

Statistica Sinica Preprint No: SS-2026-0330	
Title	Outlier Resistant Inference for Conditional Average Treatment Effect
Manuscript ID	SS-2026-0330
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0330
Complete List of Authors	Ran Mo and Honglang Wang
Corresponding Authors	Honglang Wang
E-mails	hlwang@iu.edu
Notice: Accepted author version.	

Outlier Resistant Inference for Conditional Average Treatment Effect

Ran Mo and Honglang Wang

Indiana University Indianapolis

Abstract: The estimation of causal effects based on the conditional average treatment effect (CATE) is usually vulnerable to outliers. However, to the best of our knowledge, outlier-resistant inference for the CATE has not been investigated in the literature. In this work, we propose an outlier-resistant estimation method for the CATE by incorporating M-estimation into the inverse propensity weighting (IPW) approach. The influence function and breakdown property are investigated to study the robustness of our method. In addition, we derive the asymptotic properties of the proposed estimator for inference purposes. The finite sample performance of the proposed estimator is evaluated via Monte Carlo experiments. The proposed method is compared with the IPW method and the augmented inverse probability weighting (AIPW) method, which do not account for outliers, as well as with DR-learner variants and machine-learning meta-learners, including the causal forest, X-learner, and R-learner. Finally, the proposed method is applied to the NHANES dataset to estimate the average effects of physical activity on albumin levels conditioned on age.

Key words and phrases: Conditional average treatment effect, Outlier-resistant estimators, Inverse probability weighting, M-estimator, Kernel regression.

1. Introduction

Treatment effects are widely employed across diverse fields such as healthcare, economics, education, and social sciences to assess the impact of interventions, policies, or treatments on outcomes of interest. Because we do not observe both potential outcomes for each individual simultaneously, estimating the individual treatment effect is generally impossible (Holland, 1986), a fact known as the “fundamental problem of causal inference.” A reasonable approach in the literature has been to estimate the average treatment effect (ATE) (Rosenbaum and Rubin, 1983; Hirano et al., 2003), which provides a single estimate of the treatment’s impact on an entire population, thereby offering valuable insights into the overall effectiveness of the intervention.

With the growing interest in estimating and analyzing heterogeneous treatment effects in both experimental and observational studies, researchers are increasingly focusing on the conditional average treatment effect (CATE). Traditional methods include regression adjustment (Rosenbaum and Rubin, 1983), propensity score matching (Rosenbaum and Rubin, 1983), inverse probability weighting (IPW) (Hirano et al., 2003), and augmented inverse probability weighting (AIPW) (Glynn and Quinn, 2010; Kurz, 2022). Beyond these traditional approaches, researchers have adopted a variety of methodologies, including meta-learners such as the X-learner (Künzel et al., 2019), T-learner (Athey and Imbens, 2016; Yao et al., 2018; Powers et al., 2018), R-learner (Nie and Wager, 2021), DR-learner (van der Laan, 2006; Lee et al., 2017), and S-learner (Dorie et al., 2019; Foster et al., 2011). These methods decompose CATE estimation into regression sub-problems, enabling the use of flexible nonparametric and machine learning regression methods. Additionally, there are algorithms

specifically designed for CATE estimation, such as causal forests (Athey et al., 2019) and Bayesian additive regression trees (BART) (Hill, 2011; Chipman et al., 2010).

In practical applications, researchers may only be interested in the CATE of subpopulations conditioned on a subset of all covariates. The definition of CATE has been extended by Abrevaya et al. (2015) to address a more technically challenging situation where the conditioning covariates X_1 are continuous and form a strict subset of X . Since the unconfoundedness assumption does not generally hold when conditioning solely on X_1 , it is inadequate to simply apply, for example, the CATE estimator from Lee and Whang (2009) using X_1 in place of X . Instead, one must estimate CATE as a function of X and subsequently average out the unwanted components of X not included in X_1 . However, this distribution is typically unknown and must be estimated. Besides Abrevaya et al. (2015), Lee et al. (2017) and Fan et al. (2022) also extended the doubly robust (DR) estimator conditioning on a subset of all covariates while also allowing $X_1 = X$. Further work in this direction includes Luedtke and van der Laan (2016) and Luedtke and van der Laan (2017), who developed inference strategies for the mean outcome under the optimal treatment strategy based on the estimated CATE in epidemiologic settings, and VanderWeele et al. (2019), who studied selection of optimal subgroups for treatment.

In real-world scenarios, treatment effect estimation often encounters data contaminated by outliers—extreme points that significantly deviate from the majority of observations. When outliers are present in the response variable, estimators based on the sample mean tend to be heavily influenced by these outliers. While robust techniques for estimating ATE have been studied (Firpo, 2007; Zhang et al., 2012; Harada and Fujisawa, 2021), there is a

gap in the literature regarding outlier-resistant estimators for the CATE. Existing methods such as those proposed by Abrevaya et al. (2015), Lee et al. (2017), and Fan et al. (2022) exhibit limited outlier resistance due to their reliance on the L_2 losses.

To mitigate the impact of outliers in the response variable, robust statistical methods, such as M-estimators, are commonly employed. First introduced by Huber (1964) as an outlier-resistant location estimator and later adapted for kernel regression by Härdle (1984), M-estimators estimate parameters by minimizing a robust loss function rather than relying solely on the mean or median. This approach reduces sensitivity to outliers while preserving estimation efficiency.

Building on traditional robust methods, this paper extends the second step of the semi-parametric method in Abrevaya et al. (2015) by replacing the Nadaraya-Watson (NW) estimator with a more general class of local M-estimators (Härdle, 1984). The use of M-estimators allows us to relax the requirement of finite conditional moments of the potential outcomes, as imposed in Abrevaya et al. (2015). Instead, we assume the existence of finite conditional moments for a robust loss function of the potential outcomes—a condition typically satisfied for bounded robust loss functions, such as the Huber loss. This modification ensures that asymptotic properties hold even in the presence of heavy-tailed distributions or outlier contamination, provided the robust loss function is appropriately chosen.

To delineate our contribution relative to Abrevaya et al. (2015) explicitly, we inherit from that work the IPW identification strategy, the partially conditional estimand $\tau(x_1)$ with a low-dimensional continuous conditioning covariate, and the reliance on a correctly specified parametric propensity score model. What is new in this paper is the robustifica-

tion of the second-stage local estimation together with the accompanying inference theory: (i) we replace the local least-squares smoother with a general class of local M-estimators and characterize the resulting estimator's influence function and finite-sample breakdown point; (ii) we establish consistency and asymptotic normality under conditional moment conditions on the robust loss that are weaker than the finite conditional moment requirements in Abrevaya et al. (2015) and that accommodate heavy-tailed and contaminated outcome distributions; and (iii) we develop a consistent variance estimator and pointwise confidence intervals whose validity is retained under contamination. Our contribution is therefore an extension of the IPW-CATE framework to outlier-contaminated data—incremental in its identification strategy but new in its robustness and inference theory—rather than a new identification approach.

We establish the robustness of our estimator by analyzing its influence function (IF) and breakdown point. Additionally, we derive asymptotic properties and construct confidence intervals using mean and variance estimators based on asymptotic distributions. To verify our theoretical results, we conduct Monte Carlo simulations, evaluating the performance of our estimators in terms of bias, variance, and mean squared error (MSE) under data contamination. The validity of the asymptotic distribution and variance is further confirmed through the constructed confidence intervals. Beyond the IPW and AIPW benchmarks, our simulation studies also compare the proposed estimators with robustified DR-learner variants and with machine-learning meta-learners (causal forest, X-learner, and R-learner). These comparisons show that robustifying the DR pipeline reduces estimation error but fails to deliver valid inference under contamination, that the forest-based meta-learners offer no

protection against heavy tails and come without pointwise inference theory for the partially conditional CATE, and that robustness should therefore be applied directly to the outcome rather than to a derived pseudo-outcome.

Finally, we apply our robust CATE estimator to investigate the effect of vigorous work activity on urinary albumin levels using data from the 2007–2008 National Health and Nutrition Examination Survey (NHANES). The robust method effectively accounts for outliers, and our findings align with existing studies, indicating a protective effect of vigorous work activity on albumin levels, particularly among individuals over 60. Moreover, the narrower confidence intervals from the robust estimators provide stronger statistical evidence regarding the impact of physical activity on albumin excretion.

The rest of the article is organized as follows. Section 2 introduces the outlier-resistant CATE estimator and discusses its identification and estimation. The algorithm for our proposed estimator is also presented in this section. In Section 3, we demonstrate the outlier resistance of our estimator by examining its behavior from the perspectives of the influence function and the breakdown point. Section 4 develops the asymptotic properties of the proposed estimators. Section 5 and 6 present the simulation results and real data application respectively. Section 7 concludes the paper.

2. Outlier-Resistant CATE Estimation

2.1 Definition of the Reduced Dimensional CATE

Under the potential outcome framework (Rubin, 1974; Neyman, 1923), let $Y^{(1)}$ and $Y^{(0)}$ denote the potential outcomes with and without treatment, respectively. For each individual,

the observed outcome Y is defined by

$$Y = TY^{(1)} + (1 - T)Y^{(0)},$$

where $T \in \{0, 1\}$ is an indicator variable of the treatment state, with $T = 1$ if an individual receives treatment and $T = 0$ otherwise. Let X be a p -dimensional vector of covariates with $p \geq 2$. The observations are given by $\{(T_i, Y_i, X_i), i = 1, 2, \dots, n\}$ from the independent and identically distributed joint distribution of the vector (T, Y, X) . The reduced dimensional conditional average treatment effect is defined as

$$\tau(x_1) = E(Y^{(1)} - Y^{(0)} \mid X_1 = x_1) = \mu_1(x_1) - \mu_0(x_1), \quad (2.1)$$

where X_1 is a fixed d -dimensional subvector of the covariates X , with $d \leq p$, and $\mu_1(x_1) = E(Y^{(1)} \mid X_1 = x_1)$, $\mu_0(x_1) = E(Y^{(0)} \mid X_1 = x_1)$ are the conditional means of the potential outcomes given X_1 .

2.2 Identification of CATE

Since the potential outcomes in the treatment and control groups cannot be observed simultaneously, we identify our parameter of interest using the observed outcome Y in place of the potential outcomes.

Following Abrevaya et al. (2015), we consider the situation where the unconfoundedness assumption does not generally hold when conditioning on the low-dimensional covariates X_1 . The identification of the estimand $\mu_1(x_1)$ and $\mu_0(x_1)$ is based on the common assumptions (Rosenbaum and Rubin, 1983):

Assumption 1. (a) (Unconfoundedness) $(Y^{(1)}, Y^{(0)}) \perp T \mid X$.

- (b) (Positivity) With $\pi(x) = P(T = 1 \mid X = x)$ as the propensity score, we assume there exists $C > 0$ such that $P(C \leq \pi(X) \leq 1 - C) = 1$.

Based on Assumption 1, the IPW method proposed by Abrevaya et al. (2015) is based on the following one-model estimating equation:

$$0 = E \left[\left(\frac{TY}{\pi(X)} - \frac{(1-T)Y}{1-\pi(X)} - \tau(x_1) \right) \mid X_1 = x_1 \right],$$

and the estimator $\hat{\tau}(x_1)$ is found using its empirical form:

$$0 = \frac{1}{n} \sum_{i=1}^n \left[\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)} - \hat{\tau}(x_1) \right] K \left(\frac{X_{i1} - x_1}{h} \right), \quad (2.2)$$

where $K(\cdot)$ is a kernel function with bandwidth h and $\hat{\pi}(x)$ is the estimator of propensity score.

When outliers are present in the response, a direct approach to incorporate (2.2) with the M-estimator might involve truncating the residual

$$\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)} - \hat{\tau}(x_1)$$

with a robust function such as the Huber loss function.

However, this naive approach does not correctly target the contamination. To see why, observe that the contamination occurs in the original response Y_i : when Y_i is an outlier with value $y_0 \rightarrow \infty$, the IPW-reweighted outcome $T_i Y_i / \hat{\pi}(X_i)$ for a treated observation inflates to $y_0 / \hat{\pi}(X_i)$, which conflates the outlier magnitude with the inverse of the propensity score. Truncating the combined IPW residual therefore over- or under-penalizes the outlier depending on $\hat{\pi}(X_i)$, rather than cleanly downweighting the contaminated observation. To see this formally, the influence function of an estimator based on truncating the IPW residual

2.3 Algorithm for Robust CATE

directly would involve $\psi(y_0/\pi(X) - \dots)$, which is still unbounded in y_0 for most ψ unless $\pi(X)$ is bounded away from zero. In contrast, the correct approach downweights the original potential outcome residual $(Y_i - \mu_1(x_1))$ through $\psi(Y_i - \mu_1(x_1))$, which remains bounded for bounded ψ regardless of $\pi(X_i)$. To derive an outlier-resistant method based on the idea of M-estimator, we need to truncate the direct difference of potential outcomes and their mean functions, $(Y_i - \mu_1(x_1))$ when $T = 1$ or $(Y_i - \mu_0(x_1))$ when $T = 0$ and the corresponding estimator $\hat{\tau}(x_1)$ can be calculated by $\hat{\tau}(x_1) = \hat{\mu}_1(x_1) - \hat{\mu}_0(x_1)$. In this study, our demonstration will focus solely on $\mu_1(x_1)$, as $\mu_0(x_1)$ can be addressed in a similar manner.

Here we provide the identification of $\mu_1(x_1)$ based on Assumption 1. Under the unconfoundedness and positivity assumptions, it is easy to see

$$0 = E \left[\frac{T}{\pi(X)} (Y - \mu_1(x_1)) \mid X_1 = x_1 \right]. \quad (2.3)$$

With this, the solution $\hat{\mu}_1(x_1)$ to its empirical form can be regarded as a local constant estimator with additional weight $T_i/\hat{\pi}(X_i)$:

$$0 = \sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)} (Y_i - \hat{\mu}_1(x_1)) K \left(\frac{X_{i1} - x_1}{h} \right).$$

2.3 Algorithm for Robust CATE

To derive a robust estimator based on the estimating equation, we replace the term $(Y - \mu_1(x_1))$ with an anti-symmetric function $\psi(Y - \mu_1(x_1))$ and modify the estimating equation as follows:

$$0 = E \left[\frac{T}{\pi(X)} \psi(Y - \mu_1(x_1)) \mid X_1 = x_1 \right]. \quad (2.4)$$

2.3 Algorithm for Robust CATE

However, to ensure that (2.4) holds, additional assumptions are required. The symmetry assumption is widely used as a prerequisite for applying classical robust methods (Rousseeuw et al., 1986), which tend to down weight or ignore extreme observations on both sides of the mean.

Assumption 2. $\psi(\cdot)$ is antisymmetric, and the conditional density $f(y^{(1)} | x_1)$ is symmetric with respect to $\mu_1(x_1)$.

Remark 1. Assumption 2 requires that the conditional density of $Y^{(1)}$ given $X_1 = x_1$, obtained by marginalizing over the remaining covariates $X_2^c = (X_2, \dots, X_p)$, is symmetric around $\mu_1(x_1)$. Crucially, this is a condition on the *marginalized* density $f(y^{(1)} | x_1) = \int f(y^{(1)} | x)g(x_2^c | x_1)dx_2^c$, not on the full conditional $f(y^{(1)} | x)$. Even if the full conditional density is mildly asymmetric, integration over X_2^c can restore approximate symmetry in certain settings, particularly when X_2^c is independent of X_1 and the asymmetry averages out. In practice, near-symmetry of the response distribution—as assessed, for example, by inspecting residuals after removing the treatment group mean—is a sufficient empirical guide. The Huber loss is relatively robust to mild violations of this symmetry, as illustrated in our simulation results under the log-normal outlier setting.

Theorem 1. Under Assumptions 1 and 2, we have $E\left[\frac{T}{\pi(X)}\psi(Y - \mu_1(x_1)) \middle| X_1 = x_1\right] = 0$, i.e., equation (2.4) is valid.

With Theorem 1, we propose to estimate $\hat{\mu}_1(x_1)$ by solving the empirical form of the estimating equation:

$$0 = \sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)} \psi(Y_i - \hat{\mu}_1(x_1)) K\left(\frac{X_{i1} - x_1}{h}\right). \quad (2.5)$$

2.3 Algorithm for Robust CATE

The solution to equation (2.5) can be viewed as a weighted M-estimator with weight $T_i K((X_{i1} - x_1)/h)/\pi(X_i)$. Therefore, we can solve (2.5) using the iteratively reweighted least squares (IRLS) algorithm (Holland and Welsch, 1977) with the following iterative procedure. With $\hat{\mu}_1^{(t)}(x_1)$ from t -th step ($t \geq 0$) we solve $\hat{\mu}_1^{(t+1)}(x_1)$ from

$$0 = \sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)} K\left(\frac{X_{i1} - x_1}{h}\right) W(Y_i - \hat{\mu}_1^{(t)}(x_1))(Y_i - \hat{\mu}_1^{(t+1)}(x_1)),$$

where $W(x) = \psi(x)/x$ for $x \neq 0$ and $W(x) = \psi'(x)$ for $x = 0$, i.e. a linear approximation of $\psi(x)$ for x close to 0.

Furthermore, the propensity score is typically unknown in observational studies. Similar to Abrevaya et al. (2015), we assume a correctly specified parametric model for the propensity score such as logistic regression model.

Remark 2. Assumption 3 below formalizes the requirement that the propensity score is correctly specified. This is a substantively strong assumption: the validity of our estimator depends on the parametric model for $\pi(x)$ being correctly specified, and misspecification of the propensity score model generally leads to inconsistency of $\hat{\mu}_1(x_1)$. We note this in contrast to doubly robust (DR) methods (Robins et al., 1994; Robins and Rotnitzky, 1995), which remain consistent if either the propensity score or the outcome regression model is correctly specified, but not necessarily both. In practice, the adequacy of the fitted propensity model can be probed with standard diagnostics, such as checking covariate balance after inverse-probability weighting and assessing the calibration of the fitted propensities against empirical treatment frequencies (Cole and Hernán, 2008). An outlier-resistant extension that incorporates a doubly robust augmentation term into the estimating equation (2.4) is a natural avenue for future work that would relax this requirement.

3. Robustness Analysis

In this section, we demonstrate the outlier resistance property of the proposed estimator by examining its behavior from the perspectives of the influence function and the breakdown point.

3.1 Influence Function

Introduced by Hampel (1974), the influence function (IF) of an estimator provides an approximation of how the estimator behaves when the sample contains a small fraction ϵ of identical outliers. Following Harada and Fujisawa (2021), we demonstrate the outlier-resistant property of our proposed estimator $\hat{\mu}_1(x_1)$ from the viewpoint of the influence function.

Let $\theta[F]$ denote the parameter of interest θ viewed as a functional of the distribution F —that is, θ is uniquely determined by F and does not otherwise depend on the data except through F . The influence function of an estimator $\hat{\theta}$ at distribution F is defined as:

$$IF_{\hat{\theta}(x_0, F)} = \lim_{\epsilon \rightarrow 0} \frac{\hat{\theta}[(1 - \epsilon)F + \epsilon\delta_{x_0}] - \theta[F]}{\epsilon} = \frac{\partial}{\partial \epsilon} \left\{ \hat{\theta}[(1 - \epsilon)F + \epsilon\delta_{x_0}] \right\} \Big|_{\epsilon=0},$$

where δ_x denotes the Dirac delta function for point mass at x .

Back to our problem, we denote $\tilde{g}$ as the contaminated density function of $Y \mid (X = x, T = 1)$ given by $\tilde{g}(y|x) = (1 - \epsilon)g(y|x) + \epsilon\delta_{y_0}$ where $g(y|x)$ denotes the uncontaminated density function of $Y \mid (X = x, T = 1)$. Under the contaminated distribution $\tilde{g}(y|x)$, denote $\mu_\epsilon(x_1)$ as the solution to our estimating equation:

$$0 = E_{\tilde{g}} \left[\frac{T}{\pi(X)} \psi(Y - \mu_\epsilon(x_1)) \mid X_1 = x_1 \right] = E_{\tilde{g}} \left[\psi(Y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1 \right].$$

Taking the partial derivative with respect to ϵ for the above equation,

$$\begin{aligned} 0 = & -E_g [\psi(Y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1] \\ & - (1 - \epsilon) \frac{\partial \mu_\epsilon(x_1)}{\partial \epsilon} E_g [\psi'(Y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1] \\ & + \psi(y^{(1)} - \mu_\epsilon(x_1)) \delta_{y_0} g(t|x) g(x|x_1) dy^{(1)} dt dx_1^c \\ & - \epsilon \frac{\partial \hat{\mu}_\epsilon}{\partial \epsilon} \iiint \psi'(y^{(1)} - \mu_1(x_1)) \delta_{y_0} g(y^{(1)}|t, x) g(t|x) g(x|x_1) dy^{(1)} dt dx_1^c, \end{aligned}$$

here we use x_1^c to denote the other components in x except x_1 . As $\epsilon \rightarrow 0$, the first and last terms vanish, and $\partial \mu_\epsilon(x_1)/\partial \epsilon$ corresponds to the influence function. Taking the limit $\epsilon \rightarrow 0$, we obtain:

$$0 = -IF_{(y_0, \tilde{g})} E_g [\psi'(y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1] + \psi(y_0 - \mu_\epsilon(x_1)).$$

Therefore,

$$IF_{(y_0, \tilde{g})} = \{E_g [\psi'(y^{(1)} - \mu_\epsilon(x_1)) \mid X_1 = x_1]\}^{-1} \psi(y_0 - \mu_\epsilon(x_1)).$$

When $y_0 \rightarrow \infty$, the value of the influence function is bounded if $\psi(\cdot)$ is bounded (e.g., Tukey's biweight, where $\psi(u) \rightarrow 0$ as $|u| \rightarrow \infty$, is an example of a re-descending loss function) or approaches 0 for certain re-descending loss functions. These properties are advantageous for outlier resistance.

3.2 Breakdown Properties

In this subsection, we're going to discuss the breakdown point of a location estimator with an estimating equation of the form:

$$0 = \sum_{i=1}^n \psi(Y_i - \mu_1(x_1)) W_i(x_1),$$

where the weights are defined as $W_i(x_1) = T_i K((X_{i1} - x_1)/h)/\pi(X_i)$.

To imitating the contamination of a sample, there are two approaches: replacement contamination and addition contamination (Donoho and Huber, 1983). In this study, we adopt the replacement contamination approach to define the finite sample breakdown point. Specifically, we consider a sample of size n and randomly replace m of the observations with corrupted ones. We denote S_{Y_n} as the set of the outcomes in noncontaminated observations and S_{Y_c} as the set of the outcomes of contaminated observations. Then the estimating equation can thus be rewritten as:

$$0 = \sum_{Y_i \in S_{Y_n}} \psi(Y_i - \mu_1(x_1)) W_i(x_1) + \sum_{Y_i \in S_{Y_c}} \psi(Y_i - \mu_1(x_1)) W_i(x_1).$$

Since the weights are notably related to x_1 , a fixed global breakdown point cannot be determined. Following the existing study in Giloni and Simonoff (2005), we define the breakdown point of our estimator in a local and conditional manner:

$$\varepsilon(x_1|X, T) = \min_{0 \leq m^* \leq n/2} \left\{ \frac{m^*}{n} \left| \sup_{Y_i \in S_{Y_c}} |\hat{\mu}_1(S_{Y_n} \cup S_{Y_c}) - \hat{\mu}_1(Y)| = \infty \right. \right\},$$

where the supremum is taken over all the choices of S_{Y_c} by replacing $m^* = |S_{Y_c}|$ of the n original Y with arbitrary new data.

To study the breakdown property, Li and Zhang (1998) categorized the general class of loss functions ρ into three different types: $B\rho$, $U\rho$, and $C\rho$. Since our ψ function can be regarded as the derivative of a differentiable ρ function, we can follow the categories in Li and Zhang (1998) and study the breakdown property of our estimator $\hat{\mu}_1(x_1)$ in the following conditions.

Condition 1: ($B\rho$) $\rho(x)$ attains its minimum -1 at $x = 0$; ρ is non-increasing for $x < 0$ and non-decreasing for $x > 0$. Furthermore, $\rho(x) \rightarrow 0$ as $|x| \rightarrow \infty$.

3.2 Breakdown Properties

Condition 2: ($U\rho$) $\rho(x)$ attains its minimum 0 at $x = 0$; $\rho(x)$ is symmetric about 0; $\rho(x)$ is non-decreasing for $x > 0$ and $\lim_{|x| \rightarrow \infty} \rho(x) = \infty$; $\psi(x) = \rho'(x)$ is continuous in $\mathbb{R}^1$, and there exists $x_0 \geq 0$ such that ψ is non-decreasing in $(0, x_0]$ and non-increasing in (x_0, ∞) .

Condition 3: ($C\rho$) $\rho(x)$ is convex, $\psi(x) := \rho'(x)$ exists everywhere and $\psi(-\infty) < 0 < \psi(+\infty)$.

The three conditions outlined above include many commonly used loss functions, such as L_2 loss, L_1 loss, Huber loss (Huber, 1973), and Tukey's biweighted loss (Beaton and Tukey, 1974).

Under Condition 1, the derivative of the loss function approaches 0 as x goes to infinity, and the loss function remains bounded. All bounded redescending M-estimators, such as the M-estimator based on Tukey's biweighted loss, can be rescaled to satisfy this condition. Condition 2 encompasses symmetric unbounded redescending M-estimators. Specifically, it includes M-estimators where the derivative of the loss function is bounded and bounded away from zero, while the loss function itself is unbounded. This class of M-estimators includes those based on the Huber loss, the L_1 loss, and other loss functions that tend to infinity at a slower rate than the Huber and L_1 losses. Finally, Condition 3 includes the Huber loss, the L_1 loss, and other loss functions that diverge more rapidly than the Huber and L_1 losses.

Based on these conditions, we have the following theorem:

Theorem 2. Denote $W_n = \sum_{Y_i \in S_{Y_n}} W_i(x_1)$, $W_c = \sum_{Y_i \in S_{Y_c}} W_i(x_1)$, $W_{n/c} = W_n/W_c$ and $A = -\sum_{Y_i \in S_{Y_n}} \rho(Y_i - \mu_1(x_1))W_i(x_1) / \sum_{Y_i \in S_{Y_n}} W_i(x_1)$. We have

(1) Under $(B\rho)$, $\hat{\mu}_1(x_1)$ will be unbounded if $W_{n/c} < (1 - A)/A$ and will be finite when

$$W_{n/c} > (1 - A)/A.$$

3.2 Breakdown Properties

(2) Under $(U\rho)$, $\hat{\mu}_1(x_1)$ will be bounded if $W_{n/c} > 1$ and will be unbounded if $W_{n/c} < 1$.

(3) Under $(C\rho)$, assume that $\psi(-\infty) = k_1$ and $\psi(+\infty) = k_2$. $\hat{\mu}_1(x_1)$ will be bounded if $W_{n/c} \geq \max\{k_1/k_2, k_2/k_1\}$, and will be unbounded if $W_{n/c} < \max\{k_1/k_2, k_2/k_1\}$.

Theorem 2 indicates that the breakdown properties of the estimator are closely related to the ratio $W_{n/c} = \sum_{Y_i \in S_{Y_n}} W_i(x_1) / \sum_{Y_i \in S_{Y_c}} W_i(x_1)$. For each case, the estimator will break down when this ratio falls below a certain threshold or be finite when this ratio falls over a certain threshold. This shows that the breakdown point is closely related to the minimum number m of the greatest weights chosen into the S_{Y_c} group that satisfies the breakdown condition corresponding to its loss function, that is:

$$\varepsilon(x_1|X, T) = \min_{0 \leq m^* < \frac{n}{2}} \left\{ \frac{m^*}{n} \mid \sup_{S_{Y_c}} W_{n/c} < C \right\},$$

where C is a constant specified in Theorem 2.

A concrete worked illustration of the breakdown point under a Gaussian kernel, showing that breakdown occurs roughly when the number of contaminated observations exceeds half the effective local sample size, is provided in Section S12 of the supplementary material.

In real-world applications, even though contaminated observations are randomly chosen, it is reasonable to consider the worst-case scenario where the observations with the highest weights are contaminated. When some of the $W_i(x_1)$'s are significantly larger than others, a few contaminated observations might be sufficient to cause the estimator to break down. Given the expression of $W_i(x_1) = T_i K((X_{i1} - x_1)/h) / \pi(X_i)$'s, to mitigate breakdown issues, we can consider either increasing the bandwidth h in the kernel function or truncating observations with extreme estimated propensity scores (Robins et al., 2007; Cole and Hernán,

2008). Note that increasing h beyond the asymptotically optimal rate (Assumption 8) reduces estimation efficiency by increasing bias; this is a practical trade-off rather than a recommendation for inference.

4. Asymptotic Properties of the Robust Estimator

In this section, we develop the asymptotic theory for our proposed estimator. We begin by establishing the asymptotic properties of $\hat{\mu}_1(x_1)$ and then extend our results to derive the asymptotic properties of $\hat{\tau}(x_1)$. Additionally, we provide the necessary variance estimator for constructing a point-wise confidence interval.

For simplicity, we assume that the propensity score converges at a rate fast enough for its estimation error to be dominated by the error of the parameter of interest as $n \rightarrow \infty$. Assumption 3 is a substantively strong parametric assumption: it requires not only that a parametric model for $\pi(x)$ is adopted, but also that this model is correctly specified. As noted in Remark 2, doubly robust methods (Robins et al., 1994) can circumvent this by allowing misspecification of either the propensity score or the outcome model. Assumption 3 typically holds for standard parametric estimation methods, such as logistic regression or maximum likelihood with a probit model, under reasonably mild regularity conditions.

Assumption 3. The propensity score is correctly specified and estimated by parametric method with $\sup_{x \in \mathbb{R}^p} |\hat{\pi}(x) - \pi(x)| = O_p(1/\sqrt{n})$.

With the estimated propensity score, our asymptotic theorems are established under the following underlying assumptions:

Assumption 4. $\psi(x)$ is monotone, bounded and having two continuous bounded derivatives

with $\psi'(0) > 0$.

Assumption 5. The conditional probability density function $f(y^{(1)}|x_1)$ has bounded partial derivative on $x_1 \in \mathbb{R}^d$. The density $g(x_1)$ of X_1 is positive on $\mathbb{R}^d$ and $s \geq 2$ times continuously differentiable.

Remark 3. Assumption 5 requires the marginal density $g(x_1)$ of the conditioning covariate X_1 to be $s \geq 2$ times continuously differentiable. This regularity condition on g is standard in local kernel regression (Härdle, 1984) and is needed for the uniform convergence of the kernel density estimator of g and for the Taylor expansion in the bias analysis. In practice, this is a mild condition satisfied whenever the distribution of X_1 is absolutely continuous and has a smooth density, as is typical for continuous covariates such as age or BMI.

Assumption 6. For $j = 0, 1$, $\sup_{x_1 \in \mathbb{R}^d} E[\psi(Y^{(j)} - \mu_j(x_1))^2 | X_1 = x_1] < \infty$ and $m_j(x_1) = E[\psi(Y^{(j)} - \mu_j(x_1)) | X_1 = x_1]$ is $s \geq 2$ times continuously differentiable.

Assumption 7. $K(u)$ is a kernel of order s , symmetric around 0 with bounded support.

Assumption 8. The bandwidths h satisfy the conditions: $nh^{2s+d} \rightarrow 0$ and $nh^d \rightarrow \infty$ as $n \rightarrow \infty$.

Remark 4. Here d denotes the dimension of the conditioning covariate X_1 , not the full covariate vector X . Since X_1 is chosen to be low-dimensional (e.g., $d = 1$ for conditioning on age in our application), the bias condition $nh^{2s+d} \rightarrow 0$ is much easier to satisfy than it would be if d were the full covariate dimension p . For example, with $d = 1$ and a kernel of order $s = 2$, the condition reduces to $nh^5 \rightarrow 0$, which permits $h \sim n^{-1/5-\delta}$ for small

$\delta > 0$. Higher-order kernels (larger s) further relax the smoothness requirements at the cost of allowing negative kernel values.

For convenience, we apply the notation and results from Härdle (1984). Let $\delta_n(\cdot)$ denote the delta function sequence (DFS), which satisfies (D1) $\int |\delta_n(u)| du < \infty$, for all n ; (D2) $\int \delta_n(u) du = 1$; (D3) $\delta_n(u) \rightarrow 0$, uniformly on $|u| > \eta, \eta > 0$ as $n \rightarrow \infty$; (D4) $\int_{|u| > \eta} \delta_n(u) du \rightarrow 0$, for each $\eta > 0$ as $n \rightarrow \infty$. Our choice of kernel $\delta(u) = K(u/h)/h^d$ can be regarded as DFS.

4.1 Consistency

Based on the definition of DFS, the estimating equation can be written as:

$$h(x_1, \mu_1) := \frac{1}{n} \sum_{i=1}^n \psi(Y_i - \mu_1(x_1)) \frac{T_i}{\hat{\pi}(X_i)} \delta_n(x_1 - X_{i1}) = 0.$$

Lemma 1. *Given Assumptions (1-8), we have $h(x_1, \mu_1) \xrightarrow{P} H(x_1, \mu_1)$ where*

$$H(x_1, \mu_1) = E \left[\psi(Y_i - \mu_1(x_1)) \frac{T_i}{\pi(X_i)} \mid X_{i1} = x_1 \right] g(x_1).$$

Building on Lemma 1 about the consistency of the estimating equation, we can establish the consistency of $\hat{\mu}_1(x_1)$.

Theorem 3. *Suppose $\psi(\cdot)$ is strictly increasing, and there exists a constant c such that $H(x_1, s) > 0$ for $s > c$ and $H(x_1, s) < 0$ for $s < c$. Then, under the assumptions in Lemma 1, $\hat{\mu}_1(x_1) \xrightarrow{P} \mu_1(x_1)$.*

Remark 5. For continuous but non-monotone ψ functions, such as Tukey's loss, we have multiple solutions to $h(x_1, \mu_1) = 0$. We choose the one which is closest to the estimator

$\hat{\mu}_1(x_1)$ from some monotone ψ functions as our estimator $\tilde{\mu}_1(x_1)$, i.e., $\tilde{\mu}_1(x_1) = \arg \inf_{\mu_1} \{|\mu_1 - \hat{\mu}_1(x_1)| : h(x_1, \mu_1) = 0\}$. The consistency of $\tilde{\mu}_1(x_1)$ can also be established following Härdle (1984).

4.2 Asymptotic Normality

To formulate the asymptotic normality, let us define:

$$Z_n(x_1) = \frac{C_1(x_1) \{\hat{\mu}_1(x_1) - \mu_1(x_1) - B_n(x_1)/[C_1(x_1)g(x_1)]\}}{[(nh^d)^{-1}\|K\|_2^2\sigma^2(x_1)g^{-1}(x_1)]^{1/2}},$$

where $\sigma^2(x_1) = E[\psi^2(Y - \mu_1(x_1))T/\pi^2(X) | X_1 = x_1]$, $C_1(x_1) = E(\psi'(Y - \mu_1(x_1)) | X_1 = x_1)$, $B_n(x_1) = E(h(x_1, \mu_1(x_1)) | X_1 = x_1)$ and $\|K\|_2^2 = \int K^2(u)du$.

Theorem 4. *Suppose the conditions in Theorem 3 hold, we have $Z_n(x_1) \xrightarrow{D} N(0, 1)$ if the following conditions are further satisfied:*

- (1) $\gamma_n = \int |\delta(u)|^{2+\eta} du < \infty$ for some $\eta > 0$,
- (2) $\gamma_n = o(n^{\eta/2}h^{-d(1+\eta/2)})$ as $n \rightarrow \infty$.

To eliminate the bias term from the limiting distribution, we must ensure that

$$B_n(x_1) / \sqrt{n^{-1}\alpha_n(2)\sigma^2(x_1)g(x_1)} \rightarrow 0$$

as $n \rightarrow \infty$, which means we need $B_n(x_1) = o((\alpha_n(2)/n)^{1/2})$. Since K is a kernel of order s , we have $\int u^t K(u)du = 0$ for $t \in \{1, 2, \dots, s\}$. By Taylor expansion, we have:

$$B_n(x_1) = E \left[\psi(Y^{(1)} - \mu_1(x_1)) \mid X_1 = x_1 \right] g(x_1) + O(h^s).$$

The first term can be interpreted as the bias due to the application of function ψ which equals 0 by the unbiasedness of the estimating equation as we proved in Theorem 1, we have $B_n(x_1) = O(h^s)$.

4.3 Asymptotic properties of $\hat{\tau}(x_1)$

We only need $O(h^s) = o\left(1/\sqrt{nh^d}\right)$, which is equivalent to $nh^{2s+d} \rightarrow 0$, which coincides with Assumption 8. Under these conditions, the bias term can be dropped, and a confidence interval can be provided accordingly.

Theorem 5. *When the conditions of Theorem 4 hold, we have*

$$Z_n^*(x_1) := \frac{\sqrt{nh^d}C_1(x_1)(\hat{\mu}_1(x_1) - \mu_1(x_1))}{[\|K\|_2^2\sigma^2(x_1)g^{-1}(x_1)]^{1/2}} \xrightarrow{D} N(0, 1).$$

4.3 Asymptotic properties of $\hat{\tau}(x_1)$

We can similarly extend the asymptotic properties to $\hat{\mu}_0(x_1)$, assuming that the same kernel and bandwidth are used for both $\mu_1(x_1)$ and $\mu_0(x_1)$. Due to the covariance between $\hat{\mu}_0(x_1)$ and $\hat{\mu}_1(x_1)$, we have the asymptotic normality for $\hat{\tau}(x_1)$ as following

$$\frac{\sqrt{nh^d g(x_1)}}{\|K\|_2 \sigma_\tau(x_1)} (\hat{\tau}(x_1) - \tau(x_1)) \xrightarrow{D} N(0, 1),$$

where

$$\sigma_\tau^2(x_1) = E \left[\left(\frac{H^{*1}(x_1)}{D^1(x_1)} - \frac{H^{*0}(x_1)}{D^0(x_1)} \right)^2 \mid X_{i1} = x_1 \right],$$

with

$$D^0(x_1) = E \left[(1 - T_i)(1 - \pi(X_i))^{-1} \psi'(Y_i - \mu_0(x_1)) \mid X_{i1} = x_1 \right],$$

$$D^1(x_1) = E \left[T_i \pi^{-1}(X_i) \psi'(Y_i - \mu_1(x_1)) \mid X_{i1} = x_1 \right],$$

$$H^{*0}(x_1) = \psi(Y_i - \mu_0(x_1))(1 - T_i)/(1 - \pi(X_i)) \text{ and } H^{*1}(x_1) = \psi(Y_i - \mu_1(x_1))T_i/\pi(X_i).$$

4.4 The Variance Estimator and Confidence Interval

4.4 The Variance Estimator and Confidence Interval

Based on the asymptotic distribution, we can get the standard error by

$$SE(\hat{\tau}(x_1)) = \frac{\|K\|_2 \hat{\sigma}_\tau(x_1)}{\sqrt{nh^d \hat{g}(x_1)}},$$

where we used local kernel estimation for both of $\sigma_\tau^2(x_1)$ and density function $g(x_1)$ of X_1 .

In particular,

$$\hat{\sigma}_\tau^2(x_1) = \sum_{i=1}^n K\left(\frac{X_{i1} - x_1}{h_1}\right) \left(\frac{\hat{\psi}_i^1(x_1)}{\hat{D}^1(x_1)} - \frac{\hat{\psi}_i^0(x_1)}{\hat{D}^0(x_1)} \right)^2 \bigg/ \sum_{i=1}^n K\left(\frac{X_{i1} - x_1}{h_1}\right),$$

where $\hat{\psi}_i^1(x_1) = \psi(Y_i - \hat{\mu}_1(x_1))T_i/\hat{\pi}(X_i)$ and $\hat{\psi}_i^0(x_1) = \psi(Y_i - \hat{\mu}_0(x_1))(1 - T_i)/(1 - \hat{\pi}(X_i))$, are plug-in quantities. And we again use local kernel estimation for $D^0(x_1)$ and $D^1(x_1)$ as follows,

$$\begin{aligned} \hat{D}^1(x_1) &= \sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)} \psi'(Y_i - \hat{\mu}_1(x_1)) K\left(\frac{X_{i1} - x_1}{h_2}\right) \bigg/ \sum_{i=1}^n K\left(\frac{X_{i1} - x_1}{h_2}\right), \\ \hat{D}^0(x_1) &= \sum_{i=1}^n \frac{1 - T_i}{1 - \hat{\pi}(X_i)} \psi'(Y_i - \hat{\mu}_0(x_1)) K\left(\frac{X_{i1} - x_1}{h_3}\right) \bigg/ \sum_{i=1}^n K\left(\frac{X_{i1} - x_1}{h_3}\right). \end{aligned}$$

The bandwidths h_1 , h_2 , and h_3 are specific to the kernel regressions in the estimation unknown quantities involved in the variance of robust CATE estimator and must be chosen independently in practice to balance bias and variance trade-offs effectively. With the above standard error of the proposed point estimator $\hat{\tau}(x_1)$, we are able to conduct inference for the heterogeneous treatment effect via the Wald-type confidence interval.

Practical guidance on selecting h_1 , h_2 , and h_3 is provided in Section S12 of the supplementary material. In our simulations and application, we set $h_1 = h_2 = h_3 = h$ and find adequate coverage; a sensitivity analysis varying these bandwidths separately showed minimal impact on the coverage probabilities.

5. Simulation Studies

In this section, we assess the finite sample performance of the proposed estimators using Monte Carlo simulations. The results are compared with the existing non-robust IPW-based methods (Abrevaya et al., 2015; Lee et al., 2017). In addition, Section 5.3 summarises comparisons of the proposed estimators with three DR-learner variants and with three machine-learning meta-learners (causal forest, X-learner, and R-learner); the full comparison studies are provided in the supplementary material. Furthermore, the validity and accuracy of the asymptotic normality are demonstrated with detailed outputs provided in the supplementary material.

5.1 Data Generating Process

In this simulation study, the observations are generated based on a simple causal framework. The covariates (X_1, X_2, X_3, X_4) are independently sampled from a uniform distribution over the interval $[-0.5, 0.5]$. The treatment indicator T follows a Bernoulli distribution with a success probability given by $\Lambda((X_1 + X_2 + X_3 + X_4)/2)$, where Λ denotes the logistic function. The responses are generated according to $Y = \mu^{(1)}T + \mu^{(0)}(1 - T) + (1 - \lambda)\epsilon_1 + \lambda\epsilon_2$, where $\mu^{(1)} = 10 + 5X_1 + \sin(10X_1) - X_2 + 3 \tanh(0.5X_3)$, $\mu^{(0)} = 0$ and the outlier indicator λ is generated from Bernoulli distribution with probability of outliers $p \in \{0, 0.1, 0.2, 0.3, 0.4, 0.45\}$. We generate the non-contaminated error distribution term by $\epsilon_1 \sim N(0, 0.25^2)$, and the contaminated error distribution by ϵ_2 with the following 4 different settings:

- (a) Symmetric light-tail: $N(0, 10)$;
- (b) Symmetric heavy-tail: $\text{Cauchy}(0, 0.5)$;

- (c) Asymmetric mean 0: $\log N(0, 1) - \exp(1/2)$;
- (d) Asymmetric mean $\neq 0$: Dirac delta function at point 20.

The first two settings are used to assess the finite sample performance of our estimator under the assumption of symmetric error distributions. And the remaining two asymmetric settings are employed to evaluate the robustness of our estimators in scenarios where the assumption of symmetry is not met.

5.2 Simulation Results

We estimate the CATE using our proposed estimator with three types of loss functions: L_2 , Huber, and Tukey's loss functions. The L_2 loss function represents an example of an unbounded loss function, Huber's loss function illustrates a bounded loss function, and Tukey's loss function serves as an example of a redescending loss function. Bandwidths for our proposed estimators are selected via cross-validation.

To verify the pointwise asymptotic properties, we estimate the CATE at five fixed points, $x_1 \in \{-0.4, -0.2, 0, 0.2, 0.4\}$, for sample sizes $n \in \{500, 5000\}$. Across 1,000 repetitions, we report the simulation bias, variance, mean squared error (MSE), mean of the estimated variances, and the coverage probabilities of the 90% and 95% confidence intervals constructed using the estimated variances. The detailed numerical results for all four outlier distributions and both sample sizes are provided in the supplementary material.

To compare with existing methods, we estimate the CATE using the IPW Abrevaya et al. (2015) and AIPW Lee et al. (2017) approaches in addition to our proposed methods. Following the recommendations in the references, bandwidths for IPW and AIPW are

5.2 Simulation Results

selected via the plug-in method proposed by Ruppert et al. (1995). CATE estimates are computed at 100 equally spaced points between -0.5 and 0.5 , with 1,000 repetitions. The mean integrated squared error (MISE), mean confidence interval width, and mean coverage probabilities of the resulting confidence intervals are summarized in Figure 1.

From the simulation results, we observe the following:

1) Comparing $n = 500$ and $n = 5000$, we find that a larger sample size results in reduced bias, variance, and MSE, and improves the probability of achieving coverage rates closer to 95% and 90%. This indicates that the estimator approaches its theoretical normal distribution as the sample size increases.

2) Under symmetric contamination (Normal and Cauchy), all three proposed estimators achieve comparable MISE when no outliers are present. As the outlier ratio increases, the Huber and Tukey estimators substantially outperform the L_2 estimator in MISE, with the advantage most pronounced under Cauchy contamination where L_2 MISE grows roughly 70-fold while Huber and Tukey remain nearly flat. All three proposed estimators maintain coverage rates close to or above the nominal 95% level across all contamination levels, confirming both the asymptotic normality of the estimators and the accuracy of the variance estimator. The IPW benchmark achieves lower coverage (around 80–88%) at $n = 500$, consistent with its known finite-sample bias; this improves substantially at $n = 5000$. The AIPW benchmark achieves coverage around 92–94% at $n = 500$, reflecting the efficiency gain from the doubly-robust augmentation.

3) Compared to the existing IPW and AIPW methods, our proposed estimators achieve finite-sample performance similar to the AIPW method and outperform the IPW method

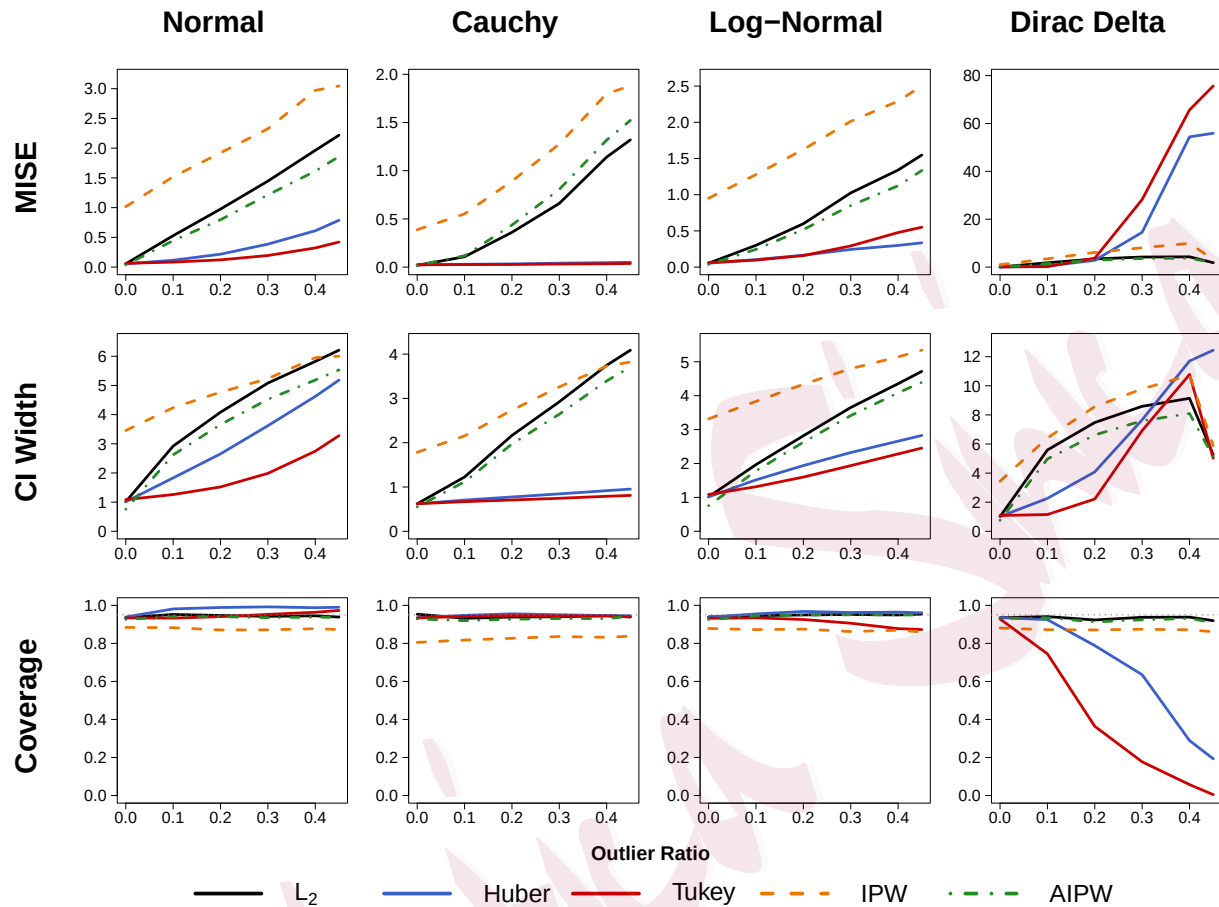


Figure 1: Comparison of MISE, width of confidence intervals, and coverage probability between the proposed method with L_2 , Huber, and Tukey's loss functions, as well as IPW and AIPW methods, across varying error distributions and outlier ratios.

in non-contaminated cases. We note that theoretically, IPW-based estimators are known to not achieve the semiparametric efficiency bound, while AIPW achieves it (Robins et al., 1994; Robins and Rotnitzky, 1995). The observed finite-sample comparability between our robust estimators and AIPW under our specific DGP should therefore be understood as a simulation-specific phenomenon, not as a theoretical equivalence. In the presence of sym-

metric outliers, the robust loss function-based methods consistently outperform the existing non-robust estimators including AIPW, demonstrating that the robustness gain from M-estimation more than compensates for the efficiency loss relative to AIPW in contaminated settings.

4) In asymmetric cases, the performance of our proposed estimators degrades as expected due to violation of the symmetry assumption, but the pattern differs markedly between the two settings. Under log-normal contamination, where the outlier distribution is asymmetric but the outliers are positively skewed with relatively moderate leverage, Huber achieves the best MISE among all estimators (roughly one-fifth of the L_2 MISE at 45% contamination) and maintains coverage between 94% and 97%, near or slightly above the nominal 95%. Tukey's coverage declines gradually from around 93% to 87% as the outlier ratio increases to 45%, while L_2 maintains nominal coverage but suffers a roughly 29-fold MISE increase. Under Dirac Delta contamination—a point mass at 20, far from the data bulk and severely violating symmetry—the picture reverses: L_2 maintains coverage between 92% and 94% at all contamination levels because the shifted mean is consistently estimated and the variance estimator captures the resulting variability. In contrast, Huber and Tukey attempt to down-weight these extreme but systematic outliers, inducing a bias that is not reflected in the estimated variance; coverage for Huber collapses from 94% to 19% and for Tukey from 93% to below 1% as the outlier ratio increases from 0 to 45%. These results confirm that the Huber and Tukey estimators are suited to symmetric or mildly asymmetric contamination, while severely asymmetric, non-centered contamination can cause coverage breakdown for M-estimators.

5.3 Comparisons with DR-Learner Variants and Meta-Learners

To further contextualise the proposed methods, we compare them with three variants of the DR-learner (Kennedy, 2023), which represents a natural class of doubly-robust competitors, and with three state-of-the-art machine-learning meta-learners: the causal forest (Athey et al., 2019), the X-learner (Künzel et al., 2019), and the R-learner (Nie and Wager, 2021). The complete comparison studies and implementation details are provided in Sections S9–S11 of the supplementary material; here we summarise the main findings.

The three DR-learner variants share the same logistic-regression propensity score but progressively robustify the pipeline: the standard DR-learner smooths the AIPW pseudo-outcome by Nadaraya–Watson kernel regression, the Robust DR-learner replaces this smoother with Huber IRLS kernel regression, and the Fully Robust DR-learner additionally replaces the least-squares outcome regressions with Huber M-estimation. Robustifying the DR pipeline reduces MISE substantially, but it fails to deliver valid inference under contamination: the pseudo-outcome channels contamination from two sources—biased outcome regressions and amplification by the inverse propensity weights—neither of which is remedied by downstream M-estimation, with coverage dropping to 64–76% under Cauchy contamination. In contrast, the proposed estimators apply robustness directly to the outcome Y before any propensity amplification occurs, and maintain near-nominal coverage (93–96%) with nearly flat MISE across all contamination levels.

The forest-based meta-learners, whose predictions we marginalize to the partially conditional CATE $\tau(x_1)$ by kernel smoothing on X_1 , are competitive in the absence of contamination and degrade gracefully under finite-variance contamination, but they break down

under heavy-tailed Cauchy contamination, with MISE exceeding even the non-robust L_2 estimator while the proposed Huber and Tukey estimators remain essentially flat. Under the Dirac delta contamination, the point mass enters both treatment arms additively and largely cancels in the treatment-effect contrast, so the meta-learners degrade gracefully while the Huber and Tukey estimators break down at high contamination, consistent with the findings of Section 5.2. Moreover, to the best of our knowledge, no valid pointwise inference theory is available for the marginalized CATE obtained from these pipelines. Overall, the proposed estimators are the only methods in the comparison that combine low MISE under symmetric contamination with theoretically justified inference.

6. Application to NHANES Data

The application involves data from the 2007-2008 National Health and Nutrition Examination Survey (NHANES), a publicly available dataset provided by the United States Centers for Disease Control and Prevention (CDC). The goal is to estimate the effect of physical activity on albumin level changes with age, while controlling for other confounders.

Albumin levels are obtained from laboratory data, measured in milligrams of albumin per gram of creatinine (mg/g). Given the right-skewed nature of the data and the lack of observations close to 0 (unlike Robinson et al. (2010)), we apply a log transformation to the response variable Y . Despite the transformation, the scatter plot indicates that the error distribution remains slightly asymmetric, with some observations deviating significantly from the majority as shown in Figure 2. This motivates the use of our robust method for this dataset. Our simulation results suggest that our proposed method performs well under

slightly skewed and low contamination conditions.

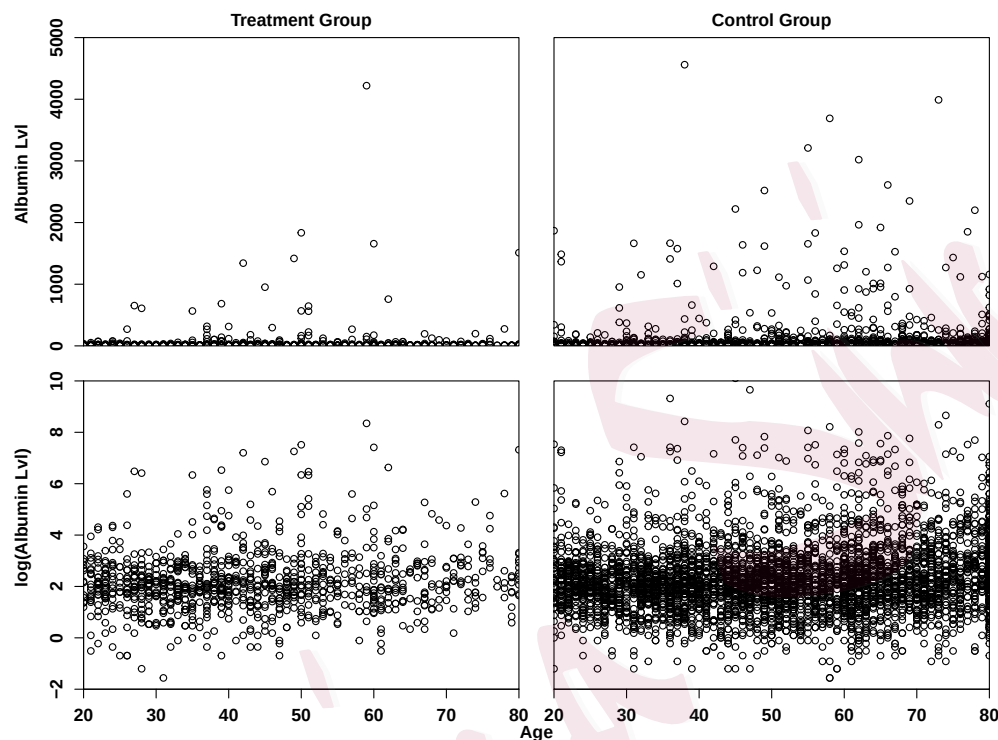


Figure 2: Scatter plot of albumin levels (first row) and log-transformed albumin level (second row) conditioned on age, separated by whether the individual engages in vigorous work activities.

The treatment indicator is derived from the physical activity questionnaire (prefix PAQ), specifically using vigorous work activity as the treatment indicator T . We select observations with $T = 1$ and $T = 0$ for our estimation and exclude other values.

The covariates X include age, ratio of family income to poverty (from 0 to 5), gender, race, education level, BMI, diabetes, history of kidney disease, and blood pressure. These variables are commonly used as confounders or in experimental designs (Chang et al., 2013; Lang et al., 2018). We assume unconfoundedness given this set of covariates and use age as

the conditioned covariate.

After merging the dataset and removing observations with missing values, the final sample size is $n = 5021$, with 982 observations in the treatment group and 4039 in the control group.

Elevated urinary albumin excretion is a predictor of future cardiovascular disease, hypertension, and chronic kidney disease (Robinson et al., 2010; Comper and Osicka, 2005; Solbu et al., 2008; Miller et al., 2009). Physical activity is often negatively associated with albuminuria, indicating potential protective effects on cardiovascular and renal health (Kuo et al., 2022). In diabetic populations, physical activity is linked to lower albumin excretion, and interventions have shown promise in reducing albuminuria (Hambrecht et al., 1998, 2000; Leon et al., 1987).

However, findings in non-diabetic individuals are mixed. Some studies have found a significant association between physical activity and reduced albumin excretion (Robinson et al., 2010), while others have reported no such association (Chang et al., 2013). These inconsistencies highlight the need to explore the heterogeneity of the effect across different subpopulations, a topic that has not been extensively studied in the literature. These mixed findings underscore the importance of investigating how physical activity impacts albuminuria across various subpopulations.

As shown in Figure 3, compared to the IPW and DR methods, the overall trend and value of our CATE estimators are similar to the doubly robust estimator. The estimated functions are in general all lower than 0, which is consistent with the existing results (Kuo et al., 2022; Hambrecht et al., 1998, 2000; Leon et al., 1987). However, the widths of the

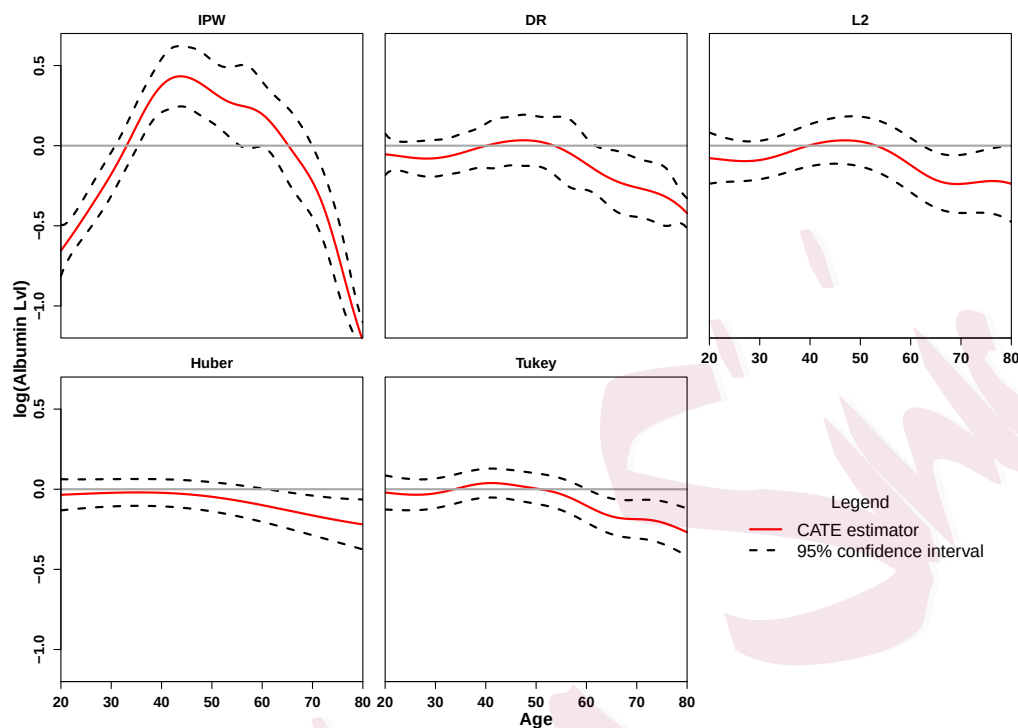


Figure 3: CATE of the effect of vigorous work activity on the log-transformed urinary albumin levels, conditioned on age, along with the 95% confidence intervals.

confidence interval of our estimator with robust loss functions (Huber's and Tukey's) are smaller and hence we're able to conclude that with age over 60, we're 95% confident that Vigorous work activities decrease the logarithm of albumin level. Furthermore, with the redescending loss function the Tukey's biweighted loss, the estimated functions are close to 0 before age 45 without an obvious increasing trend. While in the Huber loss and other nonrobust methods, the estimated functions are negative with an increasing trend before 40 years old, which is not supported by existing studies.

7. Conclusions and Future Directions

In this paper, we propose an outlier-resistant IPW method for estimating the CATE where a relatively high-dimensional vector of covariates is required to identify the average treatment effect, but the covariates of primary interest are of much lower dimension.

Our proposed estimators are shown to be consistent and asymptotically normal, demonstrating strong finite sample performance, which ensures valid inference and prediction. When no contamination is present, our method performs similarly to IPW and DR estimators in finite samples, though it sacrifices some efficiency relative to the AIPW estimator, which is known to achieve the semiparametric efficiency bound (Robins et al., 1994). This efficiency trade-off is the price of robustness. In the presence of symmetric or mildly asymmetric contamination, our method outperforms both the IPW and DR methods in accuracy while retaining near-nominal coverage; for severely asymmetric, non-centered contamination, the L_2 variant of our estimator remains the reliable choice, as discussed in Section 5. We establish the asymptotic properties of our estimators and present a pointwise confidence interval, illustrating their utility through Monte Carlo experiments and an application to the age modified effects of physical activity on albumin levels.

Several directions for future research are evident. First, exploring the impact of asymmetric density functions could provide additional insights. Second, model misspecification issues may arise since our approach relies on a parametric method to estimate the propensity score. A promising extension is to develop a doubly robust estimator that incorporates a robust M-estimating equation augmented with an outcome regression term, analogous to how AIPW augments IPW. Such an estimator would be consistent even if the propensity score

or outcome regression is misspecified, and could potentially achieve the semiparametric efficiency bound while maintaining outlier resistance. Third, developing a fully nonparametric method could potentially improve efficiency, as indicated by previous work (Abrevaya et al., 2015; Hirano et al., 2003). Fourth, extending the framework to right-censored outcomes and to settings with unmeasured confounding is of practical interest; non-robust CATE learners already exist for these settings, such as causal survival forests (Cui et al., 2023) for right-censored outcomes, and instrumental forests (Athey et al., 2019) and the P-learner of proximal causal inference (Sverdrup and Cui, 2023) under unmeasured confounding. For right-censoring, inverse-probability-of-censoring weights (Robins and Finkelstein, 2000) can be incorporated into the estimating equation (2.5), preserving the M-estimation structure, although the interaction between censoring and the symmetry condition in Assumption 2, as well as the amplification of contamination by small censoring weights, requires new robustness analysis. For unmeasured confounding, robust analogues of proximal (Miao et al., 2018) or instrumental (Angrist et al., 1996) moment equations raise identification questions beyond the scope of this paper. Finally, given the high-dimensional nature of covariates X , incorporating techniques such as variable selection or dimension reduction could enhance the method's performance.

Supplementary Materials

Technical proofs, detailed simulation results, and the full DR-learner and machine-learning meta-learner comparison studies summarised in Section 5.3, together with their implementation details, are provided in the supplementary material.

Acknowledgements

This research was supported in part by the National Science Foundation under Grant DMS-2212928 and by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

The study utilizes data from the National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics, part of the Centers for Disease Control and Prevention (CDC).

References

- Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33(4), 485–505.
- Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91(434), 444–455.
- Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* 113(27), 7353–7360.
- Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Ann. Statist.* 47(2), 1148–1178.
- Beaton, A. E. and J. W. Tukey (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* 16(2), 147–185.
- Chang, A., L. Van Horn, D. R. J. Jacobs, K. Liu, P. Muntner, and e. a. Newsome, Beth (2013). Lifestyle-related factors, obesity, and incident microalbuminuria: the cardia study. *Am. J. Kidney Dis.* 62(2), 267–275.
- Chipman, H. A., E. I. George, and R. E. McCulloch (2010). Bart: Bayesian additive regression trees. *Ann. Appl. Stat.* 4(1), 266–298.

REFERENCES

- Cole, S. R. and M. A. Hernán (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 168(6), 656–664.
- Comper, W. D. and T. M. Osicka (2005). Detection of urinary albumin. *Advances in chronic kidney disease* 12(2), 170–176.
- Cui, Y., M. R. Kosorok, E. Sverdrup, S. Wager, and R. Zhu (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 85(2), 179–211.
- Donoho, D. L. and P. J. Huber (1983). The notion of breakdown point. *A festschrift for Erich L. Lehmann* 157184.
- Dorie, V., J. Hill, U. Shalit, M. Scott, and D. Cervone (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci.* 34(1), 43–68.
- Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *J. Bus. Econom. Statist.* 40(1), 313–327.
- Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* 75(1), 259–276.
- Foster, J. C., J. M. Taylor, and S. J. Ruberg (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine* 30(24), 2867–2880.
- Giloni, A. and J. S. Simonoff (2005). The conditional breakdown properties of least absolute value local polynomial estimators. *Journal of Nonparametric Statistics* 17(1), 15–30.
- Glynn, A. N. and K. M. Quinn (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis* 18(1), 36–56.
- Hambrecht, R., E. Fiehn, C. Weigl, S. Gielen, C. Hamann, R. Kaiser, J. Yu, V. Adams, J. Niebauer, and G. Schuler (1998). Regular physical exercise corrects endothelial dysfunction and improves exercise capacity in patients with chronic heart failure. *Circulation* 98(24), 2709–2715.

REFERENCES

- Hambrecht, R., A. Wolf, S. Gielen, A. Linke, J. Hofer, S. Erbs, N. Schoene, and G. Schuler (2000). Effect of exercise on coronary endothelial function in patients with coronary artery disease. *New England Journal of Medicine* 342(7), 454–460.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association* 69(346), 383–393.
- Harada, K. and H. Fujisawa (2021). Outlier-resistant estimators for average treatment effect in causal inference. *arXiv preprint arXiv:2106.13946*.
- Härdle, W. (1984). Robust regression function estimation. *Journal of Multivariate Analysis* 14(2), 169–180.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1), 217–240.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161–1189.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81(396), 945–960.
- Holland, P. W. and R. E. Welsch (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods* 6(9), 813–827.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* 35(4), 73–101.
- Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, 799–821.
- Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17(2), 3008–3049.
- Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment

REFERENCES

- p>effects using machine learning.
- Proceedings of the national academy of sciences*
- 116(10), 4156–4165.
- Kuo, H.-Y., Y.-H. Huang, S.-W. Wu, F.-H. Chang, Y.-W. Tsuei, H.-C. Fan, and et al. (2022). The effects of exercise habit on albuminuria and metabolic indices in patients with type 2 diabetes mellitus: A cross-sectional study. *Medicina* 58(5), 577.
- Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making* 42(2), 156–167.
- Lang, J., R. Katz, J. H. Ix, O. M. Gutierrez, C. A. Peralta, C. R. Parikh, S. Satterfield, S. Petrovic, P. Devarajan, M. Bennett, et al. (2018). Association of serum albumin levels with kidney function decline and incident chronic kidney disease in elders. *Nephrology Dialysis Transplantation* 33(6), 986–992.
- Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics* 32(7), 1207–1225.
- Lee, S. and Y.-J. Whang (2009). Nonparametric tests of conditional treatment effects. (No. 1740).
- Leon, A. S., J. Connett, D. R. Jacobs, and R. Rauramaa (1987). Leisure-time physical activity levels and risk of coronary heart disease and death: the multiple risk factor intervention trial. *Jama* 258(17), 2388–2395.
- Li, G. and J. Zhang (1998). Breakdown properties of location M-estimators. *The Annals of Statistics* 26(3), 1170–1189.
- Luedtke, A. R. and M. J. van der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics* 44(2), 713–742.
- Luedtke, A. R. and M. J. van der Laan (2017). Evaluating the impact of treating the optimal subgroup. *Statistical Methods in Medical Research* 26(4), 1630–1640.
- Miao, W., Z. Geng, and E. J. Tchetgen Tchetgen (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105(4), 987–993.

REFERENCES

- Miller, W. G., D. E. Bruns, G. L. Hortin, S. Sandberg, K. M. Aakre, M. J. McQueen, Y. Itoh, J. C. Lieske, D. W. Seccombe, G. Jones, et al. (2009). Current issues in measurement and reporting of urinary albumin excretion. *Clinical chemistry* 55(1), 24–38.
- Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10(1), 1–51.
- Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108(2), 299–319.
- Powers, S., J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine* 37(11), 1767–1787.
- Robins, J. M. and D. M. Finkelstein (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 56(3), 779–788.
- Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90(429), 122–129.
- Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89(427), 846–866.
- Robins, J. M., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science* 22(4), 544–559.
- Robinson, E. S., N. D. Fisher, J. P. Forman, and G. C. Curhan (2010). Physical activity and albuminuria. *American journal of epidemiology* 171(5), 515–521.
- Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55.
- Rousseeuw, P. J., F. R. Hampel, E. M. Ronchetti, and W. A. Stahel (1986). Robust statistics: the approach based on influence functions.

REFERENCES

- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5), 688.
- Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90(432), 1257–1270.
- Solbu, M. D., J. Kronborg, B. O. Eriksen, T. G. Jenssen, and I. Toft (2008). Cardiovascular risk-factors predict progression of urinary albumin-excretion in a general, non-diabetic population: a gender-specific follow-up study. *Atherosclerosis* 201(2), 398–406.
- Sverdrup, E. and Y. Cui (2023). Proximal causal learning of conditional average treatment effects. In *Proceedings of the 40th International Conference on Machine Learning*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 33285–33298.
- van der Laan, M. J. (2006). Statistical inference for variable importance. *International Journal of Biostatistics* 2(1).
- VanderWeele, T. J., A. R. Luedtke, M. J. van der Laan, and R. C. Kessler (2019). Selecting optimal subgroups for treatment using many covariates. *Epidemiology* 30(3), 334–341.
- Yao, L., S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang (2018). Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems* 31.
- Zhang, Z., Z. Chen, J. F. Troendle, and J. Zhang (2012). Causal inference on quantiles with an obstetric application. *Biometrics* 68(3), 697–706.

Department of Mathematical Sciences, Indiana University Indianapolis, Indianapolis, IN 46202, USA.

E-mail: (ranmo@iu.edu)

Department of Mathematical Sciences, Indiana University Indianapolis, Indianapolis, IN 46202, USA.

E-mail: (hlwang@iu.edu) ORCID: 0000-0003-3140-5177