^*\left(\boldsymbol{B}^{*\prime}\boldsymbol{X}\right)$ given $\boldsymbol{X}$ is zero, i.e.,

$$\mathbb{E}[g_T^*\left(\boldsymbol{B}^{*\prime}\boldsymbol{X}\right)|\boldsymbol{X}] = 0, \tag{2.7}$$

for model identifiability, as in the general model Eq. (2.2). Let us use the notation $\mathcal{H}^{(\boldsymbol{B})}$, for each fixed $\boldsymbol{B} \in \Theta_q$, to denote the space of square-integrable functions over $\boldsymbol{B}'\boldsymbol{X} \in \mathbb{R}^q$, and assume $g_t^*\left(\boldsymbol{B}^{*\prime}\boldsymbol{X}\right) \in \mathcal{H}^{(\boldsymbol{B}^*)}$ in Eq. (2.6), for each $t \in \mathcal{T}$. To suppress the treatment level T-specific intercepts (only to simplify the illustration), we assume, without loss of generality, $\mathbb{E}[Y|T = t] = 0$, i.e., the outcome Y is centered within each treatment level t ($t \in \mathcal{T}$), which can be satisfied by removing the treatment level T-specific means from Y.

The following Theorem 2.1 and Corollary 2.1 indicate that if our interest is in the estimation of $S_{C|\boldsymbol{X}}$, we can focus on the estimation of $\boldsymbol{B}^*$ of the dimension reduction model Eq. (2.6).

Theorem 2.1 *For the mean model of form Eq. (2.1), the low-rank representation of $g(\boldsymbol{X}, T)$ with $g_t^*\big(\boldsymbol{B}^{*\prime}\boldsymbol{X}\big)$ ($t \in \mathcal{T}$) in Eq. (2.6) implies the sufficiency of the predictor reduction $R(\boldsymbol{X}) = \boldsymbol{B}^{*\prime}\boldsymbol{X}$ for the central mean subspace $\mathcal{S}_{C|\boldsymbol{X}}$.*

Corollary 2.1 *The set of columns of $\boldsymbol{B}^*$ in model Eq. (2.6) is a basis of the central mean subspace $S_{C|\boldsymbol{X}}$.*

The proofs of Theorem 2.1 and Corollary 2.1 are in Appendix A1 and A2.

2.4 Criterion

Under model Eq. (2.6), the treatment t-specific functions $\{g_t^*\}_{t\in\mathcal{T}}$ and the dimension reduction matrix $\boldsymbol{B}^*$ can be viewed as the solution to the following optimization:

$$\begin{aligned}(g_1^*, \ldots, g_K^*, \boldsymbol{B}^*) \quad &= \quad \underset{g_t \in \mathcal{H}^{(\boldsymbol{B})}, \boldsymbol{B} \in \Theta_q}{\operatorname{argmin}} \quad \mathbb{E}\big[\big(Y - g_T(\boldsymbol{B}'\boldsymbol{X})\big)^2\big] \\ &\qquad\quad \text{subject to} \qquad \mathbb{E}\left[g_T(\boldsymbol{B}'\boldsymbol{X})|\boldsymbol{X}\right] = 0.\end{aligned} \tag{2.8}$$

where we disregard the unspecified term $\mu^*(\boldsymbol{X})$ in specifying the components $(g_1^*, \ldots, g_K^*, \boldsymbol{B}^*)$ (see Appendix A3 for the justification behind this omission of $\mu(\boldsymbol{X})$). The constrained least squares framework Eq. (2.8) provides a class of regression approaches to estimating the subspace $S_{C|X} = \text{span}(\boldsymbol{B}^*)$ without involving the nuisance component $\mu^*(\boldsymbol{X})$. Specifically, the objective function on the right-hand side of Eq. (2.8) can be empirically approximated based on samples $(y_i, t_i, \boldsymbol{x}_i)$ ($i = 1, \ldots, n$), if the T-specific unknown functions $\{g_t\}_{t\in\mathcal{T}}$ are appropriately represented subject to the constraint in Eq. (2.8). We note that representation Eq. (2.8) extends the existing linear approaches to estimating interactions into a semiparametric framework equipped with unknown flexible functions $\{g_t\}_{t\in\mathcal{T}}$. In Sects. 3 and 4, we shall focus on the context of a randomized experiment, where the treatment $t \in \{1, \ldots, K\}$ is assigned independently of $\boldsymbol{X}$ with some randomization probabilities $(\pi_1, \ldots, \pi_K)$, $\sum_{t=1}^K \pi_t = 1$ and $\pi_t > 0$.

Remark 2.1 The constraint $\mathbb{E}\left[g_T(\boldsymbol{B}'\boldsymbol{X})|\boldsymbol{X}\right] = 0$ in Eq. (2.8) imposed on the unknown functions $\{g_t\}_{t\in\mathcal{T}}$ parallels the constraint $\sum_{t=1}^K c_t = 0$ imposed on the functions $\{c_t g_t\}_{t\in\mathcal{T}}$ of Eq. (2.5).

3 Linear models

Let us first consider a classical linear model for the $\boldsymbol{X}$-by-T interaction effect defined based on a set of the treatment t-specific (length-p) coefficient vectors $\boldsymbol{\eta}_t := \boldsymbol{\Sigma}_X^{-1}\text{cov}[Y, \boldsymbol{X}|T = t]$ $(t \in \mathcal{T})$. The model is written as:

$$\mathbb{E}[Y|\boldsymbol{X}, T = t] = \mu_0(\boldsymbol{X}) + \boldsymbol{\eta}_t'\boldsymbol{X} \quad (t \in \mathcal{T}), \tag{3.1}$$

where the first term $\mu_0(\boldsymbol{X})$ represents an unspecified main effect of $\boldsymbol{X}$ that does not depend on T. Let us first introduce the $p \times p$ "dispersion" matrix of the treatment t-specific slope coefficients $\{\boldsymbol{\eta}_t \in \mathbb{R}^p\}_{t\in\mathcal{T}}$ for the linear $\boldsymbol{X}$-by-T interaction model Eq. (3.1),

$$\boldsymbol{H} = \sum_{t=1}^{K} \pi_t(\boldsymbol{\eta}_t - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_t - \bar{\boldsymbol{\eta}})', \tag{3.2}$$

where $\bar{\boldsymbol{\eta}} := \mathbb{E}[\boldsymbol{\eta}_t] = \sum_{t=1}^{K} \pi_t\boldsymbol{\eta}_t \in \mathbb{R}^p$. Let us define $\boldsymbol{\Xi} := \left[\boldsymbol{\xi}_1; \ldots; \boldsymbol{\xi}_{K-1}\right] \in \mathbb{R}^{p\times(K-1)}$, as the matrix consisting of the eigenvectors $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{K-1})$ of the matrix $\boldsymbol{H}$ Eq. (3.2) associated with the $K - 1$ leading eigenvalues (there are only $K - 1$ nonzero eigenvalues; we assume $p > K - 1$). Then the following proposition states that when $\boldsymbol{\eta}_t$ $(t = 1, \ldots, K)$ are distinct, span$(\boldsymbol{\Xi})$ corresponds to the central mean subspace $S_{\mathcal{C}|\boldsymbol{X}}$.

Proposition 3.1 *Under the linear interaction model Eq. (3.1), we have* $\mathcal{C}(\boldsymbol{X}|\boldsymbol{c}) = \mathcal{C}\left(\boldsymbol{\Xi}'\boldsymbol{X}|\boldsymbol{c}\right)$ *for any arbitrary contrast* $\boldsymbol{c}$*, and thus span*$(\boldsymbol{\Xi})$ *provides a sufficient reduction for Eq. (2.4). Furthermore, if* $\boldsymbol{\eta}_t$ $(t = 1, \ldots, K)$ *are distinct,* $S_{\mathcal{C}|\boldsymbol{X}} =$ *span*$(\boldsymbol{\Xi})$.

The proof of Proposition 3.1 is in Appendix A4 If we consider model Eq. (3.1) within the dimension reduction model Eq. (2.6), then Proposition 3.1 implies that span$(\boldsymbol{\Xi}) =$ span$(\boldsymbol{B}^*)$ and its dimension, dim$(S_{\mathcal{C}|\boldsymbol{X}}) = K - 1$. Proposition 3.1 indicates that, to estimate vectors in $S_{\mathcal{C}|\boldsymbol{X}}$, one can soley focus on estimating the eigenvectors $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{K-1})$ of $\boldsymbol{H}$, if the $\boldsymbol{X}$-by-T interaction effects are linear Eq. (3.1). Next, we will describe how to estimate the leading eigenvector $\boldsymbol{\xi}_1$ of $\boldsymbol{H}$, utilizing the optimization framework of Eq. (2.8).

3.1 A linear dimension reduction model

A useful 1-dimensional approximation to the linear $\boldsymbol{X}$-by-T interaction model Eq. (3.1) is:

$$\mathbb{E}[Y \mid \boldsymbol{X}, T = t] \approx \mu_0(\boldsymbol{X}) + \tilde{\gamma}_t\boldsymbol{\beta}'\boldsymbol{X} \quad (t \in \mathcal{T}), \tag{3.3}$$

for a 1-dimensional (1-D) projection vector $\boldsymbol{\beta} \in \Theta_1$ (for the model identifiability). Model Eq. (3.3) can be used to approximate the basis of the subspace $S_{\mathcal{C}|\boldsymbol{X}}$, i.e., $\boldsymbol{\Xi}$, based on a rank-1 projection determined by $\boldsymbol{\beta} \in \mathbb{R}^p$. In Eq. (3.3), the $\boldsymbol{X}$-by-T

interaction effect term $\tilde{\gamma}_t \boldsymbol{\beta}'\boldsymbol{X}$ $(t \in \mathcal{T})$ captures the variability in $\boldsymbol{X}$ related to T via a 1-dimensional projection $\boldsymbol{\beta}'\boldsymbol{X}$, and its interaction with T via the t-specific slopes $\tilde{\gamma}_t \in \mathbb{R}$ $(t \in \mathcal{T})$. Petkova et al. (2016) called this projection $\boldsymbol{\beta}'\boldsymbol{X} \in \mathbb{R}$ a *generated effect-modifier*, that combines p pretreatment covariates $\boldsymbol{X}$ into a single treatment effect-modifier. As in Eq. (3.1), the term $\mu_0(\boldsymbol{X})$ in Eq. (3.3) represents an unspecified main effect of $\boldsymbol{X}$.

Let us consider the approximation model Eq. (3.3) within the framework Eq. (2.6) by centering the t-specific slopes $\tilde{\gamma}_t$ $(t \in \mathcal{T})$. Specifically, let us reparametrize $\gamma_t := \tilde{\gamma}_t - \bar{\gamma}$ $(t \in \mathcal{T})$, where $\bar{\gamma} := \sum_{t=1}^{K} \pi_t \tilde{\gamma}_t$. Then the resulting reparametrized model of Eq. (3.3) is

$$\mathbb{E}[Y \mid \boldsymbol{X}, T = t] \;\approx\; \mu^*(\boldsymbol{X}) + \gamma_t \boldsymbol{\beta}'\boldsymbol{X} \quad (t \in \mathcal{T}), \tag{3.4}$$

where the first term $\mu^*(\boldsymbol{X}) := \mu_0(\boldsymbol{X}) + \bar{\gamma}\boldsymbol{\beta}'\boldsymbol{X}$ is the reparametrized version of the $\boldsymbol{X}$ main effect $\mu_0(\boldsymbol{X})$, and the slope coefficients $\gamma_t \in \mathbb{R}$ in the second term is subject to the identifiability condition

$$\sum_{t=1}^{K} \pi_t \gamma_t = 0, \tag{3.5}$$

that characterizes this particular reparametrization Eq. (3.4) of the approximation model Eq. (3.3). Note, the constraint Eq. (3.5) implies $\mathbb{E}[\gamma_T \boldsymbol{\beta}'\boldsymbol{X} | \boldsymbol{X}] = 0$ for any arbitrary $\boldsymbol{\beta} \in \Theta_1$, which is a special case of the constraint in Eq. (2.8), where the unknown functions $\{g_t\}_{t \in \mathcal{T}}$ are replaced with the unknown slopes $\{\gamma_t\}_{t \in \mathcal{T}}$ and the dimension reduction matrix $\boldsymbol{B}$ is replaced with the vector $\boldsymbol{\beta}$.

To optimize the parameters $\{\gamma_t\}_{t \in \mathcal{T}}$ and $\boldsymbol{\beta}$ associated with the $\boldsymbol{X}$-by-T interactions wihtin the rank-1 approximation model Eq. (3.4), we employ the constrained criterion Eq. (2.8), which corresponds to solving:

$$\underset{\gamma_t \in \mathbb{R},\ \boldsymbol{\beta} \in \Theta_1}{\operatorname{argmin}} \; \mathbb{E}\big[\big(Y - \gamma_T \boldsymbol{\beta}'\boldsymbol{X}\big)^2\big], \tag{3.6}$$

subject to the constraint Eq. (3.5), where the minimization is over both the slopes $\{\gamma_t\}_{t \in \mathcal{T}}$ and the vector $\boldsymbol{\beta}$. The following proposition provides an explicit expression for the profile minimizer $\{\gamma_t\}_{t \in \mathcal{T}}$ within the constrained criterion Eq. (3.6), for each fixed $\boldsymbol{\beta} \in \Theta_1$.

Proposition 3.2 *For the linear $\boldsymbol{X}$-by-T interaction model Eq. (3.1), the minimizer $\{\gamma_t\}_{t \in \mathcal{T}}$ of the criterion Eq. (3.6) for a fixed vector $\boldsymbol{\beta}$ is given by $\gamma_t = (\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\boldsymbol{\Sigma}_X(\boldsymbol{\eta}_t - \bar{\boldsymbol{\eta}})$ $(t \in \mathcal{T})$, where $\bar{\boldsymbol{\eta}} = \sum_{t=1}^{K} \pi_t \boldsymbol{\eta}_t$.*

The proof of Proposition 3.2 is in Appendix A5 The results in Proposition 3.2 allow us to explicitly write the variance, $\mathrm{var}(\gamma_T \boldsymbol{\beta}'\boldsymbol{X})$, of the $\boldsymbol{X}$-by-T interaction effect component $\gamma_T \boldsymbol{\beta}'\boldsymbol{X}$,

$$\begin{aligned}\operatorname{var}(\gamma_T\boldsymbol{\beta}'\boldsymbol{X}) &= \sum_{t=1}^{K}\pi_t\frac{(\boldsymbol{\beta}'\boldsymbol{\Sigma}_X(\boldsymbol{\eta}_t-\bar{\boldsymbol{\eta}}))^2}{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}}\\ &= \frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\left[\sum_{t=1}^{K}\pi_t(\boldsymbol{\eta}_t-\bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_t-\bar{\boldsymbol{\eta}})'\right]\boldsymbol{\Sigma}_X\boldsymbol{\beta}}{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}}\\ &= \frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{H}\boldsymbol{\Sigma}_X\boldsymbol{\beta}}{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}} = \frac{\tilde{\boldsymbol{\beta}}'\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_X^{1/2}\tilde{\boldsymbol{\beta}}}{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}},\end{aligned} \tag{3.7}$$

where the p-by-p matrix $\boldsymbol{H}$ is defined in Eq. (3.2), and $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\beta}$, in which $\boldsymbol{\Sigma}_X^{1/2}$ is the symmetric "square root" of $\boldsymbol{\Sigma}_X$. We note that, minimizing criterion Eq. (3.6) over $\boldsymbol{\beta} \in \Theta_1$ is equivalent to maximizing the variance, $\operatorname{var}(\gamma_T\boldsymbol{\beta}'\boldsymbol{X})$, in Eq. (3.7) over $\boldsymbol{\beta} \in \Theta_1$. From expression Eq. (3.7), it is clear that the $\boldsymbol{X}$-by-T interaction variance Eq. (3.7) is maximized if $\tilde{\boldsymbol{\beta}}$ is the leading eigenvector of $\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_X^{1/2} = \boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\Xi}\boldsymbol{\Lambda}\boldsymbol{\Xi}'\boldsymbol{\Sigma}_X^{1/2}$, in which $\boldsymbol{\Lambda}$ is the diagonal matrix consist of the leading eigenvalues of $\boldsymbol{H}$. Thus, the maximizer $\tilde{\boldsymbol{\beta}}$ of the quantity Eq. (3.7) is the leading column vector of $\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\Xi}$. Since $\boldsymbol{\beta} = \boldsymbol{\Sigma}_X^{-1/2}\tilde{\boldsymbol{\beta}}$, it follows that the maximizer $\boldsymbol{\beta}$ of the variance Eq. (3.7) is the leading column vector of $\boldsymbol{\Xi}$, which is $\boldsymbol{\xi}_1$. Based on Proposition 3.2, we can now establish the following proposition that offers a closed form solution for the constrained criterion Eq. (3.6) within the context of the 1-D approximation model Eq. (3.4).

Proposition 3.3 *When the true model is Eq. (3.1), the minimizer $\boldsymbol{\beta} \in \Theta_1$ of the constrained criterion Eq. (3.6) for the 1-D approximation model Eq. (3.4) corresponds to $\boldsymbol{\beta} = \boldsymbol{\xi}_1$, which is the leading eigenvector associated with $\boldsymbol{H}$. Furthermore, the corresponding treatment t-specific slope is $\gamma_t = (\boldsymbol{\xi}_1{}'\boldsymbol{\Sigma}_X\boldsymbol{\xi}_1)^{-1}\boldsymbol{\xi}_1{}'\boldsymbol{\Sigma}_X(\boldsymbol{\eta}_t - \bar{\boldsymbol{\eta}})$ $(t \in \mathcal{T})$.*

Proposition 3.3 indicates that optimizing the constrained criterion Eq. (3.6) yields a vector ($\boldsymbol{\xi}_1$) that belongs to the central mean subspace $S_{C|\boldsymbol{X}}$ for the $\boldsymbol{X}$-by-T interactions.

Remark 3.1 The task of maximizing the ratio Eq. (3.7) over the vector $\boldsymbol{\beta}$ can be framed within the framework of generalized eigen-decomposition (GED) (Dahne et al. 2014; de Cheveigne and Parra 2014; Cohen 2022), providing insights on the optimization process. In this GED framework, the covariance of the $\boldsymbol{X} \times T$ interaction, i.e., $\boldsymbol{\Sigma}_X\boldsymbol{H}\boldsymbol{\Sigma}_X \in \mathbb{R}^{p\times p}$ (numerator), is a feature to enhance via optimization of $\boldsymbol{\beta}$, whereas the covariance of $\boldsymbol{X}$, i.e., $\boldsymbol{\Sigma}_X \in \mathbb{R}^{p\times p}$ (denominator), is a feature that acts as reference in the optimization process.

3.2 Equivalence to the modified covariate model

In the special case of $K = 2$ levels (i.e., when T is binary-valued), the "modified covariate" (MC) (Tian et al. 2014a) method of modeling the $\boldsymbol{X}$-by-T interaction effect posits the model (see also, Lu et al. (2011); Shi et al. (2016, 2018); Jeng et al. (2018), for similar linear model-based approaches to modeling the $\boldsymbol{X}$-by-T interactions):

$$\mathbb{E}[Y \mid \boldsymbol{X}, T = t] = \mu^*(\boldsymbol{X}) + \boldsymbol{\beta}^{*\prime}\boldsymbol{X}(t + \pi_1 - 2) \qquad (t = 1, 2), \tag{3.8}$$

for some coefficient vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$. The first term $\mu^*(\boldsymbol{X})$ in Eq. (3.8) represents an unspecified $\boldsymbol{X}$ main effect (as in Eq. (2.6)), and $\pi_1 = \Pr(T = 1)$.

In the dimension reduction model Eq. (2.6), if we specify the dimension reduction matrix $\boldsymbol{B}^*$ as the vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and impose the unspecified functions $\{g_t^*\}_{t \in \mathcal{T}}$ to have a pre-specified linear form:

$$g_t^*(u) = (t + \pi_1 - 2)u \quad (t = 1, 2) \tag{3.9}$$

(where $u = \boldsymbol{\beta}^{*\prime}\boldsymbol{X}$), then we obtain the MC model (3.8). Specifically, the set of t-specific functions $\{g_1^*, g_2^*\}$ in Eq. (3.9) satisfies the identifiability condition Eq. (2.7) of model Eq. (2.6), i.e., $\mathbb{E}[\boldsymbol{\beta}^{*\prime}\boldsymbol{X}(T + \pi_1 - 2)|\boldsymbol{X}] = \boldsymbol{\beta}^{*\prime}\boldsymbol{X}\mathbb{E}[T + \pi_1 - 2] = 0$. Since model model Eq. (3.8) fits within the framework Eq. (2.6), we can represent the coefficient $\boldsymbol{\beta}^*$ of Eq. (3.8) using the optimization framework Eq. (2.8):

$$\boldsymbol{\beta}^* \;=\; \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}}\, \mathbb{E}\big[\big(Y - \boldsymbol{\beta}'\boldsymbol{X}(T + \pi_1 - 2)\big)^2\big], \tag{3.10}$$

without including the term $\mu^*(\boldsymbol{X})$ in model Eq. (3.8). By solving an empirical version of Eq. (3.10) based on a sample $(y_i, t_i, \boldsymbol{x}_i)$ $(i = 1, \ldots, n)$, we obtain a consistent estimator of $\boldsymbol{\beta}^*$, where $\mu(\boldsymbol{X})$ in model Eq. (3.8) remains unspecified.

When $K = 2$ and assuming linear $\boldsymbol{X}$-by-T interactions Eq. (3.1), there exists an equivalence between the optimization Eq. (3.6) and the right-hand side of Eq. (3.10) in terms of generating vectors in the subspace $S_{\mathcal{C}|\boldsymbol{X}}$. In the case of $K = 2$, the subspace $S_{\mathcal{C}|\boldsymbol{X}}$ given from Proposition 3.1 is of rank-1, and is spanned by the eigenvector $\boldsymbol{\xi}_1$ of $\boldsymbol{H}$ in Eq. (3.2) that corresponds to the only non-zero eigenvalue. Specifically, we can express $\boldsymbol{\xi}_1 = (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)/\|\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\|$ (Petkova et al. 2016), up to a sign. The following Proposition summarizes the equivalence between the criteria Eqs. (3.6) and (3.10).

Proposition 3.4 *For the linear $\boldsymbol{X}$-by-T interaction model Eq. (3.1) with $K = 2$, the solution $\boldsymbol{\beta}^*$ of Eq. (3.10) satisfies: $\boldsymbol{\beta}^* = \boldsymbol{\xi}_1$, up to a scale constant. That is, under Eq. (3.1) with $K = 2$, there is an equivalence between Eqs. (3.6) and (3.10) in terms of producing vectors in $S_{\mathcal{C}|\boldsymbol{X}}(= span(\boldsymbol{\xi}_1))$.*

The proof of Proposition 3.4 is presented in Appendix A6 Proposition 3.4 states that the modified covariate method (the right-hand side of Eq. (3.10)) produces a vector in the subspace $S_{\mathcal{C}|\boldsymbol{X}}$ when T is binary (i.e., $K = 2$). This implies that in the special case of $K = 2$, the rank-1 approximation model Eq. (3.4) reduces to the modified covariate model Eq. (3.8) when using the optimization framework Eq. (2.8) to estimate the dimension reduction vector $\boldsymbol{\beta}$. The approximation model Eq. (3.4) is a specific instance of the dimension reduction model Eq. (2.6) where the t-specific functions $g_t^*(u)$ are linear functions, $g_t^*(u) = \gamma_t u$ $(t \in \mathcal{T})$. Therefore, we can view the modified covariate method as a special case of the approach that estimates a vector in $S_{\mathcal{C}|\boldsymbol{X}}$, where we restrict the t-specific flexible functions g_t to be linear, specifically in the case of $K = 2$.

4 A semiparametric model

4.1 A constrained single-index regression

A semiparametric generalization of the linear rank-1 approximation model Eq. (3.4) to model Eq. (2.6) can be defined based on replacing the unknown slopes $\{\gamma_t \in \mathbb{R}\}_{t\in\mathcal{T}}$ in model Eq. (3.4) with unknown flexible functions $\{g_t \in \mathcal{H}^{(\boldsymbol{\beta})}\}_{t\in\mathcal{T}}$ defined over an unknown single-index $\boldsymbol{\beta}'\boldsymbol{X} \in \mathbb{R}$. This formulation corresponds to a specific case within the optimization framework Eq. (2.8):

$$\begin{aligned} &\underset{g_t\in\mathcal{H}^{(\boldsymbol{\beta})},\boldsymbol{\beta}\in\Theta_1}{\text{argmin}} && \mathbb{E}\big[\big(Y - g_T(\boldsymbol{\beta}'\boldsymbol{X})\big)^2\big] \\ &\text{subject to} && \mathbb{E}[g_T(\boldsymbol{\beta}'\boldsymbol{X})|\boldsymbol{X}] = 0 \end{aligned} \tag{4.1}$$

for all $\boldsymbol{\beta} \in \Theta_1$, where the matrix $\boldsymbol{B} \in \Theta_q$ is replaced with a vector $\boldsymbol{\beta} \in \Theta_1$. Specifically, optimization Eq. (4.1) indicates that we utilize the working model $Y = g_T(\boldsymbol{\beta}'\boldsymbol{X}) + \epsilon$ subject to $\mathbb{E}[g_T(\boldsymbol{\beta}'\boldsymbol{X})|\boldsymbol{X}] = 0$, where ϵ is a mean zero noise with a finite variance (without loss of generality, the model intercept term was suppressed). We call this single-index model with the constraint (on g_t $(t \in \mathcal{T})$) a constrained single-index model (CSIM).

Within the underlying model Eq. (2.6), solving Eq. (4.1) provides us with a vector, denoted as $\boldsymbol{\beta}^* \in \Theta_1$, which serves as an approximation to a vector in the subspace $S_{\mathcal{C}|X} = \text{span}(\boldsymbol{B}^*)$. If $q = 1$, then we have $\boldsymbol{\beta}^* = \boldsymbol{B}^*$, and if $q > 1$, then $\text{span}(\boldsymbol{\beta}^*)$ represents the best rank-1 approximation to the $\text{span}(\boldsymbol{B}^*)$ in L^2. To illustrate this, under the assumed model Eq. (2.6), we can expand the square error criterion function in Eq. (4.1) by $\mathbb{E}\big[\big(Y - g_T(\boldsymbol{\beta}'\boldsymbol{X})\big)^2\big] = \mathbb{E}\big[Y^2 - 2g_T(\boldsymbol{\beta}'\boldsymbol{X})Y + (g_T(\boldsymbol{\beta}'\boldsymbol{X}))^2\big]$, and re-express Eq. (4.1), as follows

$$\begin{aligned} &\underset{g_t\in\mathcal{H}^{(\boldsymbol{\beta})},\boldsymbol{\beta}\in\Theta_1}{\text{argmin}} \ \mathbb{E}\Big[Y^2 - 2g_T(\boldsymbol{\beta}'\boldsymbol{X})\big(\mu^*(\boldsymbol{X}) + g_T^*(\boldsymbol{B}^{*\prime}\boldsymbol{X})\big) + \big(g_T(\boldsymbol{\beta}'\boldsymbol{X})\big)^2\Big] \\ = &\underset{g_t\in\mathcal{H}^{(\boldsymbol{\beta})},\boldsymbol{\beta}\in\Theta_1}{\text{argmin}} \ \mathbb{E}\Big[Y^2 - 2g_T(\boldsymbol{\beta}'\boldsymbol{X})g_T^*(\boldsymbol{B}^{*\prime}\boldsymbol{X}) + \big(g_T(\boldsymbol{\beta}'\boldsymbol{X})\big)^2\Big] \\ = &\underset{g_t\in\mathcal{H}^{(\boldsymbol{\beta})},\boldsymbol{\beta}\in\Theta_1}{\text{argmin}} \ \mathbb{E}\Big[\big(g_T(\boldsymbol{\beta}'\boldsymbol{X}) - g_T^*(\boldsymbol{B}^{*\prime}\boldsymbol{X})\big)^2\Big], \end{aligned} \tag{4.2}$$

subject to the constraint $\mathbb{E}[g_T(\boldsymbol{\beta}'\boldsymbol{X})|\boldsymbol{X}] = 0$ in Eq. (4.1). In Eq. (4.2), the first equality follows from the fact that the expected value of the cross-product term, $2g_T(\boldsymbol{\beta}'\boldsymbol{X})\mu^*(\boldsymbol{X})$, vanishes to zero (implying that the nuisance term $\mu^*(\boldsymbol{X})$ vanishes from the expression) due to $\mathbb{E}[g_T(\boldsymbol{\beta}'\boldsymbol{X})\mu^*(\boldsymbol{X})|\boldsymbol{X}] = \mu^*(\boldsymbol{X})\mathbb{E}[g_T(\boldsymbol{\beta}'\boldsymbol{X})|\boldsymbol{X}] = 0$ (as a result of the constraint in Eq. (4.1)).

Expression Eq. (4.2) indicates that the model component $g_T(\boldsymbol{\beta}'\boldsymbol{X})$ within Eq. (4.1) specifically targets the true $\boldsymbol{X}$-by-T interaction effect $g_T^*(\boldsymbol{B}^{*\prime}\boldsymbol{X})$, rather than the "nuisance" component $\mu^*(\boldsymbol{X})$ of the underlying model Eq. (2.6). Notably, within Eq. (4.1), it is not necessary to explicitly specify the $\boldsymbol{X}$ main effect term $\mu^*(\boldsymbol{X})$ when estimating a vector $\boldsymbol{\beta}$ belonging to $S_{\mathcal{C}|X}$. As the functions $\{g_t\}_{t\in\mathcal{T}}$ are unknown flexible functions, a closed-form solution for Eq. (4.1) is unavailable. Therefore, an iterative procedure is required to optimize $\boldsymbol{\beta}$ and $\{g_t\}_{t\in\mathcal{T}}$. We describe below our approach to obtaining a profile estimator of $\{g_t\}_{t\in\mathcal{T}}$ for each value of $\boldsymbol{\beta}$. In Sects. 4.3 and 4.4, we will describe our approach to estimating $\boldsymbol{B}$ with an additional regularization to address the potential high dimensionality of the covariates data.

The constraint in Eq. (4.1) on the functions $\{g_t\}_{t\in\mathcal{T}}$ can be absorbed into their basis construction through reparametrization, as we describe next. Suppose we have a set of points $(\boldsymbol{\beta}'\boldsymbol{x}_i, t_i)$ $(i = 1, \ldots, n)$ for a fixed $\boldsymbol{\beta}$. We can represent $g_{t_i}(\boldsymbol{\beta}'\boldsymbol{x}_i)$ $(i = 1, \ldots, n)$ based on a d-dimensional basis $\Psi(\cdot) \in \mathbb{R}^d$ (e.g., B-spline basis on evenly spaced knots on a bounded domain):

$$g_{t_i}(\boldsymbol{\beta}'\boldsymbol{x}_i) = \Psi(\boldsymbol{\beta}'\boldsymbol{x}_i)'\boldsymbol{\theta}_{t_i} \quad (i = 1, \ldots, n) \tag{4.3}$$

for a set of unknown basis coefficients $\{\boldsymbol{\theta}_t \in \mathbb{R}^d\}_{t\in\mathcal{T}}$, for each given $\boldsymbol{\beta}$. We impose the following restriction on Eq. (4.3) to satisfy the required constraint in Eq. (4.1),

$$\sum_{t=1}^{K} \pi_t \boldsymbol{\theta}_t = \boldsymbol{\pi}\boldsymbol{\theta} = \boldsymbol{0}, \tag{4.4}$$

where $\boldsymbol{\theta} := (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')' \in \mathbb{R}^{dK}$ is the vectorized version of the basis coefficients $\{\boldsymbol{\theta}_t\}_{t\in\mathcal{T}}$ in Eq. (4.3), and the matrix $\boldsymbol{\pi} := [\pi_1 \boldsymbol{I}_d; \ldots; \pi_K \boldsymbol{I}_d] \in \mathbb{R}^{d\times dK}$ is the constraint matrix with $\pi_t = \Pr(T = t)$, and $\boldsymbol{0} \in \mathbb{R}^d$ is the length-d vector of zeros. Condition Eq. (4.4) indicates $\mathbb{E}[\boldsymbol{\theta}_T] = \boldsymbol{0}$, and is a sufficient condition to satisfy the constraint in Eq. (4.1) for any set of functions of the form Eq. (4.3).

Let the $n \times d$ matrices $\boldsymbol{D}_{\boldsymbol{\beta},t}$ $(t = 1, \ldots, K)$ denote the evaluation matrices of the basis $\Psi(\cdot)$ on $\boldsymbol{\beta}'\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ specific to the treatment $T = t$ $(t = 1, \ldots, K)$, whose ith row is the $1 \times d$ vector $\Psi(\boldsymbol{\beta}'\boldsymbol{x}_i)'$ if $t_i = t$, and a row of zeros $\boldsymbol{0}_d'$ if $t_i \neq t$. Then, the column-wise concatenation of the design matrices $\{\boldsymbol{D}_{\boldsymbol{\beta},t}\}_{t\in\mathcal{T}}$, i.e., the matrix $\boldsymbol{D}_{\boldsymbol{\beta}} = [\boldsymbol{D}_{\boldsymbol{\beta},1}; \ldots; \boldsymbol{D}_{\boldsymbol{\beta},K}] \in \mathbb{R}^{n\times dK}$, defines the model matrix associated with $\boldsymbol{\theta} := (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')' \in \mathbb{R}^{dK}$ for representation Eq. (4.3). We define a P-spline penalty matrix associated with the basis coefficient $\boldsymbol{\theta} \in \mathbb{R}^{dK}$; we write $\boldsymbol{P} = (\boldsymbol{1}_K' \otimes \boldsymbol{\delta})'(\boldsymbol{1}_K' \otimes \boldsymbol{\delta}) \in \mathbb{R}^{dK\times dK}$, where $\boldsymbol{1}_K \in \mathbb{R}^K$ is the vector of ones, $\boldsymbol{\delta} \in \mathbb{R}^{(d-2)\times d}$ is the second order P-splines difference penalty (Eilers and Marx 1996), and $\otimes$ represents the Kronecker product. For a fixed $\boldsymbol{\beta}$ and a roughness penalty parameter $r \geq 0$, denoting $\boldsymbol{Y}_{n\times 1} = (y_1, \ldots, y_n)'$, an empirical criterion function associated with the constrained optimization problem Eq. (4.1) is $\|\boldsymbol{Y}_{n\times 1} - \boldsymbol{D}_{\boldsymbol{\beta}}\boldsymbol{\theta}\|^2 + r\boldsymbol{\theta}'\boldsymbol{P}\boldsymbol{\theta}$, subject to the constraint in Eq. (4.4). The linear constraint Eq. (4.4) $\boldsymbol{\pi}\boldsymbol{\theta} = \boldsymbol{0}$ can be absorbed into the model matrix $\boldsymbol{D}_{\boldsymbol{\beta}}$ and the penalty matrix $\boldsymbol{P}$ as follows. We can find a basis matrix $\boldsymbol{Q} \in \mathbb{R}^{dK\times d(K-1)}$ (that spans the *null* space of the linear constraint Eq. (4.4)), such that if we set $\boldsymbol{\theta} = \boldsymbol{Q}\tilde{\boldsymbol{\theta}}$ for any arbitrary vector $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{d(K-1)}$, then the resulting vector $\boldsymbol{\theta} \in \mathbb{R}^{dK}$ automatically satisfies the constraint Eq. (4.4), i.e., $\boldsymbol{\pi}\boldsymbol{\theta} = \boldsymbol{0}$. Such a basis matrix $\boldsymbol{Q}$ can be constructed by the "Q" component of a QR decomposition of the matrix $\boldsymbol{\pi}'$ in Eq. (4.4). By setting $\tilde{\boldsymbol{D}}_{\boldsymbol{\beta}} \leftarrow \boldsymbol{D}_{\boldsymbol{\beta}}\boldsymbol{Q}$ and $\tilde{\boldsymbol{P}} \leftarrow \boldsymbol{Q}'\boldsymbol{P}\boldsymbol{Q}$, we can then reparametrize the constrained objective function in terms of the unconstrained vector $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{d(K-1)}$, with the corresponding unconstrained objective, $\|\boldsymbol{Y}_{n\times 1} - \tilde{\boldsymbol{D}}_{\boldsymbol{\beta}}\tilde{\boldsymbol{\theta}}\|^2 + r\tilde{\boldsymbol{\theta}}'\tilde{\boldsymbol{P}}\tilde{\boldsymbol{\theta}}$, which we optimize over $\tilde{\boldsymbol{\theta}}$ in an unconstrained fashion. For each fixed $\boldsymbol{\beta}$, we choose the roughness penalty parameter $r \geq 0$ via generalized cross-validation (GCV). The corresponding unconstrained minimizer $\hat{\tilde{\boldsymbol{\theta}}} \in \mathbb{R}^{d(K-1)}$ is then transformed back to the constrained space to obtain $\boldsymbol{\theta} = (\hat{\boldsymbol{\theta}}_1', \ldots, \hat{\boldsymbol{\theta}}_K')' := \boldsymbol{Q}\hat{\tilde{\boldsymbol{\theta}}} \in \mathbb{R}^{dK}$, yielding the corresponding estimator $\hat{g}_t(\cdot) = \Psi(\cdot)'\hat{\boldsymbol{\theta}}_t$ $(t \in \mathcal{T})$ of the CSIM component Eq. (4.3), for each fixed $\boldsymbol{\beta}$.

Now, denoting the least squares criterion for $\boldsymbol{\beta}$ that corresponds to Eq. (4.1) as

$$\hat{Q}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{g}_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i)\right)^2 /2, \tag{4.5}$$

its gradient is $\nabla \hat{Q}(\boldsymbol{\beta}) = -n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{g}_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i)\right) \dot{\hat{g}}_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i)\boldsymbol{X}_i$, where $\dot{\hat{g}}_t(u)$ denotes the derivative of $\hat{g}_t(\cdot)$ evaluated at u.

4.2 Geometric intuition

In this section, we will provide some geometric intuition behind the optimization approach Eq. (4.1) to approximating the interaction effect term $g(\boldsymbol{X}, T)$ of model Eq. (2.1). We can easily confirm that the profile minimizer, which we denote in this section as $\{g_t^*\}_{t\in\mathcal{T}}$ for each fixed $\boldsymbol{\beta} \in \Theta_1$, of the constrained least square criterion Eq. (4.1) satisfies:

$$g_t^*(\boldsymbol{\beta}'\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T = t] - \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}] \quad (t \in \mathcal{T}). \tag{4.6}$$

In Eq. (4.6), the first term $\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T = t]$ is the treatment t-specific L^2 projection of Y onto $\mathcal{H}^{(\boldsymbol{\beta})}$, whereas the second term $-\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]$ acts as a "shift" to adjust the first term in order to satisfy the constraint in the criterion Eq. (4.1). This adjustment via shifting by $\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]$ in Eq. (4.6) ensures the orthogonality of the function $g_T^*(\boldsymbol{\beta}'\boldsymbol{X})$ in Eq. (4.6) with respect to the unspecified term $\mu^*(\boldsymbol{X})$ in the underlying model Eq. (2.1) in L^2 for each value of $\boldsymbol{\beta}$. As a result, any possible misspecification of the "nuisance" term $\mu^*(\boldsymbol{X})$ does not affect the optimization of the dimension reduction vector $\boldsymbol{\beta}$.

To provide a geometric illustration, let us consider a very simple example where we regress Y on the treatment variable $T \in \mathcal{T}$ without any covariates (i.e., the covariate vector $\boldsymbol{X}$ contains only the intercept term "1"). In this simple setting, the solution $\{g_t^*\}_{t\in\mathcal{T}}$ in Eq. (4.6) are just treatment t-specific constants:

$$g_t^* = \mathbb{E}[Y|T = t] - \mathbb{E}[Y] \quad (t \in \mathcal{T}). \tag{4.7}$$

Given the sample data (y_i, t_i) $(i = 1, \ldots, n)$, let $\boldsymbol{Y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ denote the (length-n) observed vector of responses. In Eq. (4.7), the second term $\mathbb{E}[Y]$ corresponds to the grand average of Y, represented by the vector $\bar{Y}\mathbf{1}_n \in \mathbb{R}^n$, where $\bar{Y} = \sum_{i=1}^{n} y_i/n$ is the grand average, and $\mathbf{1}_n = (1, 1, \ldots, 1)' \in \mathbb{R}^n$. On the other hand, the first term $\mathbb{E}[Y|T = t]$ $(t \in \mathcal{T})$ corresponds to the treatment group-specific average, represented by the vector $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)' \in \mathbb{R}^n$, where $\hat{Y}_i = \sum_{t=1}^{K} 1_{(t_i=t)}\bar{Y}^{(t)}$

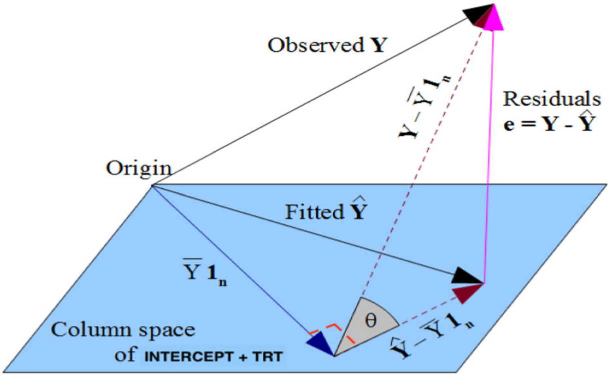

Fig. 1 In the simple linear regression $\mathbb{E}[Y|T]$ of the outcomes on the treatments, the fitted $\hat{\boldsymbol{Y}}$ for the model $\mathbb{E}[Y|T]$ is the orthogonal projection of the observed $\boldsymbol{Y}$ onto the plane of the column space spanned by the intercept and the treatments, which is represented by the blue plane. The fitted vector for the "1" (i.e., the intercept)-only model $\mathbb{E}[Y|1]$ is represented by $\bar{Y}\mathbf{1}_n$. In the picture, the magnitude of the interaction effect between T and "1" is quantified by the squared length of the vector $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$

$(i = 1, \ldots, n)$ with $\bar{Y}^{(t)} = \sum_{i=1}^{n} y_i 1_{(t_i=t)} / \sum_{i=1}^{n} 1_{(t_i=t)}$, which corresponds to the treatment t-specific average.

Then, in Fig. 1, the fitted value of g_T^* in Eq. (4.7) is represented by the adjacent side $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n \in \mathbb{R}^n$. The squared magnitude of this vector, $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2$ corresponds to the variance of g_T^*, i.e., $\text{var}[g_T^*] = \mathbb{E}[(g_T^*)^2]$. Notice that the fitted "signal" vector, $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$ (related with the T-by-"1" interaction) is orthogonal (perpendicular) to the "nuisance" vector $\bar{Y}\mathbf{1}_n$ (unrelated with T) in Fig. 1.

Intuitively, the "effect" of intercept "1" in the intercept-only model is to *average* the response $\boldsymbol{Y} \in \mathbb{R}^n$, which results in the fit $\bar{Y}\mathbf{1}_n \in \mathbb{R}^n$ in Fig. 1. The squared magnitude $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2 \sim \text{var}[g_T^*]$, where $\hat{\boldsymbol{Y}}$ is the vector of treatment t-specific averages, quantifies the extent to which the "effect" of intercept "1" (i.e., the grand averaging) is modified by the variable T, and hence the magnitude of $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2$ quantifies the intensity of the "interaction effect" between the intercept "1" and T. Analogously, within the optimization framework Eq. (4.1), given a candidate $\boldsymbol{\beta} \in \Theta_1$, the variance of Eq. (4.6), denoted as $\text{var}\left(g_T^*(\boldsymbol{\beta}'\boldsymbol{X})\right) = \mathbb{E}\left[\{g_T^*(\boldsymbol{\beta}'\boldsymbol{X})\}^2\right]$, captures the extent to which the $\boldsymbol{X}$-by-T interaction effect varies with different values of $\boldsymbol{\beta}'\boldsymbol{X}$. The goal is to maximize this variance of the $\boldsymbol{X}$-by-T interaction effect over $\boldsymbol{\beta}$, similar to the objective of maximizing the variance Eq. (3.7) in previous analysis.

When we replace the intercept "1" with the unknown index $\boldsymbol{\beta}'\boldsymbol{X}$, the blue plane in Fig. 1 represents the Hilbert space of measurable functions of $(\boldsymbol{\beta}'\boldsymbol{X}, T)$. Maximizing the variance of the $\boldsymbol{X}$-by-T interaction effect, $\text{var}\left[g_T^*(\boldsymbol{\beta}'\boldsymbol{X})\right]$, over $\boldsymbol{\beta} \in \Theta_1$ corresponds to adjusting (*tilting*) the blue plane of Fig. 1. The objective is then to minimize the angle θ formed by the hypotenuse $\boldsymbol{Y} - \bar{Y}\mathbf{1}_n$ and the adjacent side $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$ (i.e., the angle θ formed by the two dashed lines in Fig. 1). In other words, the goal is to maximize the cosine of θ (over $\boldsymbol{\beta} \in \Theta_1$), which tilts the plane to *maximize* the squared magnitude of the vector, $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2 \sim \text{var}\left(g_T^*(\boldsymbol{\beta}'\boldsymbol{X})\right)$.

Finally, we note that the two centered vectors $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n$ and $\boldsymbol{Y} - \bar{Y}\mathbf{1}_n$ (i.e., the two dashed lines in Fig. 1) correspond to the fitted ($\hat{\boldsymbol{Y}}$) and the observed ($\boldsymbol{Y}$) vectors, respectively, both of which centered by the intercept vector ($\bar{Y}\mathbf{1}_n$). Without centering the fit $\hat{\boldsymbol{Y}}$ by the intercept $\bar{Y}\mathbf{1}_n$, there would be no Pythagorean-type sum of squares decomposition:

$$\|\boldsymbol{Y} - \bar{Y}\mathbf{1}_n\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2, \tag{4.8}$$

in which the second term, $\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}_n\|^2$, quantifies the T-by-"1" interaction effect. Analogously, the "shifting" component $-\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]$ in Eq. (4.6) plays the role of an "intercept." By centering the unrestricted fit $\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T]$ by the reference function $\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]$, we can achieve the following Pythagorean-type decomposition similar to Eq. (4.8), which isolates the variance of the $\boldsymbol{X}$-by-T interaction effect in the second term of the decomposition:

$$\begin{aligned}\mathbb{E}\big[\big(Y - \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]\big)^2\big] &= \mathbb{E}\big[\big(Y - \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T]\big)^2\big] \\ &\quad + \mathbb{E}\big[\big(\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T] - \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]\big)^2\big],\end{aligned} \tag{4.9}$$

and the second term $\mathbb{E}\big[\big(g_T^*(\boldsymbol{\beta}'\boldsymbol{X})\big)^2\big] = \mathbb{E}\big[\big(\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T] - \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]\big)^2\big]$ (see Eq. (4.6) for its definition) is then maximized over $\boldsymbol{\beta} \in \Theta_1$.

4.3 L^1 regularization

One shortcoming of the formulation Eq. (4.1) is that the linear projection $\boldsymbol{\beta}'\boldsymbol{X}$ is defined in terms of all the predictors in the model, i.e., the approach forces all the predictors play a role in building an interaction term $g_T(\boldsymbol{\beta}'\boldsymbol{X})$. However, only a subset of measurements in $\boldsymbol{X}$ may be useful in determining an optimal ITR. Also, high-dimensional settings can lead to instabilities and issues of overfitting. In this section, we introduce an L^1 regularization that can both avoid overfitting and choose among multiple potential covariates by obtaining a sparse estimate of the coefficient $\boldsymbol{\beta}$ in Eq. (4.1).

Extending the L^1 penalized least squares estimation approach (Tibshirani 1996; Zou 2006; Meinshausen and Yu 2009; Schneider and Tardivel 2022), from the linear regression context (e.g., Qian and Murphy (2011)) to the single index regression context poses challenges due to the nonconvex nature of the squared error criterion function with respect to the single index coefficient ($\boldsymbol{\beta}$). Wang and Yin (2008) proposed an approach that introduces L^1 regularization into the minimum average variance estimation (MAVE) method of Xia et al. (2002), but its computational complexity grows rapidly with the sample size n, and also becomes unstable when the dimension is high. In other work, Peng and Huang (2011) estimate the single-index model by minimizing a penalized least squares criterion, performing simultaneous predictor selection; Zhu et al. (2011) use the adaptive lasso with kernel smoothing; and Wang and Wang (2015) use the smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) penalization allowing diverging number of parameters. However, Radchenko

(2015) noted that such penalization approaches may be problematic (particularly in high-dimensional settings), due to the nonconvexity of the squared error criterion function, e.g., the nonconvexity of $\hat{Q}(\boldsymbol{\beta})$ in Eq. (4.5) with respect to $\boldsymbol{\beta}$. Specifically, for a penalized criterion that appends a (convex) penalty, $p_\lambda(\boldsymbol{\beta})$ with some sparsity tuning parameter $\lambda \geq 0$, to the squared error criterion $\hat{Q}(\boldsymbol{\beta})$ in Eq. (4.5), then the solution path of the minimizer $\hat{\boldsymbol{\beta}}^{(\lambda)}$ would not be generally a continuous function of λ. For this reason, selecting an appropriate tuning parameter $\lambda(\geq 0)$ is extremely difficult.

Radchenko (2015) proposed a constrained L^1 regularization approach that handles this tuning parameter selection problem, which we will incorporate into the optimization formulation Eq. (4.1) of CSIM. It is suggested in Radchenko (2015) that the L^1 norm of $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_1 = \lambda$, is directly used as the sparsity tuning parameter for $\boldsymbol{\beta}$. Thus we can consider solving a constrained minimization problem for any $\lambda = \|\boldsymbol{\beta}\|_1$ with $\lambda \in [1, \lambda_{\max}]$, where $\lambda = 1$ represents the sparsest case, and λ increases to some specified value of $\lambda_{\max}$. The corresponding empirical version of the constrained minimization Eq. (4.1), for each choice of λ, is:

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \hat{Q}(\boldsymbol{\beta}), \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_1 = \lambda \end{aligned} \tag{4.10}$$

(where $\hat{Q}(\boldsymbol{\beta})$ is given in Eq. (4.5)), for a sparsity-inducing parameter $\lambda \in [1, \lambda_{\max}]$. For example, a small value of λ (≥ 1) in Eq. (4.10) will generate a sparse solution $\boldsymbol{\beta}$, with $\lambda = 1$ corresponding to the sparsest case for $\boldsymbol{\beta}$ (i.e., only one component of $\boldsymbol{\beta}$ equals 1 and all other components of $\boldsymbol{\beta}$ are zero).

Radchenko (2015) proved that the constrained minimizer, $\boldsymbol{\beta}^{(\lambda)}$, constructs a continuous path as a function of the tuning parameter $\lambda \in [1, \lambda_{\max}]$. Therefore, with the criterion function $\hat{Q}(\boldsymbol{\beta})$, the sparsity parameter λ can be reliably selected by minimizing an estimate of the expected value of $\hat{Q}(\boldsymbol{\beta}^{(\lambda)})$ in Eq. (4.5)), for example, a cross-validated prediction error, the Akaike information criterion (AIC, Akaike 1974), or the corrected AIC (AICc, Sugiura 1978; Hurvich and Tsai 1989). In this paper, we will select the sparsity tuning parameter $\lambda(\geq 1)$ by minimizing AICc $:= n\log(\hat{Q}(\boldsymbol{\beta}^{(\lambda)})) + 2p^* + 2p^*(p^*+1)/(n-p^*-1)$, where p^* is the number of nonzero elements in $\boldsymbol{\beta}^{(\lambda)}$. AICc behaves similarly to a cross-validated prediction error in our empirical studies, while showing some advantage over AIC in small sample applications. The algorithm for optimizing this sparse CSIM Eq. (4.10) follows closely the approach of Radchenko (2015) using a block coordinate descent (BCD) of block size 2-details are provided in Sect. 4.4.

4.4 The sparse CSIM implementation using a block-coordinate descent algorithm

In this section, we provide details of the algorithm for implementing the sparse CSIM introduced in Sect. 4.3. Note that we work with the nontrivial case that there is at least one non-zero component in $\boldsymbol{\beta}$. At the initialization step of the estimation, we choose the first non-zero component by

$$j_1 \quad = \quad \underset{j \in \{1,\ldots,p\}}{\arg\min} \quad \hat{Q}(\boldsymbol{e}_j), \tag{4.11}$$

in which $\boldsymbol{e}_j = [0, \ldots, 1, \ldots, 0]' \in \mathbb{R}^p$, where the jth component equals 1, and all other components equal 0 (i.e., the canonical basis of $\mathbb{R}^p$). Then j_1 is a component index that corresponds to the estimated best "signal" treatment effect modifier among the p covariates. Radchenko (2015) suggests fixing $\beta_{j_1} = 1$ throughout the estimation procedure, to be used as the model identifiability constraint; without loss of generality, we take $\beta_1 = 1$, as we can always arrange the j_1th covariate as the first component $\boldsymbol{X}_{i,1}$ of the covariate vector, and then can rescale to satisfy $\boldsymbol{\beta} \in \Theta_1$.

Given a new value of λ on a dense grid in $[1, \lambda_{\max}]$, say, $\lambda^{(\text{new})} \in (\lambda^{(\text{old})}, \lambda_{\max}]$, the last computed solution $\boldsymbol{\beta}^{(\text{old})}$ (which satisfies $\|\boldsymbol{\beta}^{(\text{old})}\|_1 = \lambda^{(\text{old})}$ for some $\lambda^{(\text{old})} \in [1, \lambda^{(\text{new})})$), can be used as a warm start in the search for the next one. Due to the continuity of the (L^1 norm) constraint $\|\boldsymbol{\beta}\|_1$ with respect to $\boldsymbol{\beta}$, this search only needs to be conducted locally near $\boldsymbol{\beta}^{(\text{old})}$, for a small change in λ, i.e., for a small increment $\lambda^{(\text{new})} - \lambda^{(\text{old})}$. Therefore a local (quadratic) approximation to the objective function in Eq. (4.10) near $\boldsymbol{\beta}^{(\text{old})}$ can be justified (Radchenko 2015). The approach constructs a sequence of locally approximated convex objective functions near the last computed solutions, bypassing the issue of the noncontinuity of solution $\boldsymbol{\beta}$ with respect to λ. The success of iterative algorithms depends on the initialization, i.e., solving Eq. (4.11), and this is indeed the case under the setting Eq. (4.10). Due to the constraint in Eq. (4.10), we implement a block coordinate descent algorithm as described in Algorithm 1 below, following closely that of Radchenko (2015), to obtain a sparse estimate of $\boldsymbol{\beta}$.

It is suggested in Radchenko (2015) to use a block of size two that consists of $\{j, m\}$, where m represents a fixed reference component index within each *for* loop, which gets updated appropriately in Algorithm 1. Denoting by $\mathcal{A}$ the current active component index set, the algorithm cycles through the *for* loop over j's (until convergence) which optimizes individual blocks; this procedure is repeated until $\mathcal{A}$ does not change. Within each block, a situation where a coefficient crosses zero is handled by setting that coefficient to exactly zero and correspondingly updating the other coefficient in the block. To simplify the notation, for each $\boldsymbol{\beta}$, let us denote the gradient of $\hat{Q}(\boldsymbol{\beta})$, i.e., $\nabla \hat{Q}(\boldsymbol{\beta})$ in Eq. (4.5) by ∇, and accordingly, the jth component of the gradient ∇ by ∇_j.

The expression for Δ_j of the (j, m)th block's updating rule in Algorithm 1 is given by:

$$\Delta_j = \frac{-\left(\nabla_j - S_{jm}\nabla_m\right)}{n^{-1}\sum_{i=1}^n \left(\hat{g}_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i)(X_{i,j} - S_{jm}X_{i,m})\right)^2}, \tag{4.12}$$

where $S_{jm} := \text{sign}(\beta_m\beta_j) - \text{sign}(\beta_m\nabla_j)I_{\{\beta_j=0\}}$ and Eq. (4.12) is evaluated at the current $\boldsymbol{\beta}$. Note that S_{jm} indicates the sign of $\beta_m\beta_j$, but when $\beta_j = 0$, the "sign" of β_j is simply the sign of $-\nabla_j$. The value of Δ_m is determined through the relationship:

$$\Delta_m = -\Delta_j S_{jm}, \tag{4.13}$$

to preserve the constraint $\|\boldsymbol{\beta}\|_1 = \lambda^{(\text{new})}$ within each block update. The derivation for Δ_j in Eq. (4.12) is shown below. We minimize the squared error criterion function in Eq. (4.10) (subject to the constraint $\|\boldsymbol{\beta}\|_1 = \lambda^{(\text{new})}$) in a small neighborhood of the current estimate, say, $\tilde{\boldsymbol{\beta}}$, after performing the first order Taylor approximation to the (j, m)th block portion of the regression model:

$$\hat{Q}(\boldsymbol{\beta}) \approx n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{g}_{T_i}(\tilde{\boldsymbol{\beta}}' \boldsymbol{X}_i) - \dot{\hat{g}}_{T_i}(\tilde{\boldsymbol{\beta}}' \boldsymbol{X}_i)(X_{i,j}\Delta_j + X_{i,m}\Delta_m) \right)^2 /2, \quad (4.14)$$

where $\Delta_j := \beta_j - \tilde{\beta}_j$ and $\Delta_m := \beta_m - \tilde{\beta}_m$. The update rule for the (j, m)th block will then be given by: $\left\{ \beta_j \leftarrow \tilde{\beta}_j + \hat{\Delta}_j,\ \beta_m \leftarrow \tilde{\beta}_m + \hat{\Delta}_m \right\}$, where $\left[\hat{\Delta}_j, \hat{\Delta}_m \right]'$ represents the minimizer of Eq. (4.14) over $\left[\Delta_j, \Delta_m \right]' \in \mathbb{R}^2$, subject to the constraint Eq. (4.13) which preserves the L^1 norm of the solution at $\lambda^{(\text{new})}$, i.e., $\|\boldsymbol{\beta}\|_1 = \lambda^{(\text{new})}$. After substituting Δ_m in Eq. (4.14) by $-\Delta_j S_{jm}$ in Eq. (4.13), we can take the derivative of Eq. (4.14) with respect to Δ_j and set it to 0

$$\sum_{i=1}^{n} \left(Y_i - \hat{g}_{T_i}(\tilde{\boldsymbol{\beta}}' \boldsymbol{X}_i) - \dot{\hat{g}}_{T_i}(\tilde{\boldsymbol{\beta}}' \boldsymbol{X}_i)(X_{i,j} - S_{jm}X_{i,m})\Delta_j \right) \left(-\dot{\hat{g}}_{T_i}(\tilde{\boldsymbol{\beta}}' \boldsymbol{X}_i)(X_{i,j} - S_{jm}X_{i,m}) \right) = 0,$$

and solving for Δ_j gives the expression Eq. (4.12).

At the start (i.e., when $\lambda = 1$) of the fitting procedure, set j_1 in Eq. (4.11) to 1, $\boldsymbol{\beta} = [1, 0, \ldots, 0]'$, $m = \arg\max_{j \neq 1} |\nabla_j|$, and $\mathcal{A} = \{1, m\}$. If we increase the tuning parameter $\lambda^{(\text{old})}$ to the next point $\lambda^{(\text{new})}$ on a grid $[1, \lambda_{\max}]$, then the magnitude of some coefficients of $\boldsymbol{\beta}$ needs to be increased from their current values to ensure the new constraint $\|\boldsymbol{\beta}\|_1 = \lambda^{(\text{new})}$ in Eq. (4.10) is satisfied. Without loss of generality, at the start of Algorithm 1, we will bring the increment to the mth component of $\boldsymbol{\beta}$, i.e., we set $\beta_m \leftarrow \beta_m + s_m(\lambda^{(\text{new})} - \lambda^{(\text{old})})$ with $s_m := \text{sign}(-\nabla_m)$, so that the change reduces the criterion function of Eq. (4.10). The rest of the updating procedure is given in Algorithm 1 below.

As the tuning parameter $\lambda^{(\text{old})}$ is increased to the next point $\lambda^{(\text{new})}$ on a grid $[1, \lambda_{\max}]$, Algorithm 1 employs warm-start and active-set techniques for faster computation. The loop on line 4 involves only elements within the active set, and the algorithm checks (on line 10) whether any excluded covariates should be added to the model. The active set remains unchanged for most grid points in the path, leading to substantial computational savings when p is large on the sparse part of the solution path. Throughout the paper, we use $\lambda_{\max} = 4$ and 100 equally-spaced grid points on $[1, \lambda_{\max}]$, for constructing the solution path.

Algorithm 1 Block Coordinate Descent (BCD) for optimizing $\boldsymbol{\beta}$, subject to $\|\boldsymbol{\beta}\|_1 = \lambda^{(\text{new})}$

```
1: $m \leftarrow \arg\max_{j \in \mathcal{A} \setminus 1} |\beta_j|$.
2: $s_j \leftarrow \text{sign}(\beta_j)$;  if $s_j = 0$ for $j \in \mathcal{A}$, then $s_j \leftarrow \text{sign}(-\nabla_j)$.
3: Iterate the following for loop, until convergence of $\beta_j$, $j \in \mathcal{A}$:
4: for $j \in \mathcal{A} \setminus \{1, m\}$ do the $(j, m)$th block-update:
5:     if $(\beta_j \neq 0)$ or $(s_j \nabla_j \leq 0$ and $|\nabla_j| \geq |\nabla_m|)$, then $\beta_j \leftarrow \beta_j + \Delta_j$ with $\Delta_j$ in (4.12).
6:     else $\Delta_j \leftarrow 0$.
7:     if $\beta_j$ switches sign, then $\Delta_j \leftarrow \Delta_j - \beta_j$ and $\beta_j \leftarrow 0$.
8:     $\beta_m \leftarrow \beta_m + \Delta_m$ with $\Delta_m$ in (4.13).
9:     if $\beta_m$ switches sign, then $\beta_j \leftarrow \beta_j + |\beta_m| s_j$ and $\beta_m \leftarrow 0$, and redefine $m$.
10: if $\exists j \in \mathcal{A}^c$ for which $|\nabla_j| \geq |\nabla_m|$ (using the gradient in (4.5)), then append $\mathcal{A}$ with $j$ and go to line
    1.
11: otherwise stop.
```

5 Simulation illustrations

In this section, we perform numerical studies to illustrate the performance of the proposed sparse CSIM approach for estimating ITRs in comparison with alternative approaches as well as variable selection performance compared to the MC method. We consider $K = 2$ and $K = 3$ cases in the simulation study because these settings are most commonly encountered in studying heterogeneous treatment efficacy.

5.1 ITR performance for $K = 2$ treatment case

We consider $p \in \{50, 500\}$ and $n \in \{250, 500\}$, with a varying degree in the $\boldsymbol{X}$ main effect intensity and in the nonlinearity in the $\boldsymbol{X}$-by-T interaction effect. We simulate 200 training datasets for each scenario. We generate covariates $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$, and treatments $T_i \in \{1, 2\}$ (i.e., $K = 2$) with equal probability at random, independently of $\boldsymbol{X}_i$. We generate outcomes $Y_i = \mu(\boldsymbol{X}_i) + g_{T_i}(\boldsymbol{\beta}^{*\prime}\boldsymbol{X}_i) + \epsilon_i$, with $\epsilon_i \sim N(0, 0.4^2)$. Two scenarios for the treatment-specific link function $g_T(u)$ are considered: 1) a *nonlinear* contrast function $g_T(u) = (-1)^T(\cos(u) - 0.5)$ that gives a *nonlinear* $\boldsymbol{X}$-by-T interaction effect, and 2) a *linear* contrast function $g_T(u) = (-1)^T 0.5u$ that gives a *linear* $\boldsymbol{X}$-by-T interaction effect, respectively. We set the true single-index coefficient vector as $\boldsymbol{\beta}^* = [1, 0.5, 0.25, 0.125, 0, \ldots, 0]' \in \mathbb{R}^p$, in which its elements are all zeros except for the first 4 elements, thus there are only 4 "signal" covariates among $\boldsymbol{X}$ that exhibits an interaction effect with the treatment T. For the main effect, we set $\mu(\boldsymbol{X}; \delta) = 2 + \delta\cos(\boldsymbol{\eta}'\boldsymbol{X})$, with $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_{12}, 0, \ldots, 0]' \in \mathbb{R}^p$, in which the vector $[\eta_1, \ldots, \eta_{12}]' \in \mathbb{R}^{12}$ is randomly generated from a multivariate normal distribution, and is then rescaled to have unit L^2 norm, for each simulation run. All the remaining $p - 12$ elements of $\boldsymbol{\eta}$ are set to be 0; therefore, only the first 12 covariates among $\boldsymbol{X}$ have a nonzero main effect. The scaling parameter $\delta \in \{1, 2\}$ regulates the contribution of the main effect function $\mu(\boldsymbol{X})$ on the variance of Y, in which $\delta = 1$ represents a relatively *moderate* main effect (contributing about the same variance as the interaction effect does) and $\delta = 2$ a relatively *big* main effect case (about 4 times larger than the interaction effect), respectively. For an optimal ITR specification, without loss of generality, we assume that a larger value of Y is desirable. We restrict our attention to ITRs of the form, $\hat{\mathcal{D}}(\boldsymbol{X}) = \underset{t \in \{1,\ldots,K\}}{\arg\max}\ \hat{\mathbb{E}}[Y \mid \boldsymbol{X}, T = t]$, where $\hat{\mathbb{E}}[Y \mid \boldsymbol{X}, T]$ is obtained by the following approaches.

CSIM Optimize CSIM Eq. (4.10) using the block-coordinate descent algorithm described in Sect. 4.4. Following Wang and Yang (2009), we set the number of cubic B-spline basis functions to be $d = 4 + [n^{1/5.5}]$, where $[v]$ denotes the integer part of v. The corresponding ITR is $\hat{\mathcal{D}}(\boldsymbol{X}) = \underset{t \in \{1,\ldots,K\}}{\arg\max}\ \hat{g}_t(\hat{\boldsymbol{\beta}}'\boldsymbol{X})$.

MC Estimate the MC model Eq. (3.8) with efficiency augmentation (Tian et al. 2014a) by minimizing the Lasso (Tibshirani 1996) penalized objective function with a 20-fold cross validation, under efficiency augmentation of $\mathbb{E}[Y \mid \boldsymbol{X}] = \boldsymbol{\eta}'\boldsymbol{X}$ which is fitted by the Lasso with a 20-fold cross validation, as is done in Tian et al. (2014a), implemented using R package “glmnet” (Friedman et al. 2010). This is a linear approach for estimating interactions.

K-LR Estimate a linear regression (LR) model by the Lasso for each of the K $(= 2)$ treatment groups separately, with a 20-fold cross validation to select the tuning parameter, implemented using R package “glmnet”.

K-SAM For each of the K treatment groups separately, estimate a sparse additive model (SAM) Eq. (Ravikumar et al. 2009), with the sparsity tuning parameter for the simultaneous covariate selection chosen by minimizing a 5-fold cross validation for prediction errors, implemented using R package “SAM” (Zhao et al. 2014).

One natural measure for the effectiveness of a treatment decision rule $\hat{\mathcal{D}}$ is called the “Value” (V) of a treatment decision rule $\hat{\mathcal{D}}$ (Qian and Murphy 2011), which is defined as the expected mean outcome if everyone in the population receives treatment according to that rule $\hat{\mathcal{D}}$. We can estimate the Value of $\hat{\mathcal{D}}$, i.e., $\mathbb{E}_X\left[\mathbb{E}_{Y|X}[Y \mid \boldsymbol{X}, T = \hat{\mathcal{D}}(\boldsymbol{X})]\right]$, by an inverse probability weighted estimator (Murphy 2005):

$$V(\hat{\mathcal{D}}) = \sum_{i=1}^{n} Y_i I_{T_i = d(X_i)} \Big/ \sum_{i=1}^{n} I_{T_i = d(X_i)}, \tag{5.1}$$

computed based on a testing set. Since we know the true data generating model, we can determine the optimal ITR, $\mathcal{D}_{\text{opt}}$ in (2.3), for each simulation setting. To evaluate the performance of an estimated ITR $\hat{\mathcal{D}}$ from each of the 4 methods, we report the Value ratio, $V(\hat{\mathcal{D}})/V(\mathcal{D}_{\text{opt}})$, calculated from an independent (large) testing set of size $n = 10^4$, for each simulation setting and each of the 200 simulation replications. We note that a higher value of the ratio indicates a better ITR performance.

In Fig. 2, we present the boxplots of the Value ratios of the estimated ITRs obtained from 200 simulation replications, for each combination of $n \in \{250, 500\}$, $p \in \{50, 500\}$ and the main effect intensity $\delta \in \{1, 2\}$ (which correspond to *moderate* and *big* main effects, respectively), for the *nonlinear* interaction effect cases in the top panels and the *linear* interaction effect cases in the bottom panels.

Under the *nonlinear* interaction effect, the proposed CSIM outperforms all other alternatives in all cases. Interestingly, the method significantly outperforms the K

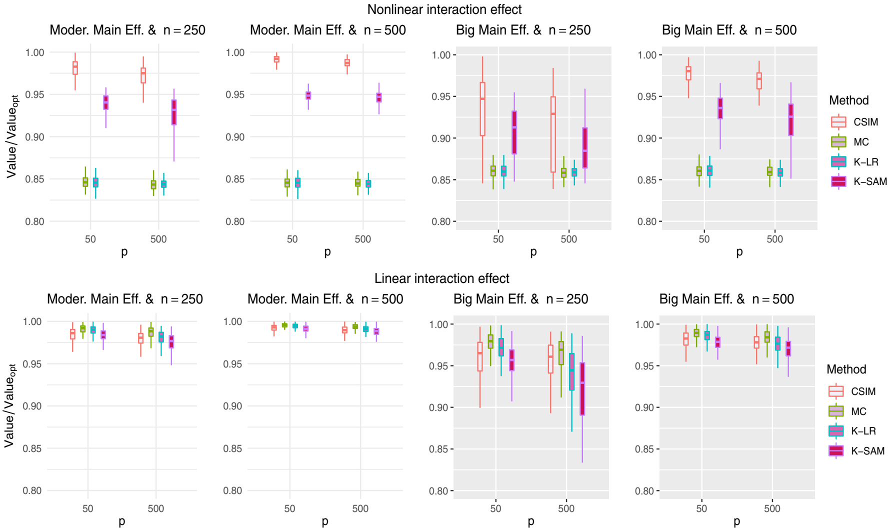

Fig. 2 Top panels: boxplots of the Value ratios of the ITRs estimated from the 4 methods (CSIM, MC, K-SAM, and K-LR) for the nonlinear $\boldsymbol{X}$-by-T interaction effect case. Lower panels: boxplots of the Value ratios of the ITRs for the linear $\boldsymbol{X}$-by-T interaction effect case

separate additive regressions (K-SAM) which is also equipped with a set of flexible additive regression functions to capture the nonlinear associations. Unlike CSIM, however, K-SAM is not robust to misspecification of the $\boldsymbol{X}$ main effect model, and thus the method suffers inconsistency in approximating the $\boldsymbol{X}$-by-T interaction effects when the $\boldsymbol{X}$ main effect is not an additive structure (as in this example). Given a *nonlinear* interaction effect structure, both MC and the K separate linear regressions (K-LR) are clearly worse than CSIM that utilizes the flexible link functions, indicating a distinct benefit of using CSIM for estimating the nonlinear interactions over the linear model-based methods. Under the *linear* interaction effect, MC outperforms the CSIM approach, but only slightly, suggesting that in the absence of prior knowledge about the interaction effects it is suitable to employ CSIM.

Next, we compare the performance of CSIM and MC in terms of the selection of treatment effect modifying variables, i.e., those associated with $\boldsymbol{X}$-by-T interaction, with the training sample size $n \in \{100, 200, \ldots, 1000\}$. We present these selection performance results in Fig. 3. For each setting, we compute the proportion of *correctly* selected treatment effect modifiers, and the proportion of *incorrectly* selected treatment effect modifiers, computed over the 200 simulation runs. The maximum possible number for the correctly selected effect modifiers (i.e., the covariates associated with nonzero elements in the estimated $\boldsymbol{\beta}$) is 4. The maximum possible number for the incorrectly selected effect modifiers is 46 for the $p = 50$ case, and 496 for the $p = 500$ case.

In Fig. 3, not surprisingly, the cases with the *nonlinear* interaction effect (the top two rows) are much more favorable to CSIM over MC. CSIM tends to correctly select the true treatment effect modifiers, whereas MC rarely selects any of the true treatment effect modifiers. The cases with the *linear* interaction effect (the bottom two rows), on the other hand, are more favorable to MC than to CSIM, although not by much. In fact, althogh the averaged proportions of correctly selected treatment effect modifiers are larger for MC, the average proportions of incorrectly selected treatment effect modifiers are actually much smaller for CSIM. Generally, due to an increasing nuisance variability unrelated to the treatment, the $\boldsymbol{X}$ main effect intensity (either moderate, $\delta = 1$, or big, $\delta = 2$) affects the treatment effect modifier selection performance. Nevertheless, as the sample size increases, CSIM tends to recover the true treatment effect modifiers. Overall, there is a clear advantage in utilizing the flexible link functions for discovery of treatment effect modifiers when there exists a *nonlinear* association between the treatment and a set of covariates, while the performance of MC and CSIM is comparable when the interaction effect is *linear*.

5.2 CSIM performance for $K = 3$ treatment case

In this section, we provide additional simulation results to investigate the performance of CSIM for estimating optimal ITRs when the number of treatment groups $K = 3$. We generate the treatment indicators T_i that take values in $\{1, 2, 3\}$ at random with equal probability, independently of $\boldsymbol{X}_i$. We generate the covariates $\boldsymbol{X}_i$ and the outcomes Y_i, using the same $\boldsymbol{X}$ main effect function $\mu(\boldsymbol{X}_i; \delta)$ and the single-index coefficient $\boldsymbol{\beta}^*$ as in the settings of Sect. 5.1, except that the treatment t-specific functions $g_t\ (u)$ $(u \in [0, 1])$, are set to be

$$\begin{cases} g_1(u) = u^1(1-u)^4/B(2,5) - f(u) \\ g_2(u) = u^1(1-u)^1/B(2,4) - f(u) \\ g_3(u) = u^4(1-u)^0/B(5,1) - f(u), \end{cases}$$

where $f(u) = (g_1(u) + g_2(u) + g_3(u))/3$. Here $B(a, b) = (\Gamma(a)\Gamma(b))/\Gamma(a+b)$ is a Beta function, and $u = F(\boldsymbol{\beta}_0'\boldsymbol{X})$, where F is the cumulative distribution function (CDF) of a re-scaled and centered $B((p^*+1)/2, (p^*+1)/2)$

$$F\ (u) = \int_{-1}^{u/r} \frac{\Gamma(p^*+1)}{\Gamma\{(p^*+1)/2\}^2\, 2^{p^*}} (1-t^2)^{(p^*-1)/2} dt, \quad u \in [-r, r], \tag{5.2}$$

in which p^* denotes the number of nonzero elements in $\boldsymbol{\beta}^*$, i.e., $p^* = 4$, and r is the maximum of the absolute values of $\{\boldsymbol{\beta}^{*\prime}\boldsymbol{X}_i, i = 1, \ldots, n\}$. The reason for employing the transformed variable $F(\boldsymbol{\beta}^{*\prime}\boldsymbol{X}_i)$ instead of $\boldsymbol{\beta}^{*\prime}\boldsymbol{X}_i$ is that $\{F(\boldsymbol{\beta}^{*\prime}\boldsymbol{X}_i), i = 1, \ldots, n\}$ is quasi-uniformly distributed on the interval $[0, 1]$ (Wang and Wang 2015).

The upper panel of Fig. 4 shows the true treatment-specific contrast functions $g_T\ (u)$, $u \in [0, 1]$, for each $T \in \{1, 2, 3\}$. The bottom panels of Fig. 4, display the boxplots of the Value ratios of the ITRs estimated from the 3 different methods (CSIM, K-LR, and K-SAM) (described in Sect. 5.1), for each combination of $n \in \{250, 500\}$, $p \in \{50, 500\}$, and the $\boldsymbol{X}$ main effect intensity parameter $\delta \in \{1, 2\}$.

The boxplots indicates that the proposed CSIM outperforms all other methods, in all cases. We note that, when $K > 2$, separately estimating $\mathbb{E}[Y \mid \boldsymbol{X}, T = t]$ for each

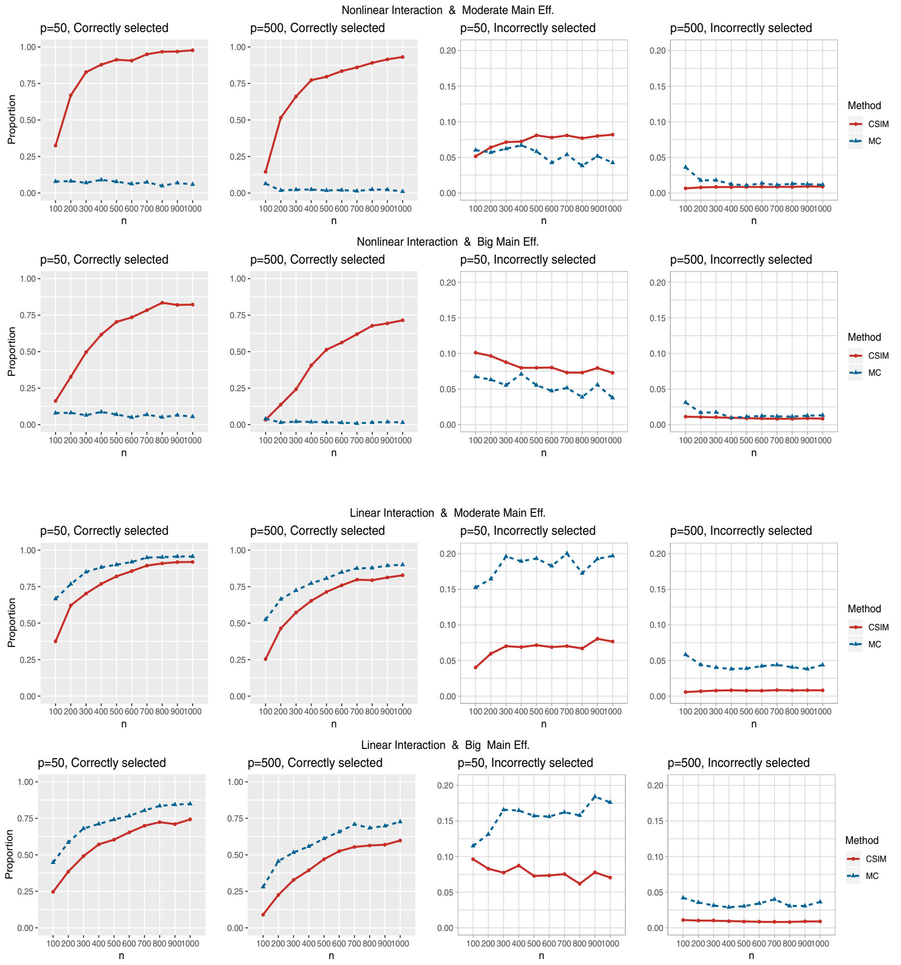

Fig. 3 The proportion of treatment effect modifiers "correctly selected" (in the left two columns), and "incorrectly selected" (in the right two columns), as the training sample size n varies from 100 to 1000, for each $p \in \{50, 500\}$. The rows correspond to the combinations of the levels of the nonlinearity in the $\boldsymbol{X}$-by-T interaction effect (nonlinear/linear interaction effect) and the levels of the intensity in the $\boldsymbol{X}$ main effect (moderate/big main effect). The results indicate that generally a more reliable variable selection is expected as the sample size n increases, under a moderate magnitude of the x main effect (1st and 3rd row) compared to a big magnitude of the X main effect (2nd and 4th row)

group $t \in \{1, \ldots, K\}$ is a typical approach of modeling the $\boldsymbol{X}$-by-T interactions. However, estimating K separate regression models lacks parsimony and interpretability (especially when $K > 2$), whereas the CSIM provides a single parsimonious projection $\boldsymbol{\beta}'\boldsymbol{X}$ to model the interaction. Moreover, for greater $\boldsymbol{X}$ main effect intensity, these K separate regressions tend to focus more on capturing the $\boldsymbol{X}$ main effect and therefore possibly fail to capture important $\boldsymbol{X}$-by-T interaction effects. On the other hand, CSIM targets only at the interaction effect. As a result, in Fig. 4, although the increased magnitude of the $\boldsymbol{X}$ main effect affects the performance of all methods,

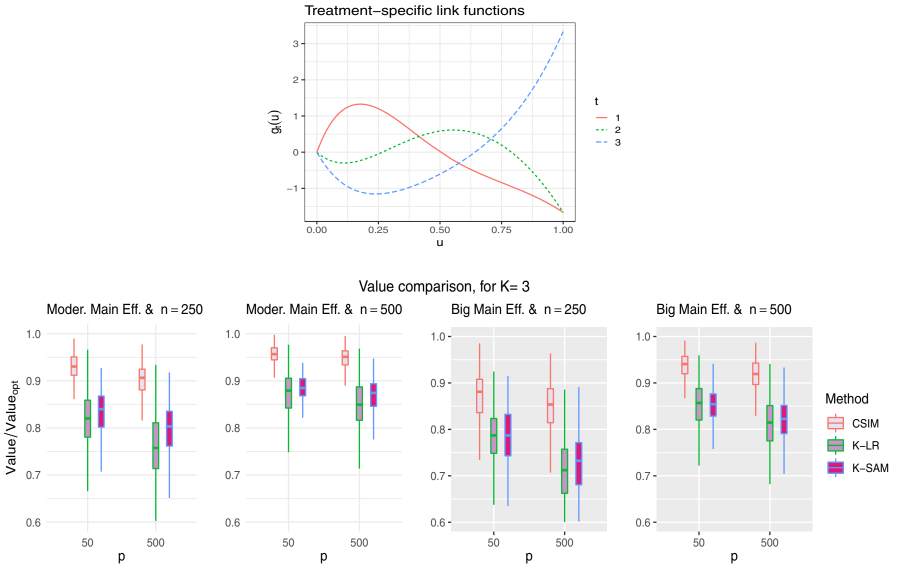

Fig. 4 Upper panel: illustration of $g_t(u)$, $t \in \{1, 2, 3\}$. Lower panels: boxplots of the Value ratios of the ITRs estimated from the 3 methods (CSIM, K-LR, and K-SAM) applied to 200 simulated datasets

it has the least influence on the performance of CSIM, and more influence on that of the K separate regression approaches.

6 Application to data from a depression RCT

The development of the CSIM method was motivated by an RCT that compares an antidepressant (sertraline) ($t = 2$) and placebo ($t = 1$) for treating major depressive disorder (MDD). The primary purpose of the study is the development of a biosignature, called a differential treatment response index (DTRI), defined as a combination of multiple biological markers which permits optimization of treatment selection for patients with MDD (Trivedi et al. 2016). In MDD, most patient characteristics have weak to non-existent modifying effects, and therefore, the proposed sparse CSIM modeling approach of parsimoniously combining treatment effect-modifiers to define a DTRI ($\boldsymbol{\beta}'\boldsymbol{x}$) that collectively exhibits a stronger, and possibly nonlinear, interaction with the treatment is a clinically significant endeavor.

Of the 166 subjects, 88 were randomized to placebo and 78 to drug. Several clinical characteristics were collected at baseline, including: (i) age; (ii) severity of depressive symptoms measured by the Hamilton Rating Scale for Depression (HRSD) at baseline (i.e., week 0); (iii) (the log-transformed) duration of the current major depressive episode (MDE) (in month); and (iv) age of onset of first MDE. In addition to these standard clinical assessments, patients underwent neuropsychiatric testing at baseline. Patients were tested on the following tasks: Flanker and Eriksen (1974), Choice reaction time (CRT; Deary et al. 2011), Word Fluency (WF; Loonstra et al. 2001), A not B working memory (AnotB; Herrera-Guzman et al. 2009), among others. The purpose of these tests is to assess psychomotor slowing, working memory, reaction time (RT) and cognitive control (e.g., post-error recovery), as these behavioral characteristics are believed to correspond to biological phenotypes related to response to antidepressants (Trivedi et al. 2016). Table 1 displays the means and the standard deviations of the $p = 13$ pretreatment patient characteristics, $X_1, \ldots, X_{13}$ under consideration. We center and scale the pretreatment covariates to have zero mean and unit variance. The response variable Y is the improvement in symptoms severity (assessed by the HRSD scores) from baseline to week 8 taken as the difference (week 0–8), and thus larger values of the response are considered desirable.

Table 1 presents the results from fitting the CSIM model. For comparison, the coefficient $\boldsymbol{\beta}$ was also estimated using the MC method Eq. (3.10) both with/without L^1 regularization. Figure 5 gives a graphic depiction of the fitted CSIM model: a plot of the responses (adjusted for the fitted main effect) against the estimated DTRI (i.e., the estimated single-index variable $\hat{\boldsymbol{\beta}}'\boldsymbol{X}$) is given in the upper panel, where $\hat{\boldsymbol{\beta}}$ is obtained by optimizing CSIM Eq. (4.10) with L^1 regularization. The treatment-specific smooth functions over the estimated single-index, fitted based on $d = 6$ cubic B-spline basis functions, are superimposed. As indicated in the third column of Table 1, the CSIM estimate $\hat{\boldsymbol{\beta}}$ has 4 nonzero coefficients, associated with the pretreatment covariates “Age at evaluation”, “Symptom severity”, “(log) Duration of MDE” and “Flanker Accuracy”. In the second row of Fig. 5, we display the marginal plots of the responses against each of these 4 pretreatment covariates individually. In each plot, the estimated treatment-specific smooth functions are overlaid to describe each predictor’s relationship with the response. As can be observed in Fig. 5, the estimated single-index variable $\boldsymbol{\beta}'\boldsymbol{X}$ exhibits a stronger interaction (in the top row), compared to each individual covariate (in the bottom row).

To evaluate the performance of the ITRs estimated from the CSIM and MC methods, we randomly split the data into a training set and a testing set using a ratio of 5 to 1, replicated 500 times, each time fitting the methods on the training set and computing the estimated Value of the ITR by Eq. (5.1) based on the testing set. In addition to CSIM and MC, we include the K separate additive models (K-SAM) and the K separate linear regression (K-LR), described in Sect. 5.1, for comparison. We also include naive decision rules that treat all patients with placebo (all PBO) only or the active drug (all DRUG) only, as well as the CSIM and the MC methods without any simultaneous variable selection (VS) procedure.

In Fig. 6, the proposed CSIM approach demonstrates superior performance in terms of the estimated Values compared to all other alternatives. In particular, the CSIM outperforms the MC and the K separate linear regressions. This suggests that the flexible link functions $g_t(u)$ $(t = 1, 2)$ utilized in CSIM may be better suited for developing ITRs compared to the restricted linear form Eq. (3.9) in this example. Another advantage of the CSIM approach is in its ability to provide a visualization of the estimated DTRI, as depicted in the top panel of Fig. 5. Furthermore, the CSIM approach enables the determination of the relative importance of each pretreatment covariate in characterizing the heterogeneous treatment effect through the coefficients

Table 1 Description of $p = 13$ pretreatment covariates and the L^1-regularized/unregularized coefficient estimates (and 95% bootstrap confidence intervals) of the interaction effect component obtained from the CSIM Eq. (4.10) and the MC Eq. (3.10) methods, respectively, estimated based on standardized covariates (mean zero and unit variance)

Baseline	Mean	CSIM coef. (95% boot. C.I.)		MC coef. (95% boot. C.I.)	
characteristics	(SD)	L^1-regul	Unregul	L^1-regul	Unregul
Age at evaluation	38.00 (13.84)	1	1	1.08 (0.64, 2.48)	1.66 (1.31, 5.56)
Symptom severity	18.8 (4.29)	0.21 (0.14, 0.54)	0.21 (0.00, 0.69)	0.50 (− 0.96, 1.32)	1.14 (0.11, 4.44)
(log) Dur. of MDE	2.84 (1.34)	0.27 (0.17, 0.66)	0.25 (0.10, 0.78)	0 (− 1.02, 0.47)	− 0.16 (− 2.53, 1.67)
Age at MDE onset	16.4 (6.09)	0 (−0.09, 0.39)	− 0.09 (− 0.44, 0.34)	− 0.34 (− 0.85, 0.89)	−1.07 (− 4.23, − 0.08)
AnotB RT (negative)	0.30 (2.13)	0 (0.00, 0.02)	0.01 (− 0.31, 0.26)	0 (0.00, 0.26)	− 1.59 (− 7.94,1.62)
AnotB RT (non-neg.)	0.32 (1.63)	0 (−0.07, 0.07)	− 0.14 (− 0.52, 0.01)	0 (0.00, 0.00)	0.43 (− 6.19, 7.79)
AnotB RT (all)	0.37 (1.77)	0 (− 0.01, 0.00)	− 0.01 (− 0.28, 0.10)	0 (0.00, 0.00)	2.27 (− 4.59, 13.03)
AnotB, total correct	0.16 (0.77)	0 (− 0.15, 0.12)	− 0.01 (− 0.34, 0.24)	0 (− 0.89, 0.98)	1.18 (0.07, 4.66)
Median choice RT	0.23 (1.45)	0 (− 0.11, 0.17)	− 0.10 (− 0.46, 0.19)	0 (− 0.71, 0.64)	− 1.19 (− 4.57, 0.13)
Word fluency	37.42 (11.68)	0 (− 0.04, 0.15)	− 0.09 (-043, 0.11)	0 (− 0.99, 0.97)	− 0.72 (− 3.56, 0.56)
Flanker accuracy	0.22 (0.15)	− 0.55(− 1.11, − 0.41)	− 0.39 (− 0.78, − 0.07)	− 1.70 (− 4.06, − 1.92)	− 2.36 (− 7.27, − 2.20)
Flanker RT	59.51 (26.63)	0 (− 0.22, 0.05)	0.08 (− 0.38, 0.34)	0.39 (− 0.64, 0.86)	0.85 (− 0.56, 4.04)
Post-conflict adjus.	0.07 (0.12)	0 (− 0.07, 0.26)	− 0.04 (− 0.31, 0.23)	0 (− 0.72,1.01)	0.95 (0.01, 3.95)

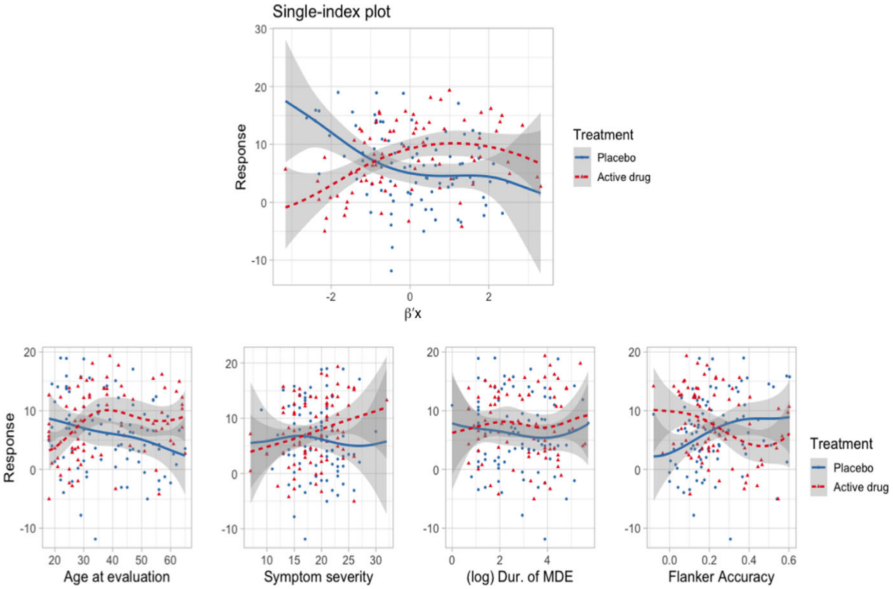

Fig. 5 Top row: A scatterplot of the responses Y_i versus the estimated single-index variable $\boldsymbol{\beta}'\boldsymbol{X}_i$, with the estimated treatment-specific smooth functions; Bottom row: Scatterplots of the response variable against each of the 4 (unscaled) pretreatment covariates, associated with the estimated nonzero coefficients of $\boldsymbol{\beta}$ along with the estimated treatment-specific smooth functions overlaid

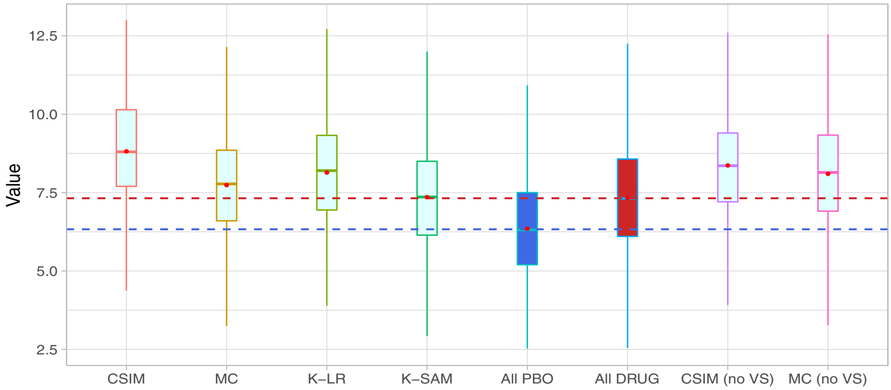

Fig. 6 Boxplots of the estimated Values Eq. (5.1) of the treatment decision rules (ITRs) determined from the 8 different approaches, obtained from 500 randomly split testing sets. Higher Values are preferred

in the estimated $\boldsymbol{\beta}$ vector. These coefficients reflect the contribution and significance of each covariate in capturing the heterogeneous treatment effects.

In addition to considering the relatively small ($p = 13$) set of pretreatment covariates setting, we considered a relative large set of pretreatment covariates case. This depression study (Trivedi et al. 2016; Petkova et al. 2017) additionally collected pretreatment structural magnetic-resonance-imaging (sMRI) measures, acquired using a Siemens (Erlangen, Germany) MAGNETOM Prisma 3T scanner with a 64-channel head coil, along with clinical measures (we refer to Trivedi et al. (2016); Almeida et al. (2018) for acquisition details). Among the $n = 166$ subjects originally considered, $n = 103$ subjects had available baseline sMRI measures. The average cortical thickness, processed using FreeSurfer (Fischl 2012), of the 148 regions (74 regions per hemisphere) defined by the Destrieux Atlas (Destrieux et al. 2010), was considered as additional pretreatment covariates. This yielded $p = 161$ (including the 13 baseline clinical measures in Table 1) for $n = 103$ subjects. CSIM selected “Age at evaluation” and “Flanker Accuracy”, with coefficients 1 and -0.06, respectively. None of the cortical thickness measures were selected as treatment effect-modifiers. MC selected three variables, “Age at evaluation,” “Symptom severity,” and “Flanker Accuracy,” with coefficients 0.92, 0.12 and -2.09, respectively. In Figure A1 in Appendix A7, we report the boxplots of the estimated Values Eq. (5.1) of the ITRs (derived from the 4 methods [CSIM, MC, K-LR, K-SAM] and the two naive rules [All PBO, All DRUG], obtained from 500 randomly split testing sets. The results indicate that CSIM exhibits superior performance in this higher dimensional setting, similar to the results in Fig. 6. The mean (SD) computation times (in seconds on the training sets) for CSIM, MC, K-LR, and K-SAM are 9.01 (3.61), 0.48 (0.05), 0.24 (0.03), and 1.50 (0.16), respectively, with CSIM exhibiting longer computation times but still manageable. CSIM can better capture the nonlinear trend in the association between the active drug response and the covariate “Age at evaluation” (see the second row in Fig. 5) in comparison to the linear model-based approaches (MC and K-LR), while achieving more targeted modeling for the $\boldsymbol{X}$-by-T interaction effect than the K separate additive models (K-SAM), exhibiting superior ITR estimation performance.

7 Discussion

Dimension reduction plays a crucial role in various statistical methods, particularly in regression analysis. It is closely linked to the concept of sufficiency, which explains the relationship between a response variable and a set of predictors through a lower-dimensional subspace in the predictor space. This paper focused on a dimension reduction framework tailored to capturing interaction effects between a variable of interest (T) and a set of predictors $\boldsymbol{X}$. Specifically, the analysis concentrated on a single-index approximation with $R(\boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{X} \in \mathbb{R}$ for these subspaces, as this single index approximation is related to linear model-based methods. Future work will focus on generalizations with more general multiple-indices, $R(\boldsymbol{X}) = \boldsymbol{B}'\boldsymbol{X} \in \mathbb{R}^q$ with $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_q$ for subspaces sufficient for modeling $\boldsymbol{X}$-by-T interaction effects, while avoiding the need to specify the $\boldsymbol{X}$ main effect, as in the framework Eq. (4.1).

In this paper, we focused on the context of an RCT where the treatment $T_i \in \mathcal{T}$ is randomized independently of pretreatment characteristics $\boldsymbol{X}_i$. However, the method can be potentially extended to the case where the treatment assignment depends on $\boldsymbol{X}_i$. To estimate individual treatment effects with observational or non-fully randomized data, we can take a “propensity method” (see, e.g., Imbens and Rubin 2015; Caron et al. 2022) upon taking an appropriate reparametrization of the proposed model, which we describe below for a more general context in which the treatment T_i takes a value $t \in \{1, \ldots, K\}$ with probability (i.e., propensity score) $P(T_i = t|\boldsymbol{X}_i) = \pi_t(\boldsymbol{X}_i)$ $(t = 1, \ldots, K)$. Let $t = 1$ be the reference (control) treatment. For each fixed $\boldsymbol{\beta}$, the constraint in Eq. (4.1), i.e., $\mathbb{E}_{T_i}[g(\boldsymbol{\beta}'\boldsymbol{X}_i)|\boldsymbol{X}_i] = \sum_{t=1}^{K} g_t(\boldsymbol{\beta}'\boldsymbol{X}_i)\pi_t(\boldsymbol{X}_i) = 0$, is equivalent to the equality, $g_1(\boldsymbol{\beta}'\boldsymbol{X}_i) = -\sum_{t=2}^{K} g_t(\boldsymbol{\beta}'\boldsymbol{X}_i)\frac{\pi_t(\boldsymbol{X}_i)}{\pi_1(\boldsymbol{X}_i)}$. Directly incorporating this constraint into the modeling component in Eq. (4.1), i.e., $g_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i) = \sum_{t=1}^{K} \mathbb{I}_{(T_i=t)} g_t(\boldsymbol{\beta}'\boldsymbol{X}_i)$, we can reparametrize $g_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i) = \sum_{t=2}^{K} g_t(\boldsymbol{\beta}'\boldsymbol{X}_i) w_t(T_i, \boldsymbol{X}_i)$, in terms of the $K - 1$ unconstrained functions $g_t(\cdot)$ $(t = 2, \ldots, K)$, where $w_t(T_i, \boldsymbol{X}_i) = \mathbb{I}_{(T_i=t)} - \frac{\pi_t(\boldsymbol{X}_i)}{\pi_1(\boldsymbol{X}_i)}\mathbb{I}_{(T_i=1)}$. The propensity score $\pi_t(\boldsymbol{X}_i)$ is incorporated through the subject i- and treatment t-specific weight $w_t(T_i, \boldsymbol{X}_i)$, and this reparametrized term $g_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i)$ is set to satisfy the constraint in Eq. (4.1), since $\mathbb{E}[w_t(T_i, \boldsymbol{X}_i)|\boldsymbol{X}_i] = 0$, indicating that $\mathbb{E}[g_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i)|\boldsymbol{X}_i] = \mathbb{E}[\sum_{t=2}^{K} g_t(\boldsymbol{\beta}'\boldsymbol{X}_i) w_t(T_i, \boldsymbol{X}_i)|\boldsymbol{X}_i] = \sum_{t=2}^{K} g_t(\boldsymbol{\beta}'\boldsymbol{X}_i)\mathbb{E}[w_t(T_i, \boldsymbol{X}_i)$ $|\boldsymbol{X}_i] = 0$, which ensures the orthogonality against the unspecified term $\mu(\boldsymbol{X}_i)$ in Eq. (2.6), thereby bypassing the need to specify $\mu(\cdot)$ in optimizing $\boldsymbol{\beta}$. In the context of an observational or a non-fully randomized study, we can proceed as follows. For each fixed $\boldsymbol{\beta}$, we can define the design matrix $\tilde{\boldsymbol{D}}_{\boldsymbol{\beta}} = [\tilde{\boldsymbol{D}}_{\boldsymbol{\beta},2}; \ldots; \tilde{\boldsymbol{D}}_{\boldsymbol{\beta},K}] \in \mathbb{R}^{n \times d(K-1)}$, where each t-specific matrix $\tilde{\boldsymbol{D}}_{\boldsymbol{\beta},t} \in \mathbb{R}^{n \times d}$ $(t = 2, \ldots, K)$ is the evaluation matrix of the basis $\psi(\cdot) \in \mathbb{R}^d$ on $\{\boldsymbol{\beta}'\boldsymbol{x}_i\}_{i=1}^n$ (as in Sect. 4.1) but multiplied by the subject i- and treatment level t-specific weight $w_{it} = w_t(t_i, \boldsymbol{x}_i) = \mathbb{I}_{(t_i=t)} - \frac{\pi_t(\boldsymbol{x}_i)}{\pi_1(\boldsymbol{x}_i)}\mathbb{I}_{(t_i=1)}$, so that its ith row corresponds to the $1 \times d$ vector, $w_{it}\psi(\boldsymbol{\beta}'\boldsymbol{x}_i)'$. The spline coefficient vector $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{d(K-1)}$ associated with the design matrix $\tilde{\boldsymbol{D}}_{\boldsymbol{\beta}}$ is then estimated as in Sect. 4.1, which yields the estimate of the component $\hat{g}_{T_i}(\boldsymbol{\beta}'\boldsymbol{X}_i)$ $(i = 1, \ldots, n)$ used in the criterion function $\hat{Q}(\boldsymbol{\beta})$ in Eq. (4.5), and the same sparse estimation procedure as in Sect. 4.4 can be employed to optimize $\boldsymbol{\beta}$.

Future work in refining and developing the proposed approach will investigate the theoretical properties of the estimation procedure. Furthermore, we can consider a generalization of the quadratic loss Eq. (4.1) to $\mathbb{E}\big[Y g_T(\boldsymbol{\beta}'\boldsymbol{X}) - b(g_T(\boldsymbol{\beta}'\boldsymbol{X}))\big]$ subject to the constraint $\mathbb{E}[g_T(\boldsymbol{\beta}'\boldsymbol{X})|X] = 0$, where $b(s) = s^2/2$ for a Gaussian Y (for which the optimization Eq. (4.1) is a special case), $b(s) = \log\{1 + \exp(s)\}$ for a Bernoulli Y, and $b(s) = \exp(s)$ for a Poisson Y, in which the constrained working model corresponds to $h(\mathbb{E}[Y|\boldsymbol{X}, T]) \approx g_T(\boldsymbol{\beta}'\boldsymbol{X})$ with the constraint $\mathbb{E}[g_T(\boldsymbol{\beta}'\boldsymbol{X})|X] = 0$, where $h(\cdot)$ is the canonical link associated with the assumed exponential family distribution. This constraint $\mathbb{E}\big[g_T(\boldsymbol{\beta}'\boldsymbol{X})|X\big] = 0$ will ensure the criterion function, $\mathbb{E}\left[Y g_T(\boldsymbol{\beta}'\boldsymbol{X}) - b(g_T(\boldsymbol{\beta}'\boldsymbol{X}))\right] = \mathbb{E}\left[\left\{\mu^*(\boldsymbol{X}) + g_T^*(\boldsymbol{\beta}^{*\top}\boldsymbol{X})\right\}g_T(\boldsymbol{\beta}'\boldsymbol{X}) - b(g_T(\boldsymbol{\beta}'\boldsymbol{X}))\right]$ $= \mathbb{E}\big[g_T^*(\boldsymbol{\beta}^{*\top}\boldsymbol{X})(g_T(\boldsymbol{\beta}'\boldsymbol{X}) - b(g_T(\boldsymbol{\beta}'\boldsymbol{X}))\big]$, to be free of the unspecified $\boldsymbol{X}$ "main" effect term $\mu^*(\boldsymbol{X})$, and thus, the optimizing component $g_T(\boldsymbol{\beta}'\boldsymbol{X})$ will target at the true "signal" component $g_T^*(\boldsymbol{\beta}^{*\top}\boldsymbol{X})$ that is orthogonal to the nuisance component $\mu^*(\boldsymbol{X})$. However, a thorough analysis is warranted particularly in terms of the theoretical development on the estimation and variable selection consistency.

Of note, directly extending Radchenko (2015)'s theoretical results to our context is not feasible due to (i) treating predictors as deterministic (which is not the case in our setting, where the treatment indicators are nost considered fixed); and (ii) the assumption of correctly specified single-index models. Radchenko (2015)'s theoretical results are based on an empirical process approach to the asymptotics of nonlinear least-squares estimation to obtain the stochastic bound for the regression function estimate. The main challenge in the theoretical development arises from the general misspecification of the working model, $Y \approx g_T(\boldsymbol{\beta}'\boldsymbol{X}) + \epsilon$. The working model is an approximation due to the omission of the unspecified term $\mu^*(\boldsymbol{X})$ present in the underlying model. In the case of such a misspecified working model, establishing the consistency of the estimators requires conducting asymptotic analysis using ideas from semiparametric M-estimation, similar to the approaches in Ichimura and Lee (2010); Wang and Yang (2009), dealing with semiparametric least squares estimation under model misspecification. More specifically, to achieve consistency, one needs to establish the estimation consistency for the function, $g_T^*(\boldsymbol{\beta}'\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T] - \mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]$ in Eq. (4.6), where its component $\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T]$ is defined as the best L^2 *approximation* based on a measurable function of $(\boldsymbol{\beta}'\boldsymbol{X}, T)$ to the response Y, rather than as an exact model given $(\boldsymbol{\beta}'\boldsymbol{X}, T)$. This involves obtaining uniformly consistent (spline) estimators of the conditional expectations $\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}, T]$ and $\mathbb{E}[Y|\boldsymbol{\beta}'\boldsymbol{X}]$ (uniformly over $\boldsymbol{\beta} \in \Theta$) using the idea considered in Wang and Yang (2009). Once these uniformly consistent estimators are obtained, the next step is to establish the consistency of the estimators of the projection directions $\boldsymbol{\beta}$, given the other model components, i.e., g_t $(t \in \{1, \ldots, K\})$. However, addressing the high dimensionality of the coefficient vector $\boldsymbol{\beta}$ adds significant complexity to the already challenging problem of developing a semiparametric estimation theory under model misspecification for the constrained estimation framework in Eq. (4.1). This will entail a significant amount of additional work, and we leave the theoretical investigation of the estimation consistency for the model components as future work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-024-01546-0.

Acknowledgements This work was supported by National Institute of Health (NIH) grant 5 R01 MH099003. The first author thanks Dr. Peter Radchenko of the University of Sydney for his comments and providing the code to implement his regression model.

References

Adragni KP, Cook DR (2009) Sufficient dimension reduction and prediction in regression. Philos Trans Royal Soc 367:4385–4405

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Almeida J, Greenberg T, Lu H, Chase H, Fournier J, Cooper C, Deckersbach T, Adams P, Carmody T, Fava M, Kurian B, McGrath P, McInnis M, Oquendo M, Parsey R, Weissman M, Trivedi M, Phillips M (2018) Est-retest reliability of cerebral blood flow in healthy individuals using arterial spin labeling: findings from the EMBARC study. Magn Reson Med 45:26–33

Bura E, Cook RD (2001) Estimating the structural dimension of regression via parametric inverse regression. J Royal Stat Soc Ser B 63:1–10

Cai T, Tian L, Wong PH, Wei LJ (2011) Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics 12:270–282

Caron A, Baio G, Manolopoulou I (2022) Estimating individual treatment effects using non-parametric regression models: a review. J Royal Stat Soc Ser A 185:1115–1149

Carroll R, Fan J, Gijbels I, Wand M (1997) Generalized partially linear single-index models. J Am Stat Assoc 1997:10

Cohen MX (2022) A tutorial on generalized eigendecomposition for denoising, contrast enhancement, and dimension reduction in multichannel electrophysiology. NeuroImage 2022:118809

Cook RD (1994) On the interpretation of regression plots. J Am Stat Assoc 89:177–189

Cook RD (1996) Graphics for regressions with a binary response. J Am Stat Assoc 91:983–992

Cook DR (1998) Regression graphics. Wiley, New York

Cook RD (2007) Fisher lecture: dimension reduction in regression. Stat Sci 22:1–26

Cook DR, Li B (2002) Dimension reduction for conditional mean in regression. Ann Stat 30:455–474

Dahne S, Meinecke FC, Haufe S, Hohne J, Tangermann M, Muller KR, Nikulin VV (2014) Spoc: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. Neuroimage 86:111–122

de Cheveigne A, Parra LC (2014) Joint decorrelation, a versatile tool for multichannel data analysis. Neuroimage 98:487–505

Deary IJ, Liewald D, Nissan J (2011) A free, easy-to-use, computer-based simple and four-choice reaction time programme: the deary-liewald reaction time task. Behav Res Methods 43:258–268

Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53:1–15

Eilers P, Marx B (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11:89–121

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Fischl B (2012) Freesurfer. Neuroimage 62:774–781

Flanker BA, Eriksen CW (1974) Effects of noise letters upon identification of a target letter in a non-search task. Percept Psychophys 16:143–149

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22

Herrera-Guzman I, Guidayol-Ferre E, Herrera-Guzman D, Guardia-Olmos J, Hinojosa-Calvo E, Herrera-Abarca JE (2009) Effects of selective serotonin reuptake and dual serotonergic-noradrenergic reuptake treatments on memory and mental processing speed in patients with major depressive disorder. Psyc Res 43:855–863

Hurvich C, Tsai C (1989) Regression and time series model selection in small samples. Biometrika 76:297–307

Ichimura H, Lee S (2010) Characterization of the asymptotic distribution of semiparametric m-estimators. J Econ 159:252–266

Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, Cambridge

Jeng X, Lu W, Peng H (2018) High-dimensional inference for personalized treatment decision. Electron J Stat 12:2074–2089

Li KC (1991) Sliced inverse regression for dimension reduction (with discussion). J Am Stat Assoc 86:316–342

Li KC (1992) On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. J Am Stat Assoc 87:1025–1039

Liu C, Zhao X, Huang J (2023) A random projection approach to hypothesis tests in high-dimensional single-index models. J Am Stat Assoc. https://doi.org/10.1080/01621459.2022.2156350

Loonstra A, Tarlow AR, Sellers AH (2001) Cowat metanorms across age, education, and gender. Appl Neuropsychol 8:161–166

Lu W, Zhang H, Zeng D (2011) Variable selection for optimal treatment decision. Stat Methods Med Res 22:493–504

Luo W, Zhu Y, Ghosh D (2017) On estimating regression-based causal effects using sufficient dimension reduction. Biometrika 104:51–65

Luo W, Wu W, Zhu Y (2018) Learning heterogeneity in causal inference using sufficient dimension reduction. J Causal Inference 7:10

Ma Y, Zhu L (2012) A semiparametric approach to dimension reduction. J Am Stat Assoc 107:168–179

Ma Y, Zhu L (2013) Efficient estimation in sufficient dimension reduction. Ann Stat 41:250–268

Meinshausen N, Yu B (2009) Lasso-type recoerty of sparse representation for high-dimensional data. Ann Stat 37:246–270

Murphy SA (2003) Optimal dynamic treatment regimes. J Royal Stat Soc Ser B (Stat Methodol) 65:331–355

Murphy SA (2005) A generalization error for q-learning. J Mach Learn 6:1073–1097

Park H, Petkova E, Tarpey T, Ogden RT (2021) A constrained single-index regression for estimating interactions between a treatment and covariates. Biometrics 77:506–518

Peng H, Huang T (2011) Penalized least squares for single index models. J Stat Plan Inference 141:1362–1379

Petkova E, Tarpey T, Su Z, Ogden RT (2016) Generated effect modifiers in randomized clinical trials. Biostatistics 18:105–118

Petkova E, Ogden R, Tarpey T, Ciarleglio A, Jiang B, Su Z, Carmody T, Adams P, Kraemer H, Grannemann B, Oquendo M, Parsey R, Weissman M, McGrath P, Fava M, Trivedi M (2017) Statistical analysis plan for stage 1 EMBARC (establishing moderators and biosignatures of antidepressant response for clinical care) study. Contemp Clin Trials Commun 6:22–30

Poon W, Wang H (2013) Bayesian analysis of generalized partially linear single-index models. Comput Stat Data Anal 68:251–261

Qian M, Murphy SA (2011) Performance guarantees for individualized treatment rules. Ann Stat 39:1180–1210

Radchenko P (2015) High dimensional single index models. J Multivar Anal 139:266–282

Ravikumar P, Lafferty J, Liu H, Wasserman L (2009) Sparse additive models. J Royal Stat Soc Ser B 71:1009–1030

Robins J (2004) Optimal structural nested models for optimal sequential decisions. Springer, New York

Rubin D (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701

Schneider U, Tardivel P (2022) The geometry of uniqueness, sparsity and clustering in penalized estimation. J Mach Learn Res 23:1–36

Shi C, Song R, Lu W (2016) Robust learning for optimal treatment decision with np-dimensionality. Electron J Stat 10:2894–2921

Shi C, Fan A, Song R, Lu W (2018) High-dimensional A-learning for optimal dynamic treatment regimes. Ann Stat 46:925–957

Stoker TM (1986) Consistent estimation of scaled coefficients. Econometrica 54:1461–1481

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat Theor Methods 7:13–26

Tian L, Alizadeh A, Gentles A, Tibshrani R (2014) A simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc 109:1517–1532

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B (Stat Methodol) 58:267–288

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

Wang G, Wang L (2015) Spline estimation and variable selection for single-index prediction models with diverging number of index parameters. J Stat Plan Inference 162:1–19

Wang L, Yang L (2009) Spline estimation of single-index models. Stat Sin 19:765–783

Wang Q, Yin X (2008) A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. Comput Stat Data Anal 52:4512–4512

Xia Y, Tong H, Li W (1999) On extended partially linear single-index models. Biometrika 86:831–842

Xia Y, Tong H, Li W, Zhu L (2002) An adaptive estimation of dimension reduction space. J Royal Stat Soc Ser B (Stat Methodol) 64:363–410

Yin X, Li B, Cook DR (2008) Successive direction extraction for estimating the central subspace in a multiple-index regression. J Multivar Anal 99:1733–1757

Zhang B, Tsiatis AA, Laber EB, Davidian M (2012) A robust method for estimating optimal treatment regimes. Biometrics 68:1010–1018

Zhao T, Li X, Liu H, Roeder K (2014) SAM: Sparse additive modelling. R Package Vers 1:5

Zhu L, Qian L, Lin J (2011) Variable selection in a class of single-index models. Ann Inst Stat Math 63:1277–1293

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.