Journal of Statistics and Computer Science
Vol. 2, Issue 1, 2023, pp. 55-77
https://doi.org/10.47509/JSCS.2023.v02i01.05

A Simple Least-square Method For Estimating Homogeneous Causal Treatment Effect

Ying Zhang[a],Yuanfang Xu[b],Giorgos Bakoyannis[c],Yuan Wu [d] and Bin Huang [e]

[a]Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, Omaha, Nebraska , U.S.A.; [b]Bristol Myers Squibb, Princeton, New Jersey, U.S.A.; [c]Department of Biostatistics and Health Data Science, Indiana University Fairbanks School of Public Health, Indianapolis, Indiana, U.S.A.; [d]Department of Biostatistics and Bioinformatics, Duke University , Durham, North Carolina,U.S.A.; [e]Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, U.S.A.

ARTICLE HISTORY
Compiled June 1, 2023

ABSTRACT
Estimating causal treatment effect with observational data is a challenging task since the underlying data-generating models for outcome and treatment assignment are unknown. Many widely used causal inference methods show poor operational characteristics from a statistical perspective. In this paper, we propose an ordinary least-squares (OLS) based approach for estimating causal treatment effect without parametric assumptions for either outcome or treatment assignment mechanism. Our model-free estimator builds on the nonparametric spline-based sieve estimates of two summary scores: the propensity score and the mean outcome score. We show that this proposed method leads to a $\sqrt{n}$-consistent and asymptotically normally distributed estimator of the causal treatment effect. Monte-Carlo simulation studies are conducted to compare our proposed method with other widely used conventional methods and demonstrate the superior performance of our model-free estimator. We apply this approach to a case study of the biologic anti-rheumatic treatment effect on children with newly onset juvenile idiopathic arthritis disease.

KEYWORDS
causal inference; empirical process theory; nonparametric estimation; potential outcomes; regression splines

1. Introduction

Observational study is increasingly used for understanding causal questions in social or biomedical science, for logistical and ethical considerations, and for its ability to efficiently use existing resources and information technology. With the causal assumptions and the concept of a potential outcome framework (Rubin, 1974; Rosenbaum and Rubin, 1983, 1985; Holland, 1986; Neyman et al., 1990), many different methods have been developed for the purpose of estimating treatment effect based on observational data (Rosenbaum and Rubin, 1983; Rosenbaum, 1987; Lunceford and Davidian,

CONTACT Ying Zhang. Email: ying.zhang@unmc.edu

Article History
Received : 22 February 2023; Revised : 30 May 2023; Accepted : 01 June 2023; Published : 30 June 2023

To cite this paper

Ying Zhang, Yuanfang Xu, Giorgos Bakoyannis, Yuan Wu & Bin Huang (2023). A Simple Least-square Method for Estimating Homogeneous Causal Treatment Effect. *Journal of Statistics and Computer Science.* 2(1), 55-77.

2004; Imbens, 2004; Bang and Robins, 2005; Austin & Mamdani, 2006). Many of these works focused mainly on obtaining an unbiased point estimator. The inferential and operational characteristics of the estimated causal treatment effect were carefully investigated by Tsiatis (2006), Kang and Schafer (2007), and Gutman and Rubin (2017). These works call for the need for a causal treatment effect estimator that is robust, consistent, and offers well-performing statistical and operational characteristics.

The introduction of potential outcomes to causal inference brings the fundamental challenge of estimating causal treatment effect: the observed outcomes comprise only half of the potential outcomes, while the other half is missing. Consider a relatively simple setting: a binary treatment A, where $A = 1$ and $A = 0$ indicate the treatment condition and control condition, respectively. If both the potential outcomes $Y(0)$ and $Y(1)$ are available, the causal treatment effect is simply the contrast of $Y(0)$ and $Y(1)$. However, only one of the two is observed depending on the status of A, i.e., $Y = AY(1) + (1 - A)Y(0)$. Thus the observed outcome (Y) is produced jointly by two correlated data-generating models: the underlying science model for the potential outcomes $(Y(0), Y(1))$ and the treatment selection model for $pr(A = 1)$. Many existing causal inference methods assume accurate knowledge of at least one of the data-generating models. For example, outcome regression methods assume that we possess knowledge about the true functional form of the science model. Propensity score based methods, such as the inverse probability weighting method (IPW) (Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983; Rosenbaum, 1987), work only when the propensity scores can be consistently estimated from a correctly specified treatment selection model. As a combination of outcome regression method and inverse probability weighting method, doubly robust methods, such as augmented inverse probability weighting (AIPW) (Robins et al., 1994; Scharfstein et al., 1999; Tsiatis, 2006), require correct model specification for at least one of the two (Lunceford and Davidian, 2004; Bang and Robins, 2005). However, in real-world applications, it is often unrealistic to assume accurate knowledge of either outcome-generating model or the treatment assignment model. If both models are misspecified, the conventional estimators of causal treatment effect can be badly biased (Carpenter et al., 2006; Kang and Schafer, 2007; Vansteelandt et al., 2012). Therefore, in causal inference practice, we frequently face the challenge of deriving a good estimator of causal treatment effect that ensures consistency without the correct specification for at least one model.

In this paper, we propose a two-stage ordinary least-squares (OLS) based model-free approach for estimating causal treatment effect and develop its asymptotic theory for causal inference. The proposed method builds on two summary scores: the mean score $E(Y|X)$ and the propensity score $pr(A = 1|X) = E(A|X)$. We adopt the widely available nonparametric regression technique, regression spline estimation (Wegman and Wright, 1983), to obtain consistent estimates of the two summary scores at the first stage. Then, at the second stage, we construct a least-squares based utility function to conduct an M-estimation with the two plugged-in consistently estimated scores, which simply yields an explicit estimate of the casual treatment effect. Without any parametric model assumption in the whole estimation procedure, this method leads to a model-free estimator of causal treatment effect that enjoys estimation consistency and robustness without worrying about the functional forms for the underlying mean and propensity scores. We use empirical process theory to show that this model-free OLS estimator of causal treatment effect is $\sqrt{n}$-consistent and asymptotically normal. Furthermore, the numerical instability issue due to inverse weights for extreme estimated propensity scores in conventional inverse probability weighting methods is

prevented in our proposed method because the estimated propensity scores are not intended to be used for creating individualized weights.

The rest of the paper is organized as follows: In Section 2, we first give the notation and background of causal treatment effect in the potential outcome framework and then describe the proposed method and the estimation procedure in detail. In Section 3, we state an asymptotic theory related to the proposed methodology. In Section 4, we use Monte-Carlo simulations to demonstrate the superior finite sample performance of the proposed model-free estimator of causal treatment effect in comparison with the conventional causal inference methods, IPW and AIPW. We illustrate the application of this proposed method by estimating the biologic anti-rheumatic treatment effect on children with newly onset juvenile idiopathic arthritis disease using data from an observational study in Section 5. The technical details for deriving the asymptotic theories are included in the appendix.

2. Method

2.1. Notation and Causal Assumptions

Consider an observational study with n subjects, in which we observed n independent and identically distributed copies of data $O = (A, Y, X)$. Y is a continuous outcome, A is a dichotomous treatment variable taking values 1 (active treatment) or 0 (control), and X is a fixed vector of measured pre-treatment covariates including potential confounders to the relationship between Y and A. We are interested in estimating the causal effect of treatment A on the outcome Y.

Using the potential outcome framework for causal inference (Rubin, 1974), the estimation of causal effects is a comparison of potential outcomes. Let $(Y_i(1), Y_i(0))$ denote a pair of potential outcomes for subject i that indicate the hypothetical outcome of subject i under treatment and control, respectively. Following Rubin's stable unit-treatment value assumption (Rubin, 1980; 1990), we assume the treatment levels are identical across all the subjects and the potential outcomes for any subject do not depend on the treatment or outcome of other subjects. We use the standard consistency assumption (Rubin, 1986; Judea, 2010) to link the potential outcomes and observed outcome for subject i as

$$Y_i = \begin{cases} Y_i(1) & \text{if } A_i = 1 \\ Y_i(0) & \text{if } A_i = 0, \end{cases}$$

which can be simply expressed as

$$Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i).$$

We denote the causal treatment effect for subject i as $\tau_i = Y_i(1) - Y_i(0)$, the contrast of two potential outcomes, and the covariate-specific average treatment effect (ATE) for a subpopulation with covariate X being x as $\tau(x) = E(Y_i(1) - Y_i(0)|X = x)$. Then the population ATE is simply $\tau = E_X\{\tau(X)\}$, where $E_X(.)$ is the expectation taking over the population with respect to X. In this paper, we are specifically interested in the estimation of ATE τ in a homogeneous population in terms of treatment effect. That is we assume that the treatment effect is homoscedastic, i.e, $\tau_i = \tau + \epsilon_i$ with ϵ_i

satisfying (i) $E(\epsilon_i) = 0$, (ii) ϵ_i is independent of X_i. In this simplified scenario, X only plays the role of an effect mediator but not an effect modifier, and $\tau(x)$ is thus equal to τ.

Another key assumption in the potential outcome framework is the identification of the causal treatment effect. It is commonly referred to as the strongly ignorable treatment assignment assumption (Rosenbaum & Rubin, 1983) and includes the following two conditions:

(1) Ignorability: $(Y_i(1), Y_i(0)) \perp A|X$, where $\perp$ denotes independence. This condition is also called no unmeasured confounder, which implies that X is sufficient to be adjusted for in order to remove all the possible confounding between the relationship of A and Y.

(2) Positivity: $0 < pr(A = 1 \mid X = x) < 1$ for all x. $pr(A = 1 \mid X = x)$ is formally termed the propensity score. This condition indicates that the probability of being treated or not treated for any subject with all possible values of X is positive.

The well-known and common approaches for causal ATE in the literature are the inverse propensity weighting (IPW) method and the augmented inverse propensity weighting (AIPW) method, which was shown to have the double robustness property (Lunceford and Davidian, 2004; Bang and Robins, 2005). Our proposed method requires a weaker version of the ignorability assumption in the form of $E(Y_i(a)|A_i = a, X_i) = E(Y_i(a)|X_i), a = 1, 0$, which is also called the mean independence assumption (Imbens, 2004). For the scenario with a homoscedastic treatment effect, ATE τ is identifiable by the observed outcome under the weak ignorability and consistency assumptions, since

$$\begin{aligned}
\tau &= E(Y(1) - Y(0)) = E(Y(1) - Y(0)|X) && \text{Homoscedastic treatment effect} \\
&= E(Y(1)|A = 1, X) - E(Y(0)|A = 0, X) && \text{Weak ignorable assumption} \\
&= E(Y|A = 1, X) - E(Y|A = 0, X) && \text{Consistency assumption}
\end{aligned}$$

2.2. Motivation

The proposed method is built on two scores: mean score and propensity score, denoted as $m(X) = E(Y|X)$ and $\pi(X) = pr(A = 1 \mid X)$, respectively. For the scenario described in Section 2.1, we have

$$\begin{aligned}
m(X) &= E(Y|X) = E(Y|A = 1, X)pr(A = 1|X) + E(Y|A = 0, X)pr(A = 0|X) \\
&= E(Y|A = 0, X) + \pi(x)\tau. \qquad (1)
\end{aligned}$$

Since

$$\begin{aligned}
E(Y|A = a, X) &= E(Y|A = 1, X)a + E(Y|A = 0, X)(1 - a) \\
&= E(Y|A = 0, X) + a\tau, \qquad (2)
\end{aligned}$$

it follows that

$$E(Y|A = a, X) = m(X) + (a - \pi(X))\tau, \qquad (3)$$

by subtracting (1) from (2), which fits exactly the framework of the partially linear semiparametric regression model studied by Robinson (1988). Equation (3) leads to a simple OLS method to estimate the ATE τ by solving for

$$\arg\min_{\tau} \sum_{i=1}^{n} \{(Y_i - m(X_i)) - (A_i - \pi(X_i))\tau\}^2. \tag{4}$$

If both the mean and propensity scores are known, the OLS estimator of τ is easily calculated by

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \{(Y_i - m(X_i))(A_i - \pi(X_i))\}}{\sum_{i=1}^{n} (A_i - \pi(X_i))^2} \tag{5}$$

2.3. Two-Stage Estimation

The proposed OLS-based estimator of τ requires knowledge of the mean and propensity scores, $m(x)$ and $\pi(x)$. We propose to estimate the causal treatment effect in a model-free manner in the following two-stage estimation procedure.

Stage 1: Nonparametric estimation of $\hat{m}(x)$ and $\hat{\pi}(x)$.

Stage 2: We plug in $(\hat{m}(X), \hat{\pi}(X))$ to the OLS estimator (5) to obtain an explicit model-free estimator of ATE τ by

$$\hat{\tau}^{mf} = \frac{\sum_{i=1}^{n} \{(Y_i - \hat{m}(X_i))(A_i - \hat{\pi}(X_i))\}}{\sum_{i=1}^{n} (A_i - \hat{\pi}(X_i))^2}.$$

Remark 1: $\hat{\tau}^{mf}$ is not a type of inverse probability weighting estimator because the denominator is the total of the squares of estimated individual propensity score. Hence, it does not suffer the numerical instability issue as frequently seen in IPW and AIPW methods due to the true or estimated propensity score being close to 0 or 1 at some covariate values X (Tsiatis, 2006; Austin and Stuart, 2015; Gutman and Rubin, 2017).

Remark 2: Hahn (1998) was the first to identify a connection between the suggested method for ATE and Robinson's partially linear regression method. However, he did not realize that the connection is only valid for the situation that the treatment effect is independent of covariates X and hence made a wrong statement about the semi-parametric inefficiency of the proposed method. As a matter of fact, if the treatment effect is non-homoscedastic, equation (3) becomes

$$E(Y|A = a, X) = m(X) + (a - \pi(X))\tau(X),$$

where $\tau(X) = E(Y(1) - Y(0)|X)$. It is clear that the ATE $\tau = E_X[\tau(X)]$ does not fit the framework of the partially linear regression model, and our proposed estimator

$\hat{\tau}^{mf}$ converges in probability to

$$\tau^* = \frac{E\left[(Y - m(X))(A - \pi(X))\right]}{E\left[(A - \pi(X))^2\right]}$$

as indicated by Hahn (1998), which actually is not $\tau = E(Y(1) - Y(0))$. Our simulation study (not included in this paper) also confirmed this fact. Nevertheless, if the homogeneous treatment effect assumption holds, $var(\tau(X)) \equiv 0$ and the asymptotic variance bound for τ given in Theorem 1 (Hahn, 1998) is the asymptotic variance of the semiparametric OLS-estimator discussed in Hahn (1998, page 323).

3. Asymptotic Properties

In this section, we provide a general asymptotic normality theorem for the proposed OLS-based estimator $\hat{\tau}^{mf}$ given that the nonparametric estimators $(\hat{m}(X), \hat{\pi}(X))$ possess some proper asymptotic properties. Let $Pf = Ef(X)$ be the probability measure on some measurable real function f on the sample space $\mathcal{X}$ and $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$ the corresponding empirical measure on a random sample from $\mathcal{X}$. Denote $\|f\|_{L_2(P)} = \left\{\int f^2(x)dP(x)\right\}^{1/2}$ the L_2-norm associated with the probability measure P.

Theorem 3.1. *Suppose the observed data consist of a random sample* $D = \{O_i = (A_i, Y_i, X_i) : \quad for \quad i = 1, 2, \cdots, n\}$ *and potential outcomes satisfy the homogeneity, weak ignorability, and consistency assumptions. Also, assume that the nonparametric estimators* $(\hat{m}(X), \hat{\pi}(X))$ *from Stage 1 satisfy the following conditions:*

C1. $\sqrt{n}(\mathbb{P}_n - P)\left\{(Y - m(X))(\hat{\pi}(X) - \pi(X))\right\} = o_P(1)$

C2. $\sqrt{n}(\mathbb{P}_n - P)\left\{(A - \pi(X))(\hat{m}(X) - m(X))\right\} = o_P(1)$

C3. $\sqrt{n}(\mathbb{P}_n - P)\left\{(\hat{m}(X) - m(X))(\hat{\pi}(X) - \pi(X))\right\} = o_P(1)$

C4. $\sqrt{n}(\mathbb{P}_n - P)\left\{(A - \pi(X))(\hat{\pi}(X) - \pi(X))\right\} = o_P(1)$

C5. $\sqrt{n}(\mathbb{P}_n - P)(\hat{\pi}(X) - \pi(X))^2 = o_P(1)$

C6. $\|\hat{m} - m\|_{L_2(P)} = O_p\left(n^{-1/4}\right)$ *and* $\|\hat{\pi} - \pi\|_{L_2(P)} = o_p\left(n^{-1/4}\right)$

Then,

$$\sqrt{n}(\hat{\tau}^{mf} - \tau) \to_d N(0, H\mathbf{\Sigma}H^\top),$$

where

$$H = \left(\frac{1}{E[A - \pi(X)]^2}, \frac{-E[Y - m(X)][A - \pi(X)]}{\left(E[A - \pi(X)]^2\right)^2}\right)$$

and

$$\mathbf{\Sigma} = Cov\begin{pmatrix} [Y - m(X)][A - \pi(X)] \\ [A - \pi(X)]^2 \end{pmatrix}.$$

Remark 3: According to Theorem 1, it appears that the conditions for the proposed model-free estimator of ATE $\hat{\tau}^{mf}$ being asymptotic normal are not strong. As long as the nonparametric estimators $(\hat{m}(X), \hat{\pi}(X))$ from Stage 1 fall into some P-Donsker classes (van der Vaart and Wellner, 1996) and are weak consistent with a convergence rate even much slower than root-n, the proposed estimator of ATE is $\sqrt{n}$-consistent and asymptotically normal. These conditions can be easily accomplished by performing the regression splines (Wegman and Wright, 1983) at Stage 1 for both the mean and propensity scores, given that they are smooth functions.

4. Simulation Study

A simulation study was conducted to evaluate the performance of our proposed model-free estimator of ATE and to compare it with estimators from other conventional methods, including IPW and AIPW in various scenarios. This Monte Carlo simulation mimicked a hypothetical retrospective cohort study with two observed baseline covariates $X = (X_1, X_2)$. X_1 was a continuous covariate generated from $N(0, 1)$ and X_2 was a binary group indicator generated from $Bernoulli(0.5)$ and independent of X_1. For each subject in a random sample, the probability of being assigned to the treatment group $A = 1$ is modeled by the following logistic model:

$$\pi(X) = pr(A = 1|X_1, X_2) = \frac{\exp(-0.8 + 0.5X_1 + 0.2X_1^2 + 0.4X_2X_1)}{1 + \exp(-0.8 + 0.5X_1 + 0.2X_1^2 + 0.4X_2X_1)}, \tag{6}$$

and the treatment assignment $A = 1$ associated with X was generated from $Bernoulli(\pi(X))$. This treatment-selection mechanism yielded a proportion of treated subjects of approximately 45%. Given A and covariates (X_1, X_2), the outcome Y was generated according to the following model:

$$Y = -2 + 0.5X_1X_2 + 0.6X_1^2 + 2e^{X_1} + 6A + \epsilon, \tag{7}$$

where the random error ϵ is independent of $X = (X_1, X_2)$ and is normally distributed with $N(0, 1)$. This results in the mean score given by

$$\begin{aligned} m(X) &= -2 + 0.5X_1X_2 + 0.6X_1^2 + 2e^{X_1} \\ &\quad + \frac{6\exp(-0.8 + 0.5X_1 + 0.2X_1^2 + 0.4X_2X_1)}{1 + \exp(-0.8 + 0.5X_1 + 0.2X_1^2 + 0.4X_2X_1)}. \end{aligned} \tag{8}$$

Our interest lies in studying the average casual effect of a dichotomous treatment A on the outcome Y, which is 6 in this setting.

We estimated the mean and propensity scores at the first stage using the cubic B-splines regression method that was only applied to X_1 since X_2 is a binary variable. For a study sample with n observations of X_1 contained in a closed interval $[a, b]$, we divided this interval into $q_n - 3$ subintervals made by a sequence of spline knots given by

$$a = \xi_1 = \xi_2 = \xi_3 = \xi_4 < \xi_5 < \cdots < \xi_{q_n} < \xi_{q_n+1} = \xi_{q_n+2} = \xi_{q_n+3} = \xi_{q_n+4} = b,$$

where the number of knots q_n was chosen to be $\left[n^{1/3}\right]$, the largest integer less than

$n^{1/3}$ and the $q_n - 4$ interior knots were placed at the quantiles of X_1. The mean and propensity scores were estimated through the regression splines by modelling

$$m(X) = \sum_{j=1}^{q_n} \alpha_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \alpha_j^{(2)} B_j(X_1) X_2$$

and

$$\log\left\{\frac{\pi(X)}{1-\pi(X)}\right\} = \sum_{j=1}^{q_n} \beta_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \beta_j^{(2)} B_j(X_1) X_2,$$

respectively, where $B_j(X)$ is the normalized B-spline basis functions at the knots ξ_j for $j = 1, \cdots, q_n$.

For the competing method IPW, we considered two scenarios: (i) one with the true parametric propensity model (IPW-pT) given in (6); (ii) another with the wrongly specified ordinary logistic regression model (IPW-pW) with covariates X_1, X_2, and X_1X_2. We particularly use the stabilized weights:

$$w_i = \frac{1}{n}\left[\frac{A_i n_1}{\hat{\pi}(X_i)} + \frac{(1-A_i)n_0}{1-\hat{\pi}(X_i)}\right] \quad i = 1, 2, \cdots, n$$

(Cole and Hernan,2008) to reduce potential influence of extreme weights resulting from the inverse probability weighting, where n_1 and n_0 are sample sizes associated with treatment and control, respectively. For the competing method, AIPW, we examine all four possible scenarios: (i) both the mean and propensity score models were specified correctly (AIPW-mT&pT) as given in (8) and (6), respectively; (ii) the mean score model was specified correctly as given in (8) but the propensity score model was specified wrongly as for the IPW-pW estimator (AIPW-mT&pW); (iii) the propensity score model was specified correctly as given in (6) but the mean score was wrongly specified as the ordinary linear regression model (AIPW-mW&pT) with covariates X_1, X_2, and X_1X_2; (iv) both the mean and propensity score models were wrongly specified as aforementioned (AIPW-mW&pW). In addition, we also included a hypothetical scenario for which we knew exactly the outcome model (7), and we implemented the maximum likelihood estimation method (MLE) to achieve an efficient estimation of ATE of 6. The result of this hypothetical analysis was used as a benchmark to evaluate the competing methods described above.

Table 1 presents the simulation results for sample sizes 200, 400, and 800 that summarize the estimation bias (Bias), Monte-Carlo standard deviation (M-C SD) based on 1000 repetitions, average standard error (ASE) based on 100 bootstrap samples with replacement, and coverage probability (CP) of the 95% Wald-confidence interval (CI). Figure 1 depicts the distributions of the competing ATE estimators with 1000 repeated samples of size $n = 400$. This simulation study clearly revealed that the IPW method with the correctly specified propensity model had very little estimation (less than 0.5% for sample size 800) but was very unstable in estimating the ATE with a larger amount of variability compared to other methods. However, when the propensity model was misspecified, the IPW method led to a very large estimation bias. As anticipated, the AIPW method performed much better than the IPW method. When both the mean and propensity score models were correctly specified, the estimation bias was virtually

Table 1. Comparison of the proposed method to IPW and AIPW methods for estimating ATE in a Monte-Carlo simulation study with 1000 repetitions. M-C SD: Monte-Carlo standard deviation; ASE: Average estimated standard error; 95% CP: Empirical coverage probability of estimated 95% Wald-confidence interval.

	$\|Bias\|$			$M-CSD$			ASE			$95\%CP$		
n	200	400	800	200	400	800	200	400	800	200	400	800
Proposed method:												
MF	0.007	0.001	<0.001	0.156	0.110	0.078	0.158	0.109	0.076	0.946	0.945	0.933
MLE method:												
MLE	0.007	<0.001	<0.001	0.150	0.108	0.077	0.151	0.106	0.075	0.951	0.944	0.937
IPW methods:												
IPW-pT	0.009	0.015	0.004	0.979	0.642	0.446	0.999	0.674	0.467	0.948	0.956	0.952
IPW-pW	3.532	3.538	3.494	1.148	0.745	0.534	1.184	0.760	0.530	0.129	0.002	0.000
AIPW methods:												
AIPW-pT&mT	0.007	0.001	<0.001	0.153	0.109	0.077	0.154	0.107	0.076	0.951	0.943	0.937
AIPW-pT&mW	0.018	0.021	0.010	0.465	0.312	0.214	0.508	0.322	0.217	0.959	0.951	0.965
AIPW-pW&mT	0.007	0.001	<0.001	0.153	0.109	0.077	0.154	0.107	0.076	0.945	0.941	0.932
AIPW-pW&mW	3.402	3.450	3.426	1.079	0.718	0.519	1.041	0.717	0.510	0.103	0.002	0.000

negligible even when the sample size was only 200; the average standard error was close to the Monte-Carlo standard deviation; and the coverage probability of the 95% CI was around the nominal value of 0.95. The double robustness property of the AIPW was also demonstrated in the settings of AIPW-mT&pW and AIPW-mW&pT for this simulation study, as the estimation bias was very close to zero. It is interesting to note that when the mean score model was correctly specified, the estimation results for the two AIPW scenarios were almost identical, regardless of whether or not the propensity score was correctly specified. However, when both scores were wrongly specified, the estimation was totally off the mark, resulting in a very large estimation bias. Compared to both IPW and AIPW methods, our proposed model-free estimator of ATE performed much better than the IPW method and almost as well as the AIPW method with the mean score model correctly specified. It was numerically stable, with virtually negligible estimation bias. The Monte-Carlo standard deviation was very small, only slightly bigger than that based on the MLE method, for which the outcome model was completely known to allow maximum likelihood estimation. This indicates that our proposed method only resulted in a minor loss of estimation efficiency. Moreover, the average standard error estimate was very close to the M-C SD even with a sample size of 200, and the coverage probability of 95% CI was also close to the nominal value of 0.95. Therefore, the asymptotic normality theory derived for this proposed estimator in Section 3 is well justified by this numerical experiment, which allows the standard statistical inference procedure to be applied for making causal inferences on the average treatment effect in a finite sample.

5. A Case Study

As one of the most common chronic rheumatic diseases in children, Juvenile Idiopathic Arthritis (JIA) refers to all forms of arthritis that begin before the age of 16 and persist for more than 6 weeks with unknown origin. It is an important cause of short-term and long-term disability and can significantly impact quality of life and mobility for children with this chronic condition. There are many treatment options available for JIA. Results from a randomized controlled trial suggest that early aggressive use of a biological disease-modifying anti-rheumatic drug (DMARD) could be efficacious (Wal-

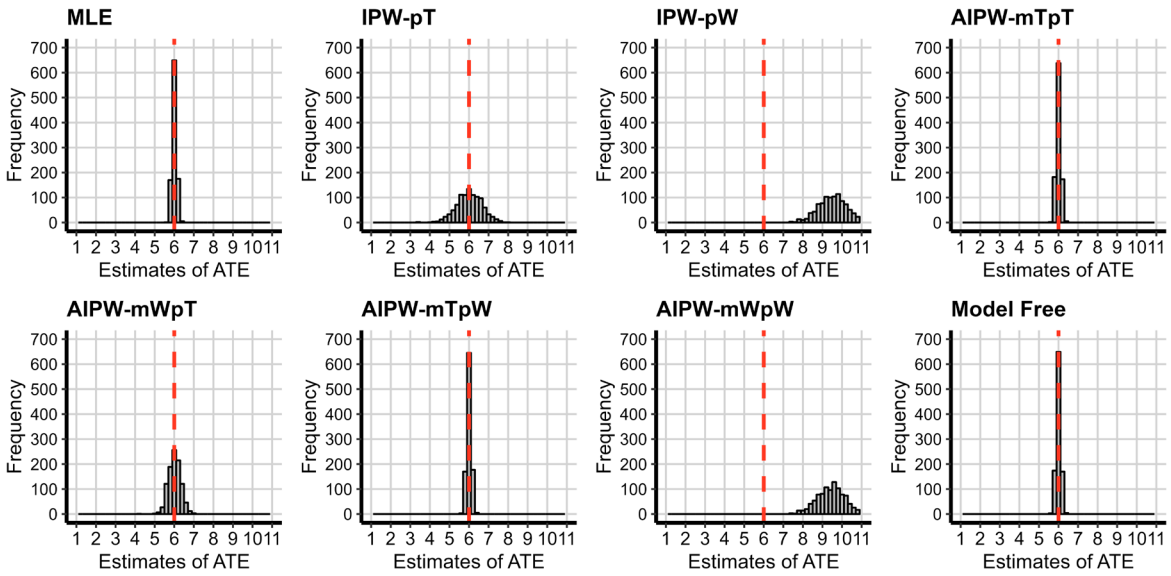

Figure 1. Comparison of the distribution of estimated ATEs in different methods.

lace et al., 2012). But how effective is the early aggressive use of biological DMARDs comparing to step-up consensus treatment plan (starting on a non-biologic DMARD followed by switching to or adding biologic DMARDs) in treating children with newly onset JIA remains an open question. In this case study, we focused on evaluating the effectiveness of early aggressive use of biologic DMARD compared to the step-up plan.

The case study data for this study are available from the authors upon reasonable request. The dataset was derived from a single institute's electronic health records (EHR) between 2009 and 2017. A total of 509 children with a newly diagnosed polyarticular form of JIA (<6 months since diagnosis) met the inclusion criteria, of which 283 patients initiated a step-up treatment plan and 124 received a combination of non-biologic and biologic DMARD. The baseline is defined as the time when the patients initialize their first DMARD. As the primary outcome of interest, cJADAs, a clinical version of the Juvenile Arthritis Disease Activity score (Consolaro et al., 2014), is a summary score derived from three core measures: the physician's global rating of overall disease activity, parent/child ratings of well-being, and counts of active joints. The score ranges from 0 to 30, with a higher score indicating higher disease severity. At 6 months following the initial treatment, defined by the clinical visit falling within the 4 to 9 month window, 327 patients (225 on the step-up plan and 102 on the combination plan) had non-missing cJADAs outcomes, and they were used in this case study. Patients from the two groups are similar in terms of demographic characteristics such as age, gender, race, etc. But the two groups differ in some important clinically relevant variables, such as baseline cJADAs, pain, etc. Patients with severe baseline disease (high cJADSAs) were more likely to be assigned to the early biologic DMARD group. Figure 2 depicts the beneficial effect of early aggressive biological DMARD usage on lowering the cJADAs in this study population: the change of cJADAs from baseline to 6 month visit is -8.4 and -6 for the early combination group and step-up group, respectively.

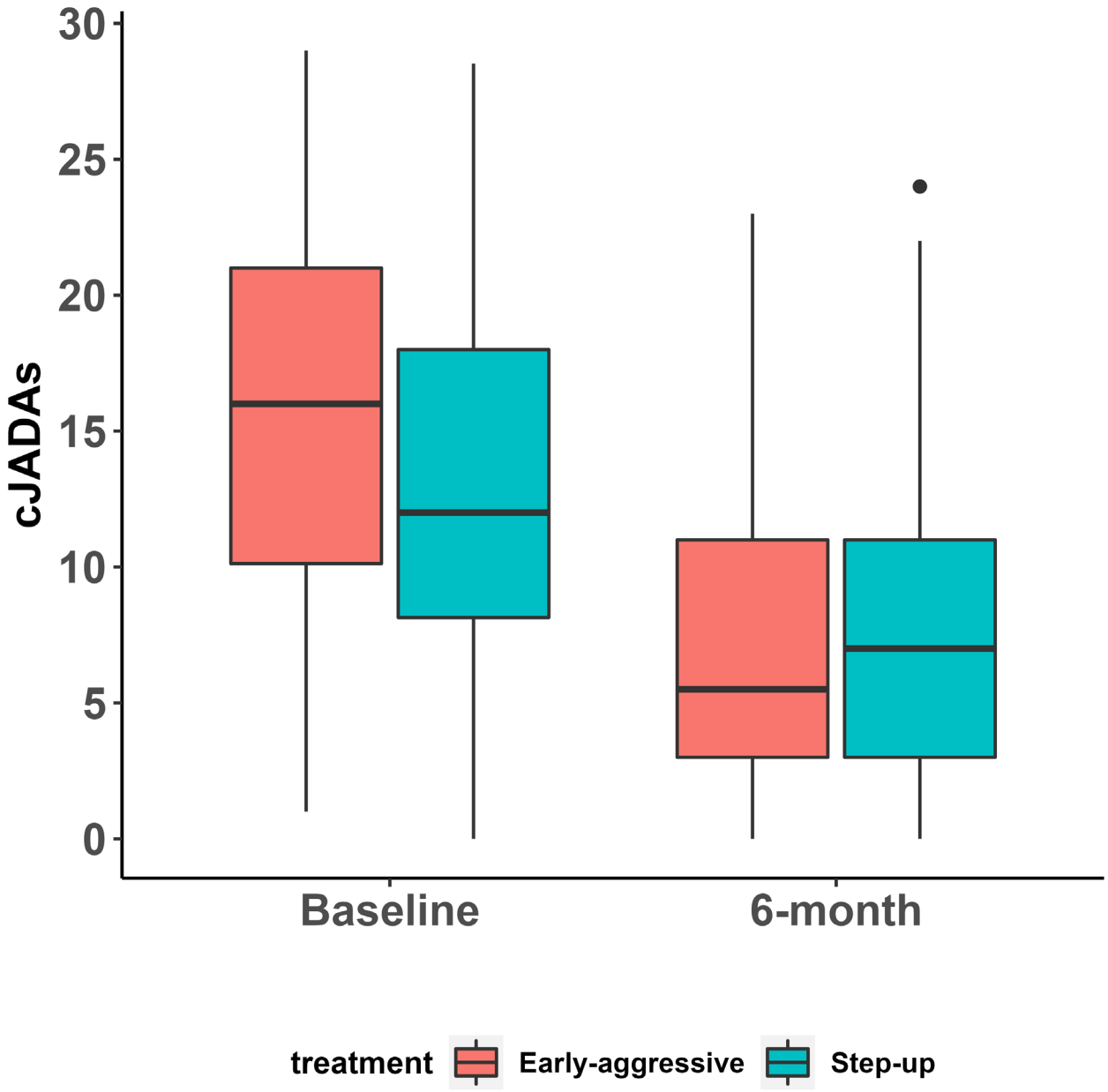

Figure 2. cJADAs at baseline and 6-month follow-up by treatment group.

Table 2. Comparison of the estimated ATEs for the early aggressive use of biologic DMARD.

Method	Estimate	Stand. Error	90%-CI
PLR	-1.362	0.762	(-2.617, -0.109)
IPW	-1.121	0.871	(-2.554, 0.312)
AIPW	-1.090	1.105	(-2.908, 0.728)
MF	-1.396	0.760	(-2.647, -0.145)

To evaluate the ATE of biologic DMARD, we considered two baseline covariates: cJADAs and pain. Pain is the most common symptom of JIA and has been suggested to be linked with disease activity. It is a score ranging from 0 to 10, with 10 indicating the highest severity of pain. The pain score was dichotomized using cutoff 3, which makes about 42% patients fall in the low-pain group with a pain score<3 in this study. Therefore, the baseline cJADAs and binary indicator of low pain take the roles of X_1 and X_2, respectively, as described in the simulation study in Section 4. We used the same estimation procedure as described in the simulation study to estimate both mean and propensity scores, applying cubic B-spline only to baseline cJADAs but including its interaction with low pain in the models. The knot number was chosen to be $q_n = 7$ with the 3 interior knots placed at the first, second, and third quartiles of the observed cJADAs values. For the purpose of comparing the estimated treatment effect with other traditional methods, we also implemented (i) partial linear regression (PLR) assuming outcome model is $Y = f(X_1, X_2) + \alpha A + \epsilon$ where $f(X_1, X_2)$ was modeled by

$$f(X_1, X_2) = \sum_{j=1}^{q_n} \eta_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \eta_j^{(2)} B_j(X_1) X_2,$$

by adopting exactly the same sieve estimation approach as used in the mean score estimation in our proposed method; (ii) IPW with the propensity score estimated by the same spline-based sieve maximum likelihood method as described in the proposed method; (iii) AIPW with the mean and propensity scores estimated by the same regression spline methods as described in the proposed method.

We present the analysis results of the estimated ATE, bootstrapping standard error estimator with 200 samples, and 90% Wald-CI in Table 2. Under the assumption of the homogeneous treatment effect, our proposed model-free method estimated the ATE at -1.396 with 90% CI of (-2.647, -0.145), in contrast to the partial linear regression estimator -1.362 (-2.617,-0.109), IPW -1.121 (-2.554, 0.312), AIPW -1.09 (-2.908, 0.728). The results implied the beneficial effect of early aggressive use of biologic DMARD using the proposed method, as it suggests that early aggressive use of biologic DMARD leads to about a significant 1.4 point reduction in cJADAs 6 months later at 0.1 level in treating children with newly diagnosed pcJIA. The partial linear regression method yielded a similar conclusion. However, both IPW and AIPW resulted in smaller estimated ATEs with larger standard errors, which led to an insignificant effect at 0.1 level.

6. Final Remarks

In this paper, we propose a nonparametric model-free estimation method for causal treatment effect in the spirit of the ordinary least-squares method, which does not need to specify any parametric functional forms for estimating both the mean and propensity scores and hence can be regarded as a robust estimation method. The most appealing feature of the proposed method is that, unlike the IPW and AIPW methods, it does not need to inversely weight the individual propensity scores and hence removes the disastrous impact caused by the estimated propensity scores that are near 0 or 1 associated with extreme observations. This methodological advantage results in numerical stability in estimating the casual treatment effect in comparison to the IPW and AIPW methods. Using empirical process theory, we showed that the proposed estimator is $\sqrt{n}$-consistent with a limiting normal distribution. We demonstrated through the simulation studies that ordinary asymptotic normality theory based inference is valid in our proposed approach for a moderate sample size. In addition to its numerical merit, the proposed method has superior finite-sample statistical properties when compared to the IPW/AIPW methods. In addition, this model-free estimation method is nearly as efficient as the MLE method when the complete stochastic mechanism for outcomes is known, which is, of course, not practically realistic. All these nice features make the proposed method a practically sound approach to making causal inferences for the treatment effect in an observational study.

Though we only considered the scenario of binary treatment for the sake of simplicity in presentation and for because it is the most common situation in biomedical studies, the proposed methodology can be readily extended to a scenario with multiple treatment levels. Suppose we have K treatment levels (a_i for $i = 1, \cdots, K$) in a study, with A indicating the treatment that a subject received. Let $Y(a_i)$ denote the potential outcome associated with treatment level a_i for $i = 1, \cdots, K$. Let a_K chosen as the reference treatment level for comparison, and $\tau_i = E(Y(a_i) - Y(a_K))$ denote the causal treatment effect for treatment level a_i in comparison to the reference level a_K for $i = 1, \cdots, K-1$. Let $\pi_i(X) = pr(A = a_i|X)$ denote the multivariate propensity scores for $i = 1, \cdots, K-1$. In the scenario of homogeneous causal treatment effects, the equation of the expected observed outcome given treatment indicator A and covariate X in a simple treatment-control case (3) can be similarly derived as

$$E(Y|A, X) = m(X) + \sum_{i=1}^{K-1} (1[A = a_i] - \pi_i(X))\, \tau_i. \tag{9}$$

Then the two-stage model-free estimator of $\tau = (\tau_1, \cdots, \tau_{K-1})$ can be explicitly obtained by solving for the least-squares problem

$$\arg\min_{\tau} \sum_{i=1}^{n} \left\{ (Y_i - m(X_i)) - \sum_{j=1}^{K-1} (1[A_i = a_j] - \pi_j(X_i))\tau_j \right\}^2,$$

after obtaining the nonparametric estimates for $m(X)$ and $\pi_i(X)$ for $i = 1, \cdots, K-1$ in the first stage. The asymptotic theory can be similarly developed with more algebraic complexity.

In the literature on the casual treatment effect, ATE was defined as the population

average of the contrast between the potential outcomes. In this paper, we focused on the estimation of ATE for the situation that X contains only confounders for the treatment effect, not effect modifiers, because ATE is not very meaningfully interpretable in practice when X modifies the treatment effect. For example, if a treatment for men has a positive effect of 5 and a negative effect of -5 for women, it would be better to interpret gender-specific treatment effects rather than report no treatment effect for the general population. This said, it would be more useful to study covariate-specific ATE defined by $\tau(X) = E(Y(1) - Y(0)|X)$ if one believes X modifies the treatment effect. The two-stage methodology can also be extended to study covariate-specific ATEs if one wishes to model $\tau(X)$ using a specific functional form such as linear regression or implement nonparametric spline-based sieve estimation for $\tau(X)$. The theoretical development of the parametric model can be similarly done with the techniques adopted in this manuscript. But it would be much more involved for the nonparametric spline-based sieve estimation and is currently under investigation by the authors.

While the proposed two-stage OLS-based estimator enjoys its numerical simplicity and stability, its validity relies on the consistency of the nonparametric estimators of the mean and propensity scores. In this paper, we considered nonparametric spline-based sieve estimators, as their asymptotic properties required for the proposed estimated ATE to have asymptotic normality are easily justified when X is a low-dimensional covariate vector. Our asymptotic theorem, however, is general for any nonparametric estimator as long as the conditions stipulated for the theorem are satisfied. Those conditions are generally weak for nonparametric estimators when X is low dimensional.

For a case with high-dimensional covariates, the spline-based sieve estimation is not only numerically inconvenient, but also possibly loses the desired statistical properties. But it is not just a problem for spline-based sieve estimation; it is a universal problem for any nonparametric estimation method. In a practice with high-dimensional confounders X, we recommend using some modern machine learning methods (McCaffrey et al., 2004; Lee et al., 2010) to estimate the mean and propensity scores in the first stage before applying this proposed method.

Acknowledgments

The authors thank the editor and two anonymous reviewers, whose constructive comments helped improve the presentation of this manuscript.

References

[1] Austin, P.C., and Mamdani, M.M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine,* 25(12), 2084–2106.

[2] Austin, P.C., and Stuart, E.A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine,* 34, 3661–3679.

[3] Bang, H., and Robins, J.M. (2005). Doubly robust estimation in missng data and causal inference models. *Biometrics,* 61(4), 962–973.

[4] Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A,* 169, 571—584.

[5] Cole, S.R., and Hernan, M.A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology,* 168(6), 656–664.

[6] Consolaro A., Negro G., Gallo, M.C., and et al. (2014). Defining criteria for disease activity states in non-systemic juvenile idiopathic arthritis based on a three-variable Juvenile Arthritis Disease Activity Score. *Arthritis Care Res (Hoboken),* 66, 1703–1709.

[7] Gutman, R., and Rubin, D.B. (2017). Estimation of causal effects of binary treatments in unconfounded studies with one continuous covariate. *Statistical Methods in Medical Research,* 26(3), 1199–1215.

[8] Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica,* 66, 315—331.

[9] Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association,* 81, 945–960.

[10] Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement From a finite universe. *Journal of the American Statistical Association,* 47, 663–685.

[11] Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics,* 86(1), 4—29.

[12] Judea, P. (2010). On the consistency rule in causal inference: axiom, definition, assumption, or theorem. *Epidemiology,* 21(6), 872–875.

[13] Kang, D.Y., and Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistic Science,* 22, 523–539.

[14] Lee, B.K., Lessler, J., and Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine,* 29, 337–346

[15] Lunceford, J.K., and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects:a comparative study. *Statistics in Medicine,* 23(19), 2937–2960.

[16] Macaffrey, D.F., Ridgeway, G., and Morral, A.R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods,* 9, 403–425.

[17] Neyman, J.S., Dabrowska, D.M., and Speed, T.P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science,* 5, 465–472.

[18] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology,* 166(5), 688–701.

[19] Rubin, D.B. (1980). Randomization analysis of experimental data: the Fisher randomization test comment. *Journal of the American Statistical Association,* 75, 591–593.

[20] Rubin, D. B. (1986). Comment: What if's have causal answers. *Journal of the American Statistical Association,* 81, 961–962.

[21] Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistic Science,* 5, 472–480.

[22] Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association,* 89, 846-866.

[23] Robinson, P.M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica,* 56, 931–954.

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

[25] Rosenbaum, P.R., and Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician,* 39, 33–38.

[26] Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association,* 82, 387–394.

[27] Scharfstein, D., Rotnitzky, A., and Robins, J.M. (1999). Adjusting for onignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association,* 94, 1096–1120.

[28] Tsiatis, A. (2006). *Semiparametric theory and missing data.* New York: Springer.

[29] van der Vaart, A.W., and Wellner, J.A. (1996). *Weak convergence and empirical processes with applications to statistics.* New York: Springer.

[30] van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge, United Kingdom: Cambridge University Press.

[31] Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research,* 21, 7–30.

[32] Wallace, C. A., E. H. Giannini, S. J. Spalding, P. J., and et al. (2012). Trial of early aggressive therapy in polyarticular juvenile idiopathic arthritis. *Arthritis Rheumatology,* 64 (6), 2012–2021.

[33] Wegman, E.J., and Wright, I.W. (1983). Splines in Statistics. *Journal of the American Statistical Association,* 78, 351–365.

Appendix

Proof of Theorem 1

We write the proposed estimated ATE as follows:

$$\hat{\tau}^{mf} = A_n(\hat{\pi})B_n(\hat{\pi}, \hat{m}),$$

where

$$A_n(\hat{\pi}) = \frac{1}{\mathbb{P}_n[A - \hat{\pi}(X)]^2} \quad \text{and} \quad B_n(\hat{\pi}, \hat{m}) = \mathbb{P}_n[Y - \hat{m}(X)][A - \hat{\pi}(X)].$$

By taking the expectation over (A, X) on (3), the population ATE can also be written as

$$\tau = A_0(\pi)B_0(\pi, m)$$

with

$$A_0(\pi) = \frac{1}{P[A - \pi(X)]^2} \quad \text{and} \quad B_0(\pi, m) = P[Y - m(X)][A - \pi(X)].$$

Straightforward algebra leads to

$$\sqrt{n}(\hat{\tau}^{mf} - \tau) = B_n(\hat{\pi}, \hat{m})\sqrt{n}[A_n(\hat{\pi}) - A_0(\pi)] + A_0(\pi)\sqrt{n}[B_n(\hat{\pi}, \hat{m}) - B_0(\pi, m)]. \quad (10)$$

First, after some algebra, it can be shown that

$$\sqrt{n}[B_n(\hat{\pi}, \hat{m}) - B_0(\pi, m)] = E_n^{(1)} + E_n^{(2)} + E_n^{(3)} + E_n^{(4)} + E_n^{(5)} + E_n^{(6)} + E_n^{(7)},$$

where

$$E_n^{(1)} = \sqrt{n}(\mathbb{P}_n - P)[Y - m(X)][A - \pi(X)], \;\; E_n^{(2)} = -\sqrt{n}(\mathbb{P}_n - P)[Y - m(X)][\hat{\pi}(X) - \pi(X)],$$

$$E_n^{(3)} = -\sqrt{n}P[Y - m(X)][\hat{\pi}(X) - \pi(X)], \;\; E_n^{(4)} = -\sqrt{n}(\mathbb{P}_n - P)[A - \pi(X)][\hat{m}(X) - m(X)],$$

$$E_n^{(5)} = -\sqrt{n}P[A - \pi(X)][\hat{m}(X) - m(X)], \;\; E_n^{(6)} = \sqrt{n}(\mathbb{P}_n - P)[\hat{m}(X) - m(X)][\hat{\pi}(X) - \pi(X)],$$

and

$$E_n^{(7)} = \sqrt{n}P[\hat{m}(X) - m(X)][\hat{\pi}(X) - \pi(X)].$$

Then we immediately obtain that $E_n^{(2)}, E_n^{(4)}$, and $E_n^{(6)}$ are all $o_P(1)$ according to Conditions C1-C3, respectively, and $E_n^{(7)} = o_P(1)$ by C6 and Cauchy-Schwarz inequality.

By taking the conditional expectation of Y and A given X first, respectively, in evaluating $E_n^{(3)}$ and $E_n^{(5)}$, it immediately follows that $E_n^{(3)} \equiv 0$ and $E_n^{(5)} \equiv 0$. Hence, we have

$$\sqrt{n}[B_n(\hat{\pi}, \hat{m}) - B_0(\pi, m)] = \sqrt{n}(\mathbb{P}_n - P)[Y - m(X)][A - \pi(X)] + o_P(1) \to_d N(0, \sigma^2)$$

with $\sigma^2 = var\{(Y - m(X))(A - \pi(X))\}$ by the ordinary central limit theorem, and it follows that $B_n(\hat{\pi}, \hat{m}) \to_p B_0(\pi, m)$.

Second, it can be shown easily that

$$\sqrt{n}\left[\mathbb{P}_n(A - \hat{\pi}(X))^2 - P(A - \pi(X))^2\right] = \sqrt{n}(\mathbb{P}_n - P)(A - \pi(X))^2 + F_n^{(1)} + F_n^{(2)} + F_n^{(3)} + F_n^{(4)},$$

where

$$F_n^{(1)} = -2\sqrt{n}(\mathbb{P}_n - P)\left[(A - \pi(X))(\hat{\pi}(X) - \pi(X))\right],$$

$$F_n^{(2)} = -2\sqrt{n}P\left[(A - \pi(X))(\hat{\pi}(X) - \pi(X))\right],$$

$$F_n^{(3)} = \sqrt{n}(\mathbb{P}_n - P)(\hat{\pi}(X) - \pi(X))^2, \quad \text{and} \quad F_n^{(4)} = \sqrt{n}P(\hat{\pi}(X) - \pi(X))^2.$$

Then, using the same arguments as above, we can show that $F_n^{(1)}$ and $F_n^{(3)}$ are $o_P(1)$ by C4 and C5, respectively; $F_n^{(2)} \equiv 0$ and $F_n^{(4)} = o_P(1)$ by the rate of convergence of $\hat{\pi}(X)$ given in C6. Hence,

$$\sqrt{n}\left[\mathbb{P}_n(A - \hat{\pi}(X))^2 - P(A - \pi(X))^2\right] = \sqrt{n}(\mathbb{P}_n - P)(A - \pi(X))^2 + o_P(1) \to_d N(0, \tilde{\sigma}^2)$$

with $\tilde{\sigma}^2 = var\{[A - \pi(X)]^2\}$, and it follows that $\mathbb{P}_n(A - \hat{\pi}(X))^2 \to_p P(A - \pi(X))^2$.

By the ordinary multivariate central limit theorem, it is easily seen

$$\sqrt{n}\begin{pmatrix} (\mathbb{P}_n - P)[Y - m(X)][A - \pi(X)] \\ (\mathbb{P}_n - P)[A - \pi(X)]^2 \end{pmatrix} \to_d N(\mathbf{0}, \mathbf{\Sigma}),$$

where

$$\mathbf{\Sigma} = Cov\begin{pmatrix} [Y - m(X)][A - \pi(X)] \\ [A - \pi(X)]^2 \end{pmatrix}$$

is the variance-covariance matrix of the vector of the two entries, and hence,

$$\begin{pmatrix} \sqrt{n}\left[B_n\left(\hat{\pi}, \hat{m}\right) - B_0(\pi, m)\right] \\ \sqrt{n}\{\mathbb{P}_n[A - \hat{\pi}(X)]^2 - P[A - \pi(X)]^2\} \end{pmatrix} \to_d N(\mathbf{0}, \mathbf{\Sigma}). \tag{11}$$

Finally, by noting that

$$\sqrt{n}[A_n(\hat{\pi}) - A_0(\pi)] = \frac{1}{\mathbb{P}_n[A-\hat{\pi}(X)]^2 P[A-\pi(X)]^2}[-\sqrt{n}\{\mathbb{P}_n[A-\hat{\pi}(X)]^2 - P[A-\pi(X)]^2\}]$$

and

$$\left(A_0(\pi), \frac{-B_n(\hat{\pi},\hat{m})}{\mathbb{P}_n[A-\hat{\pi}(X)]^2 P[A-\pi(X)]^2}\right) \to_p \left(A_0(\pi), \frac{-B_0(\pi,m)}{(P[A-\pi(X)]^2)^2}\right) \tag{12}$$

based on the arguments above, it immediately follows that

$$\begin{aligned}\sqrt{n}(\hat{\tau}^{mf} - \tau) =& B_n(\hat{\pi},\hat{m})\sqrt{n}[A_n(\hat{\pi}) - A_0(\tau)] + A_0(\pi)\sqrt{n}[B_n(\hat{\pi},\hat{m}) - B_0(\pi,m)] \\ &\to_d N(0, H\boldsymbol{\Sigma}H^\top)\end{aligned}$$

with

$$H = \left(A_0(\tau), \frac{-B_0(\pi,m)}{(P[A-\pi(X)]^2)^2}\right)$$

by (11) and (12), along with Slutsky's Theorem and the fact that a linear transformation of a multivariate normal is still a normal. This completes the proof of Theorem 1. □

In what follows, we demonstrate that, under some weak assumptions, the B-spline-based sieve nonparametric estimators for the mean and propensity scores satisfy the conditions required by Theorem 1 for the situation considered in both our simulation and case studies.

Initially, we consider the more general case of additive mean and propensity score models of the form

$$m(X) = \theta_0(X) = \sum_{j=1}^{k} \theta_{0,j}(X_j)$$

and

$$g[\pi(X)] = \phi_0(X) = \sum_{j=1}^{k} \phi_{0,j}(X_j)$$

where $X \in \mathbb{R}^k$, and $\theta_{0,j}$ and $\phi_{0,j}$ are sufficiently smooth functions of the jth component of X, and g is a known link function. The B-spline sieve spaces for the unspecified functions $\theta_{0,j}$, $j = 1, \ldots, k$, have the form

$$\Theta_{j,n} = \left\{\theta_j(x) = \sum_{l=1}^{q_{j,n}} \alpha_{j,l} B_l(x) : \alpha_{j,l} \in \mathbb{R}, l = 1, \ldots, q_{j,n}\right\}, \quad j = 1, \ldots, k$$

where $B_l(x)$ is the prespecified B-spline basis function and $\alpha_{j,l}$ the spline coefficient for $l = 1, \ldots, q_{j,n}$ and $q_{j,n} = N_{j,n} + m_j$, with $N_{j,n}$ being the number of interior knots that depends on the total sample size n and m_j is the order of the B-spline. Similarly, the B-spline sieve spaces for the unspecified functions $\phi_{0,j}$, $j = 1, \ldots, k$, have the form

$$\Phi_{j,n} = \left\{ \phi_j(x) = \sum_{l=1}^{p_{j,n}} \gamma_{j,l} B_l(x) : \gamma_{j,l} \in \mathbb{R}, l = 1, \ldots, p_{j,n} \right\}.$$

The asymptotic properties of the proposed estimator are studied under the following set of regularity conditions:

D1. The erros e have a zero mean and $E|e|^l < \infty$ for $l \geq 3$.

D2. The covariates X are bounded in the sense that $P(\|X\| \leq C) = 1$ for some $C \in (0, \infty)$. Also, e and X are independent. Moreover, $E(XX^T)$ is non-singular.

D3. The infinite-dimensional parameters have the form $\theta_0 = \sum_{j=1}^{k} \theta_{0,j}$ and $\phi_0 = \sum_{j=1}^{k} \phi_{0,j}$ with $\theta_{0,j} \in \Theta_j$ and $\phi_{0,j} \in \Phi_j$ for $j = 1, \cdots, k$. Moreover, the corresponding parameter spaces $\Theta_1, \ldots, \Theta_k, \Phi_1, \ldots, \Phi_k$ contain uniformly bounded functions, with bounded p_{θ_j}th and p_{ϕ_j}th derivative, $j = 1, \ldots, k$, for fixed $p_{\theta_j}, p_{\phi_j} \geq 1$, and with the first derivatives being continuous.

D4. The inverse of the link function for the propensity score model is continuously differentiable on compacts.

Before checking the conditions of Theorem 1 we provide a proof for a useful lemma, which will be used later.

Lemma 6.1. $\sqrt{n}(\mathbb{P}_n - P)[Y - m(X)]\,[\hat{\pi}(X) - \pi(X)] = o_P(1)$.

Proof. Corollary 19.35 in van der Vaart (1998) is used for the proof, and the conditions for this corollary are first discussed in what follows.

For $j = 1, \cdots, k$, let

$$\mathcal{F}_j = \left\{ \phi_j \in \Phi_{j,n} : \|\phi_j - \phi_{0,j}\|_{L_2(P)} \leq c n^{-p_{\phi_j}/(2p_{\phi_j}+1)} \right\}.$$

We define the function

$$\tilde{f}_{\boldsymbol{\phi}}(x, y) = \{y - m(x)\}(\pi_{\boldsymbol{\phi}}(x) - \pi_{\boldsymbol{\phi}_0}(x)) = \{y - m(x)\}[g^{-1}(\boldsymbol{\phi}(x)) - g^{-1}(\boldsymbol{\phi}_0(x))],$$

where $\boldsymbol{\phi} = \{\phi_j\}_{j=1}^{k}$ and $\boldsymbol{\phi}_0 = \{\phi_{0,j}\}_{j=1}^{k}$. And we let

$$\mathcal{F} = \left\{ \tilde{f}_{\boldsymbol{\phi}} : \phi_j \in \mathcal{F}_j, j = 1, \cdots, k \right\}.$$

By the fact that both $\phi_{0,j}$ and the spline function ϕ_j have a uniformly bounded derivative and X is continuous within a compact set and has a bounded density function, it

can be shown that for $\phi_j \in \mathcal{F}_j$

$$\|\phi_j - \phi_{0,j}\|_\infty \leq cn^{-2p_{\phi_j}/(6p_{\phi_j}+3)}.$$

Hence,

$$\|\hat{\pi} - \pi\|_\infty \leq cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)},$$

with $p_{\boldsymbol{\phi}} = \min\{p_{\phi_j} : j = 1, \cdots, k\}$. It follows that

$$|\tilde{f}_{\boldsymbol{\phi}}(x, y)| \leq cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)}|y - m(x)|.$$

So $cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)}|y - m(x)|$ is the envelop function for $\mathcal{F}$ with

$$\left\| cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)}|Y - m_0| \right\|_{L_2(P)} \leq cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)},$$

since $P\{Y - m(X)\}^2$ is bounded by the regularity conditions.

On the other hand, we know

$$N_{[\,]}(\epsilon, \mathcal{F}_j, \|\cdot\|_\infty) \leq \left\{ \frac{cn^{-p_{\phi_j}/(2p_{\phi_j}+1)}}{\epsilon} \right\}^{cn^{1/(2p_{\phi_j}+1)}}.$$

Then, with some algebra, we can show that

$$N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{L_2(P)}) \leq \left\{ \frac{cn^{-p_{\boldsymbol{\phi}}/(2p_{\boldsymbol{\phi}}+1)}}{\epsilon} \right\}^{cn^{1/(2p_{\boldsymbol{\phi}}+1)}}.$$

Now, by Corollary 19.35 in van der Vaart (1998), we have

$$\begin{aligned} E_P\|\mathbb{G}_n\|_{\mathcal{F}} &\leq cJ_{[\,]}\left\{cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)}, \mathcal{F}, L_2(P)\right\} \\ &= \int_0^{cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)}} \sqrt{1 + \log N_{[\,]}\{\epsilon, \mathcal{F}, L_2(P)\}}d\epsilon \\ &\leq \int_0^{cn^{-2p_{\boldsymbol{\phi}}/(6p_{\boldsymbol{\phi}}+3)}} cn^{1/(4p_{\boldsymbol{\phi}}+2)}n^{-p_{\boldsymbol{\phi}}/(4p_{\boldsymbol{\phi}}+2)}\epsilon^{-1/2}d\epsilon = cn^{\frac{3-5p_{\boldsymbol{\phi}}}{12p_{\boldsymbol{\phi}}+6}}. \quad (13) \end{aligned}$$

It is known that for a regression spline estimator,

$$\left\{P(\hat{\phi}_j - \phi_{0,j})^2\right\}^{1/2} = O_P\left(n^{-p_{\phi_j}/(2p_{\phi_j}+1)}\right),$$

for $j = 1, \cdots, k$. Then, by virtue of the fact that k is finite, for any small $\epsilon_0 > 0$ and any positive integer n, we can find a $M > 0$ such that

$$\Pr\left[\left\{P(\hat{\phi}_j - \phi_{0,j})^2\right\}^{1/2} \leq Mn^{-p_{\boldsymbol{\phi}}/(2p_{\boldsymbol{\phi}}+1)}, j = 1, \cdots, k\right] > 1 - \epsilon_0. \quad (14)$$

If $\left\{P(\hat{\phi}_j - \phi_{0,j})^2\right\}^{1/2} \leq Mn^{-p_\phi/(2p_\phi+1)}$ for $j = 1, \cdots, k$, then $\tilde{f}_{\hat{\boldsymbol{\phi}}} \in \mathcal{F}$ with $\hat{\boldsymbol{\phi}} = \{\hat{\phi}_j\}_{j=1}^k$. So we know that as $n \to \infty$

$$E_P\left[\left|\mathbb{G}_n\left(\tilde{f}_{\hat{\boldsymbol{\phi}}}\right)\right| : \left\{P(\hat{\phi}_j - \phi_{0,j})^2\right\}^{1/2} \leq Mn^{-p_\phi/(2p_\phi+1)}, j = 1, \cdots, k\right] \leq E_P\|\mathbb{G}_n\|_{\mathcal{F}}$$
$$\leq c_M n^{\frac{3-5p_\phi}{12p_\phi+6}} \to 0,$$

where the last inequality is by (13). Then conditional Markov's inequality implies that, for any small ϵ_1 and ϵ_2, there exists an integer $N > 0$ such that for $n > N$ we have

$$Pr\left[\left|(\mathbb{P}_n - P)\left(\frac{\tilde{f}_{\hat{\boldsymbol{\phi}}}}{n^{-1/2}}\right)\right| < \epsilon_1 : \left\{P(\hat{\phi}_j - \phi_{0,j})^2\right\}^{1/2} \leq Mn^{-p_\phi/(2p_\phi+1)}, j = 1, \cdots, k\right]$$
$$> 1 - \epsilon_2.$$

Now by (14) and the definition of conditional probability we have for $n > N$

$$\Pr\left[\left|(\mathbb{P}_n - P)\left(\frac{\tilde{f}_{\hat{\boldsymbol{\phi}}}}{n^{-1/2}}\right)\right| < \epsilon_1\right] > (1 - \epsilon_2)(1 - \epsilon_0),$$

which implies

$$\sqrt{n}(\mathbb{P}_n - P)\left(\tilde{f}_{\hat{\boldsymbol{\phi}}}\right) = o_P(1),$$

or

$$\sqrt{n}(\mathbb{P}_n - P)[Y - m(X)]\left[\pi_{\hat{\boldsymbol{\phi}}}(X) - \pi_{\boldsymbol{\phi}_0}(X)\right] = o_P(1),$$

where $\pi_{\hat{\boldsymbol{\phi}}}(X) \equiv \hat{\pi}(X)$ and $\pi_{\boldsymbol{\phi}_0}(X) \equiv \pi(X)$. □

First, by Lemma 1, we know that condition C1 of Theorem 1 is satisfied. Also, using very similar arguments to those used in the proof of Lemma 1, it can be shown that conditions C2–C5 of Theorem 1 are also satisfied. By the convergence rate calculation in the proof of Lemma 1, it follows that, even under the minimal degree of smoothness allowed by condition D3, we have that $\|\hat{m} - m\|_{L_2(P)} = O_p\left(n^{-1/3}\right)$ and $\|\hat{\pi} - \pi\|_{L_2(P)} = O_p\left(n^{-1/3}\right)$. Therefore, condition C6 of Theorem 1 is also satisfied. Thus, the model-free ATE estimator is $\sqrt{n}$-consistent and asymptotically normal, with the variance given by Theorem 1.

Finally, consider the propensity score and mean models used in the simulation and case studies. These models have the form:

$$m(X) = \theta_0(X) = \theta_{0,1}(X_1) + \theta_{0,2}(X_1)X_2$$

and

$$\text{logit}[\pi(X)] = \phi_{0,1}(X_1) + \phi_{0,2}(X_1)X_2$$

where X_2 is a Bernoulli random variable. For the above parameters, we considered B-spline sieve spaces

$$\Theta_{j,n} = \left\{ \theta_j(x) = \sum_{l=1}^{q_{j,n}} \alpha_{j,l} B_l(x) : \alpha_{j,l} \in \mathbb{R}, l = 1, \ldots, q_{j,n} \right\}, \quad j = 1, 2,$$

and

$$\Phi_{j,n} = \left\{ \phi_j(x) = \sum_{l=1}^{p_{j,n}} \gamma_{j,l} B_l(x) : \gamma_{j,l} \in \mathbb{R}, l = 1, \ldots, p_{j,n} \right\} \quad j = 1, 2.$$

Using very similar arguments as those used for the case of additive models above, it can be shown that conditions C1–C6 of Theorem 1 are satisfied. Therefore, the corresponding model-free ATE estimator is $\sqrt{n}$-consistent and asymptotically normal, with the variance given by Theorem 1.