A Robust Estimation Approach for Conditional Marginal Effects with Binary Treatments

Li Zheng* and Weiwen Yin†

Abstract

This paper proposes a nonparametric approach to estimate conditional marginal effects based on a class of orthogonal estimators (OEs) that can handle high-dimensional data and require no assumption for model specification. In addition to introducing the existing OE built on Augmented Inverse Propensity Weighting (OE-AIPW), we create another OE built on Conditional Average Treatment Effect (OE-CATE) that shares alike statistical properties and can improve estimation performance under certain data generating processes. Our approach involves a two-stage procedure in which the first-stage yields an orthogonal signal with machine learning algorithms being used to handle high-dimensional confounders, and the second-stage proceeds with a nonparametric regression of the orthogonal signal on the moderators. We establish theoretical properties including consistency, asymptotic normality, and efficiency, for OE-CATE. Monte Carlo simulations demonstrate the theoretical insights and guide the practical application of our approach. We replicate two empirical studies to illustrate the application of OEs in two important scenarios: discontinuous propensity scores and randomized experiments.

*Assistant Professor, Institute for Economic and Social Research, Jinan University.

†Corresponding author. Assistant Professor, Department of Government and Public Administration, University of Macau. Email: wwyin@um.edu.mo.

1 Introduction

Social scientists from various disciplines are increasingly interested in estimating conditional marginal effects of a treatment variable. For example, Huddy, Mason and Aarøe (2015) find that "[s]trongly identified partisans feel angrier than weaker partisans when threatened with electoral loss and more positive when reassured of victory" (abstract). In this context, the treatment effect of loss threat on the level of anger is moderated by the strength of partisan identity. Conventionally, multiplicative interaction models are widely used as the standard estimation strategy for modeling treatment effect heterogeneity (Brambor, Clark and Golder, 2006). However, recent advancement in research methods reveals multiple significant limitations of this framework and proposes alternative approaches, such as the varying coefficient model (Hainmueller, Mummolo and Xu, 2019). These alternative estimation strategies are more flexible, and thus are more adept at capturing complex data generating process (DGP) that clearly violates the assumptions required for multiplicative interaction models.

In this paper, we propose an even more flexible approach called Orthogonal Estimator (OE), which is more robust than many existing methods, to estimate conditional marginal effects. Our proposed estimation strategy has the following advantages. First and foremost, OE is free from any functional form restrictions being imposed on existing methods. Thus, it is robust to a wide range of DGP and is guarded against the risk of model misspecification. Second, it enjoys favorable statistical properties such as consistency, asymptotic normality and efficiency, and allows for valid statistical inference. Last but not least, it can handle high-dimensional data effectively by incorporating cutting-edge development in machine learning.

Here, we provide a concise overview of our proposed estimation strategy. We introduce the Orthogonal Estimators (OEs), where "orthogonal" refers to a statistical technique used to handle the bias of first-stage estimates known as nuisance functions. Of the two OEs we cover, one is the Orthogonal Estimator based on Augmented Inverse Propensity Weighting (OE-AIPW), which has been well established in statistics and econometrics but remained largely unnoticed by political scientists, see e.g., Lee, Okui and Whang (2017); Semenova and Chernozhukov (2021); Knaus (2022);

Kennedy (2023), etc. In addition, we propose a new Orthogonal Estimator based on Conditional Average Treatment Effect (OE-CATE), which may perform even better than OE-AIPW in certain empirical setups. One such circumstance is when the propensity score is discontinuous, which corresponds to the (fuzzy) regression discontinuous design. Both OEs are two-stage estimators where the first stage involves constructing an *orthogonal signal*, and the second stage is a nonparametric sieve regression of the orthogonal signal on the moderators. The orthogonal signal delivers a high-quality proxy for conditional marginal effect that is insensitive to biases from estimation of nuisance functions. The OEs also make use of an important technique called *cross-fitting*, which involves estimating the nuisance functions and calculating their predicted values using different subsets of the dataset, in order to establish their statistical properties.

The two orthogonal estimators differ in the way of constructing the orthogonal signal. The existing OE-AIPW uses the score in the Augmented Inverse Propensity Weighting estimator (Robins, Rotnitzky and Zhao, 1994; Glynn and Quinn, 2010) as the orthogonal signal, which involves estimating the nuisance functions including the propensity score and the outcome regression function for both treatment and control groups. The OE-CATE originally proposed by us, on the other hand, constructs the orthogonal signal based on conditional average treatment effects, along with the propensity score and outcome regression function. It has been show in the literature that OE-AIPW enjoys the favorable statistical properties of consistency, asymptotic normality (Semenova and Chernozhukov, 2021) and oracle efficiency (Kennedy, 2023). For OE-CATE, we first prove that its orthogonal signal is equivalent to that of OE-AIPW, and that it is indeed Neyman-orthogonal. Next, we show that OE-CATE preserves the same statistical properties–consistency, asymptotic normality, and oracle efficiency–as the OE-AIPW does, due to the Neyman-orthogonality of its signal.

More generally, there is a growing literature on how to construct orthogonal moment functions in a unified framework (Hirshberg and Wager, 2021; Chernozhukov, Newey and Singh, 2022).[1] In the context of binary treatment effect, such construction leads to OE-AIPW. Our OE-CATE method adds to this literature by offering an alternative orthogonal signal for treatment effect estimation,

[1]We thank an anonymous reviewer for pointing out this literature and the connection with this paper.

and provide theoretical underpinnings for the conditional marginal effect estimator using sieve regression in the second-stage.

To illustrate the advantages of OEs, we conduct Monte-Carlo simulation exercises against a vast array of DGPs. We compare the performance of the two OEs, the multiplicative interaction models, and the varying coefficient model advocated by Hainmueller, Mummolo and Xu (2019). Our simulation results confirm that the OEs can well approximate the conditional marginal effects under random experiments. Moreover, our proposed OE-CATE demonstrates finite sample improvement over OE-AIPW when dealing with discontinuous propensity score.

Next, we proceed to replicate two political science studies which utilize conventional estimation strategies, the (fuzzy) regression discontinuity design and multiplicative interaction models, respectively, with our proposed method. The first replication results suggest that OE-CATE is better suited in capturing the conditional marginal effects when the continuity assumption for a valid regression discontinuity design is violated. The second replication indicates that the OEs can generate estimates consistent with multiplicative interaction models when the latter's required assumption of linear interaction effect is plausibly met. In other words, we are not developing a replacement to existing methods in this paper, but offer a complement that can be used in conjunction with existing methods to provide more robust and plausible estimates of conditional marginal effects. We do not claim that our approach is superior in all situations, thus researchers should carefully consider which method is most appropriate for their specific situation.

The rest of this paper is organized as follows: In the next section, we define the parameters of interest and briefly review the existing methods to estimate conditional marginal effects. The paper proceeds to provide theoretical underpinnings of the two OEs. Monte-Carlo simulations and replication studies are then adopted to illustrate the usefulness of the two OEs in real-world scenarios. In the appendix, we provide a formal proof of the theoretical propositions, technical details for implementation, and a user's guide for applying the OEs using `R`.

2 The Conditional Marginal Effect

In this section, we first formally define the conditional marginal effects via the potential outcome framework (Rubin, 1974). Next, we review the existing prevailing approaches for estimating the conditional marginal effects, which can be classified into two broad categories: parametric and semiparametric approaches. We discuss their usefulness and potential limitations, which motivate our proposal of orthogonal estimators to be elaborated Section 3.

2.1 Definition

Suppose we have n independent and identically distributed observations labeled $i = 1, 2, \cdots, n$. Let $W_i \in \{0, 1\}$ be a binary treatment such that $W_i = 1$ when individual i is treated and $W_i = 0$ otherwise. Under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980),[2] we can denote $Y_i^{(1)}$ and $Y_i^{(0)}$ as the potential outcome of i with and without receiving the treatment, respectively. Thus, the observed outcome is $Y_i = W_i Y_i^{(1)} + (1 - W_i) Y_i^{(0)}$. Let $X_i \in \mathcal{X}$ be the moderators of interest for some theoretical reasons, and $Z_i \in \mathcal{Z}$ be other control variables.[3] We denote the full set of covariates as $V_i = (X_i, Z_i) \in \mathcal{V} \subseteq \mathbb{R}^p$. Hence, the observables are $O_i = (Y_i, W_i, V_i)$.

The treatment effect (or marginal effect) of individual i is $\xi_i = Y_i^{(1)} - Y_i^{(0)}$, and the conditional average treatment effect is given as:[4]

$$\tau_0(v) = \mathbb{E}[\xi_i | V_i = v]. \tag{2.1}$$

Our goal is to learn how the moderators X affect the marginal effect of the treatment variable.

[2] It states that the potential outcomes are not interfered by other individuals, and there are no hidden variations in treatments.

[3] The Causal Forest algorithm we utilize in constructing the OE-CATE can reveal important moderators in a data-driven manner.

[4] Throughout the paper, for some generic function $f(\cdot)$, we use $f_0(\cdot)$ to denote its true value and $\widehat{f}(\cdot)$ for its estimator.

Thus, the *conditional marginal effect of* W *on* Y *moderated by* X can be defined as

$$g_0(x) = \mathbb{E}[\xi_i | X_i = x]. \tag{2.2}$$

While $g_0(x)$ and $\tau_0(v)$ are both expectations of treatment effects conditional on a set of covariates, throughout this paper we use $g_0(x)$ to denote the conditional marginal effect which is our main interest, and $\tau_0(v)$ for the conditional average treatment effect, which is conditioned on all observed covariates.

2.2 Parametric Approach

The parametric approach typically models the relationship among the outcome, treatment, moderators, and control variables within some specific functional form. The most widely used framework is a linear regression augmented with an interaction term between the treatment and moderator. For example, Brambor, Clark and Golder (2006) consider the classical multiplicative interaction model:

$$Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 W_i + \beta(W_i \cdot X_i) + Z_i'\gamma + \epsilon_i, \tag{2.3}$$

where the coefficient of the interaction term β captures the moderating effect of X. This model and its variants have been popular in applied research in social sciences because of its ease of implementation and interpretation. However, as noted by Hainmueller, Mummolo and Xu (2019), this approach is subject to strong assumptions including *linear interaction effects* and *common support*. Violation of linear interaction effects renders the estimated conditional marginal effect inconsistent and biased, while lack of common support may give rise to misleading estimated marginal effects because interpolation or extrapolation is sensitive to the assumed model specification. Although more sophisticated parametric functional forms are available, the issue of linear interaction effects and common support remains due to the fundamental risk of model misspecification.

2.3 Semiparametric Approach

The semiparametric approach partially relaxes the functional form restrictions of the parametric approach by allowing part of the model to be nonparametric, while retaining parametric constraints on the remaining parts. For example, Hainmueller, Mummolo and Xu (2019) considers a varying coefficient model (Li and Racine, 2010):

$$Y_i = f(X_i) + g(X_i)W_i + Z_i'\gamma(X_i) + \epsilon_i. \tag{2.4}$$

In this model, the intercept $f(x)$, the treatment effect $g(x)$, and coefficients on controls $\gamma(x)$ are all allowed to be nonparametric functions of the moderator X. In particular, the flexibility on $g(x)$ admits a nonlinear estimation of the conditional marginal effect, which provides richer information of the treatment effect than the parametric approach does. Compared with multiplicative interaction model (2.3), the varying coefficient model (2.4) is more robust to the violation of the linear interaction effects and common support assumptions. Hainmueller, Mummolo and Xu (2019) propose estimating (2.4) using a local linear kernel regression.

Although (2.4) offers more flexibility on the regression function specification, it still imposes subtle restrictions: (1) the outcome regression is additive in the explanatory variables (with the varying coefficients) and the idiosyncratic shock; and (2) the coefficients $f(\cdot), \gamma(\cdot)$ and the treatment effect $g(\cdot)$ are only allowed to vary along X. In Appendix D we use simulation to show that semiparametric approaches could be biased under model misspecifications, particularly when there are other moderators of the treatment effect in addition to X. Semiparametric approaches are a compromise between functional form flexibility and high dimensionality of the control variables: if one wants to include many control variables, then traditional nonparametric methods (e.g., kernel or sieve) will suffer from the curse of dimensionality. In summary, the restrictions on model specifications of parametric and semiparametric estimation strategies call for a fully flexible nonparametric approach in the presence of a large number of control variables.

3 Orthogonal Estimators

We introduce a class of nonparametric estimators for conditional marginal effect, the orthogonal estimators (OEs). The OEs are a two-stage scheme where the first stage involves constructing an orthogonal signal to deliver a high-quality proxy for conditional marginal effect, and the second stage is a nonparametric regression of the orthogonal signal on the moderators to obtain estimates for conditional marginal effects. This class of estimators has the following strengths. First, they are fully nonparametric, and thus are robust to model specifications as opposed to the parametric and semiparametric approaches. Second, the OEs enjoy favorable statistical properties including consistency, asymptotic normality and efficiency, and allows for valid statistical inference. While the theoretical underpinnings for OE-AIPW have been established in the literature, in this paper we formally prove them for the OE-CATE estimator. Finally, they are able to handle high-dimensional control variables without functional form specification; remarkably, the OEs are insensitive to the estimation errors from the first stage which enhances our confidence of using machine learning algorithms

In this section, we first present the identification assumptions for the OEs. These assumptions, which are commonly required in the potential outcome framework, do not rely on functional form specifications for the regression function. Second, we introduce two OEs with different expressions of the orthogonal signal: one based on the Augmented Inverse Propensity Weighting (AIPW) and the other one based on the Conditional Average Treatment Effect (CATE). While OE-AIPW has been well established and widely used in econometrics and statistics, OE-CATE is new to the literature. We show that OE-CATE possesses the same favorable statistical properties as OE-AIPW, and it may outperform OE-AIPW under certain situations. Finally, we outline the algorithm for implementing both OEs.

3.1 Identification Assumptions

We impose the following assumptions to identify the conditional marginal effects, which are commonly required in the potential outcome framework.

Assumption 1 (Unconfoundedness). $(Y_i^{(1)}, Y_i^{(0)}) \perp W_i | V_i = v, \forall v \in \mathcal{V}$.

Assumption 1 posits that we have observed all the covariates that jointly affect the potential outcomes and treatment. It lends us the Neyman-orthogonality property of the orthogonal signal to be defined later, which is a necessity for the OEs to maintain the desired statistical properties. Unconfoundedness is a standard requirement to achieve identification in causal inference. Though not directly testable, its plausibility could be enhanced by including a large number of confounding variables V_i (Imbens and Rubin, 2015). This motivates us to leverage machine learning methods to handle the high dimensionality of control variables in the first stage of OE, in order to accommodate as many potential confounders as possible.

Assumption 2 (Overlapping). $0 < c \leq \mathbb{P}(W_i = 1 | V_i = v) \leq 1 - c < 1, \forall v \in \mathcal{V}$ *for some* $c > 0$.

Assumption 2 requires that for each value of the observed covariates, there is a non-zero probability of assignment to the treatment and control group. For OEs, the need for overlapping assumption is apparent because the assignment probabilities of treated and untreated group will appear in the denominator of the orthogonal signal. Hainmueller, Mummolo and Xu (2019) carefully illustrates the importance of overlapping and the jeopardy of its violation (they refer to it as the *common support* assumption).

Our proposed method requires no further identification assumption other than unconfoundedness and overlapping. For example, no model specification assumption is needed as it is fully nonparametric. Thus, our estimator is characterized by minimal identification assumptions, which enhances its applicability and robustness across diverse settings.

3.2 Orthogonal Estimator Based on Augmented Inverse Propensity Weighting (OE-AIPW)

The OEs are a two-stage estimation scheme. The first-stage is to construct an *orthogonal signal* $\Lambda(\eta)$, which is a proxy for the conditional marginal effect $g_0(x)$. The construction of $\Lambda(\eta)$ involves some *nuisance functions* $\eta = \eta(v)$ that needs to be estimated. The second-stage is to nonparametrically regress $\Lambda(\eta)$ on the moderators X to obtain an estimate for $g_0(x)$. Notably, this two-stage

scheme is implemented using *cross-fitting*. In this subsection, we will introduce the orthogonal signal for the OE-AIPW estimator as well as the the cross-fitting procedure, and then discuss its statistical properties.

3.2.1 Orthogonal Signal

A high-quality proxy $\Lambda(\eta)$ for $g_0(x)$ is crucial for a good estimator of conditional marginal effect. To obtain a good $\Lambda(\eta)$ in the first-stage, there are a couple of concerns. First, we want to allow for a nonparametric estimation of $\eta(v)$ so that it is robust to functional form specifications on the nuisance functions $\eta(v)$. Additionally, we want to include a rich set of control variables V rendering the unconfoundedness assumption more likely to hold. Therefore, machine learning algorithms are naturally recommended to approximate $\eta(v)$, which enable us to enjoy their model flexibility and prediction power with high dimensional covariates.

However, it is well-known that machine learning methods perform well by employing regularization to reduce variance, which raises concerns that the second-stage estimation could inherit the first-stage regularization bias. Fortunately, this concern could be alleviated if $\Lambda(\eta)$ is *Neyman-orthogonal*, that is, if $\Lambda(\eta)$ is an *orthogonal signal*. Roughly speaking, an orthogonal signal is insensitive to small errors on the approximation of nuisance functions $\eta(v)$. As a result, we are more confident in using machine learning methods in the first-stage without worrying that the final estimator of conditional marginal effect is biased. Formally, $\Lambda(\eta)$ is *Neyman-orthogonal* if

$$\frac{\partial}{\partial r}\mathbb{E}[\Lambda(\eta_0 + r(\eta - \eta_0))|V = v]\bigg|_{r=0} = 0, \text{ for all } v \text{ and } \eta. \tag{3.1}$$

The OE-AIPW estimator uses the following orthogonal signal:

$$\Lambda^{\text{AIPW}}(\eta) = \frac{WY}{e(V)} - \frac{(1-W)Y}{1-e(V)} - \frac{W - e(V)}{e(V)(1-e(V))}\Big[(1-e(V))\mu(1,V) + e(V)\mu(0,V)\Big], \tag{3.2}$$

where $\eta = \eta(v) = (\mu(1,v), \mu(0,v), e(v))$ are the nuisance functions with $\mu(w,v) = \mathbb{E}[Y|W = w, V = v]$ being the *treatment-specific outcome regression function*, and $e(v) = \mathbb{P}[W = 1|V = v]$ being the probability of receiving the treatment, or the *propensity score*. Equivalently, $\Lambda^{\text{AIPW}}(\eta)$ can be

written as

$$\Lambda^{\text{AIPW}}(\eta) = \mu(1,V) - \mu(0,V) + \frac{W[Y-\mu(1,V)]}{e(V)} - \frac{(1-W)[Y-\mu(0,V)]}{1-e(V)}. \tag{3.3}$$

It can be shown that under Assumptions 1 and 2, $\Lambda^{\text{AIPW}}(\eta_0)$ is unbiased for $g_0(x)$ (Glynn and Quinn, 2010). More importantly, $\Lambda^{\text{AIPW}}(\eta)$ is Neyman-orthogonal (Chernozhukov et al., 2018). The Neyman-orthogonality of $\Lambda^{\text{AIPW}}(\eta)$ gives rise to the *rate double robustness* property (Smucler, Rotnitzky and Robins, 2019), in the sense that as long as the *product* of the convergence rates of the estimators for $\mu(w,v)$ and $e(v)$ is controlled, the second-stage estimators will be consistent, asymptotically normal, and efficient. In other words, it is allowed that one part of the nuisance function (either $\mu(w,v)$ or $e(v)$) is approximated unsatisfactorily, provided the other part's error is small enough.

The use of orthogonal signal $\Lambda^{\text{AIPW}}(\eta)$ has a long history, which could date back to Robins, Rotnitzky and Zhao (1994) for the estimation of average treatment effects. This estimator is also termed as the Augmented Inverse Propensity Weighted (AIPW) estimator (Glynn and Quinn, 2010). Robins, Rotnitzky and Zhao (1994) show that the average treatment effect estimator based on $\Lambda^{\text{AIPW}}(\eta)$ is doubly robust in the sense that it remains consistent whenever either $\mu(w,v)$ or $e(v)$ is correctly specified. Moreover, $\Lambda^{\text{AIPW}}(\eta)$ is semiparametrically efficient (in the sense that the estimator attains the lowest asymptotic variance within the semiparametric family) when both $\mu(w,v)$ and $e(v)$ are correctly specified (Hahn, 1998).

3.2.2 Cross-fitting

In addition to Neyman orthogonality, the success of OEs also depends on another technique: *cross-fitting* (Chernozhukov et al., 2018). Cross-fitting is based on *sample splitting*, meaning that we estimate the nuisance functions $\eta(v)$ using subsample $\mathcal{O}^{(1)}$, while obtain the predicted values of $\eta(v)$ and construct the orthogonal signal $\Lambda(\eta)$ using another *disjoint* subsample $\mathcal{O}^{(2)}$. Sample splitting separates the model error from the estimation error and makes it easy to control the overall error, since the predictive value of $\eta(v)$ for observation i does not utilize the information of individual i.

Technically, the use of sample splitting easily ensures many favorable statistical properties to hold without imposing some high level conditions on estimators of $\eta(v)$, which is typically violated for many machine learning algorithms.[5] In spite of these advantages, sample splitting suffers from the issue of inefficiency because it only uses part of the data for estimation. To overcome this issue, we adopt *cross-fitting* (Chernozhukov et al., 2018), which requires us to swap the role $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ and repeat the estimation and prediction procedure on each subsample, to recover full efficiency.

We now formally describe the cross-fitting procedure.[6] Denote $\Lambda^{\text{AIPW}}(O;\eta) = \Lambda^{\text{AIPW}}(\eta)$ and $\Lambda_i^{\text{AIPW}}(\eta) = \Lambda^{\text{AIPW}}(O_i;\eta)$. Let $K \geq 2$ be an integer. A *K-fold cross-fitting* procedure for a random sample $(O_i)_{i=1}^n$ is as follows. First, conduct a K-fold random partition of the sample indices $\{1,2,...,n\}$ by $(J_k)_{k=1}^K$, so that each fold J_k has a sample size of n/k. Let $J_k^c = \{1,2,...,n\} \setminus J_k$. Second, for each $k = 1,...,K$, construct an estimator $\widehat{\eta}_k = \widehat{\eta}(O_{i\in J_k^c})$ of the nuisance function η_0 using only the data $\{O_i : i \in J_k^c\}$. Finally, for each observation $i \in J_k$, calculate $\widehat{\Lambda}_i^{\text{AIPW}} = \Lambda_i^{\text{AIPW}}(\widehat{\eta}_k)$.

The second stage of OE-AIPW is to nonparametrically regress $\widehat{\Lambda}_i^{\text{AIPW}}$ on the moderators, and the fitted function $\widehat{g}(x)$ is the estimated conditional marginal effects. In this paper, we consider a sieve regression in the second stage following Semenova and Chernozhukov (2021), which provides valid inference results for applied researchers in social sciences. In particular, the user specifies a basis function $p(X) \in \mathbb{R}^d$ on the moderators X, and run the regression $\widehat{\Lambda}_i^{\text{AIPW}} = p(X_i)'\beta + \nu_i$. The estimated sieve parameter is

$$\widehat{\beta}_A = \underset{\beta\in\mathbb{R}^d}{\text{argmin}} \sum_{i=1}^n \left(\widehat{\Lambda}_i^{\text{AIPW}} - p(X_i)'\beta\right)^2 = \left(\sum_{i=1}^n p(X_i)p(X_i)'\right)^{-1}\left(\sum_{i=1}^n p(X_i)\widehat{\Lambda}_i^{\text{AIPW}}\right), \tag{3.4}$$

and our final estimator for conditional marginal effect is $\widehat{g}_A(x) = p(x)'\widehat{\beta}_A$.

Notice that this second-stage model is nonparametric because the dimension of the basis function d is allowed to grow to infinity as the sample size n grows, in order to approximate any well-behaved nonparametric function $g_0(x)$. Such a model-free approximation protects applied researchers from the risk of misspecification. The choices of basis functions include splines, polynomials, wavelets, Fourier Series, etc. Besides the sieve regression, other nonparametric methods are available in

[5] These high level conditions include the Donsker property of $\widehat{\eta}(v)$. See e.g., Vaart and Wellner (2023).

[6] There are alternative cross-fitting schemes that could be applied in OEs. We describe them in Appendix A.

the literature for the second stage, for example, kernel regression (Fan et al., 2022; Zimmert and Lechner, 2019) and machine learning methods (Kennedy, 2023).

3.2.3 Statistical Properties

Semenova and Chernozhukov (2021) show that the OE-AIPW estimator $\widehat{g}_A(x)$ for conditional marginal effect is $\sqrt{n}$-consistent and asymptotically normal. Moreover, they provide pointwise and uniform confidence intervals of $\widehat{g}_A(x)$ for valid inference.[7] Remarkably, their assumptions for the asymptotic results reflect the rate double robustness mentioned previously, which results from the Neyman orthogonality of $\Lambda^{\text{AIPW}}(\eta)$. In particular, they require that $\sqrt{n}\sqrt{d}q_n m_n = o(1)$, where q_n and m_n are the L^2 convergence rate of the estimator for propensity score $\widehat{e}(v)$ and the treatment-specific outcome regression function $\widehat{\mu}(w,v)$, respectively.[8] Thus, if $e_0(v)$ and $\mu_0(w,v)$ are estimated in the same L^2 convergence rate b_n, then $b_n \lesssim (nd)^{-1/4}$ is sufficient for the $\sqrt{n}$-consistency and asymptotic normality of the estimated conditional marginal effects.[9]

Such a convergence rate can be met by using many popular machine learning methods for prediction, e.g., versions of LASSO (Belloni, Chernozhukov and Wei, 2016), L^2-Boosting (Luo, Spindler and Kück, 2016), random forests (Syrgkanis and Zampetakis, 2020), neural nets (Farrell, Liang and Misra, 2021), etc. In addition, classical nonparametric methods are also appropriate candidates, e.g., kernel, nearest neighbor, and sieve methods (Li and Racine, 2007).

Conventionally, the nuisance functions $\eta_0(v)$ are estimated using parametric approaches. When doing so, the second-stage estimator (for average treatment effect or conditional marginal effect) is *doubly robust* in the sense that it remains consistent when either $\mu_0(w,v)$ or $e_0(v)$ is correctly specified (Lee, Okui and Whang, 2017). In this paper, we consider a nonparametric model and machine learning estimation for $\eta_0(v)$ instead, and the rate double robustness of OE-AIPW is retained as it is a generalization of the conventional double robustness property.

Furthermore, Kennedy (2023) shows that OE-AIPW is oracle efficient, in the sense that $\widehat{g}_A(x)$ achieves the mean squared error as if we use the true nuisance functions η_0 to construct the orthog-

[7] Details about pointwise and uniform confidence intervals are provided in Appendix A.
[8] The L^2 convergence rate of an estimator $\widehat{f}(x)$ for $f_0(x)$ is $\mathbb{E}[(\widehat{f}(X) - f_0(X))^2]^{1/2}$.
[9] We use $a_n \lesssim b_n$ to denote $a_n/b_n \to 0$ as $n \to \infty$.

onal signal and use this *oracle* signal $\Lambda^{\text{AIPW}}(O_i; \eta_0)$ in the second-stage regression. This could be viewed as a generalization of the result that $\Lambda^{\text{AIPW}}(\eta)$ is the efficient influence function for average treatment effect (Hahn, 1998). Notably, this results also features a similar rate double robustness property, meaning that as long as the product of the convergence rate of $\widehat{e}(v)$ and $\widehat{\mu}(w, v)$ is controlled, the oracle efficiency will hold. Furthermore, the oracle efficiency is valid for many other second stage regressions, such as kernel ridge regression and some versions of random forests.

To conclude, OE-AIPW is $\sqrt{n}$-consistent, asymptotically normal, and oracle efficient for many popular machine learning methods used in the first-stage.

3.3 Orthogonal Estimator Based on Conditional Average Treatment Effect (OE-CATE)

3.3.1 Motivation and Illustrative Example

The orthogonal signal of OE-AIPW (3.3) can be regarded as augmenting the "outcome regression-based" conditional average treatment effects, namely $\mu(1, V) - \mu(0, V)$, with the propensity score weighted residuals. Suppose a researcher is able to obtain a high quality estimate for conditional average treatment effects $\tau(V)$, then using it to replace $\mu(1, V) - \mu(0, V)$ may improve the approximation quality of the orthogonal signal, and thus enhance the finite sample performance of the estimated conditional marginal effect in the second stage. In fact, as illustrated in Kennedy (2023), in many empirical scenarios we often expect the conditional average treatment effects to be "[m]ore structured and simple than the rest of the data generating process ... similar in spirit to the common belief that interaction terms are more often zero than main effects" (p.3009). Meanwhile, there are important scenarios where the treatment-specific outcome regression functions $\mu_0(1, v)$ and $\mu_0(0, v)$ are difficult to estimate. A notable example is when the propensity score is discontinuous at some cutoff point, which corresponds to the fuzzy regression discontinuity design (RDD). In this case, it is more preferable to estimate $\tau_0(v)$ rather than $\mu_0(w, v)$.

To be more specific, we consider an illustrative DGP as in Kennedy (2023). We generate a single covariate V from uniform distribution $\mathcal{U}[-1, 1]$. The propensity score is $e_0(v) = 0.1 + 0.8\mathbb{1}(v > 0)$,

which has a discontinuous point at $v = 0$. The outcome regression functions are

$$\mu_0(1, v) = \mu_0(0, v) = \begin{cases} (v+2)^2/2, \text{ if } -1 \leq v \leq -0.5, \\ v/2 + 0.875, \text{ if } -0.5 < v \leq 0, \\ -5(v-0.2)^2 + 1.075, \text{ if } 0 < v \leq 0.5, \\ v + 0.125, \text{ if } v > 0.5 \end{cases}$$

Therefore, $\tau_0(v) = \mu_0(1, v) - \mu_0(0, x) = 0$. Finally, $Y_i = \mu_0(0, V_i) + W_i\tau_0(V_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 0.1)$. In this DGP, the conditional average treatment effects are simply zero, while the treatment-specific outcome regressions are difficult to estimate because the data points in the treatment group and control groups are unevenly distributed for $v < 0$ and $v > 0$. We use random forests to estimate $\mu_0(w, v)$. We can see in Figure 1(a) that $\mu_0(1, v)$ is estimated well for $v > 0$ where there are rich data points and poorly for $v < 0$ where data are sparse. Likewise, $\mu_0(0, v)$ is approximated well at $v < 0$ and poorly at $v > 0$ as in Figure 1(b). On the other hand, we could estimate the conditional average treatment effects $\tau_0(v)$ in a satisfactory way using some off-the-shell algorithm. Figure 1(c) shows that estimating $\tau_0(v)$ using Causal Forests (Athey and Wager, 2019) leads to a small error, while using the "plug in" estimator $\widehat{\mu}(1, v) - \widehat{\mu}(0, v)$ is badly performed, due to the aforementioned reasoning.

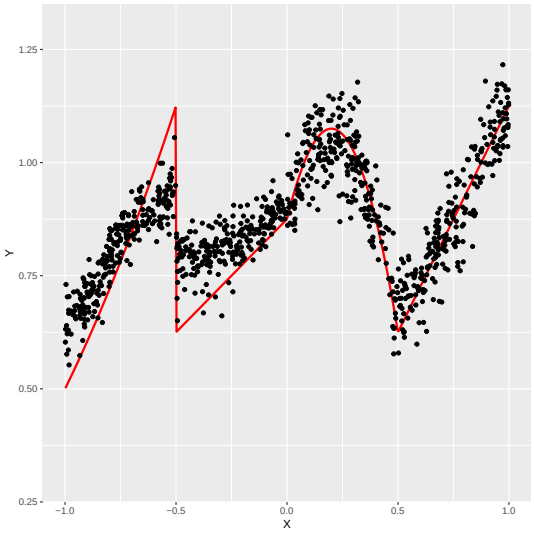
In light of such limitation of the OE-AIPW, we propose an alternative orthogonal signal based on estimating conditional average treatment effect (CATE) in the first-stage, which we term as OE-CATE:

$$\Lambda^{\text{CATE}}(\delta) = \tau(V) + \frac{W - e(V)}{e(V)(1 - e(V))}\Big(Y - m(V) - \big(W - e(V)\big)\tau(V)\Big), \tag{3.5}$$

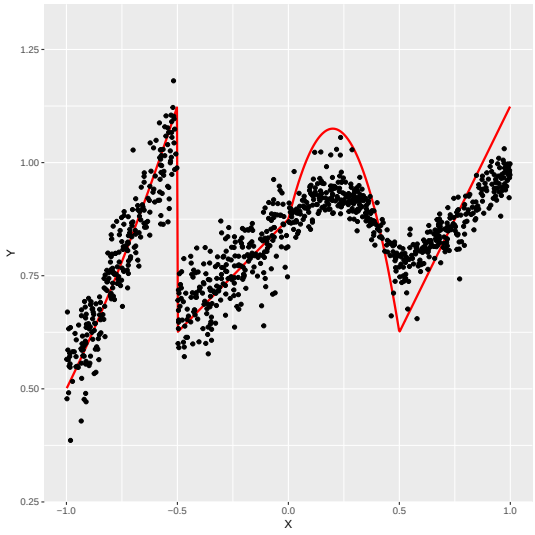
where the nuisance functions are $\delta = \delta(v) = (\tau(v), m(v), e(v))$, including the conditional average treatment effect $\tau(v)$ defined in (2.1), the *outcome regression function* without conditioning on treatment $m(v) = \mathbb{E}[Y|V = v]$, and the propensity score $e(v) = \mathbb{P}[W = 1|V = v]$. Our OE-CATE estimator differs from OE-AIPW in the construction of the first-stage signal, while the cross-fitting

Figure 1: Illustrative Example

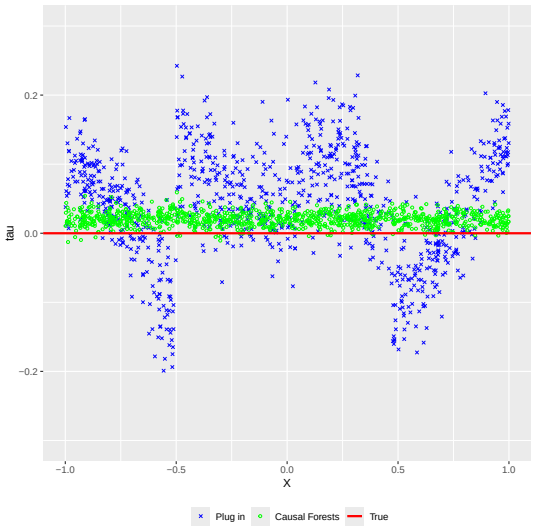
(a) $\widehat{\mu}(1, v)$



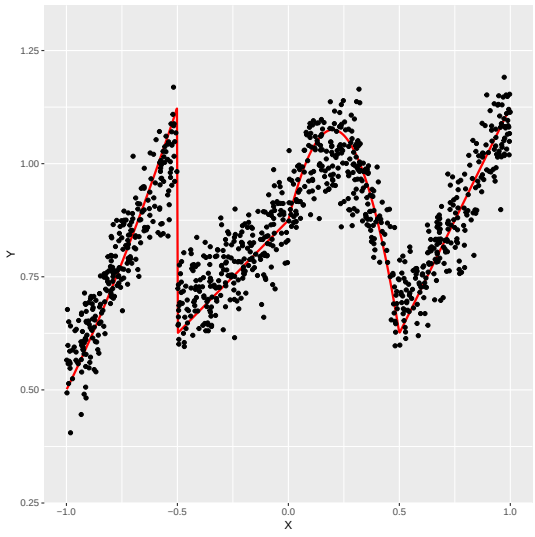
(b) $\widehat{\mu}(0, v)$



(c) $\widehat{\tau}(v)$ from Plug in and Causal Forests



(d) $\widehat{m}(v)$



In all four figures, the red lines represent the true DGP, the dots represent the predicted values.

and second-stage regression remains unchanged. Denote $\widehat{g}_C(x)$ as the OE-CATE estimator for conditional marginal effects.

Notice that in our illustrative example, $m_0(v)$ could be estimated well since it is not affected by the discontinuous propensity score. Figure 1(d) shows that $m_0(v)$ is approximated well both for $v < 0$ and $v > 0$ using random forests. Therefore, albeit sharing a common nuisance function $e(v)$, OE-CATE differs from OE-CATE in the other two nuisance functions, where $\tau(v)$ and $m(v)$ could be better estimated than $\mu(1, v)$ and $\mu(0, v)$ under some important empirical scenarios, such as discontinuous propensity scores.

3.3.2 Theoretical Analysis

Next, we establish some crucial theoretical properties for OE-CATE. The proofs and technical assumptions of the theoretical results are provided in Appendix C. First, observe that $\tau_0(v) = \mu_0(1, v) - \mu_0(0, v)$ under the unconfoundedness assumption, and $m_0(v) = \mu_0(1, v)e_0(v) + \mu_0(0, v)(1 - e_0(v))$, it is straightforward to show that the orthogonal signal of OE-CATE is identical to the one of OE-AIPW at the true nuisance functions.

Lemma 1 (Equivalence)**.** *Under Assumptions 1 and 2, we have* $\Lambda^{AIPW}(\eta_0) = \Lambda^{CATE}(\delta_0)$.

Moreover, notice that $\Lambda^{\text{CATE}}(\delta)$ involves the nuisance functions $\delta(v) = (\tau(v), m(v), e(v))$, which could be expressed as a transformation from $\eta(v) = (\mu(1, v), \mu(0, v), e(v))$. Roughly speaking, a small fluctuation in $\delta(v)$ is corresponding to a small fluctuation in $\eta(v)$. Hence, it is natural to believe that $\Lambda^{\text{CATE}}(\delta)$ inherits the Neyman-orthogonality from $\Lambda^{\text{AIPW}}(\eta)$. In fact, we formally prove that this is true:

Lemma 2 (Orthogonality)**.** $\Lambda^{CATE}(\delta)$ *is Neyman-orthogonal with respect to the nuisance function* $\delta(v) = (\tau(v), m(v), e(v))$.

The Neyman-orthogonality is essential for proving that OE-CATE enjoys the favorable statistical properties of the orthogonal estimator family, just as OE-AIPW does. In particular, we show that OE-CATE is $\sqrt{n}$-consistent, asymptotically normal, and oracle efficient.

Theorem 1 (Asymptotic Normality). *Under Assumptions C.1-C.6 presented in Appendix C, the OE-CATE estimator $\widehat{g}_C(x)$ for conditional marginal effect is $\sqrt{n}$-consistent and asymptotically normal:*

$$\frac{\sqrt{n}(\widehat{g}_C(x_0) - g_0(x_0))}{\sqrt{p(x_0)'\Omega p(x_0)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $x_0 \in \mathcal{X}$, and $\Omega = Q^{-1}\mathbb{E}[p(X)p(X)'(U + r(X))^2]Q^{-1}$. Here Q, U and $r(\cdot)$ are defined in Appendix C, and $p(x)$ is the basis function in the second stage.

Notably, Assumption C.5 regulates the convergence rate for the nuisance functions so that $\sqrt{n}\sqrt{d}q_n(r_n + s_n) = o(1)$, where q_n, r_n, s_n are the L^2 convergence rate for $\widehat{e}(v), \widehat{m}(v), \widehat{\tau}(v)$, respectively, which mirrors the convergence rate requirement for OE-AIPW. Thus, when $\widehat{e}(v), \widehat{m}(v), \widehat{\tau}(v)$ are estimated in the same L^2 convergence rate b_n, then $b_n \lesssim (nd)^{-1/4}$ is sufficient for the $\sqrt{n}$-consistency and asymptotic normality of OE-CATE, resembling the results for OE-AIPW.

For $\widehat{e}(v)$ and $\widehat{m}(v)$, we can use off-the-shell machine learning methods just as the first stage estimation in OE-AIPW. For $\widehat{\tau}(v)$, there is a growing literature of estimating conditional average treatment effects, adapting machine learning algorithms especially. Some leading methods include Bayesian additive regression trees (Hill, 2011), regression trees (Athey and Imbens, 2016), boosting (Powers et al., 2018), support vector machine (Imai, Ratkovic et al., 2013), random forests (Foster, Taylor and Ruberg, 2011), neural nets (Stadie et al., 2018), meta learners (Künzel et al., 2019; Nie and Wager, 2021), and ensembling methods (Grimmer, Messing and Westwood, 2017), to name a few. Given the convergence rate of $\widehat{\tau}(v)$, we are able to discuss how it fits our regularity condition. For example, in the Causal Forests algorithm (Athey et al., 2019) we use for simulation and empirical application, $\widehat{\tau}(x) - \tau_0(x) = O(\sqrt{s/n})$ with $s = n^\beta$ and $\beta < 1$. Suppose we choose the spline order as $d \gtrsim n^{p/2\gamma}$ as the undersmoothing condition in Newey (1997) and Belloni et al. (2015) where $g_0(x)$ is assumed to fall in the Hölder class of smoothness order γ. Then when $\sqrt{s/n} = n^{(\beta-1)/2} < (nd)^{-1/4}$, i.e., $\beta < 1 - \frac{p+2\gamma}{4\gamma}$, the Causal Forests estimator $\widehat{\tau}(v)$ satisfy this condition of convergence rate.

Following Kennedy (2023), we show that OE-CATE is also oracle efficient.

Theorem 2 (Oracle Efficiency). *Under Assumption C.7 presented in Appendix C, the OE-CATE*

estimator for conditional marginal effect is oracle efficient in the sense that it achieves the mean squared error as if we use the true nuisance functions δ_0 in the second-stage regression:

$$\frac{\widehat{g}_C(x_0) - g_0(x_0)}{\widetilde{g}(x_0) - g_0(x_0)} = o_p(1), \forall x_0 \in \mathcal{X},$$

where $\widetilde{g}(\cdot)$ is obtained by regressing $\Lambda^{CATE}(\delta_0)$ on $p(X)$.

For both Theorem 1 and 2 to hold, it is sufficient to control the product of the convergence rate of nuisance functions $q_n(r_n + s_n)$, which provides us the rate double robustness as in OE-AIPW. In other words, OE-CATE retains $\sqrt{n}$-consistency, asymptotic normality, and oracle efficiency as long as we are able to have a satisfactory estimator for either the propensity score $\widehat{e}(v)$ *or* for both $\widehat{\tau}(v)$ and $\widehat{m}(v)$. Consider the case where a given propensity score estimator is in use. Then OE-CATE may improve upon OE-AIPW when estimating $\tau_0(v)$ and $m_0(v)$ are easier tasks than estimating $\mu_0(w, v)$, as the illustrative example shows. However, holding $\widehat{e}(v)$ fixed, to achieve the same performance, OE-CATE demands both $\widehat{m}(v)$ and $\widehat{\tau}(v)$ to approximate the true functions well, while OE-AIPW only needs a single estimator $\widehat{\mu}(w, v)$ to behave satisfactorily. Therefore, it is a concern that OE-CATE requires more estimators to be good in order to fulfill the favored statistical properties.

Remarkably, there is a growing literature on the construction of orthogonal moment functions or loss functions in general setups (Ichimura and Newey, 2022; Chernozhukov, Newey and Singh, 2022; Chernozhukov et al., 2022; Hirshberg and Wager, 2021; Foster and Syrgkanis, 2023), which involves finding the Riesz representer of the original moment function. In our binary treatment effect context, such a construction leads to the orthogonal signal of OE-AIPW. The exploration of other orthogonal signals will be a promising topic for future research.

3.4 Summary of the Algorithm and Implementation

To summarize, our orthogonal estimators for conditional marginal effect proceeds as follows:

Algorithm 1 (Orthogonal Estimators for Conditional Marginal Effects)**.**

Step 1. First stage. Estimate the nuisance functions and calculate the orthogonal signal $\widehat{\Lambda}_i$ *using cross-fitting.*

(a) For OE-AIPW, the nuisance functions are $\eta(v) = (\mu(1,v), \mu(0,v), e(v))$ *and the orthogonal signal is* $\Lambda^{AIPW}(\eta)$ *as in* (3.3)*.*

(b) For OE-CATE, the nuisance functions are $\delta(v) = (\tau(v), m(v), e(v))$ *and the orthogonal signal is* $\Lambda^{CATE}(\delta)$ *as in* (3.5)*.*

Step 2. Second stage. Run a sieve regression of $\widehat{\Lambda}_i$ *on the moderators* X*:* $\widehat{\Lambda}_i = p(X_i)'\beta + \nu_i$ *and obtain* $\widehat{\beta}$*. The estimated conditional marginal effect at* $X = x$ *is* $\widehat{g}(x) = p(x)'\widehat{\beta}$*. The pointwise and uniform confidence intervals for* $\widehat{g}(x)$ *are presented in Appendix A.*

Algorithm 1 is a general procedure for OEs through which users can specify their options in several places. First, the estimation methods for the nuisance functions in the first stage could be chosen at the user's pleasure. Off-the-shell machine learning algorithms for $\widehat{e}(v), \widehat{\mu}(w,v)$ and $\widehat{m}(v)$, as well as adapted machine learning methods for $\widehat{\tau}(v)$ are discussed above. Second, practitioners are also free to select the basis function $p(x)$ in the second-stage, e.g., splines, polynomials, wavelets, Fourier series, etc. Third, there are options for the cross-fitting procedure. On the one hand, we could choose the number of folds K; on the other hand, there are alternative cross-fitting schemes that could be used, which are presented in Appendix A.

In our simulations and empirical applications, we use random forests to estimate $e_0(v)$, $\mu_0(w,v)$ and $m_0(v)$, which enjoys a good reputation for prediction preciseness. For $\tau_0(v)$, we use Causal Forests (Athey et al., 2019), because it possesses favorable theoretical properties such as unbiasedness and centered asymptotic normality. Moreover, its practical success has been demonstrated by a wide range of applications in political science (Yin, Huo and Lin, 2023; Zheng and Yin, 2023), economics (Guryan et al., 2023), and marketing (Chen et al., 2020), etc. We use B-spline regression in our second-stage, which is a piecewise polynomial function that offers remarkable local and global approximation capability with efficient computational algorithms. We use a 2-fold cross-fitting scheme as illustrated above.

To facilitate the use of the OEs, we provide easy-to-use replication codes in `R` where practitioners

can specify their choices in Algorithm 1. We discuss the details in Appendix B.

4 Monte-Carlo Experiments

4.1 Simulation Design

To evaluate the performance of our proposed Orthogonal Estimators (OEs) in comparison with existing methods, we carry out Monte Carlo experiments using a wide variety of data generating processes (DGPs) that social scientists might encounter in their modeling tasks. The general framework of our simulation design is as follows: We draw a single moderator X and covariates $Z = (Z_1, \cdots, Z_{p-1})$ from a uniform distribution $\mathcal{U}[-1, 1]$. The outcome is generated as $Y_i = \mu_0(0, V_i) + W_i\tau_0(V_i) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 1)$. Next, in each DGP framework, we specify the setting of: (1) the regression function without treatment $\mu_0(0, V)$; (2) the propensity score $e_0(V)$; and (3) the treatment effect $\tau_0(V)$.

We estimate the conditional marginal effect using four methods:(1) the Orthogonal Estimator based on Conditional Average Treatment Effect (OE-CATE); (2) the Orthogonal Estimator based on Augmented Inverse Propensity Weighting (OE-AIPW); (3) the varying coefficient model (VCM) (Hainmueller, Mummolo and Xu, 2019); and (4) the multiplicative interaction model (MIM) (Brambor, Clark and Golder, 2006).

The training sample size is set to be $n = 2000$ and the dimension of all covariates is $p = 10$. We evaluate each estimation method based on their prediction performance and coverage rate of confidence interval. For the prediction performance, we compute the root mean squared error (RMSE) $(\sum_{i=1}^{n_t}[g_0(X_i) - \widehat{g}(X_i)]^2/n_t)^{1/2}$ and bias $\sum_{i=1}^{n_t}(g_0(X_i) - \widehat{g}(X_i))/n_t$ of the conditional marginal effect on a test set with a sample size of $n_t = 100$. Letting the pointwise confidence interval be $\widehat{C}_i^p$ at a test point X_i, we calculate the expected pointwise coverage rate $\sum_{i=1}^{n_t} \mathbb{1}(g_0(X_i) \in \widehat{C}_i^p)/n_t$. While all the four methods considered here have pointwise confidence intervals, for OE-CATE and OE-AIPW we also have uniform confidence bands. For the uniform confidence band $\widehat{C}_i^u$, we focus on the uniform coverage indicator $\mathbb{1}(\widehat{g}(X_i) \in \widehat{C}_i^u, \forall i = 1, \cdots, n_t)$, and take average over all replication

trials. The simulation results are obtained from averaging $m = 100$ replications for each DGP.[10]

We consider two different setups of DGP in the main text: (A) randomized experiment, and (B) discontinuous propensity score, which correspond to the two empirical applications in the next section. Under each setup umbrella, we consider two kinds of conditional marginal effect: (1) linear conditional marginal effect, (2) non-linear continuous conditional marginal effect. The details of the considered DGPs are as follows:

(i) *Setup A: Randomized Experiment.* In this setup, we consider random experiment cases with $e_0(v) = 0.5$, and $\mu_0(0, v) = 2x + z_1$. The treatment effects are:

DGP A1: $\tau_0(v) = 0$

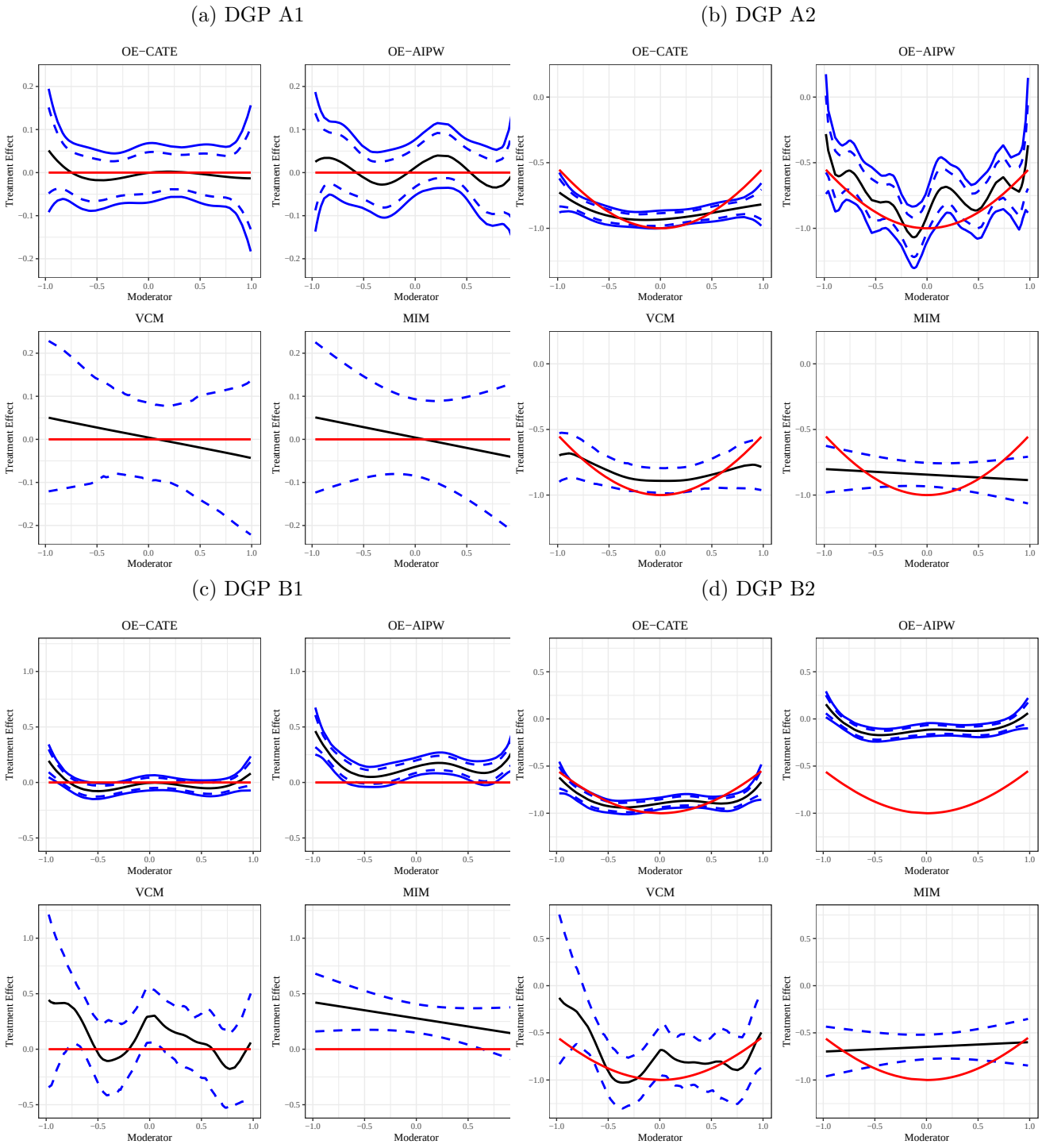
DGP A2: $\tau_0(v) = -\cos(x)$

(ii) *Setup B: Discontinuous Propensity Score.* This setup shares the similar spirit with the illustrative example in Section 3.3, and is motivated by Kennedy (2023). The propensity score is $e_0(v) = 0.1 + 0.8\mathbb{1}(x > 0)$, which is discontinuous on the moderator x. This context is often modeled using the fuzzy regression discontinuity design. The regression functions is

$$\mu_0(0, v) = \begin{cases} (x+2)^2/2, \text{ if } -1 \leq x \leq -0.5, \\ x/2 + 0.875, \text{ if } -0.5 < x \leq 0, \\ -5(x-0.2)^2 + 1.575, \text{ if } 0 < x \leq 0.5, \\ x + 0.625, \text{ if } x > 0.5 \end{cases}$$

DGP B1-B2 have the same treatment effects $\tau_0(v)$ as those specified in DGP A1-A2, respectively. It is important to note that in DGP B1-B2, not only the propensity score is discontinuous in x, but the regression function is also discontinuous in x at $x = 0$. Hence, this set-up violates the identification assumption of regression discontinuity design.

Figure 2: Conditional Marginal Effects from Simulations



The red lines represent the true DGP, black lines represent the estimates, and blue lines represents confidence intervals (solid blue lines for uniform confidence band, and dashed blue lines for pointwise confidence intervals).

Table 1: Simulation Results

		OE-CATE	OE-AIPW	VCM	MIM
DGP A1	RMSE	0.0568	0.0433	0.0615	0.0602
	Bias	0.0028	0.0009	0.0017	0.0018
	Coverage-P	0.6042	0.8330	0.9318	0.9508
	Coverage-U	0.3400	0.6500	-	-
DGP A2	RMSE	0.1128	0.1523	0.1101	0.1529
	Bias	0.0271	-0.1012	-0.0209	0.0003
	Coverage-P	0.3647	0.4965	0.6775	0.4814
	Coverage-U	0.0000	0.0600	-	-
DGP B1	RMSE	0.0984	0.1334	0.2017	0.2511
	Bias	-0.0509	-0.1021	-0.1117	-0.2357
	Coverage-P	0.3495	0.3031	0.6392	0.2603
	Coverage-U	0.0500	0.0000	-	-
DGP B2	RMSE	0.1437	0.5920	0.2185	0.2344
	Bias	0.0043	-0.5683	-0.1158	-0.1679
	Coverage-P	0.2539	0.0500	0.6131	0.4100
	Coverage-U	0.0000	0.0000	-	-

4.2 Results and Discussion

The simulation results are reported in Table 1. The estimated and true conditional marginal effects for each setup are presented in Figure 2, where the red lines represent the true DGP, black lines represent the estimates, and blue lines represents confidence intervals (solid blue lines for uniform confidence band, and dashed blue lines for pointwise confidence intervals).

Results from Setup A indicate that when the required functional form assumptions are met, the parametric and semiparametric approaches perform very well, while our OEs also deliver consistent estimates. Specifically, for DGP A1 where the true conditional marginal effects are linear, VCM and MIM work well as expected, and so are the two OEs. For DGP A2 where the true conditional marginal effects are nonlinear, VCM is consistent, and OE-CATE possesses a similar performance. OE-AIPW is also able to trace the true effects satisfactorily, though it generates a slightly larger bias. Such results meet our expectation that under correct specifications, both the conventional

[10] We report the simulation results of the two OEs with 1000 replications and evaluation on the training sample in Appendix D as a robustness check, which resembles the results in this section.

(parametric or semiparametric) and our OE approach are unbiased and consistent, while the parametric and semiparametric approaches may have efficiency gain over the nonparametric ones. In addition, the conventional methods seem to have better coverage rates than the OEs.

Results from Setup B confirms our discussion in Section 3.3. When the propensity score is discontinuous on the moderator, it is hard to approximate the treatment-specific outcome regression function $\mu_0(w, v)$, while accurately estimating the conditional average treatment effect $\tau_0(v)$ is relatively easy using some algorithms for estimating heterogeneous treatment effect, e.g., the Causal Forests algorithm (Athey et al., 2019). As a result, OE-AIPW could be improved by OE-CATE. We see that in Setup B1 and B2, OE-CATE performs better than OE-AIPW, especially for the nonlinear conditional marginal effects case. As expected, MIM exhibits bias because the outcome regression function is nonlinear. Moreover, VCM also performs poorly possibly due of the discontinuity of $\mu_0(w, v)$. Simply put, when we are faced with a discontinuous propensity score, which is often modeled using the fuzzy regression discontinuity design, we should prioritize the use of OE-CATE. We will elaborate more on this point in one of our replication exercises below.

5 Empirical Applications

To demonstrate the usefulness of the OEs, we apply them to two empirical scenarios commonly encountered by social scientists: discontinuous propensity scores and random experiments with treatment effect heterogeneity. Specifically, we use the OEs to replicate publications in leading political science journals, which employ fuzzy regression discontinuity design (RDD) and multiplicative interaction models (MIM), respectively. Note that Setup B in the preceding simulation exercises involves a DGP that appears suitable for the use of RDD, but violates the continuity assumption for a valid RDD estimation. Meanwhile, the DGP in Setup A represents random experiments with potential treatment effect heterogeneity. Hence, the simulation results based on these two setups can be informative for interpreting the replication findings presented below.

5.1 Case I: Discontinuous Propensity Scores

RDD has gained increasing popularity in empirical research, thanks to their known and observable treatment assignment mechanisms. These designs prove especially valuable when randomization is impractical or unethical. Empirical studies underscore the importance of identification assumptions for an RDD to be valid. Specifically, a credible RDD requires continuity in the regression function of pretreatment covariates as well as continuity in the density of the forcing variable, or running variable (Hartman, 2021).

Whether the continuity assumption can be met in one of the most popular applications of RDD in political science, namely the close election RDD, has been heatedly debated in the literature. The idea of a close election RDD is that if the difference in vote share between the two candidates is small enough, then any differences in outcomes between winners and losers can be attributed to the random chance of winning/losing rather than systematic differences in the characteristics or policies of the candidates. The estimated coefficient on the treatment variable represents the effect of winning/losing a close election on a outcome variable (Lee, Moretti and Butler, 2004; Lee, 2008).

However, as a growing literature points out, applying RDD to close elections can impose challenges for causal inference as it may violate the assumption of continuity in the regression function of pretreatment covariates (Hartman, 2021). This violation arises from strategic manipulation by the candidates and the voters. For example, a candidate who knows that they are close to the threshold of winning may try to change their policy positions or campaign strategies in order to appeal to voters on the other side of the threshold. A candidate may even engineer electoral fraud which clearly violates the continuity assumption needed for a valid RDD (De la Cuesta and Imai, 2016). Finally, voters may also engage in strategic manipulations in close elections by changing their voting behavior based on the characteristics of the candidates.[11] These forms of strategic manipulations can introduce bias into RDD estimates by affecting which candidates win close elections and which do not, making electoral victories not an exogenous treatment but an endogenous one. It is important to note that these concerns are relevant to both sharp and fuzzy RDD and are

[11] For example, a voter who prefers a candidate with a certain characteristic may be more likely to vote for that candidate if they know the election is close and their vote could make a difference.

difficult to fully address within the RDD modeling framework.

Our approach to addressing these concerns for close election RDD is different and more flexible. As is demonstrated in the simulation exercises, even when the assumption for RDD is violated (i.e, when the regression function of the pre-treatment covariates is discontinuous in the running variable at the cut-off), OE-CATE can still correctly identify the treatment effect and outperforms OE-AIPW for finite samples.[12] In fact, the simulation results in Setup B demonstrate that regardless of whether the regression function is discontinuous at the cut-off or elsewhere, our original OE-CATE consistently outperforms the existing OE-AIPW and other more conventional methods. Moreover, it allows the treatment effect to vary along the forcing variable, while such heterogeneity cannot be captured using RDD that only identifies the local treatment effect around the threshold. Simply put, the OEs are particularly useful when we are not certain whether the regression function is continuous, or when we are interested in heterogeneous treatment effects along the running variable.

To illustrate the use of OEs in identifying the treatment effect with a discontinuous propensity score, we replicate the study of Dasgupta and Kapur (2020) who leverages close legislative election to examine the effect of party alignment on the investment in bureaucratic resources in the context of India. The outcome variable in their study is the local resource index constructed from the the number of full time employees, contract employees, vehicles, and computers per 100,000 rural residents, which reflects the amount of investment by politicians on local bureaucracy. The unit of analysis is *rural development block*, an administrative unit that implements local development programs but has an imperfect congruence with legislative constituencies.[13] Such imperfect correspondence of block and constituency boundaries gives rise to the imperfect compliance with the treatment of narrowly electing an aligned legislator, and thus justifies for the use of a fuzzy RDD framework.

To study the effect of alignment on bureaucratic resources, in their close-election fuzzy RDD design, the authors instrument for the *share of the block population located in an aligned constituency* with the *narrow election of a ruling-party legislator in the constituency that contains the block*

[12] While De la Cuesta and Imai (2016) suggest that sorting based on pre-election candidate behavior and characteristics does not necessarily violate the continuity assumption for a valid RDD, our simulation setup above has clearly violated this assumption.

[13] Most, but not all, rural development blocks are fully contained by a single legislative constituency.

headquarters. Thus, the running variable is the win/loss vote share margin of the ruling party's candidate in such constituencies. As is presented in the top panel of Figure 6 in Dasgupta and Kapur (2020), the share of the block population located in an aligned constituency jumps sharply at the cut-off point of 0 in the running variable. According to their RDD estimation results, "increases in the share of the block population located in an aligned constituency are associated with large improvements in bureaucratic resources (Dasgupta and Kapur, 2020, p.1330)." Specifically, compared to a block with none of the population represented by an aligned legislator, the one with all of the population represented enjoys a 0.76-standard-deviation improvement in the resource index (Panel A, Table 6).

Again, there remains the question of whether the assumption of continuity in the regression function is met for the use of a close election fuzzy RDD. Hence, we employ the OEs that do not rely on the continuity assumption. The treatment variable in our estimation is the binary measure of winning or losing a close election, which probabilistically determines the status of political alignment, while the moderator of interest is the running variable in the fuzzy RDD framework, i.e., the vote share margins. We adopt the same 395 blocks as those used in Table 7 of Dasgupta and Kapur (2020) for our sample and include all 15 pre-treatment variables used for balance checks as our controls. The ability to accommodate a large set of covariates stands as an advantage of our proposed method over RDD.

Similar to the simulation results, we present the OE-CATE and OE-AIPW estimates along with those of the conventional strategies, namely the VCM based on Hainmueller, Mummolo and Xu (2019), and the MIM (see Figure 3).[14] The point estimates of the treatment effects vary substantially between the OE and existing approaches. The OE results indicate that across all ranges of vote share margins, the effect of political alignment is consistently similar. Moreover, the OE-CATE results suggest that the treatment effect is statistically distinguishable from 0 at the 95% level, a conclusion partially supported by VCM. While we do not directly observe the true DGP for

[14]In the main text's replication exercises, we use causal forests to estimate the CATE, random forests to estimate the nuisance functions for OE-CATE and OE-AIPW, and B-spline regression for the second-stage estimation. We also allow for alternative algorithms, such as the X-learner (Künzel et al., 2019) for CATE, regression forest function of the `grf` package (Athey et al., 2019) for the nuisance functions, and polynomial regression for the second-stage estimation.

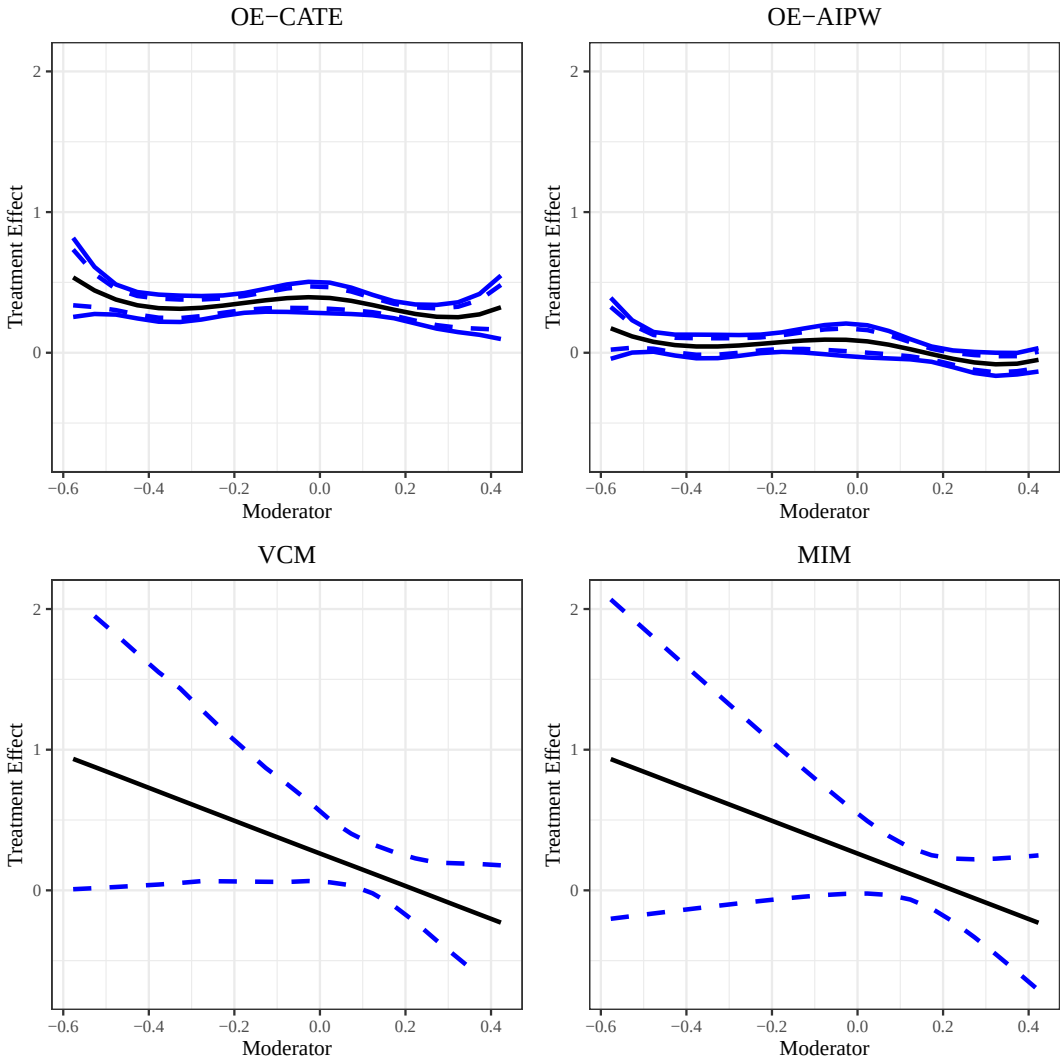
this replication exercise, there are compelling reasons to consider the OE estimates more reliable. The parametric and semi-parametric nature of MIM and VCM may result in incorrect extrapolations for the point estimates, particularly for observations with negative vote share margins, where common support is not sufficient. In contrast, the flexibility of the OEs helps avoid this issue as no functional form is assumed.

5.2 Case II: Random Experiments with Heterogeneous Treatment Effects

In our second replication exercise, we use OEs to explore and evaluate treatment effect heterogeneity in an experimental setting. We replicate the study of Huddy, Mason and Aarøe (2015), which employs MIM to estimate heterogeneous treatment effects based on experimental data. Unlike the previous subsection, where we emphasized the differences between our approach and existing methods, here we focus on the consistency between them under specific conditions. As demonstrated in our simulation results under Setup A1-A2, the OEs produce comparable outcomes to those from VCM and MIM in setups with simple conditional marginal effect functions. Therefore, our proposed method is capable of handling both simple and complex DGPs.

In the study of Huddy, Mason and Aarøe (2015), the authors investigate if party identity in the United States has an instrumental or expressive nature using experiments. Their study involves a binary treatment variable, the threat of electoral loss, which is assigned a value of 1 when participants read a fabricated blog post suggesting that their party is facing defeat in the upcoming election. The effect of this treatment variable on anger levels, an action-oriented political emotion, is the parameter of interest in Huddy, Mason and Aarøe (2015). Specifically, the authors focus on two moderating variables—*partisan identity* and *ideological issue intensity*—to gauge the expressive and instrumental aspects of partisanship, respectively, and interact these two moderators with the treatment variable, the threat of electoral loss. Based on MIMs, the authors find that the interaction term between electoral threat and *partisan identity* is positive and statistically significant at the 1% level, while the interaction term between the threat variable and *ideological issue intensity* is statistically indistinguishable from 0 at conventional levels. Their findings support the expressive model, indicating that when faced with electoral loss, individuals with a strong partisan identity

Figure 3: Replication of Dasgupta and Kapur (2020)



The horizontal axis represents the moderating variable, which is the vote share margin, while the vertical axis represents the treatment effect of political alignment from different estimators. The blue lines represents confidence intervals (solid blue lines for uniform confidence band, and dashed blue lines for pointwise confidence intervals).

experience greater anger than those with weaker partisan affiliations.

We replicate the study by Huddy, Mason and Aarøe (2015) using the same treatment variable—threat of electoral loss—and the same set of covariates.[15] We focus on the moderating effect of the strength of partisan identity, which according to Huddy, Mason and Aarøe (2015), positively moderates the treatment effect. The replication results are presented in Figure 4. The estimation outcomes generated from the OEs are similar to those from VCM and MIM, aligning with our expectations.[16] As the simulation results of DGP A2 indicates, the OEs yield point estimates akin to those from MIM and VCM when the data are generated from a random experiment with a linear conditional marginal effect. The similarity between the OE and MIM/VCM results suggests that the effect of electoral loss on anger levels does increase linearly with the moderating variable, partisan identity. In summary, our proposed OEs can accommodate both simple and complex DGPs. When a researcher is unsure about the characteristics of the underlying DGPs or the validity of the assumptions for conventional methods, we recommend the use of OEs to explore potential empirical patterns, which has the advantage of necessitating fewer restrictive assumptions.

6 Concluding Remarks

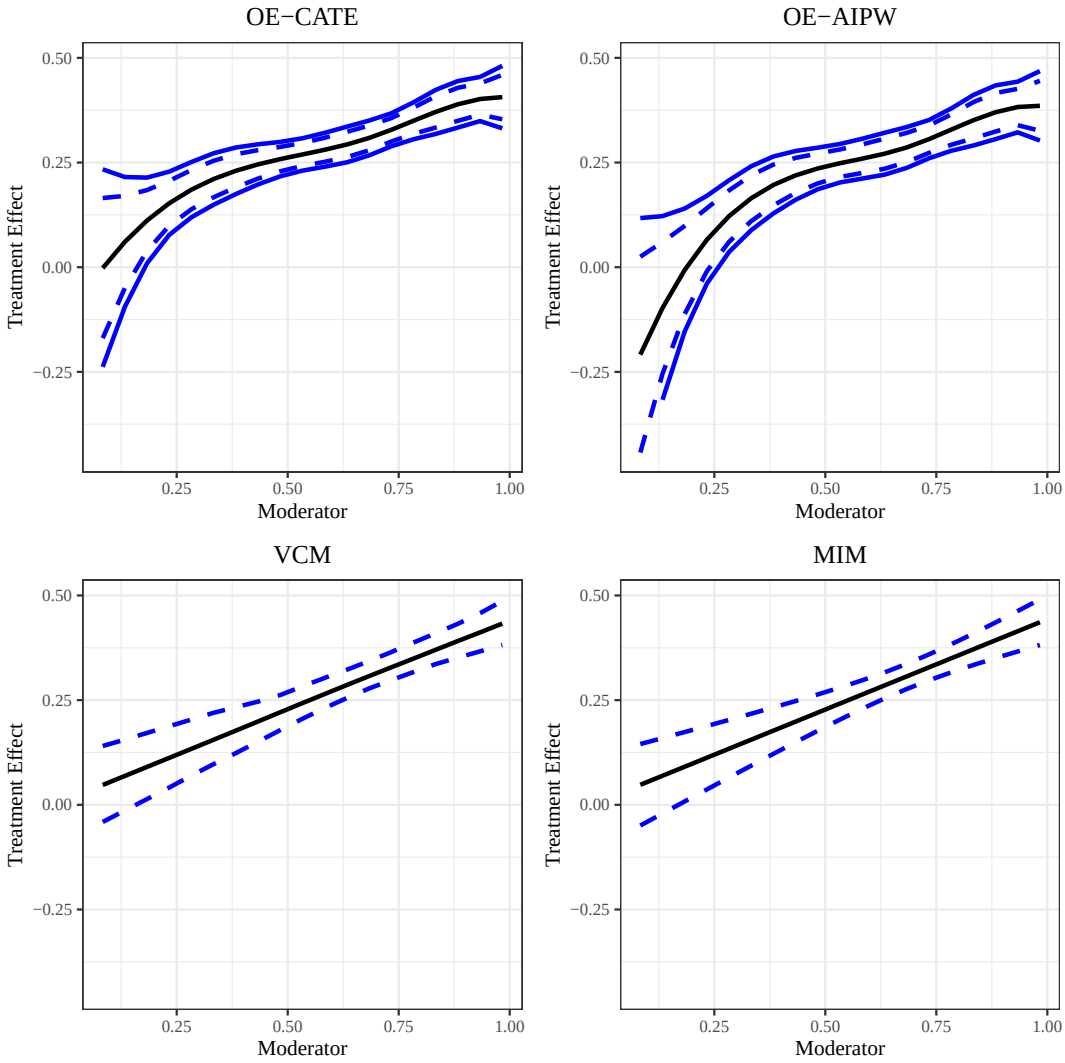
In this paper, we propose the Orthogonal Estimator (OE) as a flexible and robust approach to estimating conditional marginal effects. Our method incorporates state-of-the-art machine learning techniques to handle high dimensional control variables and allows for a fully nonparametric functional form, which requires fewer assumptions compared to many existing methods. We demonstrate the advantages of OE, including its robustness to a wide range of DGPs and its consistency with conventional approaches when their restrictive assumptions are met.

We introduce two OEs, OE-AIPW and OE-CATE, both of which are consistent, asymptotically normal, and oracle efficient. While the two estimators share many similarities, we show that the OE-CATE, which is new to the literature, may enjoy finite sample improvement over OE-AIPW

[15] Our replication is based on the model presented in Column 2, Table 5 of Huddy, Mason and Aarøe (2015).

[16] When the moderator is less than 0.25, the point estimates derived from OEs deviate from those obtained through VCM and MIM. In this replication case, we believe that the inherent flexibility of OEs facilitates a more plausible extrapolation in scenarios where common support is lacked.

Figure 4: Replication of Huddy, Mason and Aarøe (2015)



The horizontal axis represents the moderating variable, which is the strength of partisan identity, while the vertical axis represents the treatment effect of electoral loss from different estimators. The blue lines represents confidence intervals (solid blue lines for uniform confidence band, and dashed blue lines for pointwise confidence intervals).

when the treatment-specific outcome regression function cannot be well approximated while the conditional average treatment effect function can. Through Monte Carlo simulation exercises and replications of two social science publications, we illustrate the potential applications of our proposed method in addressing treatment effect heterogeneity under various scenarios. Our results confirm the superiority of the OE-CATE estimator in complex settings and its ability to generate estimates consistent with conventional methods in simple settings. In conclusion, our proposed orthogonal estimator provides a supplementary strategy to the existing techniques used for estimating conditional marginal effects. This flexible approach could be beneficial for practitioners who are interested in testing conditional theories in both observational and experimental settings. Future research could extend our proposed method to more complex scenarios, such as multivalued treatments, continuous treatments, and endogenous treatments.

References

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

Athey, Susan, Julie Tibshirani, Stefan Wager et al. 2019. "Generalized random forests." *The Annals of Statistics* 47(2):1148–1178.

Athey, Susan and Stefan Wager. 2019. "Estimating treatment effects with causal forests: An application." *arXiv preprint arXiv:1902.07409* .

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov and Kengo Kato. 2015. "Some new asymptotic theory for least squares series: Pointwise and uniform results." *Journal of Econometrics* 186(2):345–366.

Belloni, Alexandre, Victor Chernozhukov and Ying Wei. 2016. "Post-selection inference for generalized linear models with many controls." *Journal of Business & Economic Statistics* 34(4):606–619.

Brambor, Thomas, William Roberts Clark and Matt Golder. 2006. "Understanding interaction models: Improving empirical analyses." *Political Analysis* 14(1):63–82.

Chen, Yixing, Ju-Yeon Lee, Shrihari Sridhar, Vikas Mittal, Katharine McCallister and Amit G Singal. 2020. "Improving cancer outreach effectiveness through targeting and economic assessments: Insights from a randomized field experiment." *Journal of Marketing* 84(3):1–27.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters.".

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey and James M Robins. 2022. "Locally robust semiparametric estimation." *Econometrica* 90(4):1501–1535.

Chernozhukov, Victor, Whitney K Newey and Rahul Singh. 2022. "Automatic debiased machine learning of causal and structural effects." *Econometrica* 90(3):967–1027.

Dasgupta, Aditya and Devesh Kapur. 2020. "The political economy of bureaucratic overload: Evidence from rural development officials in india." *American Political Science Review* 114(4):1316–1334.

De la Cuesta, Brandon and Kosuke Imai. 2016. "Misunderstandings about the regression discontinuity design in the study of close elections." *Annual Review of Political Science* 19:375–396.

Fan, Qingliang, Yu-Chin Hsu, Robert P Lieli and Yichong Zhang. 2022. "Estimation of conditional average treatment effects with high-dimensional data." *Journal of Business & Economic Statistics* 40(1):313–327.

Farrell, Max H, Tengyuan Liang and Sanjog Misra. 2021. "Deep neural networks for estimation and inference." *Econometrica* 89(1):181–213.

Foster, Dylan J and Vasilis Syrgkanis. 2023. "Orthogonal statistical learning." *The Annals of Statistics* 51(3):879–908.

Foster, Jared C, Jeremy MG Taylor and Stephen J Ruberg. 2011. "Subgroup identification from randomized clinical trial data." *Statistics in medicine* 30(24):2867–2880.

Glynn, Adam N and Kevin M Quinn. 2010. "An introduction to the augmented inverse propensity weighted estimator." *Political analysis* 18(1):36–56.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods." *Political Analysis* 25(4):413–434.

Guryan, Jonathan, Jens Ludwig, Monica P Bhatt, Philip J Cook, Jonathan MV Davis, Kenneth Dodge, George Farkas, Roland G Fryer Jr, Susan Mayer, Harold Pollack et al. 2023. "Not too late: Improving academic outcomes among adolescents." *American Economic Review* 113(3):738–765.

Hahn, Jinyong. 1998. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects." *Econometrica* pp. 315–331.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice." *Political Analysis* 27(2):163–192.

Hartman, Erin. 2021. "Equivalence testing for regression discontinuity designs." *Political Analysis* 29(4):505–521.

Hill, Jennifer L. 2011. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Hirshberg, David A and Stefan Wager. 2021. "Augmented minimax linear estimation." *The Annals of Statistics* 49(6):3206–3227.

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive partisanship: Campaign involvement, political emotion, and partisan identity." *American Political Science Review* 109(1):1–17.

Ichimura, Hidehiko and Whitney K Newey. 2022. "The influence function of semiparametric estimators." *Quantitative Economics* 13(1):29–61.

Imai, Kosuke, Marc Ratkovic et al. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kennedy, Edward H. 2023. "Towards optimal doubly robust estimation of heterogeneous causal effects." *Electronic Journal of Statistics* 17(2):3008–3049.

Knaus, Michael C. 2022. "Double machine learning-based programme evaluation under unconfoundedness." *The Econometrics Journal* 25(3):602–627.

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

Lee, David S. 2008. "Randomized experiments from non-random selection in us house elections." *Journal of Econometrics* 142(2):675–697.

Lee, David S., Enrico Moretti and Matthew J. Butler. 2004. "Do voters affect or elect policies? evidence from the us house." *The Quarterly Journal of Economics* 119(3):807–859.

Lee, Sokbae, Ryo Okui and Yoon-Jae Whang. 2017. "Doubly robust uniform confidence band for the conditional average treatment effect function." *Journal of Applied Econometrics* 32(7):1207–1225.

Li, Qi and Jeffrey S Racine. 2010. "Smooth varying-coefficient estimation and inference for qualitative and quantitative data." *Econometric Theory* 26(6):1607–1637.

Li, Qi and Jeffrey Scott Racine. 2007. *Nonparametric econometrics: theory and practice*. Princeton University Press.

Luo, Ye, Martin Spindler and Jannis Kück. 2016. "High-dimensional l_2 boosting: Rate of convergence." *arXiv preprint arXiv:1602.08927* .

Newey, Whitney K. 1997. "Convergence rates and asymptotic normality for series estimators." *Journal of econometrics* 79(1):147–168.

Nie, Xinkun and Stefan Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika* 108(2):299–319.

Powers, Scott, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie and Robert Tibshirani. 2018. "Some methods for heterogeneous treatment effect estimation in high dimensions." *Statistics in medicine* 37(11):1767–1787.

Robins, James M, Andrea Rotnitzky and Lue Ping Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American Statistical Association* 89(427):846–866.

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

Rubin, Donald B. 1980. "Randomization analysis of experimental data: The fisher randomization test comment." *Journal of the American statistical association* 75(371):591–593.

Semenova, Vira and Victor Chernozhukov. 2021. "Debiased machine learning of conditional average treatment effects and other causal functions." *The Econometrics Journal* 24(2):264–289.

Smucler, Ezequiel, Andrea Rotnitzky and James M Robins. 2019. "A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts." *arXiv preprint arXiv:1904.03737* .

Stadie, Bradly C, Sören R Künzel, Nikita Vemuri and Jasjeet S Sekhon. 2018. "Estimating heterogeneous treatment effects using neural networks with the y-learner.".

Syrgkanis, Vasilis and Manolis Zampetakis. 2020. Estimation and inference with trees and forests in high dimensions. In *Conference on learning theory*. PMLR pp. 3453–3454.

Vaart, AW van der and Jon A Wellner. 2023. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer pp. 127–384.

Yin, Weiwen, Weidong Huo and Danyang Lin. 2023. "The effects of state coercion on voting outcome in protest movements: a causal forest approach." *Political Science Research and Methods* 11(3):645–653.

Zheng, Li and Weiwen Yin. 2023. "Estimating and evaluating treatment effect heterogeneity: A causal forests approach." *Research & Politics* 10(1):20531680231153080.

Zimmert, Michael and Michael Lechner. 2019. "Nonparametric estimation of causal heterogeneity under high-dimensional confounding." *arXiv preprint arXiv:1908.08779* .