Robust inference of conditional average treatment effects using dimension reduction

Ming-Yueh Huang

Institute of Statistical Science, Academia Sinica

Shu Yang

Department of Statistics, North Carolina State University

Abstract:

Personalized treatment aims at tailoring treatments to individual characteristics. An important step is to understand how treatment effect varies across individual characteristics, known as the conditional average treatment effect (CATE). This article concerns making robust inferences of the CATE from observational data, which becomes challenging with multivariate confounder. To reduce the curse of dimensionality while keeping the nonparametric merit, we propose double dimension reductions that achieve different goals: first, we target at identifying the central mean subspace of the CATE directly using dimension reduction in order to detect the most accurate and parsimonious structure of the CATE; second, a nonparametric regression with prior dimension reduction is used to impute counterfactual outcomes, which helps to improve the stability of the imputation. We establish the asymptotic properties of the proposed estimator taking into account the two-step double dimension reduction, and propose an effective bootstrapping procedure without bootstrapping the estimated central mean subspace to make valid inferences. Simulation and applications show that the proposed estimator outperforms existing competitors.

Key words and phrases: augmented inverse probability weighting; matching; kernel smoothing; U-statistic; weighted bootstrap.

1. Introduction

Because of patient heterogeneity in response to various aspects of treatment, the paradigm of biomedical and health policy research is shifting from the "one-size-fits-all" treatment approach to precision medicine (Hamburg and Collins, 2010). Toward that end, an important step is to understand how treatment effect varies across patient characteristics, known as the conditional average treatment effect (CATE) (Rothwell, 2005). A large body of literature focuses on modeling the treatment-specific prognostic score (e.g., Chakraborty et al., 2010; Zhao et al., 2011; Song et al., 2017), since the CATE is simply the difference between the treated and control prognostic scores. However, modeling prognostic scores may lead to an overfitting problem for the CATE, and direct modeling of the CATE may provide a more accurate characterization of treatment effects, avoiding redundancy of non-useful features; see Section 2.2. Another body of literature thus focuses on modeling and approximating the CATE parametrically (Murphy, 2003; Robins, 2004), semiparametrically (Liang and Yu, 2020) and by machine learning methods (Zhao et al., 2012; Zhang et al., 2012; Rzepakowski and Jaroszewicz, 2012; Athey and Imbens, 2016; Athey et al., 2019; Künzel et al., 2019). However, parametric and semiparametric methods are susceptible to model misspecification, and machine learning produces results that are too complicated to be interpretable. Most importantly, it is a daunting task to draw valid inferences based on machine learning methods.

In this article, we propose a nonparametric framework to make robust inferences of the CATE with multivariate confounder. To mitigate the possible curse of dimensionality, we consider the central mean subspace of the CATE, which is the smallest linear subspace spanned by a set of linear index that can sufficiently characterize the estimand of interest (Cook and Li, 2002). Under this framework, we specify the CATE nonparametrically and use a model selection procedure to determine a sufficient structural dimension. Directly targeting the central mean subspace of the CATE enables us to detect the most accurate and parsimonious structure of the CATE. However, existing dimension reduction methods are not applicable due to the fundamental problem in causal inference that not all potential outcomes are observable. To estimate the central mean subspace, we propose imputing counterfactual outcomes by kernel regression with a prior dimension reduction. The prior dimension reduction helps to improve the stability of the imputation and the subsequent estimation of the CATE. In our simulation studies, the proposed imputation method outperforms existing methods such as the nearest neighbor imputation, the inverse probability weighted adjusted outcomes (Abrevaya et al., 2015), and the augmented inverse probability weighting (Zhao et al., 2012).

Theoretically, we derive the consistency and asymptotic normality of the proposed estimator of the CATE. The main challenge is that the imputed counterfactual outcomes are not independent. To overcome this challenge, we calculate the difference between imputed and conditional counterfactual outcomes, which can be expressed as a weighted empirical average of the influence functions of the kernel regression estimator. Thus, we can show that the influence function of the proposed estimator can be approximated by a U-statistic. Invoking the properties of degenerate U-processes discussed in Nolan and Pollard (1987), we can derive the asymptotic distribution of the estimated CATE and show that the imputation step plays a non-negligible role. To make valid inference, we propose an under-smooth strategy such that the asymptotic bias is dominated by the asymptotic variance. We estimate the asymptotic variances by applying weighted bootstrap techniques and construct Wald-type confidence intervals. Interestingly, the fact that the central mean subspace is estimated does not affect the asymptotic distribution of the proposed estimator of the CATE. Thus, in our bootstrap procedure, we can safely skip the step of bootstrapping the estimated central mean subspace, which saves a lot of computation time in practice.

The remaining of the article is organized as follows. Section 2 establishes the proposed robust inference framework and the asymptotic properties. In Section 3, we conduct simulation studies to assess the finite-sample performance of the proposed inference procedure in comparison with existing competitors. In Section 4, we apply the proposed method to estimate the CATE of maternal smoking on birth weight based on two datasets. We conclude the article in Section 5 with discussions.

2. Methodology

2.1 Preliminaries

Let $X \in \mathcal{X} \subseteq \mathbb{R}^p$ be a vector of pre-treatment covariates, $A \in \mathcal{A} = \{0, 1\}$ the binary treatment, and $Y \in \mathbb{R}$ the outcome of interest. Under the potential outcomes framework (Rubin, 1974), let $Y(a)$ denote the potential outcome had the individual received treatment $a \in \mathcal{A}$. Based on the potential outcomes, the individual causal effect is $D = Y(1) - Y(0)$, and the CATE is $\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\} = \mathbb{E}(D \mid X = x)$. To link the potential outcomes with the observed outcome, we make the usual causal consistency assumption that $Y = Y(A) = AY(1) + (1-A)Y(0)$. The main goal of this article is to estimate $\tau(x)$ based on observational data $\{(A_i, Y_i, X_i) : i = 1, \ldots, n\}$, which independently and identically follow $f(A, Y, X)$.

To identify the treatment effects based on observational data, we assume the following assumptions, which are standard in causal inference with observational studies (Rosenbaum and Rubin, 1983):

Assumption 1. $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$.

Assumption 2. There exist constants c_1 and c_2 such that $0 < c_1 \leq \pi(X) \leq c_2 < 1$ almost surely, where $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$ is the propensity score.

Assumption 1 rules out latent confounding between the treatment assignment and outcome. This can be made plausible by collecting detailed information on characteristics of the units that are related to treatment assignment and outcome. Assumption 2 implies a sufficient overlap of the covariate distribution between the treatment groups. If this assumption is violated, a common approach is to trim the sample; see Yang and Ding (2018).

Let $\mu_a(x) = \mathbb{E}\{Y(a) \mid X = x\}$ $(a = 0, 1)$. Under Assumptions 1–2, $\mu_a(x) = \mathbb{E}(Y \mid A = a, X = x)$ and $\tau(x) = \mu_1(x) - \mu_0(x)$ are identifiable from $f(A, Y, X)$. This identification formula motivates a common strategy of estimating $\tau(x)$ by approximating $\mu_a(X)$ separately for $a = 0, 1$. However, this may lead to an overfitting model for $\tau(x)$, as we will discuss in the next subsection. Alternatively, we propose robust inference of $\tau(x)$ directly using dimension reduction, which requires no parametric model assumptions and can detect accurate and parsimonious structures of $\tau(x)$.

2.2 Dimension reduction on CATE

The main idea is to search for the fewest linear indices $B_\tau^{\mathrm{T}} x$ such that

$$\tau(x) = g(B_\tau^{\mathrm{T}} x), \tag{2.1}$$

where B_τ is a $p \times d_\tau$ matrix consisting of index coefficients, and g is an unknown d_τ-variate function. Since $\tau(x) = \mathbb{E}(D \mid X = x)$, the column space of B_τ is called the central mean subspace of D given X, denoted by $\mathcal{S}_{\mathbb{E}(D|X)}$ (Cook and Li, 2002).

The central mean subspace $\mathcal{S}_{\mathbb{E}(D|X)}$ is nonparametric. In other words, for any multivariate function $\tau(x)$, without particular parametric or semiparametric modeling, there always exists a central mean subspace. To illustrate, consider the single-index model $g(x^{\mathrm{T}}\beta)$ which leads to a one-dimensional central mean subspace spanned by β. Unlike the single-index model that prefixes the dimension of the central mean subspace, we leave both d_τ and B_τ unspecified, and the primary goal of dimension reduction is to estimate d_τ and B_τ. In addition, the curse of dimensionality can be avoided if d_τ is much smaller than p.

Remark 1. Recall that $\tau(x) = \mu_1(x) - \mu_0(x)$. An alternative way to employ dimension reduction is to search for two sets of linear indices $B_0^{\mathrm{T}}x$ and $B_1^{\mathrm{T}}x$ such that

$$\mu_0(x) = g_0(B_0^{\mathrm{T}}x), \qquad \mu_1(x) = g_1(B_1^{\mathrm{T}}x), \tag{2.2}$$

where g_0 and g_1 are unknown functions. That is, we can also estimate $\mathcal{S}_{\mathbb{E}\{Y(0)|X\}} = \mathrm{span}(B_0)$ and $\mathcal{S}_{\mathbb{E}\{Y(1)|X\}} = \mathrm{span}(B_1)$, and then recover $\tau(x)$ by $g_1(B_1^{\mathrm{T}}x) - g_0(B_0^{\mathrm{T}}x)$. In fact, we can show that $\mathcal{S}_{\mathbb{E}(D|X)} \subseteq \mathcal{S}_{\mathbb{E}\{Y(0)|X\}} + \mathcal{S}_{\mathbb{E}\{Y(1)|X\}}$, where the sum of two linear subspaces is $U + V = \{u + v : u \in U, v \in V\}$. In some cases $\mathcal{S}_{\mathbb{E}(D|X)}$ may be different from $\mathcal{S}_{\mathbb{E}\{Y(0)|X\}}$ and $\mathcal{S}_{\mathbb{E}\{Y(1)|X\}}$ or have a strictly smaller dimension than $\mathcal{S}_{\mathbb{E}\{Y(0)|X\}}$ and $\mathcal{S}_{\mathbb{E}\{Y(1)|X\}}$, as demonstrated by the following examples. Thus, using model (2.1) may detect more parsimonious structures of $\tau(x)$ than using model (2.2).

Example 1. Let $Y(0) = \alpha^{\mathrm{T}}X$ and $Y(1) = \beta^{\mathrm{T}}X$, where $\alpha, \beta \in \mathbb{R}^p$, and α and β are linearly independent. Then, $\tau(x) = (\beta - \alpha)^{\mathrm{T}}X$. Thus, $\mathcal{S}_{\mathbb{E}(D|X)} = \mathrm{span}(\beta - \alpha)$, which is different from $\mathcal{S}_{\mathbb{E}\{Y(0)|X\}} = \mathrm{span}(\alpha)$ and $\mathcal{S}_{\mathbb{E}\{Y(1)|X\}} = \mathrm{span}(\beta)$. Thus, nonparametric dimension reduction for $\mu_0(x)$ and $\mu_1(x)$ can detect two directions α and β separately but can not detect the central mean subspace of the CATE function.

Example 2. Let $Y(0) = \alpha^{\mathrm{T}}X + (\beta^{\mathrm{T}}X)^2$ and $Y(1) = \alpha^{\mathrm{T}}X + (\beta^{\mathrm{T}}X)^3$, where $\alpha, \beta \in \mathbb{R}^p$, and α and β are linearly independent. Then, $\tau(x) = (\beta^{\mathrm{T}}X)^3 - (\beta^{\mathrm{T}}X)^2$. Thus, $\dim(\mathcal{S}_{\mathbb{E}\{Y(0)|X\}}) = \dim(\mathcal{S}_{\mathbb{E}\{Y(1)|X\}}) = \dim\{\mathrm{span}(\alpha, \beta)\} = 2$, while $\dim(\mathcal{S}_{\mathbb{E}(D|X)}) = \dim\{\mathrm{span}(\beta)\} = 1$. In this example, detecting the smaller dimension of $\mathcal{S}_{\mathbb{E}(D|X)}$ can help estimate $\tau(x)$ with an only one-dimensional nonparametric smoothing estimator. If we recover $\tau(x)$ by estimating $\mu_1(x) - \mu_0(x)$, two-dimensional nonparametric smoothing estimators for $\mu_1(x)$ and $\mu_0(x)$ are required, and hence are more unstable in finite sample.

Remark 2. As discussed in Ma and Zhu (2013), the parameter B is not identifiable without further restrictions. To see this, suppose that Q is an invertible $d \times d$ matrix and consider $g^*(u) = g\{(Q^{\mathrm{T}})^{-1}u\}$. Then we can derive another equivalent representation of $\tau(x)$ as

$$\tau(x) = g(B^{\mathrm{T}}x) = g\{(Q^{\mathrm{T}})^{-1}Q^{\mathrm{T}}B^{\mathrm{T}}x\} = g^*\{(BQ)^{\mathrm{T}}x\}.$$

Thus, the two sets of parameters (B, g) and (BQ, g^*) correspond to the same CATE. As a result, the central subspace was introduced to make the column space invariant to these invertible linear transformations. We use a particular parametrization of the central mean subspace as used in Ma and Zhu (2013). Without loss of generality, we set the upper $d \times d$ block of B to be the identity matrix $I_{d\times d}$ and write $X = (X_u^{\mathrm{T}}, X_l^{\mathrm{T}})^{\mathrm{T}}$, where $X_u \in \mathbb{R}^d$ and $X_l \in \mathbb{R}^{p-d}$. Hence, the free parameters are the lower $(p-d) \times d$ entries of B, corresponding to the coefficients of X_l. For generic matrix B, we now denote $\mathrm{vecl}(B)$ as the vector formed by the lower $(p-d) \times d$ entries of B.

2.3 Imputation and Estimation

If D were known, existing methods can be directly applied to estimate $\mathcal{S}_{\mathbb{E}(D|X)}$. However, the fundamental problem in causal inference is that the two potential outcomes can never be jointly observed for each unit, one is factual $Y(A)$ and the other one is counterfactual $Y(1-A)$. To overcome this challenge, we propose an imputation step to impute the counterfactual outcomes. A natural choice to impute $Y(1-A)$ is using its conditional mean given X, $\mu_{1-A}(X)$. As mentioned in Section 2.1, $\mu_a(x)$ can be estimated by matching or other nonparametric smoothing techniques. To further reduce the possible curse of dimensionality, we propose a prior dimension reduction procedure to estimate $\mu_a(x)$.

The proposed imputation and estimation procedure proceeds as follows.

Step 1. Estimate the central mean subspace $\mathcal{S}_{\mathbb{E}\{Y(a)|X\}}$ $(a=0,1)$. Let $\mu_a(u;B)=\mathbb{E}(Y \mid A=a, B^{\mathrm{T}}X=u)$, where B is a $p\times d$ parameter matrix. Given B, the kernel smoothing estimator of $\mu_a(u;B)$ is

$$\widehat{\mu}_a(u;B)=\frac{\sum_{j=1}^n Y_j 1(A_j=a)\mathcal{K}_{q,h}(B^{\mathrm{T}}X_j-u)}{\sum_{j=1}^n 1(A_j=a)\mathcal{K}_{q,h}(B^{\mathrm{T}}X_j-u)}, \tag{2.3}$$

where $1(\cdot)$ is the indicator function, $\mathcal{K}_{q,h}(u)=\prod_{k=1}^d K_q(u_k/h)/h$ with $u=(u_1,\ldots,u_d)$, K_q is a qth ordered and twice continuously differentiable kernel function with bounded support, and h is a positive bandwidth. The basis matrix of $\mathcal{S}_{\mathbb{E}\{Y(a)|X\}}$ can be estimated by $\widehat{B}_a$, where $(\widehat{d}_a,\widehat{B}_a,\widehat{h}_a)$ is the minimizer of the cross-validation criterion

$$\mathrm{CV}_a(d,B,h)=\sum_{i=1}^n\{Y_i-\widehat{\mu}_a^{-i}(B^{\mathrm{T}}X_i;B)\}^2 1(A_i=a), \tag{2.4}$$

where the superscript $-i$ indicates the estimator (2.3) based on data without the ith subject. The order of the kernel function $q>\max(d/2+1,2)$ is specified for each working dimension d. This criterion (2.4) is a mean regression version of Huang and Chiang (2017), and more details as well as computation algorithms can be found therein.

Step 2. Impute the individual treatment effect by

$$\widehat{D}_i=A_i\{Y_i-\widehat{\mu}_0(\widehat{B}_0^{\mathrm{T}}X_i;\widehat{B}_0)\}+(1-A_i)\{\widehat{\mu}_1(\widehat{B}_1^{\mathrm{T}}X_i;\widehat{B}_1)-Y_i\}\ \ (i=1,\ldots,n)$$

with specified orders (q_0, q_1) of kernel functions and bandwidths (h_0, h_1) in $\widehat{\mu}_0(\widehat{B}_0^{\mathrm{T}} X_i; \widehat{B}_0)\}$ and $\widehat{\mu}_1(\widehat{B}_1^{\mathrm{T}} X_i; \widehat{B}_1)$. The choices of q_0 and q_1 will be discussed in § 2.4. The bandwidths can be chosen as estimated optimal bandwidths by nonparametric smoothing methods, such that $h_a = O_{\mathbb{P}}\{n^{-1/(2q_a+d_a)}\}$, where $d_a = \dim(\mathcal{S}_{\mathbb{E}\{Y(a)|X\}})$ $(a = 0, 1)$.

Step 3. Estimate the central mean subspace $\mathcal{S}_{\mathbb{E}(D|X)}$ based on $\{(\widehat{D}_i, X_i) : i = 1, \ldots, n\}$. Let $\tau(u; B) = \mathbb{E}\{Y(1) - Y(0) \mid B^{\mathrm{T}} X = u\}$. Given B, the kernel smoothing estimator of $\tau(u; B)$ is

$$\widehat{\tau}(u; B) = \frac{\sum_{j=1}^{n} \widehat{D}_j \mathcal{K}_{q,h}(B^{\mathrm{T}} X_j - u)}{\sum_{j=1}^{n} \mathcal{K}_{q,h}(B^{\mathrm{T}} X_j - u)}. \tag{2.5}$$

We then estimate (d_τ, B_τ) and a suitable bandwidth for $\widehat{\tau}(u; B)$ by the minimizer $(\widehat{d}, \widehat{B}, \widehat{h})$ of the following criterion:

$$\mathrm{CV}(d, B, h) = n^{-1} \sum_{i=1}^{n} \{\widehat{D}_i - \widehat{\tau}^{-i}(B^{\mathrm{T}} X_i; B)\}^2,$$

where the superscript $-i$ indicates the estimator (2.5) based on data without the ith subject. Here, $q > \max(d/2 + 1, 2)$ is also specified for each working dimension d.

Step 4. Estimate $\tau(x)$ by $\widehat{\tau}(\widehat{B}^{\mathrm{T}} x; \widehat{B})$ with some suitable choice of (q_τ, h_τ), which will be further discussed in Section 2.4.

Remark 3. There have been a lot of existing dimension reduction methods in the literature that can be applied in Steps 1 and 3. Representative approaches include the inverse regression (Li, 1991; Li and Wang, 2007; Zhu et al., 2010), average derivative methods (Xia et al., 2002; Zhu and Zeng, 2006; Xia, 2007; Wang and Xia, 2008; Yin and Li, 2011), and the semiparametric approach (Ma and Zhu, 2012, 2013). Different from these methods, the cross-validation criterion of Huang and Chiang (2017) can estimate the structural dimension, the basis matrix, and an optimal bandwidth for the link function simultaneously. In particular, all the parameters are estimated in a data-driven way and no ad-hoc tuning is required. As for the computational burden, the leave-one-out cross-validation is applied for the unknown link functions but not for the index coefficients, and hence we do not remove each subject and repeatedly calculate the criterion. Instead, we simply calculate the kernel weights $\mathcal{K}_{q,h}(B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i)$ $(i, j = 1, \ldots, n)$ and then remove the diagonal weights $\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}X_i)$ $(i = 1, \ldots, n)$ to form the link function estimates. Thus, for each fixed B, the computation of the proposed criterion only involves a kernel weight matrix of size $n \times n$ as commonly seen in nonparametric smoothing methods and is feasible in practice. Due to these properties, we adopt this method in our estimation procedure.

Remark 4. Liang and Yu (2020) considered the multiple index model with a fixed dimension of the index and proposed the semiparametric efficient score of B_τ, while our proposed estimator $\widehat{B}$ may not reach the semiparametric efficiency bound. However, as we will show in Theorem 1, the asymptotic distribution of $\widehat{B}$ does not affect the asymptotic distribution of the estimated CATE as long as $\widehat{B}$ is root-n consistent. Therefore, it is not necessary to pursue the semiparametric efficiency estimation of the central mean subspace in our context.

Remark 5. An alternative method of imputing the counterfactual outcomes is matching (Yang and Kim, 2019, 2020). To fix ideas, we consider matching without replacement and with the number of matches fixed at one. Then the matching procedure becomes nearest neighbor imputation (Little and Rubin, 2002). Without loss of generality, we use the Euclidean distance to determine neighbors; the discussion applies to other distances (Abadie and Imbens, 2006). Let $\mathcal{J}_i$ be the index set for the matched subject of ith subject. Define the imputed missing outcome as $\widetilde{Y}_i(A_i) = Y_i$ and $\widetilde{Y}_i(1-A_i) = \sum_{j\in\mathcal{J}_i} Y_j$. Then the individual causal effect can be estimated by $\widehat{D}_{\mathrm{MAT},i} = \widetilde{Y}_i(1) - \widetilde{Y}_i(0)$. Matching uses the full vector of confounders to determine the distance and corresponding neighbors. When the number of confounders gets larger, this distance may be too conservative to determine proper neighbors due to the curse of dimensionality. In the simulation studies, we find that the estimation of $\mathcal{S}_{\mathbb{E}(D|X)}$ based on $\widehat{D}_{\mathrm{MAT},i}$ has poorer performance compared with our proposed method.

Remark 6. Instead of imputing the counterfactual outcomes, weighting can also be used to estimate D_i directly. Several authors have considered an adjusted outcome $\widehat{D}_{\mathrm{IPW},i} = \{A_i - \pi(X_i)\}Y_i/[\pi(X_i)\{1 - \pi(X_i)\}]$ by inverse propensity score weighting. The adjusted outcome is unbiased of $\tau(X_i)$ due to

$$\mathbb{E}(\widehat{D}_{\mathrm{IPW},i} \mid X_i) = \mathbb{E}\left\{\frac{A_iY_i}{\pi(X_i)} - \frac{(1-A_i)Y_i}{1-\pi(X_i)} \mid X_i\right\} = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X_i\} = \tau(X_i).$$

This approach is attractive in clinical trials, where $\pi(X_i)$ is known by trial design. In observational studies, $\pi(X_i)$ is usually unknown and needs to be estimated. Abrevaya et al. (2015) considered kernel regression to estimate $\pi(X_i)$. To avoid possible curse of dimensionality and keep the nonparametric merit, we can perform a prior dimension reduction to find B_π such that $\pi(X_i) = \mathbb{P}(A_i = 1 \mid B_\pi^{\mathrm{T}} X_i)$. Then an improved estimator of $\pi(X_i)$ is

$$\widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi) = \frac{\sum_{j=1}^n A_j \mathcal{K}_{q,h}(\widehat{B}_\pi^{\mathrm{T}} X_j - \widehat{B}_\pi^{\mathrm{T}} X_i)}{\sum_{j=1}^n \mathcal{K}_{q,h}(\widehat{B}_\pi^{\mathrm{T}} X_j - \widehat{B}_\pi^{\mathrm{T}} X_i)},$$

where $\widehat{B}_\pi$ can be obtained similarly following Step 1 in § 2.3 by changing the outcome to A. However, the estimator $\widehat{D}_{\mathrm{IPW},i} = \{A_i - \widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\}Y_i/[\widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\{1 - \widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\}]$ still suffers from the instability due to the inverse weighting, especially when $\widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)$ is close to zero or one. It is well known that the augmented inverse propensity weighted estimator reduces this instability by combining inverse propensity weighting and outcome regressions. Specifically, the corresponding estimator of D_i is

$$\widehat{D}_{\mathrm{AIPW},i} = \{A_i - \widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\}\frac{Y_i - \{1 - \widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\}\widehat{\mu}_1(\widehat{B}_1^{\mathrm{T}} X_i; \widehat{B}_1) - \widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\widehat{\mu}_0(\widehat{B}_0^{\mathrm{T}} X_i; \widehat{B}_0)}{\widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\{1 - \widehat{\pi}(\widehat{B}_\pi^{\mathrm{T}} X_i; \widehat{B}_\pi)\}}.$$

One can easily show that $\mathbb{E}(\widehat{D}_{\mathrm{AIPW},i} \mid X_i)$ is asymptotically unbiased of $\tau(X_i)$. The estimator $\widehat{D}_{\mathrm{AIPW},i}$ is a refined version of Lee et al. (2017), in which the propensity scores are estimated without a prior dimension reduction. Our simulation shows that the estimated central mean subspace and CATE based on $\widehat{D}_i$ and $\widehat{D}_{\mathrm{AIPW},i}$ are comparable and both outperform those based on $\widehat{D}_{\mathrm{MAT},i}$ and $\widehat{D}_{\mathrm{IPW},i}$. Since $\widehat{D}_{\mathrm{AIPW},i}$ requires an extra dimension reduction on $\pi(x)$ and, hence, more computational time, our proposed $\widehat{D}_i$ is more computationally efficient in practice.

2.4 Inference

In this subsection, we derive the large sample properties of $\widehat{B}$ and $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B})$, and an inference procedure for $\tau(x)$ based on these large sample properties is proposed. Based on the notations and regularity conditions in the online supplementary materials, we first establish the following theorem for the prior sufficient dimension reduction for $\mu_a(x)$ $(a = 0, 1)$.

Theorem 1. *Suppose that Assumptions 1 and 2 and Conditions A1–A5 are satisfied. Then,* $\mathbb{P}(\widehat{d}_a = d_a) \to 1$, $\widehat{h}_a = O_{\mathbb{P}}\{n^{-1/(2q+d_a)}\}$, *and*

$$n^{1/2}\mathrm{vecl}(\widehat{B}_a - B_a)1(\widehat{d}_a = d_a) = n^{1/2}\sum_{i=1}^{n} \xi_{B_a,i} + o_{\mathbb{P}}(1) \xrightarrow{d} \mathrm{N}(0, \Sigma_{B_a})$$

as $n \to \infty$, *where* $\xi_{B_a} = \{V_a(B_a)\}^{-1}S_a(B_a)$ *and* $\Sigma_{B_a} = \{V_a(B_a)\}^{-1}\mathbb{E}\{S_a^{\otimes 2}(B_a)\}\{V_a(B_a)\}^{-1}$ *for* $a = 0, 1$.

Exact forms of $V_a(B_a)$ and $S_a(B_a)$ are presented in the online supplementary materials. Theorem 1 as well as Conditions A1–A5 are modifications of results in Huang and Chiang (2017) and hence we omit the proof. Generally speaking, we require the prognostic scores and the joint density functions of $B^{\mathrm{T}}X$ to be smooth enough so that the nonparametric smoothing estimators for these parameter functions are consistent. The constraints on the rate of bandwidths are used to ensure the $n^{1/2}$-consistency of the estimated central mean subspaces, which can be automatically satisfied by the proposed estimated bandwidths. Coupled with the identifiability of $\mathrm{vecl}(B_a)$, the cross-validation type criterion can successfully estimate the true parameters. Theorem 1 serves as a stepping stone to deriving the asymptotic distributions of the estimated central mean space and the proposed estimator for $\tau(x)$, taking into account the fact that D_i is imputed.

Theorem 2. *Suppose that Assumptions 1 and 2 and Conditions A1–A8 are satisfied. Then* $\mathbb{P}(\widehat{d}=d_\tau)\to 1$, $\widehat{h}=O_{\mathbb{P}}\{n^{-1/(2q+d_\tau)}\}$, *and*

$$n^{1/2}\mathrm{vecl}(\widehat{B}-B_\tau)1(\widehat{d}=d_\tau)=n^{1/2}\sum_{i=1}^{n}\xi_{B_\tau,i}+o_{\mathbb{P}}(1)\xrightarrow{d}\mathrm{N}(0,\Sigma_{B_\tau})$$

as $n\to\infty$, *where* $\xi_{B_\tau}=\{V(B_\tau)\}^{-1}S(B_\tau)$ *and* $\Sigma_{B_\tau}=\{V(B_\tau)\}^{-1}\mathbb{E}\{S^{\otimes 2}(B_\tau)\}\{V(B_\tau)\}^{-1}$.

Theorem 3. *Suppose that Assumptions 1 and 2 and Conditions A1–A10 are satisfied. Then,*

$$(nh_\tau^{d_\tau})^{1/2}\{\widehat{\tau}(\widehat{B}^{\mathrm{T}}x;\widehat{B})-\tau(x)-h_\tau^{q_\tau}\gamma(x)\}\xrightarrow{d}\mathrm{N}\{0,\sigma_\tau^2(x)\}$$

as $n\to\infty$, *where*

$$\gamma(x)=\kappa\frac{\partial_u^{q_\tau}\{\mathbb{E}(Z\mid B_\tau^{\mathrm{T}}X=u)f_{B_\tau^{\mathrm{T}}X}(u)\}-\mathbb{E}(Z\mid B_\tau^{\mathrm{T}}X=u)\partial_u^{q_\tau}f_{B_\tau^{\mathrm{T}}X}(u)}{f_{B_\tau^{\mathrm{T}}X}(u)}\bigg|_{u=B_\tau^{\mathrm{T}}x},$$

$$\sigma_\tau^2(x)=\frac{\left\{\int K_{q_\tau}^2(s)\mathrm{d}s\right\}^{d_\tau}\mathbb{V}[Z+\{1-\pi(X)\}\varepsilon_1-\pi(X)\varepsilon_0\mid B_\tau^{\mathrm{T}}X=B_\tau^{\mathrm{T}}x]}{f_{B_\tau^{\mathrm{T}}X}(B_\tau^{\mathrm{T}}x)},$$

$\kappa=\int s^{q_\tau}K_{q_\tau}(s)\mathrm{d}s/q_\tau!$, $Z=(2A-1)\{Y-\mu_{1-A}(B_{1-A}^{\mathrm{T}}X;B_{1-A})\}$, *and* $\varepsilon_a=\{Y-\mu_a(X)\}1(A=a)$ *for* $a=0,1$.

The exact forms of $V(B_\tau)$ and $S(B_\tau)$ as well as the proofs of Theorems 2–3 are given in the online supplementary materials. Similar to Conditions A1–A5, we require the smoothness of $\tau(x)$ and the identifiability of $\mathrm{vecl}(B_\tau)$ to guarantee the results of Theorems 2–3.

The constraints on the bandwidth h_τ are satisfied by our suggested bandwidths, which will be discussed later. The proof of Theorem 2 is similar to that of Theorem 1. The main difference is that the outcome contributing to the asymptotic distribution is now Z instead of the counterfactual D. The proof of Theorem 3 mainly focuses on approximating the influence function coupled with the difference between imputed and non-imputed counterfactual outcomes.

Remark 7. One should note that the asymptotic bias of $\widehat{\mu}_a(u; B)$ is not involved in the asymptotic distribution of $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B})$. This is an important result of Condition A6, which ensures that the convergence rate of $\widehat{\mu}_a(u; B) - \mu_a(u; B)$ is always faster than that of $\widehat{\tau}(u; B) - \mathbb{E}(Z \mid B^{\mathrm{T}}X = u)$.

Remark 8. The most important feature of Theorem 3 is that the asymptotic variance of $\widehat{B}$ is not involved in the asymptotic variance of $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B})$. More precisely speaking, $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B})$ has the same asymptotic variance as the one of $\widehat{\tau}(B_\tau^{\mathrm{T}}x; B_\tau)$. The reason is that $\|\widehat{B} - B_\tau\| = O_{\mathbb{P}}(n^{-1/2})$, which is much faster then the convergence rate $O_{\mathbb{P}}[h_\tau^{q_\tau} + \{\log n/(nh_\tau^{d_\tau})\}^{1/2}]$ of $\widehat{\tau}(B_\tau^{\mathrm{T}}x; B_\tau) - \tau(x)$.

Based on Theorem 3, we can make inference of $\tau(x)$ by estimating the asymptotic bias and variance. However, in practice, direct estimates of $\gamma(x)$ and $\sigma_\tau^2(x)$ are usually unstable, especially when the imputed counterfactual outcomes are involved. For a pre-specified q_τ that satisfies Condition A10, we propose a under-smooth strategy such that the asymptotic bias is dominated by the asymptotic variance. We propose to choose an optimal bandwidth $h_{\tau,\mathrm{opt}} = O\{n^{-1/(2q_\tau + d_\tau)}\}$ by using standard cross-validation for $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B})$ and use $h_\tau = h_{\tau,\mathrm{opt}} n^{-\delta_\tau}$ for some small positive value δ_τ in the inference procedure. We then use a bootstrapping method to estimate the asymptotic distribution of $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B}) - \tau(x)$.

Let ξ_i $(i = 1, \ldots, n)$ be independent and identically distributed from a certain distribution with mean μ_ξ and variance σ_ξ^2. Then $w_i = \xi_i / \sum_{j=1}^n \xi_j$ $(i = 1, \ldots, n)$ are exchangeable random weights. The bootstrapped estimator $\widehat{\tau}^*(x)$ is calculated as

$$\widehat{\tau}^*(x) = \frac{\sum_{j=1}^n w_j \widehat{D}_j^* \mathcal{K}_{q_\tau, h_\tau}(\widehat{B}^{\mathrm{T}} X_j - \widehat{B}^{\mathrm{T}} x)}{\sum_{j=1}^n w_j \mathcal{K}_{q_\tau, h_\tau}(\widehat{B}^{\mathrm{T}} X_j - \widehat{B}^{\mathrm{T}} x)},$$

where

$$\widehat{D}_i^* = A_i\{Y_i - \widehat{\mu}_0^*(\widehat{B}_0^{\mathrm{T}} X_i; \widehat{B}_0)\} + (1 - A_i)\{\widehat{\mu}_1^*(\widehat{B}_1^{\mathrm{T}} X_i; \widehat{B}_1) - Y_i\},$$

$$\widehat{\mu}_a^*(u; B) = \frac{\sum_{j=1}^n w_j Y_j 1(A_j = a) \mathcal{K}_{q_a, h_a}(\widehat{B}_a^{\mathrm{T}} X_j - u)}{\sum_{j=1}^n w_j 1(A_j = a) \mathcal{K}_{q_a, h_a}(\widehat{B}_a^{\mathrm{T}} X_j - u)} \quad (a = 0, 1).$$

According to Remark 8, $\widehat{B}$, $\widehat{B}_0$, and $\widehat{B}_1$ require no bootstrapping in the inference, which highly reduces the computational burden in practice.

The asymptotic variance of $\widehat{\tau}(\widehat{B}^{\mathrm{T}} x; \widehat{B})$ is estimated by $[\mathrm{se}\{\widehat{\tau}^*(x)\}\mu_\xi/\sigma_\xi]^2$, where $\mathrm{se}(\cdot)$ denote the standard error of N bootstrapped estimators. The confidence region of $\tau(x)$ with $1 - \alpha$ confidence level can then be constructed as

$$\widehat{\tau}(\widehat{B}^{\mathrm{T}} x; \widehat{B}) \pm \mathcal{Z}_{1-\alpha/2} \mathrm{se}\{\widehat{\tau}^*(x)\} \frac{\mu_\xi}{\sigma_\xi},$$

where $\mathcal{Z}_p$ is the pth quantile of the standard normal distribution.

3. Simulation study

3.1 Data generating processes

In this section, we present a Monte Carlo exercise aimed at evaluating the finite-sample accuracy of the asymptotic approximations given in the previous section. The covariates $X = (X_1, \ldots, X_{10})$ are generated from independent and identical $\mathrm{Unif}(-3^{1/2}, 3^{1/2})$. The propensity score is $\mathrm{logit}\{\pi(X)\} = 0.5(1 + X_1 + X_2 + X_3)$. The percentage of treated is about 60%. The potential outcomes are designed as following two settings:

M1. $Y(0) = X_1 - X_2 + \varepsilon(0)$ and $Y(1) = 2X_1 + X_3 + \varepsilon(1)$, where $\varepsilon(0)$ and $\varepsilon(1)$ independently follow $N(0, 0.02^2)$. Hence, the CATE is $\tau(x) = x_1 + x_2 + x_3$, and the central mean subspace is $\text{span}\{(1, 1, 1, 0, \dots, 0)^{\mathrm{T}}\}$.

M2. $Y(0) = (X_1 + X_3)(X_2 - 1) + \varepsilon(0)$ and $Y(1) = 2X_2(X_1 + X_3) + \varepsilon(1)$, where $\varepsilon(0)$ and $\varepsilon(1)$ independently follow $N(0, 0.02^2)$. Hence, the CATE is $\tau(x) = (x_1 + x_3)^2(x_2 + 1)^2$, and the central mean subspace is $\text{span}\{(1, 0, 1, 0, \dots, 0)^{\mathrm{T}}, (0, 1, 0, \dots, 0)^{\mathrm{T}}\}$.

The sample size ranges from $n = 250$ and $n = 500$. All the results are based on 1000 replications.

3.2 Competing estimators and simulation results

First, we compare the finite-sample performance of the estimated central mean subspaces using different imputed or adjusted outcomes. In addition to our proposed $\widehat{D}_i$, the nearest neighbor imputation $\widehat{D}_{\mathrm{MAT},i}$, the inverse weighted outcome $\widehat{D}_{\mathrm{IPW},i}$ as well as $\widehat{D}_{\mathrm{AIPW},i}$, we also consider $\widehat{D}_{X,i} = (2A_i - 1)\{Y_i - \widehat{\mu}_{1-A_i}(X_i; I_p)\}$, which is the imputed outcome without any dimension reduction. To compare the information loss for counterfactual outcomes and prior dimension reduction, we further perform the dimension reduction based on the true individual effect D_i and the imputed outcome $\widehat{D}_{\mathrm{OR},i} = (2A_i - 1)\{Y_i - \widehat{\mu}_{1-A_i}(X_i; B_{1-A_i})\}$ based on true oracle central mean subspaces of the prognostic scores. The proportions of estimated structural dimension, the mean squared errors $\|\widehat{B}(\widehat{B}^{\mathrm{T}}\widehat{B})^{-1}\widehat{B}^T - B_\tau(B_\tau^{\mathrm{T}}B_\tau)^{-1}B_\tau^{\mathrm{T}}\|^2$ of the estimated central mean subspaces, and the computing time in seconds are displayed in Table 1. In general, all the proportions of selecting the correct structural dimension tend to one and the mean squared errors tend to zero as sample size increases. Moreover, our proposed estimator outperforms the others and is comparable with respect to the simulated estimators based on $\widehat{D}_{\mathrm{OR},i}$.

Second, we compare the finite-sample performance of the estimated CATE, which include our proposed estimator $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x;\widehat{B})$, the estimator $\widehat{\tau}_X(x)$ based on imputed outcome $\widehat{D}_{X,i}$, the estimator $\widehat{\tau}_{\mathrm{MAT}}(x)$ based on the imputed outcome $\widehat{D}_{\mathrm{MAT},i}$, the estimator $\widehat{\tau}_{\mathrm{IPW}}(x)$ based on the adjusted outcome $\widehat{D}_{\mathrm{IPW},i}$, and the estimator $\widehat{\tau}_{\mathrm{AIPW}}(x)$ based on the adjusted outcome $\widehat{D}_{\mathrm{AIPW},i}$. In addition, we also estimate the CATE by using the difference of two estimated prognostic scores $\widehat{\tau}_{\mathrm{prog}}(x) = \widehat{\mu}_1(\widehat{B}_1^{\mathrm{T}}x;\widehat{B}_1) - \widehat{\mu}_0(\widehat{B}_0^{\mathrm{T}}x;\widehat{B}_0)$. The smoothing estimator $\widehat{\tau}_0(x)$ based on D_i is also considered as a reference to demonstrate the information loss. The CATEs are evaluated at $x = (0,\ldots,0)^{\mathrm{T}}$. The means, standard deviations, and the mean squared errors are displayed in Table 2. In general, our proposed estimator and the $\widehat{\tau}_{\mathrm{AIPW}}$ have comparable performance, and both of them outperform the others.

Finally, we construct confidence intervals and inference for the CATEs by using bootstrapping. Here naive bootstrapping is adopted. That is, $(w_1,\ldots,w_n)$ follows a multinomial distribution with number of trials being n and event probabilities $(1/n,\ldots,1/n)$. Table 3 includes the standard deviations, bootstrapped standard errors and 95% quantile intervals of estimated CATEs, as well as the normal-type 95% confidence intervals with corresponding coverage probabilities and quantile-type 95% confidence intervals with corresponding coverage probabilities for true CATE. As expected, the standard errors get close to the standard deviations, and the coverage probabilities tend to the nominal level when the sample size gets larger.

4. Empirical examples

4.1 The effect of maternal smoking on birth weight

We apply our proposed method to two existing datasets to estimate the effect of maternal smoking on birth weight conditional on different levels of confounders. In the literature, many studies documented that mother's health, education, and labor market status have important effects on child birth weight (Currie and Almond, 2011). In particular, maternal smoking is considered as the most important preventable negative cause (Kramer, 1987). Lee et al. (2017) studied the CATE of smoking given mother's age. In this work, our goal is to fully characterize the CATE of smoking on child birth weight given a vector of important confounding variables while maintaining the interpretability.

4.2 Pennsylvania data

The first dataset consists of observations collected in 2002 from mothers in Pennsylvania, available from the STATA website (http://www.stata-press.com/data/r13/cattaneo2.dta). Following Lee et al. (2017), we focus on white and non-Hispanic mothers, leading to the sample size 3754. The outcome Y of interest is infant birth weight measured in grams. The treatment variable A is equal to 1 if the mother is a smoker and 0 otherwise. The set of covariates X includes the number of prenatal care visits (X_1), mother's educational attainment (X_2), age (X_3), an indicator for the first baby (X_4), an indicator for alcohol consumption during pregnancy (X_5), an indicator for the first prenatal visit in the first trimester (X_6), and an indicator for whether there was a previous birth where the newborn died (X_7). In Lee et al. (2017), parametric models for the prognostic and propensity scores are considered to recover counterfactual outcomes. Here we relax these stringent assumptions and use the proposed nonparametric estimation procedure to provide more detailed structures for the CATE function.

The estimated central mean subspace has dimension one. The coefficients of estimated linear index and corresponding standard errors are displayed in Table 4. Figure 1 shows the estimated CATE at different levels of linear index values along with corresponding normal-type confidence intervals. In general, smoking has significant negative effects on low birth weights, as detected in the existing studies. In the estimated linear index, our method selects X_4 as the baseline covariate and, compared to this baseline covariate, gives a significantly negative coefficient -0.668 with standard error of 0.065 for the number of prenatal care visits. Coupled with the fact that the estimated CATE decreases when the linear index value increases, smoking has significantly more negative effects for mothers who had a non-first baby and more frequent prenatal care visits. This result shows that more frequent prenatal care visits and whether it is a first pregnancy mitigate the effect of smoking on low birth weights.

4.3 North Carolina data

The second dataset is based on the records between 1988 and 2002 by the North Carolina Center Health Services. The dataset was analyzed by Abrevaya et al. (2015) and can be downloaded from Prof. Leili's website. To make a comparison with the Pennsylvania data, we focus on white and first-time mothers and form a random sub-sample with sample size $n = 3754$ among the subjects collected in 2002. The outcome Y and the treatment variable A remain the same as for the Pennsylvania data. The set of covariates includes those used in the analysis of Pennsylvania data but the indicator for the first baby and the indicator for whether there was a previous birth where the newborn died. Besides, it includes indicators for gestational diabetes (X_8), hypertension (X_9), amniocentesis (X_{10}), and ultrasound exams X_{11}). In the analysis of Abrevaya et al. (2015), only the CATE of the mother's age is estimated, and a multi-dimensional kernel smoothing without dimension reduction is used in the estimation procedure. In our analysis, we estimate the CATE of all collected confounding variables, and the dimension reduction techniques are applied to reduce the possible curse of dimensionality.

The estimated central mean subspace has dimension one. The coefficients of estimated linear index and corresponding standard errors are also displayed in Table 4. Figure 1 shows the estimated CATE at different levels of linear index values along with corresponding normal-type confidence intervals. Similar to the results from the Pennsylvania data, smoking has significantly negative effects on low birth weights. Differently, the estimated linear index includes the amniocentesis as the baseline covariate and mothers educational attainment, mothers age, and hypertension as significant covariates. According to the signs of the estimated coefficients and the fact that the estimated CATE decreases when the level of estimated linear index values decreases, smoking has larger detrimental effects for older mothers with lower educational attainment, no hypertension, and amniocentesis. A practical implication is that mothers with such characteristics should quit smoking to prevent low birth weight.

5. Discussion

We propose a nonparametric framework for making inferences about the CATE with multivariate confounder. Our approach is based on the sufficient dimension reduction technique.

The key insight is that $\mathcal{S}_{\mathbb{E}(D|X)}$ may be a strict subspace of $\mathcal{S}_{\mathbb{E}\{Y(0)|X\}} + \mathcal{S}_{\mathbb{E}\{Y(1)|X\}}$. Thus, we target directly for estimating the central mean space of the CATE based on imputed potential outcomes. The contribution of this work is multifold. First, a dimension reduction technique is applied to detect a parsimonious structure of the CATE. This approach is non-parametric in nature, and therefore does not require stringent parametric or semiparametric model assumptions. Second, a kernel regression imputation with prior dimension reduction is proposed to impute the counterfactual outcomes from observational studies, which has better finite sample performances and more efficient computation than existing methods. Third, we derive the asymptotic distribution of the estimated CATE given the estimated central mean space, allowing for transparent interpretation and valid inference, in sharp contrast to usual machine learning methods. In this regard, the proposed approach is the middle ground between simple parametric model approaches and flexible machine learning approaches. Forth, in the theoretical development, the asymptotic distribution of the estimated central mean subspace is not involved in the asymptotic distribution of the estimated CATE. With this observation, the inference procedures on conditional average treatment effects can be done by treating the estimated central mean subspace as the true central mean subspace. This helps save a lot of computational time in our proposed bootstrap procedure. Overall, we believe our method can be a valuable tool for causal inference with a reasonable number of confounders.

However, our proposed estimator is not a panacea with limitations. First, like most causal inference literature, our method is reliant on the key ignorability assumption which is not verifiable based on existing data. Sensitivity analysis is often recommended to assess the robustness of the conclusion the non-testable assumptions (Yang and Lok, 2018). Second, our proposal cannot handle cases with ultra high-dimensional confounders. Regularization techniques may be coupled with the dimension reduction to deal with these cases. The proposed framework of robust inference of the CATE can also be generalized in the following directions. We use under-smoothing to avoid the asymptotic bias of the CATE estimator. Without under-smoothing, the asymptotic bias is not negligible but may be estimated empirically as in Cheng and Chen (2019). We will investigate the finite sample and asymptotic properties of possible bias-corrected estimators in the future. Moreover, we can extend to estimate the CATE with continuous treatment. In this case, the first-stage dimension reduction applies to the potential outcomes for a given treatment level and a reference treatment level, and the second-stage searches the central space for the contrast between the two prognostic scores under the two levels. Third, the first-stage dimension reduction is not confined to the central mean space but can be applied to a transformation of the outcome $g\{Y(a)\}$ for any function $g(\cdot)$. This allows the estimation of the general type of conditional treatment effects such as conditional distribution effects, quantile treatment effects, or survival treatment effects (Yang et al., 2020). Similar to the main paper, we can also derive robust estimators for these causal estimands.

Supplementary Materials

Additional information for this article is available in online supplementary materials, including additional notation and the regularity conditions and the proofs of Theorems 2–3.

Acknowledgements

Dr. Huang is partially supported by MOST grant 108-2118-M-001-011-MY2. Dr. Yang is partially supported by NSF grant DMS 1811245, NCI grant P01 CA142538, NIA grant 1R01AG066883, and NIEHS grant 1R01ES031651.

References

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *J. Bus. Econom. Statist. 33*(4), 485–505.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA 113*(27), 7353–7360.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Ann. Statist. 47*(2), 1148–1178.

Chakraborty, B., S. Murphy, and V. Strecher (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res. 19*(3), 317–343.

Cheng, G. and Y.-C. Chen (2019). Nonparametric inference via bootstrapping the debiased estimator. *Electron. J. Stat. 13*(1), 2194–2256.

Cook, R. D. and B. Li (2002). Dimension reduction for conditional mean in regression. *Ann. Statist. 30*(2), 455–474.

Currie, J. and D. Almond (2011). Human capital development before age five. In *Handbook of labor economics*, Volume 4, pp. 1315–1486. Elsevier.

Hamburg, M. A. and F. S. Collins (2010). The path to personalized medicine. *New England Journal of Medicine 363*(4), 301–304.

Huang, M.-Y. and C.-T. Chiang (2017). An effective semiparametric estimation approach for the sufficient dimension reduction model. *J. Amer. Statist. Assoc. 112*(519), 1296–1310.

Kramer, M. S. (1987). Intrauterine growth and gestational duration determinants. *Pediatrics 80*(4), 502–511.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences 116*(10), 4156–4165.

Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*(7), 1207–1225.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc. 102*(479), 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc. 86*(414), 316–342. With discussion and a rejoinder by the author.

Liang, M. and M. Yu (2020). A semiparametric approach to model effect modification. *Journal of the American Statistical Association 00*(0), 1–13.

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data* (Second ed.). Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc. 107*(497), 168–179.

Ma, Y. and L. Zhu (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist. 41*(1), 250–268.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol. 65*(2), 331–366.

Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *Ann. Statist. 15*(2), 780–799.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, Volume 179 of *Lect. Notes Stat.*, pp. 189–326. Springer, New York.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet 365*(9454), 176–186.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Rzepakowski, P. and S. Jaroszewicz (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems 32*(2), 303–327.

Song, R., S. Luo, D. Zeng, H. H. Zhang, W. Lu, and Z. Li (2017). Semiparametric single-index model for estimating optimal individualized treatment strategy. *Electron. J. Stat. 11*(1), 364–384.

Wang, H. and Y. Xia (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc. 103*(482), 811–821.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist. 35*(6), 2654–2690.

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64*(3), 363–410.

Yang, S. and P. Ding (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika 105*(2), 487–493.

Yang, S. and J. K. Kim (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. In *The Econometrics of Complex Survey Data*. Emerald Publishing Limited.

Yang, S. and J. K. Kim (2020). Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scand. J. Stat. 47*(3), 839–861.

Yang, S. and J. J. Lok (2018). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statist. Sinica 28*(4), 1703–1723.

Yang, S., K. Pieper, and F. Cools (2020). Semiparametric estimation of structural failure time models in continuous-time processes. *Biometrika 107*(1), 123–136.

Yin, X. and B. Li (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist. 39*(6), 3392–3416.

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

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc. 107*(499), 1106–1118.

Zhao, Y., D. Zeng, M. A. Socinski, and M. R. Kosorok (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics 67*(4), 1422–1433.

Zhu, L.-P., L.-X. Zhu, and Z.-H. Feng (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc. 105*(492), 1455–1466.

Zhu, Y. and P. Zeng (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Assoc. 101*(476), 1638–1651.

Institute of Statistical Science, Academia Sinica

E-mail: myh0728@stat.sinica.edu.tw

Department of Statistics, North Carolina State University

E-mail: syang24@ncsu.edu

Table 1: The proportions of $\widehat{d}$, the mean squared errors (MSE) of $\widehat{B}$, and the computing time in seconds under different model settings, sample sizes (n), and imputation of D_i

			proportions of $\widehat{d}$						
model	n		0	1	2	3	≥4	MSE	time
M1	250	$\widehat{D}_i$	0.000	0.976	0.024	0.000	0.000	0.0293	134
		$\widehat{D}_{X,i}$	0.000	0.716	0.246	0.037	0.001	0.5840	94
		$\widehat{D}_{\mathrm{MAT},i}$	0.000	0.833	0.148	0.018	0.001	0.2927	119
		$\widehat{D}_{\mathrm{IPW},i}$	0.000	0.680	0.229	0.087	0.004	0.7143	130
		$\widehat{D}_{\mathrm{AIPW},i}$	0.000	0.955	0.045	0.000	0.000	0.0555	157
		D_i	0.000	0.999	0.001	0.000	0.000	0.0013	142
		$\widehat{D}_{\mathrm{OR},i}$	0.000	0.979	0.021	0.000	0.000	0.0267	64
	500	$\widehat{D}_i$	0.000	0.985	0.015	0.000	0.00	0.0171	634
		$\widehat{D}_{X,i}$	0.000	0.676	0.295	0.029	0.00	0.5392	327
		$\widehat{D}_{\mathrm{MAT},i}$	0.000	0.897	0.097	0.006	0.00	0.1588	288
		$\widehat{D}_{\mathrm{IPW},i}$	0.000	0.615	0.256	0.119	0.01	0.6744	1517
		$\widehat{D}_{\mathrm{AIPW},i}$	0.000	0.980	0.020	0.000	0.00	0.0236	1367
		D_i	0.000	0.999	0.001	0.000	0.00	0.0012	448
		$\widehat{D}_{\mathrm{OR},i}$	0.000	0.985	0.015	0.000	0.00	0.0171	497
M2	250	$\widehat{D}_i$	0.000	0.000	0.995	0.005	0.000	0.0237	136
		$\widehat{D}_{X,i}$	0.000	0.062	0.883	0.053	0.002	0.3222	110
		$\widehat{D}_{\mathrm{MAT},i}$	0.000	0.050	0.894	0.052	0.004	0.3608	104
		$\widehat{D}_{\mathrm{IPW},i}$	0.000	0.269	0.610	0.110	0.011	0.9581	298
		$\widehat{D}_{\mathrm{AIPW},i}$	0.000	0.008	0.978	0.014	0.000	0.0616	362
		D_i	0.000	0.000	0.995	0.005	0.000	0.0119	94
		$\widehat{D}_{\mathrm{OR},i}$	0.000	0.003	0.992	0.004	0.001	0.0243	126
	500	$\widehat{D}_i$	0.000	0.000	0.997	0.003	0.000	0.0139	710
		$\widehat{D}_{X,i}$	0.000	0.008	0.955	0.035	0.002	0.1858	338
		$\widehat{D}_{\mathrm{MAT},i}$	0.000	0.013	0.963	0.021	0.003	0.2040	493
		$\widehat{D}_{\mathrm{IPW},i}$	0.000	0.165	0.714	0.109	0.012	0.7532	1019
		$\widehat{D}_{\mathrm{AIPW},i}$	0.000	0.001	0.995	0.004	0.000	0.0224	1334
		D_i	0.000	0.000	1.000	0.000	0.000	0.0090	573
		$\widehat{D}_{\mathrm{OR},i}$	0.000	0.000	1.000	0.000	0.000	0.0027	687

Table 2: The mean squared errors of estimated CATEs under different model settings and sample sizes (n)

model	n		$\widehat{\tau}(\widehat{B}^{\mathrm{T}}x;\widehat{B})$	$\widehat{\tau}_X(x)$	$\widehat{\tau}_{\mathrm{MAT}}(x)$	$\widehat{\tau}_{\mathrm{IPW}}(x)$	$\widehat{\tau}_{\mathrm{AIPW}}(x)$	$\widehat{\tau}_{\mathrm{prog}}(x)$	$\widehat{\tau}_0(x)$
M1	250	mean	0.003	-0.025	0.094	0.008	0.002	0.003	-0.000
		s.d.	0.0493	0.2203	0.2325	0.5903	0.0532	0.0545	0.0258
		MSE	0.0024	0.0492	0.0629	0.3485	0.0028	0.0030	0.0007
	500	mean	-0.000	0.006	0.065	-0.005	-0.000	0.003	-0.001
		s.d.	0.0300	0.1474	0.1417	0.3642	0.0311	0.0310	0.0159
		MSE	0.0009	0.0218	0.0243	0.1327	0.0010	0.0010	0.0003
M2	250	mean	-0.029	-0.091	-0.180	-0.035	-0.007	-0.048	0.001
		s.d.	0.1006	0.2072	0.3103	0.3803	0.1074	0.1399	0.0639
		MSE	0.0110	0.0512	0.1288	0.1459	0.0116	0.0219	0.0041
	500	mean	-0.015	-0.104	-0.157	-0.010	-0.002	-0.024	0.001
		s.d.	0.0651	0.1418	0.2024	0.2463	0.0566	0.0926	0.0410
		MSE	0.0045	0.0309	0.0655	0.0607	0.0032	0.0092	0.0017

Table 3: The standard deviations (s.d.), bootstrapped standard errors (s.e.), and 95% quantile intervals (Q.I.) of estimated CATEs, and normal-type 95% confidence intervals (N.C.I.) with corresponding coverage probabilities (N.C.P.) and quantile-type 95% confidence intervals (Q.C.I.) with corresponding coverage probabilities (Q.C.P.) for true conditional treatment treatment effect

model	n	s.d.	s.e.	Q.I.	N.C.I	N.C.P.	Q.C.I.	Q.C.P.
M1	250	0.0493	0.0621	(-0.095,0.107)	(-0.119,0.125)	0.966	(-0.119,0.124)	0.975
	500	0.0300	0.0365	(-0.066,0.062)	(-0.072,0.071)	0.965	(-0.074,0.067)	0.972
M2	250	0.1006	0.0998	(-0.226,0.159)	(-0.225,0.166)	0.944	(-0.224,0.167)	0.921
	500	0.0651	0.0645	(-0.132,0.109)	(-0.142,0.111)	0.951	(-0.140,0.112)	0.937

Table 4: The estimated coefficients of linear indices and corresponding standard errors (s.e.) for the Pennsylvania and North Carolina data: $*$ indicates the estimated coefficient is statistically significant at 0.05 level

		Pennsylvania data		North Carolina data	
	covariate	coefficient	s.e.	coefficient	s.e.
X_1	prenatal visit number	-0.668^*	0.0645	0.043	0.0719
X_2	education	-0.059	0.2101	-0.271^*	0.0477
X_3	age	-0.210	0.3076	0.243^*	0.0485
X_4	first baby	1			
X_5	alcohol	0.142	0.6103	-0.101	0.2122
X_6	first prenatal visit	0.275	0.3224	-0.104	0.1556
X_7	previous newborn death	0.169	0.1257		
X_8	diabetes			-0.129	0.1268
X_9	hypertension			-0.333^*	0.1084
X_{10}	amniocentesis			1	
X_{11}	ultrasound			-0.006	0.1612

Figure 1: The estimated CATEs at different levels of linear index values with corresponding confidence intervals.