Statistica Sinica **34** (2024), 133-155
doi:https://doi.org/10.5705/ss.202021.0254

OUTLIER-RESISTANT ESTIMATORS FOR AVERAGE TREATMENT EFFECT IN CAUSAL INFERENCE

Kazuharu Harada*[1] and Hironori Fujisawa[1,2,3]

[1] *The Graduate University for Advanced Studies (SOKENDAI),*
[2] *The Institute of Statistical Mathematics*
and [3] *RIKEN Center for Advanced Intelligence Project*

Abstract: The inverse probability weighting (IPW) and doubly robust (DR) estimators are often used to estimate the average treatment effect (ATE), but are vulnerable to outliers. The IPW/DR median can be used to provide an outlier-resistant estimation of the ATE, but this resistance is limited, and is not sufficiently resistant to heavy contamination. We propose extending the IPW/DR estimators using density power weighting, which eliminates the effects of outliers almost completely. The resistance of the proposed estimators to outliers is evaluated using the unbiasedness of the estimating equations. Unlike the median-based methods, our estimators are resistant to outliers, even under heavy contamination. Interestingly, the naive extension of the DR estimator requires a bias correction to maintain its double robustness, even under the most tractable form of contamination. In addition, the proposed estimators are found to be highly resistant to outliers in more difficult settings in which the contamination ratio depends on the covariates. The resistance of our estimators to outliers from the viewpoint of the influence function is also favorable. We verify our theoretical results using Monte Carlo simulations and a real-data analysis. The proposed methods are shown to have greater resistance to outliers than the median-based methods do, and we estimate the potential mean with a smaller error than that of the median-based methods.

Key words and phrases: Causal inference, doubly robust, missing data, propensity score, robust statistics.

1. Introduction

Statistical causal inference provides various estimators for causal quantities such as the average treatment effect (ATE). To estimate such quantities, the propensity score is widely applied in, for example, stratification, matching, inverse probability weighting (IPW), and the doubly robust (DR) estimator (Robins, Rotnitzky and Zhao (1994); Rosenbaum and Rubin (1983); Bang and Robins (2005)). These estimators are designed to control confounding, and are consistent with the target quantity, under some assumptions.

*Corresponding author.

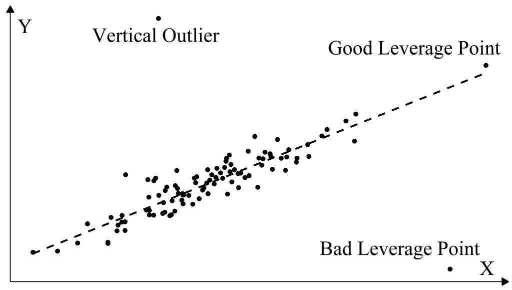

Figure 1. Three types of outliers.

As discussed later, the IPW and DR estimators are vulnerable to outliers, because they partially use the sample mean. An outlier in a multivariate setting is classified as one of three types: a vertical outlier, a good leverage point, or a bad leverage point (Rousseeuw and van Zomeren (1990)). Figure 1 illustrates the three types of outliers. Canavire-Bacarreza, Castro Peñarrieta and Ugarte Ontiveros (2021) investigated how these types of outliers affect the estimators of the ATE, including the IPW, using exhaustive Monte Carlo simulations, finding that vertical outliers in the outcome variable lead to serious bias in the ATE estimation. Therefore, we focus on reducing this bias caused by vertical outliers.

Although there has been numerous research on outlier-resistant statistics, most work do not consider a causal setting (Huber (2004); Hampel et al. (2011); Maronna et al. (2019)). In many causal settings, the established methods for outlier-resistant statistics cannot be applied. The median-based estimators are the only ones that can be used to estimate the ATE under outlier contamination (Firpo (2007); Zhang et al. (2012); Díaz (2017); Sued, Valdora and Yohai (2020)). It is well known that the sample median is more resistant to outliers than is the sample mean, but it is still affected. In particular, when the contamination ratio is not small and the outliers lie on one side of the data-generating density, the effect becomes so large that it cannot be ignored (Fujisawa and Eguchi (2008)).

In this paper, we propose extensions of the IPW and DR estimators for the mean of the potential outcome that are more resistant to outliers than the median-based methods are. We discuss the outlier resistance of these estimators from the viewpoint of the unbiasedness of the estimating equation and influence function (IF). In most works on outlier-resistant statistics, the contamination ratio is assumed to be small and independent of the covariates. Here, we discuss the outlier resistance of the proposed estimators under more general assumptions, including the case in which the contamination ratio is not small and is related to the covariates. Interestingly, a straight extension of the DR estimator loses its robustness to a model misspecification under contamination. Thus, we also propose a bias-corrected version of the extended DR estimator that maintains its double robustness under contamination. Furthermore, we verify the theoretical advantages of our estimators using Monte Carlo simulations and a real-data analysis.

The remainder of this paper is organized as follows. In Section 2, we introduce the potential outcome framework for causal inference and the basic concept of outliers. In Section 3, we propose our novel estimators, and discuss their resistance to outliers from the viewpoint of the unbiasedness of the estimating equations. In Section 4, we evaluate the outlier resistance of the proposed estimators in terms of the IF. In Section 5, we discuss their asymptotic properties, and in Sections 6 and 7, we present the results of our experiments.

2. Preliminaries

2.1. Potential outcome and treatment effect

Let (Y, T, X) be the observable random variables, where X is the outcome, T is the treatment, and X is the confounder. We assume that Y is continuous and that T is binary; it is straightforward to extend T to multiple discrete treatments. We have the observations $(Y_i, T_i, X_i)_{i=1}^n$ drawn from the distribution of (Y, T, X) in an independently and identically distributed (i.i.d.) manner. Denote the potential outcome under $T = t$ by $Y^{(t)}$, and let $\mu^{(t)} = \mathbb{E}[Y^{(t)}]$. Here, $Y^{(t)}$ is uniquely defined for every treatment as a random variable, that is, it is well-defined. Note that the i.i.d. sampling and the well-definedness of the potential outcome are collectively called the stable unit treatment value assumption (SUTVA; Imbens and Rubin (2015)). The ATE is defined as $\mu^{(1)} - \mu^{(0)}$. The ATE cannot be estimated directly, because we cannot observe $Y^{(1)}$ and $Y^{(0)}$ simultaneously. Instead, we use the observed variables under the following common assumptions (e.g., Imbens and Rubin (2015)):

1. *Conditional Unconfoundedness*: $Y^{(t)} \perp\!\!\!\perp T|X$, for all $t \in \{0, 1\}$;
2. *Consistency*: $Y = Y^{(t)}$ if $T = t$;
3. *Positivity*: $P(T = 1|X) > c$, for some constant $c > 0$.

The ATE is identifiable in that it can be estimated from the observed variables under these assumptions. Hereafter, we assume the triple assumption holds and focus on the estimation of $\mu^{(1)}$, for simplicity. We estimate $\mu^{(0)}$ in a similar way. Then, the ATE is estimated as the difference between the estimates of $\mu^{(1)}$ and $\mu^{(0)}$.

We introduce three consistent estimators of the potential mean. The IPW estimator (Rosenbaum and Rubin (1983)) is based on the propensity score (PS). Let $\pi(x; \alpha) \in (0, 1)$ be the PS, which models $P(T = 1|x)$. We assume the PS is correctly specified, in other words, there exists α^* such that $\pi(x; \alpha^*) = P(T = 1|x)$, for every x. The IPW estimator has several forms (Lunceford and Davidian

(2004)), but we use the weighted average form: $\hat{\mu}^{(1)}_{IPW} = \{\sum_{i=1}^{n} T_i Y_i/\pi(X_i;\hat{\alpha})\}/\{\sum_{i=1}^{n} T_i/\pi(X_i;\hat{\alpha})\}$, where $\hat{\alpha}$ is an estimate of α obtained in a consistent manner, for example, using the maximum likelihood estimation (MLE). The IPW estimator can be viewed as the root of the following estimating equation:

$$\sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i - \mu) = 0. \tag{2.1}$$

Outcome regression (OR) is also popular. To construct the OR estimator, we model $\mathbb{E}[Y|T = 1, X]$ by some function $m_1(X;\beta)$. Then, the OR estimator is obtained as $n^{-1}\sum_{i=1}^{n} m_1(X_i;\hat{\beta})$, where $\hat{\beta}$ is a consistent estimate of β. The IPW and OR estimators are asymptotically consistent with $\mu^{(1)}$ when the model used in each estimator is specified correctly, but this does not hold if the model is misspecified. The DR estimator (Scharfstein, Rotnitzky and Robins (1999); Bang and Robins (2005)) combines the IPW and OR estimators. Because the DR estimator is consistent with $\mu^{(1)}$ when either the PS or OR model is specified correctly, it is said to be "doubly robust." Furthermore, if both models are specified correctly, the DR estimator is semiparametrically efficient (Robins and Rotnitzky (1995); Tsiatis (2006)). Although many estimators are equipped with double robustness, we refer to the root of the following special case of the augmented IPW estimator as the DR estimator $\hat{\mu}^{(1)}_{DR}$:

$$\sum_{i=1}^{n} \left[\frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i - \mu) - \frac{T_i - \pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\{m_1(X_i;\hat{\beta}) - \mu\} \right] = 0. \tag{2.2}$$

2.2. IPW/DR M-estimators

Let $\sum_{i=1}^{n} \psi(Y_i,\theta) = 0$ be an estimating equation, where ψ is a known vector-valued map, and θ is the parameter of interest. An estimator $\hat{\theta}$ that solves the estimating equation is called an M-estimator. M-estimators form a large class of estimators, including the MLE, IPW, OR, and DR. If the estimating equation is unbiased, say $\mathbb{E}_\theta[\psi(Y,\theta)] = 0$, the M-estimator is consistent with the truth, under some regularity conditions (e.g., Chap. 5 of Van der Vaart (2000)).

By replacing $Y_i - \mu$ in (2.2) with an estimating function $\psi(Y_i;\theta)$, the IPW and DR estimators can be expanded to a general M-estimator. If we are interested in the same parameter θ with respect to $Y^{(1)}$, we can use the following IPW and DR M-estimators (Tsiatis (2006)):

$$\sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})}\psi(Y_i;\theta) = 0, \tag{2.3}$$

$$\sum_{i=1}^{n} \left\{ \frac{T_i}{\pi(X_i;\hat{\alpha})}\psi(Y_i;\theta) - \frac{T_i - \pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\mathbb{E}_{\hat{q}}[\psi(Y_i;\theta)|T = 1, X_i] \right\} = 0. \tag{2.4}$$

The conditional expectation $\mathbb{E}_{\hat{q}}[\psi(Y_i;\theta)|T = 1, X_i]$ is calculated using the parametric OR model $q(y|T=1,x;\hat{\beta})$ by direct calculation or by using a Monte Carlo approximation (Hoshino (2007)). When the original M-estimating equation is unbiased, the IPW/DR estimating equations are unbiased under a proper model specification. The asymptotic properties of the IPW and DR M-estimators follow from the standard theory of M-estimators.

2.3. Outlier-resistant estimation

In this section, we review the outlier-resistant estimation of a mean in a one-variable and non-causal setting. Let $\tilde{g}$ be the density function of a random variable $Z \in \mathbb{R}$. Assume that the density is contaminated as $\tilde{g}(z) = (1-\varepsilon)f_{\mu^*}(z) + \varepsilon\delta(z)$, where f_{μ^*} is the density of Z without contamination equipped with the mean μ^*, ε is the contamination ratio, and δ is the density of outliers. Our goal is to estimate μ^* from i.i.d. observations $\{Z_1,\ldots,Z_n\}$ drawn from $\tilde{g}$. If we model the contamination in this way, the sample mean converges to $(1-\varepsilon)\mu^* + \varepsilon\mathbb{E}_\delta[Z]$; if the mean of the outliers is far from μ^*, the sample mean is asymptotically biased. Numerous M-estimators have been proposed to deal with contamination. The unbiasedness of the estimating equation does not usually hold under contamination because

$$\mathbb{E}_{\tilde{g}}[\psi(Z,\mu^*)] = (1-\varepsilon)\underbrace{\mathbb{E}_{f_{\mu^*}}[\psi(Z,\mu^*)]}_{=0} + \varepsilon\mathbb{E}_\delta[\psi(Z,\mu^*)] \neq 0. \tag{2.5}$$

By designing ψ to eliminate or bound $\mathbb{E}_\delta[\psi(Z,\mu^*)]$, we can reduce the influence of outliers. Let θ^*_ψ denote a root of $\mathbb{E}_{\tilde{g}}[\psi(Z,\theta)] = 0$. Then, the latent bias is defined as $\theta^*_\psi - \theta^*$. If δ is Dirac's delta and ε is sufficiently small, the latent bias is approximated by the IF. The IF-based discussion in Section 4 provides some insights into the outlier resistance of the estimators when the contamination ratio is small. For a detailed discussion of the latent bias and M-estimators, see Huber (2004), Fujisawa (2013), and Fujisawa and Eguchi (2008), among others.

2.4. IPW and DR under contamination

Next, we move to a causal setting with vertical outliers. In other words, we assume that only the outcome Y may be contaminated, and that the contamination does not affect the causal mechanism among (Y,T,X). A typical example is the contamination of laboratory values in medical research with foreign substances. Let $\delta_{Y|TX}$ be the conditional density of the outliers given (T,X), and let $\varepsilon_t(x)$ be the contamination ratio. Then, the contaminated conditional density of Y given (T,X) is defined as

$$\tilde{g}_{Y|TX}(y|t,x) = \{1-\varepsilon_t(x)\}g_{Y|TX}(y|t,x) + \varepsilon_t(x)\delta_{Y|TX}(y|t,x), \tag{2.6}$$

where g without the tilde denotes the density without contamination; the tilde indicates that the distribution is contaminated. To simplify the notation, we often drop the subscripts of the density functions, as long as this does not cause any confusion, and write $\delta_t(y|x) = \delta_{Y|TX}(y|t,x)$. The contamination ratio and the density depend on the treatment T and the confounder X. Because we estimate $\mu^{(t)}$ for each treatment separately, the dependence on T is tractable. In contrast, the dependence on X is critical in our analysis. The X-dependent contamination is referred to as heterogeneous contamination. We also discuss the special case in which ε and δ are not dependent on X, called homogeneous contamination. Note that we do not assume that $\varepsilon_t(x)$ is small enough to be negligible, except in Section 4.

We are interested in the marginal mean of $Y^{(1)}$. Let $f_{Y^{(1)}}(y;\mu^{(1)})$ be the true marginal density of $Y^{(1)}$, obtained by integrating X out from $g_{Y|TX}(y|T,X)$ under $T=1$:

$$f_{Y^{(1)}}(y;\mu^{(1)}) = \int g_{Y^{(1)}|X}(y|x) g_X(x) dx = \int g_{Y|TX}(y|1,x) g_X(x) dx. \tag{2.7}$$

The second equality holds from the triple assumption in Section 2.1. We often write $f_{Y^{(1)}}(y;\mu^{(1)})$ as $f_1(y)$, for simplicity.

Under contamination, the IPW estimating equation is severely biased, even if the true PS is obtained as $\pi(X|\alpha^*) = P(T=1|X)$:

$$\mathbb{E}_{\tilde{g}}\left[\frac{T}{\pi(X|\alpha^*)}(Y-\mu^{(1)})\right] = \mathbb{E}_g\left[\varepsilon_1(X)\mathbb{E}_{-g+\delta}\left[(Y-\mu^{(1)})|X\right]\right] \neq 0. \tag{2.8}$$

The remaining term contains the expectation of Y with respect to δ, which implies that the estimating equation is severely affected by outliers. The DR estimating equation is similarly biased. To estimate $\mu^{(1)}$ accurately, we have to remove the influence of contamination.

3. Outlier-Resistant Extensions of the IPW and DR

Before we propose novel estimators, we introduce an assumption on outliers. Intuitively, we assume that the outliers are sufficiently far from the main outcome density. Figure 2 shows real examples of outliers that satisfy this assumption, where the outliers are far from the main body of the density, both conditionally and marginally.

To formalize this assumption, we introduce density power weighting. The density power weight is used to enhance the outlier resistance in noncausal settings (Windham (1995); Basu et al. (1998); Jones et al. (2001); Fujisawa and Eguchi (2008)). Let $h(y;\mu)^\gamma$ $(\gamma > 0)$ be a density power weight for $Y^{(1)}$, where $h(y;\mu)$ is a symmetric density function with location parameter μ. The density $h(y;\mu^{(1)})$ is not necessarily equal to the true marginal density $f_1(y)$, but

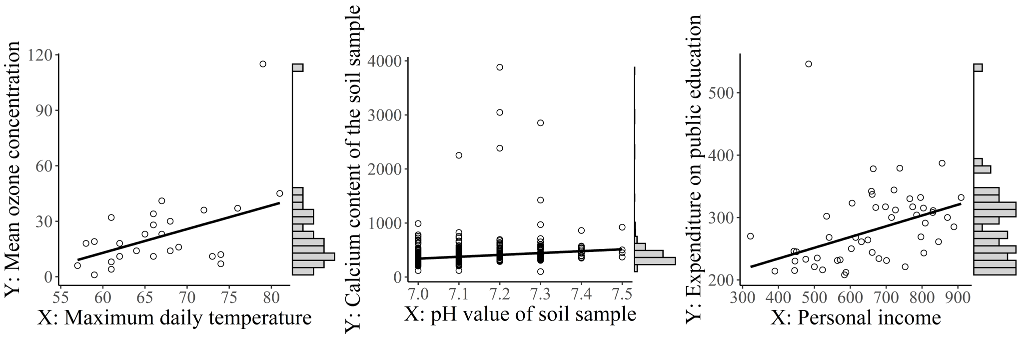

Figure 2. Real examples of outliers that satisfy Assumption 1. All data sets are included in the R package "robustbase" (Maechler et al. (2021)): airmay (left), condroz (center), education (right).

we assume that both h and the true density $f_1(y)$ are symmetric about $\mu^{(1)}$. The assumption of symmetry is common in outlier-resistant estimation, and is a prerequisite for using the sample median as an estimator of the population mean. Any symmetric density can be used for h, as long as it satisfies Assumption 1. Typically, we assume h is Gaussian. The tuning parameter γ controls the variability of the weight, leading to a trade-off between outlier resistance and asymptotic efficiency. Assumption 1 formally describes the assumption on outliers.

Assumption 1. *Let $h(y;\mu)$ be a weighting density symmetric about μ. Then, there exists $\gamma > 0$ such that*

$$\xi_1(X,\gamma) = \int \delta_1(y|X)h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})dy \approx 0 \qquad a.e. \tag{3.1}$$

Denote an arbitrary bounded function by $\phi(x)$. Assumption 1 implies that

$$\nu_1(\phi) := \mathbb{E}[\phi(X)\xi_1(X,\gamma)] = \int \phi(x)\xi_1(x,\gamma)g_X(x)dx \approx 0. \tag{3.2}$$

In particular, let $\phi(x) = 1$. Then, the outliers are marginally negligible:

$$\nu_1(1) = \mathbb{E}[\xi_1(X,\gamma)] = \int \delta_1(y)h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})dy \approx 0. \tag{3.3}$$

Throughout this paper, we assume that γ is sufficiently large that Assumption 1 holds. Assumption 1 reduces to a simpler form when $\delta_1(y|X)$ is Dirac's delta at y_0. This is one of the core assumptions in Section 4.

Assumption 1a. *Let $h(y;\mu)$ be a weighting density that is symmetric about μ, and assume that the density of the outliers is Dirac's delta at y_0 ($\neq \mu^{(1)}$), say*

$\delta_{y_0}(y)$. *Then, there exists* $\gamma > 0$ *such that*

$$\int \delta_{y_0}(y)h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})dy = h(y_0;\mu^{(1)})^{\gamma}(y_0-\mu^{(1)}) \approx 0. \tag{3.4}$$

For example, if h is a Gaussian density with mean $\mu^{(1)}$ and fixed variance, (3.4) tends to zero as $|y_0| \to \infty$, for fixed $\gamma > 0$, because $h(y_0;\mu^{(1)})^{\gamma}(y_0-\mu^{(1)}) \propto \exp\{-\gamma(y_0-\mu^{(1)})^2\}(y_0-\mu^{(1)})$.

3.1. IPW-type estimator

First, we introduce an extension of the IPW estimator, called the density power inverse probability weighting (DP-IPW) estimator. The DP-IPW estimator is defined as a root of the following estimating equation:

$$\sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})} h(Y_i;\mu)^{\gamma}(Y_i-\mu) = 0. \tag{3.5}$$

In the case of no contamination, the DP-IPW estimating equation is unbiased.

Theorem 1. *Assume that the true propensity score* $\pi(X;\alpha^*)$ *is given. Then, under no contamination, we have*

$$\mathbb{E}_g\left[\frac{T}{\pi(X;\alpha^*)} h(Y;\mu^{(1)})^{\gamma}(Y-\mu^{(1)})\right] = 0. \tag{3.6}$$

In practice, we often only have an estimate $\pi(X;\hat{\alpha})$, but the asymptotic consistency of (DP-)IPW still holds if the model $\pi(X;\alpha)$ is correctly specified.

Now, we consider the contaminated case. The bias of the DP-IPW estimating equation takes a different form from (2.8).

Theorem 2. *Assume* Y *is contaminated as* (2.6). *Under the same assumptions as those in Theorem* 1, *the expectation of the DP-IPW estimating equation is expressed as*

$$\mathbb{E}_{\tilde{g}}\left[\frac{T}{\pi(X;\alpha^*)} h(Y;\mu^{(1)})^{\gamma}(Y-\mu^{(1)})\right] = B_1 + \nu_1(\varepsilon_1), \tag{3.7}$$

$$\text{where} \quad B_1 = -\int \varepsilon_1(x) \int h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})g(y|x)dy\ g(x)dx.$$

In particular, under homogeneous contamination, B_1 *reduces to zero.*

The DP-IPW estimating equation is still biased, even if $\nu_1(\varepsilon_1)$ is small. Because we assume that $\nu_1(\varepsilon_1)$ is negligible, B_1 is dominant. However, compared with (2.8), the dominant bias of DP-IPW does not contain δ_1, implying that the bias of DP-IPW is not strongly affected by the absolute values of the outliers. Under homogeneous contamination, the dominant term disappears, and so the bias is negligible.

3.2. DR-type estimator

Next, we introduce the density power doubly robust (DP-DR) estimator. The DP-DR estimator is a straight application of the DR M-estimator, and is defined as a root of the following estimating equation:

$$\sum_{i=1}^{n}\left\{\frac{T_i h(Y_i;\mu)^\gamma}{\pi(X_i;\hat{\alpha})}(Y_i-\mu)-\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]\right\}=0. \tag{3.8}$$

As discussed in Section 2.1, $\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]$ is obtained by direct calculation or using a Monte Carlo approximation based on the parametric OR model $\hat{q}:=q(y|T=1,X;\hat{\beta})$. In the Appendix, we present the explicit forms of $\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]$ when h and q are assumed to be Gaussian. The parameter β is usually estimated in an outlier-resistant manner; for example, see the Huber regression (Huber (2004, Chap. 7)), MM estimator (Yohai (1987)), density power regression (Basu et al. (1998); Kanamori and Fujisawa (2015)), and γ-regression (Fujisawa and Eguchi (2008); Kawashima and Fujisawa (2017)), among others.

The DP-DR estimator is doubly robust under no contamination, as with the general DR M-estimator.

Theorem 3. *Assume either the true PS or the true OR model is given. Then, if there is no contamination, the DP-DR estimating equation is unbiased.*

Now, we evaluate the bias of the DP-DR estimating equation under contamination.

Theorem 4. *Assume that Y is contaminated as* (2.6)*. If the true PS model is given, the expectation of the DP-DR estimating equation is expressed as*

$$-\int\varepsilon_1(x)\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy\ g(x)dx+\nu_1(\varepsilon_1). \tag{3.9}$$

In particular, under homogeneous contamination, (3.9) *reduces to $\nu_1(\varepsilon_1)$. If the true OR model is given, the expectation of the DP-DR estimating equation is expressed as*

$$-\int\varepsilon_1(x)\frac{P(T=1|x)}{\pi(x;\alpha)}\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy\ g(x)dx+\nu_1\left\{\frac{\varepsilon_1 P(T=1|\cdot)}{\pi(\cdot;\alpha)}\right\}. \tag{3.10}$$

Under homogeneous contamination, (3.10) *becomes*

$$-\varepsilon_1\int\frac{P(T=1|x)}{\pi(x;\alpha)}\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy\ g(x)dx+\nu_1\left\{\frac{\varepsilon_1 P(T=1|\cdot)}{\pi(\cdot;\alpha)}\right\}. \tag{3.11}$$

Assuming that $\pi(\cdot;\alpha)$ is bounded away from zero and one, we find that $P(T = 1|\cdot)/\pi(\cdot;\alpha)$ is bounded. Then, from Assumption 1, $\nu_1\{\varepsilon_1 P(T = 1|\cdot)/\pi(\cdot;\alpha)\}$ is negligible. As with the DP-IPW, the dominant bias is independent of δ, indicating that the influence of outliers is reduced. Unfortunately, DP-DR is still biased in the PS-incorrect and OR-correct cases, even under homogeneous contamination, because the dominant term of (3.11) is not eliminated.

In the OR-correct case, the DP-DR is biased under homogeneous contamination because, under Assumption 1, the expectation of the DP-DR estimating function becomes

$$\mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha)}\left\{\mathbb{E}_{\tilde{g}}[\psi(Y^{(1)};\mu^{(1)})|X] - \mathbb{E}_g[\psi(Y^{(1)};\mu^{(1)})|X]\right\}\right]$$
$$\approx \mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha)}\left\{(1-\varepsilon_1)\mathbb{E}_g[\psi(Y^{(1)};\mu^{(1)})|X] - \mathbb{E}_g[\psi(Y^{(1)};\mu^{(1)})|X]\right\}\right],$$

where we denote the density power estimating function by ψ. In the last formula, the terms in curly brackets do not cancel, because the first term is reduced by $1-\varepsilon_1$. As such, we propose a bias-corrected version of the DP-DR, called the εDP-DR estimator, that is designed to cancel the dominant bias under homogeneous contamination. The εDP-DR estimator is a root of the following estimating equation:

$$\sum_{i=1}^{n}\left\{\frac{T_i h(Y_i;\mu)^\gamma}{\pi(X_i;\hat{\alpha})}(Y_i-\mu) - \frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}(1-\hat{\varepsilon}_1)\mathbb{E}_{\hat{q}}\left[h(Y;\mu)^\gamma(Y-\mu)|T=1,X\right]\right\}$$
$$=0, \tag{3.12}$$

where $\hat{\varepsilon}_1$ is a consistent estimator of the expected contamination ratio $\overline{\varepsilon}_1 = \int \varepsilon_1(x)g(x)dx$. Here, $\hat{\varepsilon}_1$ can be obtained simultaneously with the parametric OR model using unnormalized modeling with the density power score (Kanamori and Fujisawa (2015)), for example. While the DP-DR is a special case of the DR M-estimator, the εDP-DR goes beyond this framework, owing to the bias correction. Under no contamination, the εDP-DR estimating equation is asymptotically identical to the DP-DR estimating equation. The εDP-DR estimating equation is also biased under heterogeneous contamination; however, the bias takes a different form.

Corollary 1. *If the true PS model is given, the expectation of the εDP-DR estimating equation is equal to* (3.9). *If the true OR model is given, the expectation of the εDP-DR estimating equation is expressed as*

$$\mathbb{E}_g\left[(\overline{\varepsilon}_1 - \varepsilon_1(X))\frac{P(T=1|X)}{\pi(X;\alpha)}\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right]$$
$$+\nu_1\left\{\frac{\varepsilon_1 P(T=1|\cdot)}{\pi(\cdot;\alpha)}\right\}. \tag{3.13}$$

Table 1. Summary of the biases of the proposed estimating equations. The function $\phi(X)$ differs in each cell. PS-correct means that the PS model is correctly specified and the OR model may not be; OR-correct means the opposite.

Contamination	model	DP-IPW	DP-DR	εDP-DR
No contamination	PS-correct	0	0	0
	OR-correct	-	0	0
homogeneous: ε	PS-correct	≈ 0	≈ 0	≈ 0
	OR-correct	-	$\approx \varepsilon\mathbb{E}[\phi(X)]$	≈ 0
heterogeneous: $\varepsilon(X)$	PS-correct	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$
	OR-correct	-	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[(\bar{\varepsilon} - \varepsilon(X))\phi(X)]$

The first term disappears under homogeneous contamination.

Proof. The derivation is the same as that of Theorem 4. If $\varepsilon_1(X)$ is a constant ε_1, the first term disappears, because $\bar{\varepsilon}_1 = \varepsilon_1 \int g(x)dx = \varepsilon_1$.

Similarly to (3.11), the second term of (3.13) is approximately zero if we assume that $\pi(\cdot;\alpha)$ is bounded away from zero and one.

Remark 1. Note that "$\varepsilon(X)$"DP-DR may work better than εDP-DR under heterogeneous contamination. In fact, the bias (3.13) disappears if we replace $\bar{\varepsilon}$ with $\varepsilon(X)$. However, it is necessary to model $\varepsilon(X)$ correctly for a consistent estimation of "$\varepsilon(X)$"DP-DR. To the best of our knowledge, no easy method is available for this purpose.

3.3. Summary

We have proposed three types of outlier-resistant semiparametric estimators: DP-IPW, DP-DR, and εDP-DR. Table 1 shows the bias of the estimating equations under the conditions discussed above. Under heterogeneous contamination, all estimators are biased, but the bias is little affected by the absolute values of the outliers. Furthermore, as discussed in Section 4, outliers have negligible influence if the contamination ratio is sufficiently small. Here, the εDP-DR improves on the DP-DR in the OR-correct case under homogeneous contamination, but we continue to discuss DP-DR for the following three reasons: the contamination ratio is sometimes difficult to estimate, the bias (3.11) is not serious if $\pi(X;\alpha)$ is close to $P(T=1|X)$, and the simulation results presented in Section 6 indicate that the DP-DR outperforms existing methods, even in the OR-correct case.

4. Influence-function-based Analysis of Outlier Resistance

As discussed in the previous section, the proposed estimators suffer less from outliers than ordinary estimators do in terms of the unbiasedness of the estimating equation. In this section, we demonstrate that they are outlier-resistant from the viewpoint of the IF.

Here, we briefly review the IF for the univariate M-estimator, and then expand it to evaluate our estimators. Let G be the distribution of $Z \in \mathbb{R}$, and let $T(G)$ be a functional of G, which is the parameter of interest. The IF of $T(G)$ is defined as

$$IF(z_0; G) := \lim_{\varepsilon \to 0} \frac{T\{(1-\varepsilon)G + \varepsilon\Delta_{z_0}\} - T(G)}{\varepsilon}$$
$$= \frac{\partial}{\partial\varepsilon}[T\{(1-\varepsilon)G + \varepsilon\Delta_{z_0}\} - T(G)]\bigg|_{\varepsilon=0}, \tag{4.1}$$

where Δ_{z_0} is a degenerate distribution at z_0. Furthermore, the latent bias $T\{(1-\varepsilon)G + \varepsilon\Delta_{z_0}\} - T(G)$ is approximated by $\varepsilon IF(z_0; G)$. Therefore, the behavior of the IF approximates that of the latent bias. In the population, the M-estimator $T_M(G)$ satisfies $\int \psi\{z, T_M(G)\}dG(z) = 0$. Then, the IF for $T_M(G)$ is obtained by differentiating $\int \psi[z, T_M\{(1-\varepsilon)G + \varepsilon\Delta_{z_0}\}]d\{(1-\varepsilon)G + \varepsilon\Delta_{z_0}\}(z) = 0$ with respect to ε. This yields

$$IF(z_0; G) = -\mathbb{E}\left[\frac{\partial}{\partial\eta}\psi(Z,\eta)\bigg|_{\eta=T_M(G)}\right]^{-1}\psi\{z_0, T_M(G)\}. \tag{4.2}$$

The function ψ is said to have a redescending property if $\psi\{z_0, T_M(G)\}$ approaches zero as the outlier $|z_0|$ increases. Therefore, when ψ has the redescending property and z_0 is an outlier, the latent bias is sufficiently small. This is favorable for outlier resistance.

Because ε_1 depends on X, we cannot apply the IF directly to our estimators. To overcome this issue, we consider an IF with fixed covariates $\{X_i\}_{i=1}^n$; this approach is similar to the fixed carrier model in Hampel et al. (2011, Chap. 6). Consider the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{g}}\left[\psi(Y,T,X_i;\mu)|\, X_i\right] = 0. \tag{4.3}$$

If the fixed sample $\{X_i\}_{i=1}^n$ consists of i.i.d. observations, then the left-hand side of (4.3) converges to $\mathbb{E}_{\tilde{g}}[\psi(Y,T,X;\mu)]$ as $n \to \infty$. Let $\tilde{\mu}_n^{(1)}$ denote a root of (4.3), and let $\tilde{\mu}^{(1)}$ be a root of $\mathbb{E}_{\tilde{g}}[\psi(Y,T,X;\mu)]$. Then, $\tilde{\mu}_n^{(1)}$ also converges to $\tilde{\mu}^{(1)}$. Therefore, $\tilde{\mu}_n^{(1)}$ exhibits roughly the same behavior as that of the target estimator $\tilde{\mu}^{(1)}$. The contaminated density $\tilde{g}$ is defined as (2.6), and $\delta_1(y|X_i)$ is assumed to be Dirac's delta at y_0. The IF of $T_n(\tilde{G})$ at X_i is obtained by differentiating (4.3) with respect to $\varepsilon_1(X_i)$ at $\varepsilon_1(X_i) = 0$.

To conserve space, we discuss only the εDP-DR. Assume that $\bar{\varepsilon}_1 = (1/n) \sum_{i=1}^{n} \varepsilon_1(X_i)$. Then, the IF of the εDP-DR is

$$- \mathbb{E}_g \left[\left. \frac{\partial \psi}{\partial \mu} \right|_{\mu=\mu_n^{(1)}} \middle| X_i \right]^{-1} \left\{ \frac{P(T=1|X_i)}{\pi(X_i;\alpha)} h(y_0 - \mu_n^{(1)})^\gamma (y_0 - \mu_n^{(1)}) \right.$$
$$\left. - \frac{n-1}{n} \frac{P(T=1|X_i) - \pi(X_i;\alpha)}{\pi(X_i;\alpha)} \mathbb{E}_{\hat{q}} \left[h(Y;\mu_n^{(1)})^\gamma (Y - \mu_n^{(1)}) | T=1, X \right] \right\}. \quad (4.4)$$

In the PS-correct case, the second term in square brackets is equal to zero, and the IF tends to zero as $|y_0| \to \infty$. In the OR-correct case, the second term does not disappear. Considering the limit of $|y_0| \to \infty$, the IF converges to

$$\frac{n-1}{n} \mathbb{E}_g \left[\left. \frac{\partial \psi}{\partial \mu} \right|_{\mu=\mu_n^{(1)}} \middle| X_i \right]^{-1}$$
$$\left\{ \frac{P(T=1|X_i) - \pi(X_i;\alpha)}{\pi(X_i;\alpha)} \mathbb{E}_{\hat{q}} [h(Y;\mu_n^{(1)})^\gamma (Y - \mu_n^{(1)}) | T=1, X_i] \right\}. \quad (4.5)$$

Thus, the εDP-DR estimator has the redescending property only in the PS-correct case. In the OR-correct case, the effects of outliers cannot be eliminated, but the IF tends to a constant when $|y_0|$ tends to infinity, implying that these effects are not serious. The DP-DR has an IF similar to that of the εDP-DR, and the DP-IPW has an IF similar to that of the εDP-DR, which has a correct PS. The derivations of all IFs are presented in the Appendix.

Under homogeneous contamination, the ordinary IF is applicable, and the proposed estimators have the redescending property in the PS-correct case. In addition, the εDP-DR has the redescending property even in the OR-correct case. This result is consistent with Corollary 1. An IF-based analysis under homogeneous contamination is presented in the Appendix.

5. Asymptotic Properties

In this section, we discuss the asymptotic properties of the εDP-DR estimator. For the other proposed estimators, we obtain similar results, with small changes. The asymptotic properties can be obtained as in Hoshino (2007). Assume that the PS and OR models are regular and are estimated consistently if the models are correctly specified. Furthermore, the contamination ratio ε_1 is known. Note that when the contamination ratio is consistently estimated simultaneously using the OR model of Kanamori and Fujisawa (2015), we can replace β with $(\varepsilon_1, \beta^T)^T$ in the following discussion.

We write (3.12) as $(1/n)\sum_{i=1}^n \psi_i(\mu;\hat{\alpha},\hat{\beta})$, and let $(1/n)\sum_{i=1}^n s_i^{PS}(\alpha) = 0$ and $(1/n)\sum_{i=1}^n s_i^{OR}(\beta) = 0$ be the estimating equations for the PS and OR models, respectively. Let $\lambda = (\mu, \alpha^T, \beta^T)^T$ be the parameter vector, and let the full estimating equation be defined as

$$\sum_{i=1}^{n} S_i(\lambda) = \sum_{i=1}^{n} \begin{pmatrix} \psi_i(\mu;\alpha,\beta) \\ s_i^{PS}(\alpha) \\ s_i^{OR}(\beta) \end{pmatrix} = \mathbf{0}. \tag{5.1}$$

Let $\lambda^* = (\mu^*, \alpha^{*T}, \beta^{*T})^T$ be a root of (5.1) in the population. Note that, in this section, a $*$ does not necessarily mean that the model is specified correctly. With the results presented in Van der Vaart (2000, Chap. 5), the following theorem holds under some regularity conditions.

Theorem 5. *Under the regularity conditions presented in the Appendix, the following asymptotic properties hold:*

$$\hat{\lambda} \xrightarrow{p} \lambda^*, \tag{5.2}$$

$$\sqrt{n}(\hat{\lambda} - \lambda^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{\tilde{g}}(\lambda^*)\right), \tag{5.3}$$

where $\mathbf{V}^{\tilde{g}}(\lambda^*) = \mathbf{J}^{\tilde{g}}(\lambda^*)^{-1}\mathbf{K}^{\tilde{g}}(\lambda^*)\{\mathbf{J}^{\tilde{g}}(\lambda^*)^T\}^{-1}$, $\mathbf{J}^{\tilde{g}}(\lambda^*) = \mathbb{E}_{\tilde{g}}\left[\partial S_i(\lambda^*)/\partial\lambda^T\right]$, *and* $\mathbf{K}^{\tilde{g}}(\lambda^*) = \mathbb{E}_{\tilde{g}}\left[S_i(\lambda^*)S_i(\lambda^*)^T\right]$.

Using this and applying the results presented in Section 3.2, we find that the limit μ^* is in the neighborhood of $\mu^{(1)}$.

Theorem 6. *Let* $\lambda^{**} = (\mu^{(1)}, \alpha^{*T}, \beta^{*T})^T$ *and assume that* $\mathbf{J}_{11}^{\tilde{g}}(\lambda)$ *is nonzero within the interval* $[\lambda^*, \lambda^{**}]$. *Under Assumption* 1 *and homogeneous contamination, if either the PS or the OR model is correct, it then holds that*

$$\mu^* = \mu^{(1)} + \mathcal{O}\{\nu_1(\phi)\}, \tag{5.4}$$

where $\phi(\cdot) = \varepsilon_1$ *(constant) in the PS-correct case, and* $\phi(\cdot) = \varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha)$ *in the OR-correct case.*

The proof of Theorem 6 and further discussions on the asymptotic variance are available in the Appendix.

6. Monte Carlo Simulation

We conduct Monte Carlo simulations to evaluate the performance of the proposed estimators. Here, we compare our methods with the naive IPW and DR estimators, as well as some existing outlier-resistant methods (Firpo (2007); Zhang et al. (2012); Díaz (2017); Sued, Valdora and Yohai (2020)). Because these methods focus on the median of the potential outcome, they are resistant to outliers, up to a point, but are not resistant to heavy contamination. To the best of our knowledge, the proposed method is the first to offer resistance to outliers that is greater than that of the median. Firpo's IPW estimator (Firpo (2007)) is defined as

$$\hat{\mu}_{\text{Firpo}} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{T_i}{\pi(X_i; \hat{\alpha})} (Y_i - \mu)\{0.5 - \mathbb{I}(Y_i \le \mu)\}, \tag{6.1}$$

where the function $\mathbb{I}$ is an indicator function. Zhang's IPW median (Zhang et al. (2012)) is based on the IPW-empirical distribution. Firpo's IPW and Zhang's IPW are almost equivalent, except for a slight difference in their computation. Zhang's and Sued's DR methods (Zhang et al. (2012); Sued, Valdora and Yohai (2020)) estimate the empirical distribution in a doubly robust way by incorporating an IPW-type estimator into the first term. The remaining term of Zhang's DR is based on the Gaussian cumulative distribution function of Y given X. In contrast, Sued's DR constructs the remaining term in a nonparametric manner. Diaz's DR median (Díaz (2017)) offers a different approach by using the targeted maximum likelihood estimator (TMLE) (Van der Laan and Rubin (2006)). We implemented our methods, Zhang's IPW/DR, and Sued's DR in R. For Firpo's IPW and the TMLE, we used the *causalquantile* package (`https://github.com/idiazst/causalquantile`; updated on August 31, 2017).

6.1. Numerical algorithm for the proposed methods

Because the proposed estimating equations cannot be solved explicitly, we develop an iterative algorithm. Various algorithms are available, but we propose a standard algorithm for M-estimators (Huber (2004); Hampel et al. (2011)). A detailed version of the algorithm is available in the Appendix. Hereafter, we suppose h and q are Gaussian, and we provide explicit updating formulae in this case. Note that some additional parameters of h should be estimated in a roughly unbiased and outlier-resistant way.

6.2. Simulation model

We simulate random observations based on a simple causal setting. The confounders (X_1, X_2) are drawn independently from a Gaussian or uniform distribution with mean zero and unit variance. The treatment T is assigned with the conditional probability $P(T = 1|X_1, X_2)$, which is defined as a sigmoid function of $0.8X_1+0.2X_2$. The potential outcomes $(Y^{(1)}, Y^{(0)})$ are generated from linear functions of (X_1, X_2) with a Gaussian error: $Y^{(1)} = \mu^{(1)}+1.2X_1+0.3X_2+e$ and $Y^{(0)} = \mu^{(0)} + 1.2X_1 + 0.3X_2 + e$. Here, $\mu^{(1)}$ and $\mu^{(0)}$ are set to three and zero, respectively. The standard deviation (SD) of e is set to $\sqrt{0.72}$; then, $\mathrm{SD}[Y^{(1)}] = \mathrm{SD}[Y^{(0)}] = 1.5$. When the confounders are not Gaussian, the potential outcomes are not Gaussian. The observed outcome Y is defined as $Y = TY^{(1)} + (1 - T)Y^{(1)}$ under no contamination. Outliers are drawn from $\mathcal{N}(\mu^{(t)} + 10\sigma^{(t)}, 1)$, with $\sigma^{(t)} = \mathrm{SD}[Y^{(t)}] = 1.5$. For the homogeneous contamination settings, the contamination ratio is set to a constant ε_t. For the heterogeneous contamination settings, the contamination ratio is set to $1.5\varepsilon_t$ if

$X_1 + X_2 \leq 0$, and $0.5\varepsilon_t$ if $X_1 + X_2 > 0$. The average contamination ratio is set to $\varepsilon_t \in \{0, 0.05, 0.1, 0.2\}$. Then, the observations of Y are replaced randomly with outliers according to the contamination ratio. The sample size is fixed to $n = 100$ throughout the simulations. Furthermore, we generate data sets in which the outcome follows a symmetric and heavy-tailed distribution. We draw the error term of $Y^{(t)}$ from the standard Cauchy distribution, rather than inserting outliers.

6.3. Results

First, we perform a comparative study. The potential mean $\mu^{(1)}$ is estimated using various aforementioned methods. In this experiment, we use all settings described in the previous section. The propensity score is estimated using a logistic regression. The parametric OR is conducted in two ways: a Gaussian MLE with nonoutliers, and unnormalized Gaussian modeling (the tuning parameter is set to 0.5) (Kanamori and Fujisawa (2015)). For the DR estimators, we investigate three patterns of model misspecification: PS-correct/OR-correct, PS-correct/OR-incorrect, and PS-incorrect/OR-correct. For the model-correct case, we include an intercept and (X_1, X_2) as covariates. For the model-incorrect case, we include only an intercept and X_2. We perform 10,000 simulations for every setting and method. Table 2 and 3 show the results of the comparative study of the IPW-type estimators and that of the DR-type estimators, respectively. To save space, only the results when the covariates are Gaussian are presented. The estimation error is measured as the root mean squared error (RMSE). The mean and SD of all estimates, mean computation time, and results for the other settings are provided in the Appendix. In Table 2, the naive IPW estimator shows a significantly larger RMSE under contamination. The median-based methods and DP-IPW both dramatically reduce the RMSE. The RMSE increases as the contamination ratio increases. The RMSE tends to be larger for heterogeneous contamination than for homogeneous contamination. With the optimal γ, the proposed method outperforms the comparative methods and has the smallest RMSE for all settings. In Table 3, the results for the DR-type estimators are similar to those for the IPW estimators. The proposed method with a proper γ again outperforms the comparative methods and has the smallest RMSE in all settings. The DP-DR and εDP-DR perform similarly, although the εDP-DR is slightly superior in many settings. The TMLE performs best of the median-based methods, but it takes much longer than the other methods, including the proposed methods, and occasionally ($< 1\%$) failed to converge.

Table 4 shows the RMSE of each method for the data with the Cauchy error. As before, the proposed method outperforms the comparative methods. In this setting, we use unnormalized Gaussian modeling for the OR for the DR-type estimators. Only in the PS-correct/OR-incorrect case does, the median (TMLE) perform slightly better than the proposed method.

Table 2. RMSE of the IPW-type estimators. X is drawn from a Gaussian distribution.

	Homogeneous				**Heterogeneous**		
ε	0.00	0.05	0.10	0.20	0.05	0.10	0.20
Naive	0.222	0.957	1.683	3.153	0.993	1.752	3.253
median (Firpo)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
median (Zhang-IPW)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
DP-IPW ($\gamma = 0.1$)	0.218	0.276	0.531	2.263	0.293	0.609	2.377
DP-IPW ($\gamma = 0.5$)	0.227	0.249	0.272	0.639	0.245	0.287	0.726
DP-IPW ($\gamma = 1.0$)	0.261	0.271	0.275	0.413	0.262	0.281	0.498

Next, we conduct a γ-sensitivity study, and estimate $\mu^{(1)}$ using the proposed methods with different γ. Here, X follows a Gaussian distribution, and the contamination ratio varies in $\{0, 0.05, 0.1, 0.2\}$ under homogeneous contamination. For the DR-type estimators, we perform the OR using the Gaussian MLE with nonoutliers. We simulated 10,000 data sets for every setting and method. Table 5 shows the results of the γ-sensitivity study. As in the comparative study, the bias increases with the ratio of the outliers. Larger γ results in increased variance. When the contamination ratio is small, it is sufficient to use a small γ, such as $\gamma = 0.1$ or 0.2, to remove the adverse effects of outliers. Even in highly contaminated cases, $\gamma > 1.0$ is not needed. Comparing the DP-DR and εDP-DR estimates in the PS-incorrect/OR-correct case, we find that the DP-DR estimates are biased, especially when ε is large. In contrast, the εDP-DR estimates are not biased, demonstrating that the bias correction by $1 - \hat{\varepsilon}$ works well in our experiments.

As in many other outlier-resistant statistical methods, parameter tuning is challenging. We suggest a possible method based on the solution paths of the proposed estimators, provided in the Appendix. The effects of the outliers on the paths decrease as γ increases, and the paths became stable around the true value after reaching a certain γ. Thus, we suggest using the smallest γ for which the estimate is stable.

7. Real-Data Analysis

In this section, we use the proposed method to estimate the ATE on a real data set. We use data from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). The NHEFS is a national longitudinal study performed by U.S. public agencies. We use a processed data set, available online (Hernán and Robins (2020, `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`). The NHEFS data set contains 1,566 observations of smokers who enrolled in the study from 1971 to 1975. By the follow-up visit in 1982, 403 (25.7%) participants had quit smoking. The goal of the study was to evaluate the treatment effect of smoking

Table 3. RMSE of the DR-type estimators. X is drawn from a Gaussian distribution. The OR model is obtained using Gaussian MLE using nonoutliers.

		Homogeneous			**Heterogeneous**		
ε	0.00	0.05	0.10	0.20	0.05	0.10	0.20
(PS-correct/OR-correct)							
Naive	0.184	0.957	1.684	3.154	0.997	1.758	3.265
median (Zhang-DR)	0.239	0.317	0.391	0.733	0.330	0.452	0.905
median (Sued)	0.238	0.316	0.388	0.693	0.329	0.450	0.869
median (TMLE)	0.237	0.280	0.359	0.603	0.295	0.402	0.701
DP-DR ($\gamma = 0.1$)	0.183	0.302	0.564	2.262	0.318	0.649	2.394
DP-DR ($\gamma = 0.5$)	0.202	0.285	0.326	0.697	0.274	0.349	0.834
DP-DR ($\gamma = 1.0$)	0.240	0.288	0.307	0.524	0.287	0.336	0.669
εDP-DR ($\gamma = 0.1$)	0.183	0.296	0.554	2.255	0.314	0.636	2.385
εDP-DR ($\gamma = 0.5$)	0.202	0.264	0.302	0.669	0.271	0.323	0.793
εDP-DR ($\gamma = 1.0$)	0.240	0.287	0.299	0.513	0.286	0.335	0.648
(correct/incorrect)							
Naive	0.237	0.963	1.686	3.156	1.001	1.758	3.262
median (Zhang-DR)	0.275	0.342	0.408	0.741	0.350	0.465	0.912
median (Sued)	0.275	0.342	0.407	0.699	0.350	0.464	0.872
median (TMLE)	0.242	0.284	0.363	0.622	0.297	0.404	0.719
DP-DR ($\gamma = 0.1$)	0.237	0.314	0.561	2.267	0.330	0.644	2.393
DP-DR ($\gamma = 0.5$)	0.247	0.319	0.349	0.714	0.319	0.361	0.839
DP-DR ($\gamma = 1.0$)	0.280	0.334	0.347	0.581	0.329	0.372	0.709
εDP-DR ($\gamma = 0.1$)	0.237	0.311	0.557	2.264	0.328	0.640	2.388
εDP-DR ($\gamma = 0.5$)	0.247	0.317	0.344	0.694	0.313	0.356	0.817
εDP-DR ($\gamma = 1.0$)	0.280	0.333	0.338	0.551	0.327	0.369	0.708
(incorrect/correct)							
Naive	0.181	0.879	1.591	3.026	0.826	1.490	2.813
median (Zhang-DR)	0.237	0.263	0.316	0.503	0.269	0.337	0.548
median (Sued)	0.236	0.272	0.346	0.599	0.277	0.364	0.627
median (TMLE)	0.234	0.260	0.309	0.478	0.265	0.328	0.522
DP-DR ($\gamma = 0.1$)	0.182	0.192	0.345	2.057	0.191	0.299	1.681
DP-DR ($\gamma = 0.5$)	0.199	0.206	0.218	0.366	0.203	0.209	0.283
DP-DR ($\gamma = 1.0$)	0.230	0.232	0.239	0.273	0.230	0.233	0.242
εDP-DR ($\gamma = 0.1$)	0.182	0.193	0.381	2.207	0.194	0.335	1.839
εDP-DR ($\gamma = 0.5$)	0.199	0.203	0.208	0.376	0.203	0.212	0.318
εDP-DR ($\gamma = 1.0$)	0.230	0.230	0.231	0.243	0.231	0.237	0.260

cessation ($T = 1$) on weight gain (Y). Other than the treatment and outcome, several baseline variables were collected, including sex, age, race, education level, intensity and duration of smoking, physical activity in daily life, recreational exercise, and baseline weight. We use all of these to control for confounding in a similar manner to that of Hernán and Robins (2020). We include linear

Table 4. RMSE of the comparative study using heavy-tailed data. The covariate X is drawn from a Gaussian distribution. The OR model for the DR-type estimators is obtained using unnormalized Gaussian modeling.

	IPW
Naive	274.024
median (Firpo)	0.414
median (Zhang-IPW)	0.414
DP-IPW ($\gamma = 0.1$)	0.443
DP-IPW ($\gamma = 0.5$)	0.367
DP-IPW ($\gamma = 1.0$)	0.380

		DR	
PS/OR	correct/correct	correct/incorrect	incorrect/correct
Naive	275.447	275.446	263.629
median (Zhang-DR)	0.415	0.456	0.390
median (Sued)	0.408	0.436	0.373
median (TMLE)	0.392	0.394	0.389
DP-DR ($\gamma = 0.1$)	0.501	0.514	0.390
DP-DR ($\gamma = 0.5$)	0.363	0.404	0.358
DP-DR ($\gamma = 1.0$)	0.372	0.418	0.364
εDP-DR ($\gamma = 0.1$)	0.487	0.503	0.377
εDP-DR ($\gamma = 0.5$)	0.361	0.399	0.328
εDP-DR ($\gamma = 1.0$)	0.370	0.412	0.334

and quadratic terms for all continuous covariates (age, intensity and duration of smoking, and baseline weight), and dummy terms for the discrete covariates. We estimate the propensity score using a logistic regression, and perform the outcome regression using unnormalized Gaussian modeling (the tuning parameter is set to 0.2). The original data set does not contain obvious outliers. Therefore, we randomly replace 10% of the observations with outliers drawn from $\mathcal{N}(100, 5^2)$. Then, we estimated $\mu^{(1)}$, $\mu^{(0)}$, and the ATE using the same methods as those in the Monte Carlo simulations. This process is repeated 10,000 times. The results are summarized in Table 6. For reference, we estimated every target quantity using the naive IPW/DR from the original data.

For the IPW-type estimators, the median-based methods give larger estimates of $\mu^{(1)}$ and $\mu^{(0)}$ than those in the case of IPW (no outliers), particularly for $\mu^{(0)}$. As a result, using the median-based methods, the ATE is estimated to be smaller than that in the case of IPW (no outliers). In contrast, DP-IPW overestimates $\mu^{(1)}$ with $\gamma = 0.05$, and underestimates $\mu^{(1)}$ with $\gamma \geq 0.10$. It overestimates $\mu^{(0)}$ compared with the case of IPW (no outliers), and this tendency strengthened with increasing γ. However, because the overestimation of $\mu^{(0)}$ is smaller than that of the median-based methods, the estimate of the ATE by

Table 5. Results of the γ-sensitivity study. Each figure is the mean (sd) of 10,000 simulations for each setting. In the second column, "T" and "F" denote correct and incorrect modeling, respectively.

	PS/OR	ε	$\gamma = 0.0$	0.1	0.2	0.5	1.0	1.5	2.0
DP-IPW	T/-	0.00	3.004 (0.22)	2.998 (0.22)	2.994 (0.22)	2.986 (0.23)	2.980 (0.26)	2.974 (0.30)	2.970 (0.34)
		0.05	3.749 (0.59)	3.030 (0.27)	2.999 (0.26)	2.987 (0.25)	2.978 (0.27)	2.970 (0.30)	2.963 (0.33)
		0.10	4.493 (0.78)	3.142 (0.51)	3.015 (0.32)	2.989 (0.27)	2.977 (0.27)	2.969 (0.30)	2.963 (0.33)
		0.20	5.983 (1.02)	4.492 (1.70)	3.536 (1.39)	3.052 (0.64)	2.990 (0.41)	2.978 (0.39)	2.971 (0.40)
DP-DR	T/T	0.00	2.999 (0.18)	2.998 (0.18)	2.997 (0.19)	2.996 (0.20)	2.992 (0.24)	2.989 (0.28)	2.985 (0.31)
		0.05	3.745 (0.60)	3.029 (0.30)	3.002 (0.27)	2.997 (0.29)	2.991 (0.29)	2.985 (0.31)	2.980 (0.34)
		0.10	4.489 (0.79)	3.140 (0.55)	3.017 (0.36)	3.000 (0.33)	2.992 (0.31)	2.986 (0.32)	2.981 (0.33)
		0.20	5.979 (1.04)	4.465 (1.72)	3.532 (1.41)	3.060 (0.69)	3.009 (0.52)	2.999 (0.51)	2.994 (0.51)
	T/F	0.00	3.004 (0.24)	2.998 (0.24)	2.994 (0.24)	2.986 (0.25)	2.979 (0.28)	2.974 (0.32)	2.968 (0.36)
		0.05	3.750 (0.60)	3.033 (0.31)	3.001 (0.29)	2.989 (0.32)	2.978 (0.33)	2.970 (0.36)	2.963 (0.39)
		0.10	4.494 (0.78)	3.150 (0.54)	3.020 (0.37)	2.992 (0.35)	2.979 (0.35)	2.970 (0.37)	2.963 (0.39)
		0.20	5.984 (1.03)	4.490 (1.71)	3.546 (1.41)	3.059 (0.71)	3.001 (0.58)	2.985 (0.55)	2.975 (0.54)
	F/T	0.00	2.999 (0.18)	2.999 (0.18)	2.999 (0.18)	3.001 (0.20)	3.005 (0.23)	3.010 (0.26)	3.014 (0.29)
		0.05	3.725 (0.50)	2.997 (0.19)	2.976 (0.19)	2.975 (0.20)	2.978 (0.23)	2.982 (0.26)	2.986 (0.29)
		0.10	4.451 (0.65)	3.051 (0.34)	2.956 (0.21)	2.950 (0.21)	2.953 (0.23)	2.956 (0.26)	2.960 (0.28)
		0.20	5.902 (0.86)	4.326 (1.57)	3.301 (1.15)	2.907 (0.35)	2.895 (0.25)	2.897 (0.26)	2.900 (0.28)
εDP-DR	T/T	0.00	2.999 (0.18)	2.998 (0.18)	2.997 (0.19)	2.996 (0.20)	2.992 (0.24)	2.989 (0.28)	2.985 (0.31)
		0.05	3.745 (0.60)	3.028 (0.29)	3.002 (0.27)	2.997 (0.26)	2.991 (0.29)	2.985 (0.31)	2.980 (0.34)
		0.10	4.489 (0.78)	3.138 (0.54)	3.017 (0.35)	2.999 (0.30)	2.991 (0.30)	2.985 (0.32)	2.980 (0.33)
		0.20	5.978 (1.03)	4.464 (1.72)	3.531 (1.40)	3.058 (0.67)	3.007 (0.51)	2.998 (0.50)	2.993 (0.51)
	T/F	0.00	3.004 (0.24)	2.998 (0.24)	2.994 (0.24)	2.986 (0.25)	2.979 (0.28)	2.974 (0.32)	2.968 (0.36)
		0.05	3.750 (0.60)	3.033 (0.31)	3.001 (0.29)	2.989 (0.32)	2.978 (0.33)	2.970 (0.36)	2.963 (0.39)
		0.10	4.493 (0.78)	3.149 (0.54)	3.020 (0.36)	2.992 (0.34)	2.978 (0.34)	2.970 (0.37)	2.963 (0.39)
		0.20	5.983 (1.02)	4.489 (1.71)	3.543 (1.40)	3.057 (0.69)	2.998 (0.55)	2.984 (0.54)	2.976 (0.54)
	F/T	0.00	2.999 (0.18)	2.999 (0.18)	2.999 (0.18)	3.001 (0.20)	3.005 (0.23)	3.010 (0.26)	3.014 (0.29)
		0.05	3.746 (0.50)	3.020 (0.19)	2.998 (0.19)	2.998 (0.20)	3.001 (0.23)	3.005 (0.26)	3.009 (0.29)
		0.10	4.493 (0.66)	3.108 (0.37)	3.004 (0.20)	2.998 (0.21)	3.001 (0.23)	3.004 (0.26)	3.007 (0.28)
		0.20	5.986 (0.87)	4.541 (1.58)	3.486 (1.24)	3.020 (0.38)	3.003 (0.24)	3.005 (0.25)	3.008 (0.27)

Table 6. Results of the NHEFS data analysis. Each figure shows the mean (sd) of 10,000 estimates.

	Target Quantities		
	$\mu^{(1)}$	$\mu^{(0)}$	ATE
IPW (no outliers)	5.221 (-)	1.780 (-)	3.441 (-)
IPW	14.718 (1.57)	11.607 (0.87)	3.111 (1.78)
median (Firpo)	5.439 (0.21)	2.753 (0.10)	2.686 (0.24)
median (Zhang-IPW)	5.439 (0.21)	2.753 (0.10)	2.686 (0.24)
DP-IPW ($\gamma = 0.05$)	5.597 (0.30)	1.851 (0.07)	3.746 (0.31)
DP-IPW ($\gamma = 0.10$)	5.157 (0.15)	1.819 (0.07)	3.338 (0.17)
DP-IPW ($\gamma = 0.20$)	5.089 (0.15)	1.875 (0.06)	3.215 (0.16)
DP-IPW ($\gamma = 0.50$)	4.949 (0.15)	2.007 (0.06)	2.941 (0.16)
DR (no outliers)	5.136 (-)	1.772 (-)	3.364 (-)
DR	14.574 (1.57)	11.589 (0.90)	2.985 (1.81)
median (Zhang-DR)	5.352 (0.20)	2.743 (0.10)	2.609 (0.22)
median (Sued)	5.353 (0.20)	2.744 (0.10)	2.609 (0.23)
median (TMLE)	5.363 (0.21)	2.739 (0.10)	2.624 (0.23)
DP-DR ($\gamma = 0.05$)	5.478 (0.27)	1.842 (0.07)	3.636 (0.28)
DP-DR ($\gamma = 0.10$)	5.057 (0.16)	1.810 (0.07)	3.248 (0.17)
DP-DR ($\gamma = 0.20$)	4.983 (0.16)	1.865 (0.06)	3.119 (0.17)
DP-DR ($\gamma = 0.50$)	4.834 (0.16)	1.997 (0.06)	2.837 (0.17)
εDP-DR ($\gamma = 0.05$)	5.574 (0.29)	1.851 (0.07)	3.723 (0.30)
εDP-DR ($\gamma = 0.10$)	5.148 (0.15)	1.819 (0.07)	3.330 (0.17)
εDP-DR ($\gamma = 0.20$)	5.080 (0.15)	1.874 (0.06)	3.206 (0.17)
εDP-DR ($\gamma = 0.50$)	4.937 (0.15)	2.007 (0.06)	2.930 (0.16)

the DP-IPW is closer to that obtained using IPW (no outliers) than when using the median-based methods. The DR-type estimators show similar results. The median-based methods overestimate $\mu^{(1)}$ and $\mu^{(0)}$, and the DP-DR and εDP-DR underestimate $\mu^{(1)}$ and overestimate $\mu^{(0)}$. The DP-DR and εDP-DR estimate the ATE better than the median-based methods do. In addition, the DP-DR and εDP-DR exhibit the same tendency of estimation bias and γ; a larger value of γ increases the bias.

Supplementary Material

The online Supplementary Material contains appendices and additional tables.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 17K00 065.

References

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

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559.

Canavire-Bacarreza, G., Castro Peñarrieta, L. and Ugarte Ontiveros, D. (2021). Outliers in Semi-Parametric estimation of treatment effects. *Econometrics* **9**, 19.

Díaz, I. (2017). Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference* **190**, 39–51.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75**, 259–276.

Fujisawa, H. (2013). Normalized estimating equation for robust parameter estimation. *Electronic Journal of Statistics* **7**, 1587–1606.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis* **99**, 2053–2081.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (2011). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton.

Hoshino, T. (2007). Doubly robust-type estimation for covariate adjustment in latent variable modeling. *Psychometrika* **72**, 535–549.

Huber, P. J. (2004). *Robust Statistics*. John Wiley & Sons.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jones, M., Hjort, N. L., Harris, I. R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88**, 865–873.

Kanamori, T. and Fujisawa, H. (2015). Robust estimation under heavy contamination using unnormalized models. *Biometrika* **102**, 559–572.

Kawashima, T. and Fujisawa, H. (2017). Robust and sparse regression via γ-divergence. *Entropy* **19**, 608.

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

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M. et al. (2021). *robustbase: Basic Robust Statistics, R package* (R package version 0.93.9). Web: `https://robustbase.R-forge.R-project.org/`.

Maronna, R. A., Martin, R. D., Yohai, V. J. and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

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

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

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* **85**, 633–639.

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

Sued, M., Valdora, M. and Yohai, V. (2020). Robust doubly protected estimators for quantiles with missing data. *TEST* **63**, 819–843.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer New York.

Van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Windham, M. P. (1995). Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 599–609.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 642–656.

Zhang, Z., Chen, Z., Troendle, J. F. and Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics* **68**, 697–706.

Kazuharu Harada

Department of Health Data Science, Tokyo Medical University, Shinjuku-ku, Tokyo 160-8402, Japan.

E-mail: haradak@tokyo-med.ac.jp

Hironori Fujisawa

Department of Statistical Inference and Mathematics, The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan.

E-mail: fujisawa@ism.ac.jp

(Received July 2021; accepted April 2022)