Statistica Sinica Preprint No: SS-2020-0409	
Title	Robust Inference of Conditional Average Treatment Effects Using Dimension Reduction
Manuscript ID	SS-2020-0409
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0409
Complete List of Authors	Ming-Yueh Huang and Shu Yang
Corresponding Author	Shu Yang
E-mail	syang24@ncsu.edu

Statistica Sinica
Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 a treatment effect varies across individual characteristics, known as the conditional average treatment effect (CATE). In this study, we make robust inferences of the CATE from observational data, which becomes challenging with a multivariate confounder. To reduce the curse of dimensionality, while keeping the nonparametric advantages, we propose double dimension reductions that achieve different goal. First, we identify 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, we use a nonparametric regression with a prior dimension reduction 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. A 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

1. Introduction

Because of patient heterogeneity in response to various aspects of treatment, the paradigm of biomedical and health policy research is shifting from a "one-size-fits-all" treatment approach to one of precision medicine (Hamburg and Collins, 2010). Toward that end, an important step is to understand how a 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), because the CATE is simply the difference between the treated and the control prognostic scores. However, modeling prognostic scores may lead to an overfitting problem for the CATE. Thus, 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 focuses on modeling and approximating the CATE parametrically (Murphy, 2003; Robins, 2004), semiparametrically (Liang and Yu, 2020) and using 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 for making robust inferences of the CATE with a 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 indices that sufficiently characterize the estimand of interest (Cook and Li, 2002). Under this framework, we specify the CATE nonparametrically, and use a

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

model selection procedure to determine a sufficient structural dimension. Directly targeting this central mean subspace enables us to detect the most accurate and parsimonious structure of the CATE. However, existing dimension reduction methods are not applicable, owing 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 using a 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, inverse probability weighted adjusted outcomes (Abrevaya et al., 2015), and augmented inverse probability weighting (Zhao et al., 2012).

We derive the theoretical 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 the imputed and the conditional counterfactual outcomes, which can be expressed as a weighted empirical average of the influence functions of the kernel regression estimator. Thus, we 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 derive the asymptotic distribution of the estimated CATE and show that the imputation step plays a non-negligible role. To make a 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 remainder of this paper 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 data sets. We conclude the paper in Section 5.

2. Methodology

2.1 Preliminaries

Let $X \in \mathcal{X} \subseteq \mathbb{R}^p$ be a vector of pre-treatment covariates, $A \in \mathcal{A} = \{0, 1\}$ be the binary treatment, and $Y \in \mathbb{R}$ be 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 study is to estimate $\tau(x)$ based on the 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 make the following assumptions, which are standard in causal inference with observational studies (Rosenbaum and Rubin, 1983):

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 the outcome. This can be made plausible by collecting detailed information on characteristics of the units related to the 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. As an alternative, we propose a 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. Because $\tau(x) = \mathbb{E}(D \mid X = x)$, the column space of B_τ is called the central mean subspace of D given X, and is denoted by $\mathcal{S}_{\mathbb{E}(D|X)}$ (Cook and Li, 2002).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 $g(x^{\mathrm{T}}\beta)$ that 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 the 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 when 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, a $\mathcal{S}_{\mathbb{E}(D|X)} = \mathrm{span}(\beta-\alpha)$, 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 cannot detect the central mean subspace of the CATE function.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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, $\dim(\mathcal{S}_{\mathbb{E}\{Y(0)|X\}}) = \dim(\mathcal{S}_{\mathbb{E}\{Y(1)|X\}}) = \dim\{\mathrm{span}(\alpha, \beta)\} = 2$, and $\dim(\mathcal{S}_{\mathbb{E}(D|X)}) = \dim\{\mathrm{span}(\beta)\} = 1$. 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 samples.

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 the following equivalent representation of $\tau(x)$:

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

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 the parametrization of the central mean subspace 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 the 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 could be applied directly to estimate $\mathcal{S}_{\mathbb{E}(D|X)}$. However, the fundamental problem in causal inference is that the two potential outcomes can never be

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

jointly observed for each unit; one is factual $Y(A)$, and the other 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 using 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 and computation algorithms can be found therein.

Step 2. Impute the individual treatment effect by

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 are discussed in § 2.4. The bandwidths can be chosen as estimated optimal bandwidths using 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 $\widehat{\tau}(u; B)$ using the minimizer $(\widehat{d}, \widehat{B}, \widehat{h})$ of the following criterion:

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

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. Many existing dimension reduction methods in the literature 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). In contrast to these methods, the cross-validation criterion of Huang and Chiang (2017) estimates the structural dimension, the basis matrix, and an optimal bandwidth for the link function simultaneously. In particular, all of the parameters

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

are estimated in a data-driven way and no ad-hoc tuning is required. In terms of the computational burden, leave-one-out cross-validation is applied for the unknown link functions, but not for the index coefficients. 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. Owing 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. In contrast, our proposed estimator $\widehat{B}$ may not reach the semiparametric efficiency bound. However, as we 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 a 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). We consider matching without replacement and with the number of matches fixed at one. Then, the matching procedure becomes a nearest neighbor imputation (Little and Rubin, 2002). Without loss of generality, we use the Euclidean distance to determine the neighbors; however, the discussion applies to other distances as well (Abadie and Imbens, 2006). Let $\mathcal{J}_i$ be the index set for the matched subject of the 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 as $\widehat{D}_{\mathrm{MAT},i} = \widetilde{Y}_i(1) - \widetilde{Y}_i(0)$. Matching uses

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

the full vector of confounders to determine the distance and corresponding neighbors. When the number of confounders increases, this distance may be too conservative to determine proper neighbors, owing to the curse of dimensionality. In our simulation studies, we find that the performance of the estimation of $\mathcal{S}_{\mathbb{E}(D|X)}$ based on $\widehat{D}_{\mathrm{MAT},i}$ is worse than that of 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 $\widehat{D}_{\mathrm{IPW},i} = \{A_i - \pi(X_i)\}Y_i/[\pi(X_i)\{1-\pi(X_i)\}]$ using inverse propensity score weighting. The adjusted outcome is unbiased of $\tau(X_i)$ because

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

This approach is attractive in clinical trials, where $\pi(X_i)$ is known by the trial design. In observational studies, $\pi(X_i)$ is usually unknown and needs to be estimated. Abrevaya et al. (2015) considered using a kernel regression to estimate $\pi(X_i)$. To avoid the possible curse of dimensionality and keep the nonparametric advantages, we 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 to 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 instability owing 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

regressions. Specifically, the corresponding estimator of D_i is

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

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 $\widehat{D}_{\mathrm{IPW},i}$. Because $\widehat{D}_{\mathrm{AIPW},i}$ requires an extra dimension reduction on $\pi(x)$, and hence a longer computation 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 propose an inference procedure for $\tau(x)$ based on these properties. Using the notation and regularity conditions in the online Supplementary Material, 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* $a = 0, 1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

Exact forms of $V_a(B_a)$ and $S_a(B_a)$ are presented in the Supplementary Material. Theorem 1 and Conditions A1–A5 are modifications of the 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 rates of the bandwidths 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*

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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

The exact forms of $V(B_\tau)$ and $S(B_\tau)$ and the proofs of Theorems 2–3 are given in the Supplementary Materials. Similarly 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 are 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 focuses on approximating the influence function, coupled with the difference between the imputed and the non-imputed counterfactual outcomes.

Remark 7. 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, $\widehat{\tau}(\widehat{B}^{\mathrm{T}} x; \widehat{B})$ has the same asymptotic variance as that 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 an 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 prespecified q_τ that satisfies Condition A10, we propose an under-smooth strategy in which the asymptotic bias is

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

dominated by the asymptotic variance. We propose choosing an optimal bandwidth $h_{\tau,\text{opt}} = O\{n^{-1/(2q_\tau+d_\tau)}\}$ using standard cross-validation for $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x;\widehat{B})$, and using $h_\tau = h_{\tau,\text{opt}}n^{-\delta_\tau}$ 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 (i.i.d.) 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

$$\widehat{D}_i^* = A_i\{Y_i - \widehat{\mu}_0^*(\widehat{B}_0^{\mathrm{T}}X_i;\widehat{B}_0)\} + (1-A_i)\{\widehat{\mu}_1^*(\widehat{B}_1^{\mathrm{T}}X_i;\widehat{B}_1) - Y_i\},$$
$$\widehat{\mu}_a^*(u;B) = \frac{\sum_{j=1}^n w_j Y_j 1(A_j = a)\mathcal{K}_{q_a,h_a}(\widehat{B}_a^{\mathrm{T}}X_j - u)}{\sum_{j=1}^n w_j 1(A_j = a)\mathcal{K}_{q_a,h_a}(\widehat{B}_a^{\mathrm{T}}X_j - u)} \ (a = 0,1).$$

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

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 an i.i.d. Unif$(-3^{1/2}, 3^{1/2})$. The propensity score is $\text{logit}\{\pi(X)\} = 0.5(1 + X_1 + X_2 + X_3)$. The treated percentage is about 60%. The potential outcomes are designed as the 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 $\text{span}\{(1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}\}$.

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 $\text{span}\{(1, 0, 1, 0, \ldots, 0)^{\mathrm{T}}, (0, 1, 0, \ldots, 0)^{\mathrm{T}}\}$.

The sample sizes are $n = 250$ and $n = 500$. All 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 $\widehat{D}_{\text{MAT},i}$, and the inverse weighted outcome $\widehat{D}_{\text{IPW},i}$ and $\widehat{D}_{\text{AIPW},i}$, 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 the counterfactual outcomes and prior dimension reduction, we further perform the dimension reduction based on the true

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 the true oracle central mean subspaces of the prognostic scores. The proportions of the estimated structural dimension, 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 computing time in seconds are displayed in Table 1. In general, all proportions of selecting the correct structural dimension tend to one and the mean squared errors tend to zero as the 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 for our proposed estimator $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B})$, the estimator $\widehat{\tau}_X(x)$ based on the 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 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 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 mean squared errors are displayed in Table 2. In general, our proposed estimator and the $\widehat{\tau}_{\mathrm{AIPW}}$ have comparable performance, and both outperform the others.

Finally, we construct confidence intervals and inference for the CATEs using bootstrapping. Here, naive bootstrapping is adopted. That is, $(w_1, \ldots, w_n)$ follows a multinomial distribution with number of trials n and event probabilities $(1/n, \ldots, 1/n)$. Table 3 includes the standard deviations, bootstrapped standard errors, and 95% quantile intervals of the estimated CATEs, as well as the normal-type 95% confidence intervals with corresponding

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

coverage probabilities and quantile-type 95% confidence intervals with corresponding coverage probabilities for the 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 increases.

4. Empirical examples

4.1 The effect of maternal smoking on birth weight

We apply our proposed method to two existing data sets to estimate the effect of maternal smoking on birth weight, conditional on different levels of confounders. In the literature, many studies have documented that a 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 the most important preventable negative cause (Kramer, 1987). Lee et al. (2017) studied the CATE of smoking, given a 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 data set 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, yielding sample size of 3754. The outcome Y of interest is infant birth weight, measured in grams. The treatment variable A is equal to one if the mother is a smoker, and zero otherwise. The set of covariates X includes the number of prenatal care visits (X_1), mother's educational attainment (X_2),

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 the estimated linear index and the 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 a significant negative effect on low birth weight, as detected in 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 a 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 greater 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 weight.

4.3 North Carolina data

The second data set is based on records between 1988 and 2002 by the North Carolina Center Health Services. The data set was analyzed by Abrevaya et al. (2015), and can be

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 the Pennsylvania data, except for the indicator for the first baby and the indicator for whether there was a previous birth where the newborn died. In addition, 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 the estimated linear index and the 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. Similarly to the results from the Pennsylvania data, smoking has a significantly negative effect on low birth weight. However, the estimated linear index includes amniocentesis as the baseline covariate, and the mother's educational attainment, mother's age, and hypertension as significant covariates. According to the signs of the estimated coefficients and the fact that the estimated CATE decreases when the estimated linear index values decrease, 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

weight.

5. Discussion

We propose a nonparametric framework for making inferences about the CATE with a 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 directly estimate 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 nonparametric in nature, and therefore does not require stringent parametric or semiparametric model assumptions. Second, a kernel regression imputation with a prior dimension reduction is proposed to impute the counterfactual outcomes from observational studies, which has better finite-sample performance and a more efficient computation than those of 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. Fourth, 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 the conditional average treatment effects can be done by treating the estimated central mean subspace as the true central mean subspace. This reduces the computation time in our proposed bootstrap procedure. Overall, we believe our method can be a valuable tool for causal inference with a reasonable number

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

of confounders.

However, our proposed estimator does have limitations. First, like most approaches in the causal inference literature, our method relies on the key ignorability assumption, which is not verifiable based on existing data. A sensitivity analysis is often recommended to assess the robustness of the conclusion based on the non-testable assumptions (Yang and Lok, 2018). Second, our proposal cannot handle cases with ultrahigh-dimensional confounders. Regularization techniques may be coupled with the dimension reduction to deal with these cases. The proposed framework for a robust inference of the CATE can be generalized in the following ways. 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 future research. Moreover, we can extend our work to estimate the CATE for continuous treatments. 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 us to estimate general-types conditional treatment effects, such as conditional distribution effects, quantile treatment effects, or survival treatment effects (Yang et al., 2020). We can also derive robust estimators for these causal estimands.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

Supplementary Material

Additional information is available in the online Supplementary Material, including additional notation and the regularity conditions and the proofs of Theorems 2–3.

Acknowledgments

Dr. Huang was partially supported by MOST grant 108-2118-M-001-011-MY2. Dr. Yang was 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. *Journal of Business and Economic Statistics 33*(4), 485–505.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Chakraborty, B., S. Murphy, and V. Strecher (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research 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. *The Annals of Statistics 30*(2), 455–474.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

REFERENCES

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. *Journal of the American Statistical Association 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. *Journal of the American Statistical Association 102*(479), 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–342.

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*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ.

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association 107*(497), 168–179.

Ma, Y. and L. Zhu (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics 41*(1), 250–268.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

REFERENCES

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B 65*(2), 331–366.

Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *The Annals of Statistics 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, 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. *Journal of the American Statistical Association 103*(482), 811–821.

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

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

REFERENCES

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. *Scandinavian Journal of Statistics 47*(3), 839–861.

Yang, S. and J. J. Lok (2018). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica 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. *The Annals of Statistics 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. *Journal of the American Statistical Association 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. *Journal of the American Statistical Association 105*(492), 1455–1466.

Zhu, Y. and P. Zeng (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association 101*(476), 1638–1651.

Institute of Statistical Science, Academia Sinica

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

Department of Statistics, North Carolina State University

E-mail: syang24@ncsu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

Table 1: The proportions of $\widehat{d}$, mean squared errors (MSE) of $\widehat{B}$, and 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	$\geq$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
		$\widehat{D}_{\text{MAT},i}$	0.000	0.833	0.148	0.018	0.001	0.2927	119
		$\widehat{D}_{\text{IPW},i}$	0.000	0.680	0.229	0.087	0.004	0.7143	130
		$\widehat{D}_{\text{AIPW},i}$	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
		$\widehat{D}_{\text{OR},i}$	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
		$\widehat{D}_{\text{MAT},i}$	0.000	0.897	0.097	0.006	0.00	0.1588	288
		$\widehat{D}_{\text{IPW},i}$	0.000	0.615	0.256	0.119	0.01	0.6744	1517
		$\widehat{D}_{\text{AIPW},i}$	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
		$\widehat{D}_{\text{OR},i}$	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
		$\widehat{D}_{\text{MAT},i}$	0.000	0.050	0.894	0.052	0.004	0.3608	104
		$\widehat{D}_{\text{IPW},i}$	0.000	0.269	0.610	0.110	0.011	0.9581	298
		$\widehat{D}_{\text{AIPW},i}$	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
		$\widehat{D}_{\text{OR},i}$	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
		$\widehat{D}_{\text{MAT},i}$	0.000	0.013	0.963	0.021	0.003	0.2040	493
		$\widehat{D}_{\text{IPW},i}$	0.000	0.165	0.714	0.109	0.012	0.7532	1019
		$\widehat{D}_{\text{AIPW},i}$	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
		$\widehat{D}_{\text{OR},i}$	0.000	0.000	1.000	0.000	0.000	0.0027	687

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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 the true conditional treatment effect

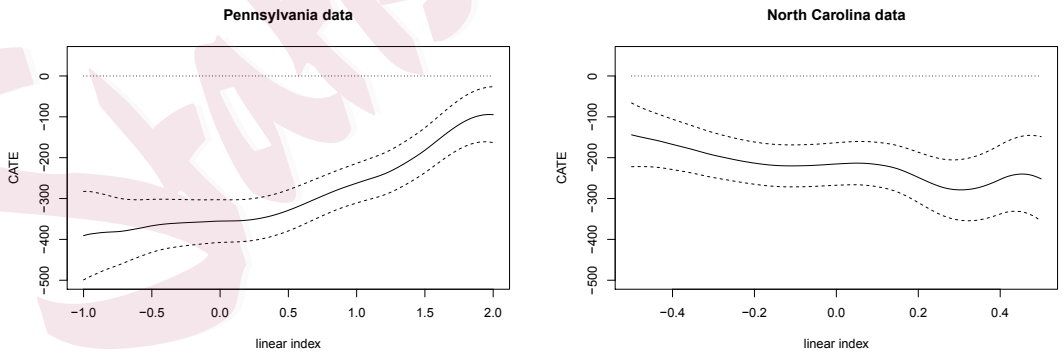
model	n	s.d.	s.e.	Q.I.	N.C.I	N.C.P.	Q.C.I.	Q.C.P.
M1	250	0.0493	0.0621	(-0.095,0.107)	(-0.119,0.125)	0.966	(-0.119,0.124)	0.975
	500	0.0300	0.0365	(-0.066,0.062)	(-0.072,0.071)	0.965	(-0.074,0.067)	0.972
M2	250	0.1006	0.0998	(-0.226,0.159)	(-0.225,0.166)	0.944	(-0.224,0.167)	0.921
	500	0.0651	0.0645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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0409

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



Supplementary materials for "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

1. Additional Notation and Regularity Conditions

Let $(\cdot)^{\otimes}$ denote the Kronecker power of a vector and let $\|\cdot\|$ represent the Frobenius norm of a matrix. Denote $f_{B^{\mathrm{T}}X}(u)$ as the marginal density of $B^{\mathrm{T}}X$,

$$f^{[m]}(x,u;B) = \partial_u^m[\mathrm{E}\{(X_l - x_l)^{\otimes m} \mid B^{\mathrm{T}}X = u\} f_{B^{\mathrm{T}}X}(u)],$$

$$E_a^{[m]}(x,u;B) = \partial_u^m[\mathrm{pr}(A = a \mid B^{\mathrm{T}}X = u)\mathrm{E}\{(X_l - x_l)^{\otimes m} \mid B^{\mathrm{T}}X = u\} f_{B^{\mathrm{T}}X}(u)],$$

$$F_a^{[m]}(x,u;B) = \partial_u^m[\mathrm{E}\{Y1(A = a) \mid B^{\mathrm{T}}X = u\}\mathrm{E}\{(X_l - x_l)^{\otimes m} \mid B^{\mathrm{T}}X = u\} f_{B^{\mathrm{T}}X}(u)],$$

$$G^{[m]}(x,u;B) = \partial_u^m[\mathrm{E}(Z \mid B^{\mathrm{T}}X = u)\mathrm{E}\{(X_l - x_l)^{\otimes m} \mid B^{\mathrm{T}}X = u\} f_{B^{\mathrm{T}}X}(u)], \ \ (a = 0, 1, \ \ m = 0, 1, 2),$$

where $Z = (2A-1)\{Y - \mu_{1-A}(B_{1-A}^{\mathrm{T}}X; B_{1-A})\}$. We will show that

$$\partial_{\mathrm{vecl}(B)}^m \widehat{\mu}_a(B^{\mathrm{T}}x; B) \to \mu^{[m]}(x;B) = \sum_{\ell=0}^{m} \binom{m}{\ell} F_a^{[\ell]}(x, B^{\mathrm{T}}x; B) E_{a,\mathrm{inv}}^{[m-\ell]}(x, B^{\mathrm{T}}x; B),$$

and

$$\partial_{\mathrm{vecl}(B)}^m \widehat{\tau}(B^{\mathrm{T}}x; B) \to \tau^{[m]}(x;B) = \sum_{\ell=0}^{m} \binom{m}{\ell} G^{[\ell]}(x, B^{\mathrm{T}}x; B) f_{\mathrm{inv}}^{[m-\ell]}(x, B^{\mathrm{T}}x; B),$$

uniformly as $n \to \infty$, where

$$f^{[0]}_{\mathrm{inv}}(x,u;B) = 1/f_{B^{\mathrm{T}}X}(u), \qquad E^{[0]}_{a,\mathrm{inv}}(x,u;B) = 1/E^{[0]}_{a}(x,u;B),$$

$$f^{[1]}_{\mathrm{inv}}(x,u;B) = -\frac{f^{[1]}(x,u;B)}{f^{2}_{B^{\mathrm{T}}X}(u)}, \qquad f^{[2]}_{\mathrm{inv}}(x,u;B) = \frac{2\{f^{[1]}(x,u;B)\}^{2}}{f^{3}_{B^{\mathrm{T}}X}(u)} - \frac{f^{[2]}(x,u;B)}{f^{2}_{B^{\mathrm{T}}X}(u)},$$

$$E^{[1]}_{a,\mathrm{inv}}(x,u;B) = -\frac{E^{[1]}_{a}(x,u;B)}{\{E^{[0]}_{a}(x,u;B)\}^{2}}, \qquad E^{[2]}_{a,\mathrm{inv}}(x,u;B) = \frac{2\{E^{[1]}_{a}(x,u;B)\}^{2}}{E^{[0]}_{a}(x,u;B)} - \frac{E^{[2]}_{a}(x,u;B)}{\{E^{[0]}_{a}(x,u;B)\}^{2}}.$$

According to the notation, we can define the corresponding score vectors and information matrices of $\mathrm{CV}_a(d,B,h)$ and $\mathrm{CV}(d,B,h)$:

$$S_a(B) = -1(A=a)\{Y - \mu_a(B^{\mathrm{T}}X;B)\}\mu^{[1]}(X;B),$$

$$V_a(B) = \mathrm{E}(1(A=a)[\{\mu^{[1]}(X;B)\}^{\otimes 2} - \{Y - \mu_a(B^{\mathrm{T}}X;B)\}\mu^{[2]}(X;B)]),$$

$$S(B) = -\{Z - \mathrm{E}(Z \mid B^{\mathrm{T}}X)\}\tau^{[1]}(X;B),$$

$$V(B) = \mathrm{E}[\{\tau^{[1]}(X;B)\}^{\otimes 2} - \{Z - \mathrm{E}(Z \mid B^{\mathrm{T}}X)\}\tau^{[2]}(X;B)].$$

In addition, let $B_{d,a}$ be the minimizer of $b^2_a(B) = \mathrm{E}[\{\mu_a(B^{\mathrm{T}}X;B) - \mu(X)\}^2]$ and let $B_{d,\tau}$ be the minimizer of $b^2_\tau(B) = \mathrm{E}[\{\mathrm{E}(Z \mid B^{\mathrm{T}}X) - \tau(X)\}^2]$ over all $p \times d$ matrices B. Then, $b^2_a(B) \to b^2_a(B_{d,a})$ implies $B \to B_{d,a}$ for $\mathrm{span}(B) \not\supseteq \mathrm{span}(B_a)$, and $b^2_\tau(B) \to b^2_\tau(B_{d,\tau})$ implies $B \to B_{d,\tau}$ for $\mathrm{span}(B) \not\supseteq \mathrm{span}(B_\tau)$. The following regularity conditions are imposed for our theorems:

A1 $\partial_u^{q+m}\mathrm{E}\{(X_l - x_l)^{\otimes m} \mid B^{\mathrm{T}}X = u\}$, $\partial_u^{q+2} f_{B^{\mathrm{T}}X}(u)$, $\partial_u^{q+2}\mathrm{pr}(A=a \mid B^{\mathrm{T}}X = u)$, $\partial_u^{q+2}\mathrm{E}\{Y1A=a \mid B^{\mathrm{T}}X = u\}$, and $\partial_u^{q+2}\mathrm{E}(Z \mid B^{\mathrm{T}}X = u)$ $(a = 0,1,\ m = 1,2)$, are Lipschitz continuous in u with the Lipschitz constants being independent of (x,B).

A2 $\inf_{(x,B)} f_{B^{\mathrm{T}}X}(B^{\mathrm{T}}x) > 0$ and $\inf_{(x,B)} \mathrm{pr}(A=a \mid B^{\mathrm{T}}X = B^{\mathrm{T}}x) > 0$ $(a = 0,1)$.

A3 For each working dimension $d > 0$, h falls in the interval $H_{\delta,n} = [h_l n^{-\delta}, h_u n^{-\delta}]$ for some positive constants h_l and h_u and $\delta \in (1/(4q), 1/\max\{2d+2, d+4\})$. In particular, this requires $q > \max(d/2+1, 2)$.

A4 $\inf_{\{B:d<d_a\}} b^2_a(B) > 0$ and $b^2_a(B) = 0$ if and only if $B = B_a$ when $d = d_a$ $(a = 0,1)$.

A5 $V_a(B_{d,a})$ is non-singular for $d \geq d_a$ $(a = 0,1)$.

A6 For each working dimension d, $q_a > qd_a/d$ $(a = 0,1)$.

A7 $\inf_{\{B:d<d_\tau\}} b_\tau^2(B) > 0$ and $b_\tau^2(B) = 0$ if and only if $B = B_\tau$ when $d = d_\tau$.

A8 $V(B_{d,\tau})$ is non-singular for $d \geq d_\tau$.

A9 $h_\tau \to 0$ and $nh_\tau^{d_\tau} \to \infty$.

A10 For each working dimension d, $q_\tau > qd_\tau/d$.

Conditions A1–A2 are the smoothness and boundedness conditions for the population functions to ensure the uniform convergence of kernel estimators. Moreover, to remove the remainder terms in the approximation of $\textsc{cv}(d, B, h)$ and $\textsc{cv}(d, B, h)$ to their target functions, the constraints for the orders of kernel functions and the bandwidths are drawn in Conditions A3 and A6. Conditions A4–A5 and A7–A8 ensure the identifiability of B_a $(a = 0, 1)$ and B_τ, respectively. The requirements of h_τ and q_τ used in $\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B})$ are given in Condition A9–A10.

2. Preliminary Lemmas

The proofs of the main theorems rely on the following lemma:

Lemma 1. *Suppose that Assumption 1 and Conditions A1–A6 are satisfied. Then,*

$$\widehat{\tau}(u; B) - \mathrm{E}(Z \mid B^{\mathrm{T}}X = u) = \frac{1}{n}\sum_{i=1}^{n}[Z_i - \mathrm{E}(Z \mid B^{\mathrm{T}}X = u) + \{1 - \pi(X_i)\}\varepsilon_{1,i} - \pi(X_i)\varepsilon_{0,i}]\omega_{h,i}(u; B) + r_n(u; B),$$

where $\varepsilon_{a,i} = \{Y_i - \mu_a(X_i)\}1(A_i = a)$, $(a = 0, 1)$, $\omega_{h,i}(u; B) = \mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - u)/\sum_{j=1}^{n}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_j - u)$, *and* $\sup_{(u,B)}|r_n(u, B)| = o_{\mathbb{P}}[h^q + \{\log n/(nh^d)\}^{1/2}]$.

Proof. First note that

$$\begin{aligned}\widehat{\tau}(u; B) - \mathrm{E}(Z \mid B^{\mathrm{T}}X = u) &= \frac{1}{n}\{\widehat{D}_i - \mathrm{E}(Z \mid B^{\mathrm{T}}X = u)\}\omega_{h,i}(u; B) \\ &= \frac{1}{n}\{Z_i - \mathrm{E}(Z \mid B^{\mathrm{T}}X = u)\}\omega_{h,i}(u; B) + \frac{1}{n}\{\widehat{D}_i - Z_i\}\omega_{h,i}(u; B).\end{aligned}$$

Further,

$$
\begin{aligned}
&\frac{1}{n}(\widehat{D}_i - Z_i)\omega_{h,i}(u;B) \\
=\ &\frac{1}{n}\sum_{i=1}^{n}[(1-A_i)\{\widehat{\mu}_1(\widehat{B}_1^{\mathrm{T}}X_i;\widehat{B}_1)-\mu_1(X_i)\} - A_i\{\widehat{\mu}_0(\widehat{B}_0^{\mathrm{T}}X_i;\widehat{B}_0)-\mu_0(X_i)\}]\omega_{h,i}(u;B) \\
=\ &\frac{1}{n}\sum_{i=1}^{n}(1-A_i)\{\widehat{\mu}_1(B_1^{\mathrm{T}}X_i;B_1)-\mu_1(X_i)\}\omega_{h,i}(u;B) \\
&-\frac{1}{n}\sum_{i=1}^{n}A_i\{\widehat{\mu}_0(B_0^{\mathrm{T}}X_i;B_0)-\mu_0(X_i)\}\omega_{h,i}(u;B) + O_{\mathbb{P}}(n^{-1/2}) \\
\triangleq\ &I_1 + I_2 + O_{\mathbb{P}}(n^{-1/2}),
\end{aligned} \tag{21}
$$

because of $\|\mathrm{vecl}(\widehat{B}_a - B_a)\| = O_{\mathbb{P}}(n^{-1/2})$ by Theorem 1. Now let $\kappa_{a,h,i}(u) = \mathcal{K}_{q_a,h}(B_a^{\mathrm{T}}X_i - u)/\sum_{j=1}^{n}1(A_j = a)\mathcal{K}_{q_a,h}(B_a^{\mathrm{T}}X_j - u)$. Then, we decompose I_1 into

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(1-A_i)\widehat{\mu}_1(B_1^{\mathrm{T}}X_i;B_1)-\mu_1(X_i)\}\omega_{h,i}(u;B) \\
=\ &\frac{1}{n}\sum_{i=1}^{n}\{1-\pi(X_i)\}\omega_{h,i}(u;B)\sum_{j=1}^{n}\{Y_j-\mu_1(X_i)\}1(A_j=1)\kappa_{1,h_1,j}(B_1^{\mathrm{T}}X_i) \\
&+\frac{1}{n}\sum_{i=1}^{n}\{\pi(X_i)-A_i\}\omega_{h,i}(u;B)\sum_{j=1}^{n}\{Y_j-\mu_1(X_i)\}1(A_j=1)\kappa_{1,h_1,j}(B_1^{\mathrm{T}}X_i) \\
=\ &\frac{1}{n}\sum_{i=1}^{n}\{1-\pi(X_i)\}\varepsilon_{1,i}\omega_{h,i}(u;B) \\
&+\frac{1}{n}\sum_{i=1}^{n}\{1-\pi(X_i)\}\left\{\sum_{j=1}^{n}\varepsilon_{1,j}\kappa_{1,h_1,j}(B_1^{\mathrm{T}}X_i)-\varepsilon_{1,i}\right\}\omega_{h,i}(u;B) \\
&+\frac{1}{n}\sum_{i=1}^{n}\{1-\pi(X_i)\}\left[\sum_{j=1}^{n}\{\mu_1(X_j)-\mu_1(X_i)\}1(A_j=1)\kappa_{1,h_1,j}(B_1^{\mathrm{T}}X_i)\right]\omega_{h,i}(u;B) \\
&+\frac{1}{n}\sum_{i=1}^{n}\{\pi(X_i)-A_i\}\omega_{h,i}(u;B)\sum_{j=1}^{n}\{Y_j-\mu_1(X_i)\}1(A_j=1)\kappa_{1,h_1,j}(B_1^{\mathrm{T}}X_i) \\
\triangleq\ &J_0 + J_1 + J_2 + J_3.
\end{aligned} \tag{22}
$$

To bound J_1, we re-write it as

$$
J_1 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{1,i}\left\{\sum_{j=1}^{n}\{1-\pi(X_j)\}\omega_{h,j}(u;B)\kappa_{1,h_1,j}(B_1^{\mathrm{T}}X_i) - \{1-\pi(X_i)\}\omega_{h,i}(u;B)\right\}.
$$

Since $\mathrm{E}(\varepsilon_{1,i} \mid X_i) = 0$, we can show that J_1 is a degenerate U-process indexed by (u,B). An application of Theorem 6 in Nolan and Pollard (1987) ensures that $\mathrm{E}(\sup_{(u,B)}|J_1|) \le C/(n^2h_1^{d_1}h^d)$. Thus, by selecting h_1 in an optimal rate $O\{n^{-1/(2q_1+d_1)}\}$ and coupled with Conditions A3 and A6, we have

$$
\sup_{(u,B)}|J_1| = o_{\mathbb{P}}\left\{h^q + \left(\frac{\log n}{nh^d}\right)^{1/2}\right\}. \tag{23}
$$

Second, similar to the proofs in Huang and Chiang (2017), standard arguments in kernel smoothing estimation show that

$$\begin{aligned}&\sup_i |\sum_{j=1}^n \{\mu_1(X_j) - \mu_1(X_i)\} 1(A_j = 1)\kappa_{1,h_1,j}(B_1^{\mathrm{T}} X_i)| \\ &= O_{\mathbb{P}}\left\{ h_1^{q_1} + \left(\frac{\log n}{n h_1^{d_1}}\right)^{1/2} \right\} = O_{\mathbb{P}}\{n^{-q_1/(2q_1+d_1)}\}\end{aligned}$$

by selecting h_1 in an optimal rate $O\{n^{-1/(2q_1+d_1)}\}$. Under Conditions A3 and A6, one can further show that this rate is $o_{\mathbb{P}}[h^q + \{\log n/(nh^d)\}^{1/2}]$. Thus, we have

$$\sup_{(u,B)} |J_2| = o_{\mathbb{P}}\left\{ h^q + \left(\frac{\log n}{n h^d}\right)^{1/2} \right\}. \tag{24}$$

Finally, note that J_3 is also a degenerate U-process indexed by (u, B). Thus, by the same argument for J_1, we can show that

$$\sup_{(u,B)} |J_3| = o_{\mathbb{P}}\left\{ h^q + \left(\frac{\log n}{n h^d}\right)^{1/2} \right\}. \tag{25}$$

By substituting (23)–(25) into (22), we then have

$$\sup_{(u,B)} |I_1 - \frac{1}{n}\sum_{i=1}^n (1 - A_i)\varepsilon_{1,i}\omega_{h,i}(u; B)| = o_{\mathbb{P}}\left\{ h^q + \left(\frac{\log n}{n h^d}\right)^{1/2} \right\}. \tag{26}$$

Following the same arguments above, we can also show that

$$\sup_{(u,B)} |I_2 - \frac{1}{n}\sum_{i=1}^n A_i\varepsilon_{0,i}\omega_{h,i}(u; B)| = o_{\mathbb{P}}\left\{ h^q + \left(\frac{\log n}{n h^d}\right)^{1/2} \right\}. \tag{27}$$

Substituting (26)–(27) into (21) completes the proof. □

Now we derive the independent and identically distributed representations of $\widehat{\tau}(B^{\mathrm{T}}x; B) - \tau^{[0]}(x; B)$ and $\partial_{\mathrm{vecl}(B)}\widehat{\tau}(B^{\mathrm{T}}x; B) - \tau^{[1]}(x; B)$.

Lemma 2. *Suppose that Assumption 1 and Conditions A1–A6 are satisfied. Then,*

$$\sup_{(x,B)} |\widehat{\tau}(B^{\mathrm{T}}x; B) - \tau^{[0]}(x; B) - \frac{1}{n}\sum_{i=1}^n \eta_{h,i}^{[0]}(x; B)| = o_{\mathbb{P}}\left(h^{2q} + \frac{\log n}{n h^d} \right), \tag{28}$$

$$\sup_{(x,B)} \|\partial_{\mathrm{vecl}(B)}\widehat{\tau}(B^{\mathrm{T}}x; B) - \tau^{[1]}(x; B) - \frac{1}{n}\sum_{i=1}^n \eta_{h,i}^{[1]}(x; B)\| = o_{\mathbb{P}}\left(h^{2q} + \frac{\log n}{n h^{d+1}} \right), \tag{29}$$

where

$$\eta_{h,i}^{[0]}(x;B) = \frac{\xi_i(x;B)}{f_{B^{\mathrm{T}}X}(B^{\mathrm{T}}x)}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x),$$
$$\eta_{h,i}^{[1]}(x;B) = \frac{\xi_i(x;B)}{f_{B^{\mathrm{T}}X}(B^{\mathrm{T}}x)}\partial_{\mathrm{vecl}(B)}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x)$$
$$- \tau^{[1]}(x;B)\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x) - \frac{f^{[1]}(x,B^{\mathrm{T}}x;B)}{f_{B^{\mathrm{T}}X}(B^{\mathrm{T}}x)}\eta_{h,i}^{[0]}(x;B),$$

and $\xi_i(x;B) = Z_i - \mathrm{E}(Z \mid B^{\mathrm{T}}X = B^{\mathrm{T}}x)$.

Proof. First, (28) is a direct result of Lemma 1. As for (29), note that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\partial_{\mathrm{vecl}(B)}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x) - G^{[1]}(x,B^{\mathrm{T}}x;B)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\xi_i(x;B)\partial_{\mathrm{vecl}(B)}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x) + r_{1n}(x;B), \quad (210)$$

where $\sup_{(x,B)}|r_{1n}(x,B)| = o_{\mathbb{P}}[h^q + \{\log n/(nh^{d+1})\}^{1/2}]$, by paralleling the proof steps of Lemma 1. Now by using the Taylor expansion, we have

$$\begin{aligned}
&\partial_{\mathrm{vecl}(B)}\widehat{\tau}(B^{\mathrm{T}}x;B) - \tau^{[1]}(x;B)\\
=\ & \frac{\sum_{i=1}^{n}\widehat{D}_i\partial_{\mathrm{vecl}(B)}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x)/n - \tau^{[0]}(x;B)\sum_{i=1}^{n}\partial_{\mathrm{vecl}(B)}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x)/n}{f_{B^{\mathrm{T}}X}(B^{\mathrm{T}}x)}\\
&- \frac{\tau^{[1]}(x;B)}{n}\sum_{i=1}^{n}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x) - \frac{f^{[1]}(x,B^{\mathrm{T}}x;B)}{f_{B^{\mathrm{T}}X}(B^{\mathrm{T}}x)}\{\widehat{\tau}(B^{\mathrm{T}}x;B) - \tau^{[0]}(x;B)\}\\
&+ r_{2n}(x;B), \qquad (211)
\end{aligned}$$

where

$$\begin{aligned}
r_{2n}(x,B) \quad = \quad & O_{\mathbb{P}}\{|\widehat{\tau}(B^{\mathrm{T}}x;B) - \tau^{[0]}(x;B)|^2\\
& + \|\sum_{i=1}^{n}\widehat{D}_i\partial_{\mathrm{vecl}(B)}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_i - B^{\mathrm{T}}x)/n - G^{[1]}(x,B^{\mathrm{T}}x;B)\|^2\}.
\end{aligned}$$

Finally, substituting the result in Lemma 1 and (210) into (211) completes the proof. □

Corollary 1. *Suppose that Assumption 1 and Conditions A1–A6 are satisfied. Then,*

$$\sup_{(x,B)}|\widehat{\tau}(B^{\mathrm{T}}x;B) - \tau^{[0]}(x;B)| = O_{\mathbb{P}}\left\{h^q + \left(\frac{\log n}{nh^d}\right)^{1/2}\right\},$$
$$\sup_{(x,B)}\|\partial_{\mathrm{vecl}(B)}\widehat{\tau}(B^{\mathrm{T}}x;B) - \tau^{[1]}(x;B)\| = O_{\mathbb{P}}\left\{h^q + \left(\frac{\log n}{nh^{d+1}}\right)^{1/2}\right\}.$$

3. Proofs of Theorems 2 and 3

3.1 Proof of Theorem 2

Proof. Let $\bar{\tau}^{-i}(B^{\mathrm{T}}X_i; B) = \sum_{j\neq i} Z_j \mathcal{K}_{q,h}(B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i)/\sum_{j\neq i}\mathcal{K}_{q,h}(B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i)$. We can decompose $\mathrm{CV}(d,B,h)$ into

$$\begin{aligned}
\mathrm{CV}(d,B,h) &= \frac{1}{n}\sum_{i=1}^{n}\{Z_i - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)\}^2 + \frac{1}{n}\sum_{i=1}^{n}(\widehat{D}_i - Z_i)^2 \\
&+ \frac{1}{n}\sum_{i=1}^{n}\{\widetilde{\tau}^{-i}(B^{\mathrm{T}}X_i;B) - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)\}^2 \\
&+ \frac{2}{n}\sum_{i=1}^{n}(\widehat{D}_i - Z_i)\{\widetilde{\tau}^{-i}(B^{\mathrm{T}}X_i;B) - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)\} \\
&+ \frac{2}{n}\sum_{i=1}^{n}(\widehat{D}_i - Z_i)\{Z_i - \tau(X_i)\} + \frac{2}{n}\sum_{i=1}^{n}(\widehat{D}_i - Z_i)\{\tau(X_i) - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)\} \\
&+ \frac{2}{n}\sum_{i=1}^{n}\{Z_i - \tau(X_i)\}\{\widetilde{\tau}^{-i}(B^{\mathrm{T}}X_i;B) - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)\} \\
&+ \frac{2}{n}\sum_{i=1}^{n}\{\tau(X_i) - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)\}\{\widetilde{\tau}^{-i}(B^{\mathrm{T}}X_i;B) - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)\} \\
&\triangleq SS_1 + SS_2 + SS_3 + SC_1 + SC_2 + SC_3 + SC_4 + SC_5.
\end{aligned}$$

Note that

$$\sup_i |\widehat{D}_i - Z_i| \leq \sum_{a=0}^{1}\sup_{(u,B)}|\widehat{\mu}_a(u;B) - \mu_a(u;B)| = o_{\mathbb{P}}\left\{h^q + \left(\frac{\log n}{nh^d}\right)^{1/2}\right\}, \tag{312}$$

$$\sup_{(i,B)}|\widetilde{\tau}^{-i}(B^{\mathrm{T}}X_i;B) - \bar{\tau}^{-i}(B^{\mathrm{T}}X_i;B)| \leq C\sum_{a=0}^{1}\sup_{(u,B)}|\widehat{\mu}_a(u;B) - \mu_a(u;B)| = o_{\mathbb{P}}\left\{h^q + \left(\frac{\log n}{nh^d}\right)^{1/2}\right\} \tag{313}$$

for some positive constant C, by using Conditions A1–A3, Condition A6, and standard arguments in kernel smoothing estimation.

When $\mathrm{span}(B) \supseteq \mathrm{span}(B_\tau)$, Theorem 1 of Huang and Chiang (2017) implies that $SS_1 = \sigma_\tau^2 + O_{\mathbb{P}}\{h^{2q} + \log n/(nh^d)\}$, where $\sigma_\tau^2 = \mathrm{E}[\{Z - \tau(X)\}^2]$. From (312)–(313), $\sup_B |SS_3|$ and $\sup_B |SC_1|$ are of order $o_{\mathbb{P}}\{h^{2q} + \log n/(nh^d)\}$. Further, by using $\sup_{(x,B)}|\bar{\tau}(B^{\mathrm{T}}x;B) - \tau(x)| = O_{\mathbb{P}}[h^q + \{\log n/(nh^d)\}^{1/2}]$, $\sup_B |SC_3|$ and $\sup_B |SC_5|$ are also of order $o_{\mathbb{P}}\{h^{2q} + \log n/(nh^d)\}$. Now note that SC_4 can be expressed a U-process indexed by B asymptotically. By using the same proof steps for the cross term in Theorem 1 of Huang and Chiang (2017), one can immediately conclude that $\sup_B |SC_4| = o_{\mathbb{P}}\{h^{2q} + \log n/(nh^d)\}$. Combining the results above, we have $\mathrm{CV}(d,B,h) = SS_1 + SS_2 + SC_2 + o_p(SS_1)$ uniformly in B. When $\mathrm{span}(B) \not\supseteq \mathrm{span}(B_\tau)$, Theorem 1 of Huang and Chiang (2017) implies that

$SS_1 = \sigma_\tau^2 + b_\tau^2(B) + o_{\mathbb{P}}(1)$. By using (312)–(313) again, we have $\text{CV}(d, B, h) = SS_1 + SS_2 + SC_2 + o_{\mathbb{P}}(1)$ uniformly in B. Finally, since SS_2 and SC_2 are independent of B, the minimizer of $\text{CV}(d, B, h)$ has the same asymptotic distribution as the minimizer of SS_1. Thus, Theorem 2 is a direct result of Theorem 2 in Huang and Chiang (2017). □

3.2 Proof of Theorem 3

Proof. By using first-ordered Taylor expansion, we have

$$\begin{aligned}\widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B}) - \tau(x) &= \widehat{\tau}(\widehat{B}^{\mathrm{T}}x; \widehat{B}) - \widehat{\tau}(B_\tau^{\mathrm{T}}x; B_\tau) + \widehat{\tau}(B_\tau^{\mathrm{T}}x; B_\tau) - \tau(x) \\ &= \partial_{\mathrm{vecl}(B)}\widehat{\tau}(\bar{B}^{\mathrm{T}}x; \bar{B})\mathrm{vecl}(\widehat{B} - B_\tau) + \widehat{\tau}(B_\tau^{\mathrm{T}}x; B_\tau) - \tau(x),\end{aligned}$$

where $\bar{B}$ lies on the line segment between $\widehat{B}$ and B_τ. From Theorem 2, $\mathrm{vecl}(\widehat{B} - B_\tau) = O_{\mathbb{P}}(n^{-1/2})$. Coupled with Corollary 1 and continuous mapping theorem, $\partial_{\mathrm{vecl}(B)}\widehat{\tau}(\bar{B}^{\mathrm{T}}x; \bar{B}) = O_{\mathbb{P}}(1)$. Moreover, from (28), we have

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

in distribution as $n \to \infty$. Combining the results above completes the proof of Theorem 3. □

References

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.

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