Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Robust Inference Using Inverse Probability Weighting

Xinwei Ma & Jingshen Wang

To cite this article: Xinwei Ma & Jingshen Wang (2020) Robust Inference Using Inverse Probability Weighting, Journal of the American Statistical Association, 115:532, 1851-1860, DOI: 10.1080/01621459.2019.1660173

To link to this article: https://doi.org/10.1080/01621459.2019.1660173

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 1851–1860: Theory and Methods
https://doi.org/10.1080/01621459.2019.1660173

Taylor & Francis
Taylor & Francis Group

Robust Inference Using Inverse Probability Weighting

Xinwei Ma[a] and Jingshen Wang[b]

[a]Department of Economics, University of California, San Diego, La Jolla, CA; [b]Division of Biostatistics, University of California, Berkeley, Berkeley, CA

ABSTRACT

Inverse probability weighting (IPW) is widely used in empirical work in economics and other disciplines. As Gaussian approximations perform poorly in the presence of "small denominators," trimming is routinely employed as a regularization strategy. However, ad hoc trimming of the observations renders usual inference procedures invalid for the target estimand, even in large samples. In this article, we first show that the IPW estimator can have different (Gaussian or non-Gaussian) asymptotic distributions, depending on how "close to zero" the probability weights are and on how large the trimming threshold is. As a remedy, we propose an inference procedure that is robust not only to small probability weights entering the IPW estimator but also to a wide range of trimming threshold choices, by adapting to these different asymptotic distributions. This robustness is achieved by employing resampling techniques and by correcting a non-negligible trimming bias. We also propose an easy-to-implement method for choosing the trimming threshold by minimizing an empirical analogue of the asymptotic mean squared error. In addition, we show that our inference procedure remains valid with the use of a data-driven trimming threshold. We illustrate our method by revisiting a dataset from the National Supported Work program. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2018
Accepted August 2019

KEYWORDS
Bias correction; Heavy tail; Inverse probability weighting; Robust inference; Trimming

1. Introduction

Inverse probability weighting (IPW) is widely used in empirical work in economics and other disciplines. In practice, it is common to observe small probability weights entering the IPW estimator. This renders inference based on standard Gaussian approximations invalid, even in large samples, because these approximations rely crucially on the probability weights being well-separated from zero. In a recent study, Busso, DiNardo, and McCrary (2014) investigated the finite sample performance of commonly used IPW treatment effect estimators, and documented that small probability weights can be detrimental to statistical inference. In response to this problem, observations with probability weights below a certain threshold are often excluded from subsequent statistical analysis. The exact amount of trimming, however, is usually ad hoc and will affect the performance of the IPW estimator and the corresponding confidence interval in nontrivial ways.

In this article, we show that the IPW estimator can have different (Gaussian or non-Gaussian) asymptotic distributions, depending on how "close to zero" the probability weights are and on how large the trimming threshold is. We propose an inference procedure that adapts to these different asymptotic distributions, making it robust not only to small probability weights, but also to a wide range of trimming threshold choices. This "two-way robustness" is achieved by combining subsampling with a novel bias correction technique. In addition, we propose an easy-to-implement method for choosing the trimming threshold by minimizing an empirical analogue of the asymptotic mean squared error (MSE), and show that our inference procedure remains valid with the use of a data-driven trimming threshold.

To understand why standard inference procedures are not robust to small probability weights, and why their performance can be sensitive to the amount of trimming, we first study the large-sample properties of the IPW estimator

$$\hat{\theta}_{n,b_n} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{\hat{e}(X_i)} \mathbb{1}_{\hat{e}(X_i)\geq b_n},$$

where $D_i \in \{0,1\}$ is binary, Y_i is the outcome of interest, $e(X_i) = \mathbb{P}[D_i = 1|X_i]$ is the probability weight conditional on the covariates with $\hat{e}(X_i)$ being its estimate, and b_n is the trimming threshold (the untrimmed IPW estimator is a special case with $b_n = 0$). The asymptotic framework we employ is general and allows, but does not require that the probability weights have a heavy tail near zero. Specifically, if the tail is relatively thin, the asymptotic distribution will be Gaussian; otherwise a slower-than-$\sqrt{n}$ convergence rate and a non-Gaussian asymptotic distribution can emerge, and they will depend on the trimming threshold b_n. In the latter case,

$$\frac{n}{a_{n,b_n}}\left(\hat{\theta}_{n,b_n} - \theta_0 - \mathbf{B}_{n,b_n}\right) \xrightarrow{\mathrm{d}} \mathcal{L}(\gamma_0, \alpha_+(\cdot), \alpha_-(\cdot)), \quad (1)$$

where θ_0 is the parameter of interest, $a_{n,b_n} \to \infty$ is a sequence of normalizing factors, and $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution.

CONTACT Xinwei Ma x1ma@ucsd.edu Department of Economics, 9500 Gilman Dr. #0508, La Jolla, CA 92093-0508.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

First, a trimming bias $\mathbf{B}_{n,b_n}$ emerges. This bias has order $\mathbb{P}[e(X) \leq b_n]$, and hence it will vanish asymptotically if the trimming threshold shrinks to zero. What matters for inference, however, is the asymptotic bias, defined as the trimming bias scaled by the convergence rate: $\frac{n}{a_{n,b_n}}\mathbf{B}_{n,b_n}$. This asymptotic bias may not vanish even in large samples, and can be detrimental to statistical inference, as it shifts the asymptotic distribution away from the target estimand. Second, the asymptotic distribution, $\mathcal{L}(\cdot)$, depends on three parameters. The first parameter γ_0 is related to tail behaviors of the probability weights near zero, and the other two parameters characterize shape and tail properties of the asymptotic distribution. In particular, $\mathcal{L}(\cdot)$ does not need to be symmetric. Third, the convergence rate, $n/a_{n,b_n}$, is usually unknown, and depends again on how "close to zero" the probability weights are and how large the trimming threshold is.

As the large-sample properties of the IPW estimator are sensitive to small probability weights and to the amount of trimming, it is important to develop an inference procedure that automatically adapts to the relevant asymptotic distributions. However, the presence of additional nuisance parameters makes it challenging to base inference on estimating the asymptotic distribution. In addition, the standard nonparametric bootstrap is known to fail in our setting (Athreya 1987; Knight 1989). We instead propose the use of subsampling (Politis and Romano 1994). We show that subsampling provides valid approximations to the asymptotic distribution in (1). With self-normalization (i.e., subsampling a Studentized statistic), it also overcomes the difficulty of having a possibly unknown convergence rate. Subsampling alone does not suffice for valid inference due to the bias induced by trimming. To make our inference procedure also robust to a wide range of trimming threshold choices, we combine subsampling with a novel bias correction method based on local polynomial regressions (Fan and Gijbels 1996). To be precise, our method regresses the outcome variable on a polynomial basis of the probability weight in a region local to the origin, and estimates the trimming bias with the regression coefficients.

We also address the question of how to choose the trimming threshold. One extreme possibility is fixed trimming ($b_n = b > 0$). Although fixed trimming helps restore asymptotic Gaussianity by forcing the probability weights to be bounded away from zero, this practice is difficult to justify, unless one is willing to reinterpret the estimation and inference result completely. We instead propose to determine the trimming threshold by taking into account the bias and variance of the (trimmed) IPW estimator. We suggest an easy-to-implement method to choose the trimming threshold by minimizing an empirical analogue of the asymptotic MSE.

This article relates to a large body of literature on program evaluation and causal inference (Imbens and Rubin 2015; Abadie and Cattaneo 2018; Hernán and Robins 2019). Estimators with inverse weighting are widely used in missing data models (Robins, Rotnitzky, and Zhao 1994; Wooldridge 2007) and treatment effect estimation (Hirano, Imbens, and Ridder 2003; Cattaneo 2010). They also feature in settings such as instrumental variables (Abadie 2003), difference-in-differences (Abadie 2005), and counterfactual analysis (DiNardo, Fortin, and Lemieux 1996). Khan and Tamer (2010) showed that, depending on tail behaviors of the probability weights, the variance bound of the IPW estimator can be infinite, which leads to a slower-than-$\sqrt{n}$ convergence rate. Sasaki and Ura (2018) proposed a trimming method and a companion sieve-based bias correction technique for conducting inference for moments of ratios, which complement our article. Chaudhuri and Hill (2016) proposed a different trimming strategy based on $|DY/e(X)|$ rather than the probability weight, and an inference procedure relying on asymptotic Gaussianity. Crump et al. (2009) also studied the problem of trimming threshold selection, the difference is that their method is based on minimizing a variance term, and hence can lead to a much larger trimming threshold than what we propose.

The untrimmed IPW estimator (i.e., $b_n = 0$) is a special case of (1), and the asymptotic distribution is known as the Lévy stable distribution. Stable convergence has been established in many contexts. For example, Vaynman and Beare (2014) showed that stable convergence may arise for the variance targeting estimator, and hence the tail trimming of Hill and Renault (2012) can be crucial for establishing asymptotic Gaussianity. Khan and Nekipelov (2013) also established a stable convergence result for the untrimmed IPW estimator. However, they do not discuss the impact of trimming or how the trimming threshold should be chosen in practice. Hong, Leung, and Li (2018) consider a different setting where observations fall into finitely many strata. They demonstrate that for estimating treatment effects the effective sample size can be much smaller as a result of disproportionately many treated or control units (a.k.a. limited overlap), and relate the rate of convergence to how fast the probability weight approaches an extreme.

With the IPW estimator as a special case, Cattaneo and Jansson (2018) and Cattaneo, Jansson, and Ma (2019) show how an asymptotic bias can arise in a two-step semiparametric setting as a result of overfitting the first step. Chernozhukov et al. (2018) develop robust inference procedures against underfitting bias. The trimming bias we document in this article is both qualitatively and quantitatively different, as it will be present even when the probability weights are observed, and certainly will not disappear with model selection or machine learning methods (Farrell 2015; Athey, Imbens, and Wager 2018; Belloni et al. 2018; Farrell, Liang, and Misra 2018).

The rest of the article is structured as follows. In Section 2, we state and discuss the main assumptions, and study the large-sample properties of the IPW estimator. In Section 3, we discuss in detail our robust inference procedure, including how the bias correction and the subsampling are implemented. A data-driven method to choose the trimming threshold is also proposed. Section 4 showcases our methods with an empirical example. Section 5 concludes. To conserve space, we collect auxiliary lemmas, additional results, simulation evidence, and all proofs in the online supplementary materials. We also discuss in the supplementary materials how our IPW framework can be generalized to provide robust inference for treatment effect estimands and parameters defined by general (nonlinear) estimating equations.

2. Large-Sample Properties

Let (Y_i, D_i, X_i), $i = 1, 2, \ldots, n$ be a random sample from $Y \in \mathbb{R}$, $D \in \{0, 1\}$ and $X \in \mathbb{R}^{d_x}$. Recall that the probability weight is

defined as $e(X) = \mathbb{P}[D = 1|X]$. Define the conditional moments of the outcome variable as

$$\mu_s(e(X)) = \mathbb{E}[Y^s|e(X), D = 1], \quad s > 0,$$

then the parameter of interest is $\theta_0 = \mathbb{E}[DY/e(X)] = \mathbb{E}[\mu_1(e(X))]$. At this level of generality, we do not attach specific interpretations to the parameter and the random variables in our model. To facilitate understanding, one can think of Y as an observed outcome variable and D as an indicator of treatment status, hence the parameter is the population average of one potential outcome.

As previewed in Section 1, large-sample properties of the IPW estimator $\hat{\theta}_{n,b_n}$ depend on the tail behavior of the probability weights near zero: if $e(X)$ is bounded away from zero, the IPW estimator is $\sqrt{n}$-consistent and asymptotically Gaussian; in the presence of small probability weights, however, non-Gaussian distributions can emerge. In this section, we first discuss the assumptions and formalize the notion of "probability weights being close to zero" or "having a heavy tail." Then we characterize the asymptotic distribution of $\hat{\theta}_{n,b_n}$, and show how it is affected by the trimming threshold b_n.

2.1. Tail Behavior

For an estimator that takes the form of a sample average (or more generally can be linearized into such), distributional approximation based on the central limit theorem only requires a finite variance. The problem with IPW with "small denominators," however, is that the estimator may not have a finite variance. In this case, distributional convergence relies on tail features, which we formalize in the following assumption.

Assumption 1 (Regular variation). For some $\gamma_0 > 1$, the probability weights have a regularly varying tail with index $\gamma_0 - 1$ at zero:

$$\lim_{t \downarrow 0} \frac{\mathbb{P}[e(X) \le tx]}{\mathbb{P}[e(X) \le t]} = x^{\gamma_0 - 1}, \quad \text{for all } x > 0.$$

Assumption 1 only imposes a local restriction on the tail behavior of the probability weights, and is common when dealing with sums of heavy-tailed random variables. It is equivalent to $\mathbb{P}[e(X) \le x] = c(x)x^{\gamma_0 - 1}$ with $c(x)$ being a slowly varying function (see the supplementary materials or Feller 1991, chap. XVII for a definition). A special case of Assumption 1 is "approximately polynomial tail," which requires $\lim_{x \downarrow 0} c(x) = c > 0$. To see how the tail index γ_0 features in data, we illustrate in Section 4 with estimated probability weights from an empirical example, and it is clear that the probability weights exhibit a heavy tail near zero. Later in Theorem 1, we show that $\gamma_0 = 2$ is the knife-edge case that separates the Gaussian and the non-Gaussian asymptotic distributions for the (untrimmed) IPW estimator. With $\gamma_0 = 2$, the probability weights are approximately uniformly distributed, a fact that can be used in practice as a rough guidance on the magnitude of this tail index.

Remark 1 (Implied tail of X). To see how the tail behavior of the probability weights is related to that of the covariates X, we consider a Logit model: $e(X) = \exp(X^{\mathrm{T}}\pi_0)/(1 + \exp(X^{\mathrm{T}}\pi_0))$, which implies $\mathbb{P}[e(X) \le x] = \mathbb{P}[X^{\mathrm{T}}\pi_0 \le -\log(x^{-1} - 1)]$. As a result, Assumption 1 is equivalent to that, for all x large enough, $\mathbb{P}[X^{\mathrm{T}}\pi_0 \le -x] \approx e^{-(\gamma_0 - 1)x}$, meaning that the (left) tail of $X^{\mathrm{T}}\pi_0$ is approximately sub-exponential.

Assumption 1 characterizes the tail behavior of the probability weights. However, it alone does not suffice for the IPW estimator to have an asymptotic distribution. The reason is that, for sums of random variables without finite variance to converge in distribution, one needs not only a restriction on the shape of the tail, but also a "tail balance condition." For this purpose, we impose the following assumption.

Assumption 2 (Conditional distribution of Y). (i) For some $\varepsilon > 0$, $\mathbb{E}\left[|Y|^{(\gamma_0 \vee 2)+\varepsilon} \middle| e(X) = x, D = 1\right]$ is uniformly bounded. (ii) There exists a probability distribution F, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y)|e(X) = x, D = 1] \to \int_{\mathbb{R}} \ell(y)F(\mathrm{d}y)$ as $x \downarrow 0$.

This assumption has two parts. The first part requires the tail of Y to be thinner than that of $D/e(X)$, therefore the tail behavior of $DY/e(X)$ is largely driven by the "small denominator $e(X)$." As our primary focus is the implication of small probability weights entering the IPW estimator rather than a heavy-tailed outcome variable, we maintain this assumption. The second part requires convergence of the conditional distribution of Y given $e(X)$ and $D = 1$. Together, they help characterize the tail behavior of $DY/e(X)$.

Lemma 1 (Tail property of $DY/e(X)$). Under Assumptions 1 and 2,

$$\lim_{x \to \infty} \frac{x\mathbb{P}[DY/e(X) > x]}{\mathbb{P}[e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0}\alpha_+(0),$$
$$\lim_{x \to \infty} \frac{x\mathbb{P}[DY/e(X) < -x]}{\mathbb{P}[e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0}\alpha_-(0),$$

where $\alpha_+(x) = \lim_{t \to 0} \mathbb{E}[|Y|^{\gamma_0}\mathbb{1}_{Y>x}|e(X) = t, D = 1]$ and $\alpha_-(x) = \lim_{t \to 0} \mathbb{E}[|Y|^{\gamma_0}\mathbb{1}_{Y<x}|e(X) = t, D = 1]$.

Assuming the distribution of the outcome variable is non-degenerate conditional on the probability weights being small (i.e., $\alpha_+(0) + \alpha_-(0) > 0$), Lemma 1 shows that $DY/e(X)$ has regularly varying tails with index $-\gamma_0$. As a result, γ_0 determines which moment of the IPW estimator is finite: for $s < \gamma_0$, $\mathbb{E}[|DY/e(X)|^s] < \infty$. We compare to a common assumption made in the IPW literature, which requires the probability weights to be bounded away from zero. This assumption is sufficient but not necessary for asymptotic Gaussianity. In fact, the IPW estimator is asymptotically Gaussian as long as $\gamma_0 \ge 2$. Intuitively, small denominators appear so infrequently that they will not affect the large-sample properties. For $\gamma_0 \in (1, 2)$, the IPW estimator no longer has a finite variance, as the distribution of $e(X)$ does not approach zero fast enough (or equivalently, the density of $e(X)$, if it exists, diverges to infinity). This scenario represents the empirical difficulty of dealing with small probability weights entering the IPW estimator, for which regular asymptotic analysis no longer applies.

Thanks to Assumption 2(ii), Lemma 1 also implies that $DY/e(X)$ has balanced tails: the ratio $\frac{\mathbb{P}[DY/e(X)>x]}{\mathbb{P}[|DY/e(X)|>x]}$ tends to

a finite constant. It turns out that without a finite variance, the asymptotic distribution of the IPW estimator is non-Gaussian, and the asymptotic distribution depends on both the left and right tail of $DY/e(X)$. Thus, tail balancing (and Assumption 2(ii)) is indispensable for developing a large-sample theory allowing for small probability weights entering the IPW estimator.

2.2. Asymptotic Distribution

The following theorem characterizes the asymptotic distribution of the IPW estimator, both with and without trimming. To make the result concise, we assume the oracle (rather than estimated) probability weights are used, making the IPW estimator a one-step procedure. We extend the theorem to estimated probability weights in the following subsection. In the supplementary materials, we also discuss how our IPW framework can be generalized to provide robust inference for treatment effect estimands and parameters defined by general (nonlinear) estimating equations.

Theorem 1 (Asymptotic distribution). Assume Assumptions 1 and 2 hold with $\alpha_+(0) + \alpha_-(0) > 0$, $b_n \to 0$, and let a_n be defined such that

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{DY}{e(X)} - \theta_0\right|^2 \mathbb{1}_{|DY/e(X)|\le a_n}\right] \to 1.$$

(i) If $\gamma_0 \ge 2$, (1) holds with $a_{n,b_n} = a_n$, and the asymptotic distribution is standard Gaussian.
(ii.1) No trimming, light trimming, and moderate trimming: if $\gamma_0 < 2$ and $b_n a_n \to t \in [0,\infty)$, (1) holds with $a_{n,b_n} = a_n$, and the asymptotic distribution is infinitely divisible with characteristic function

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x)\right\},$$
$$\text{where } M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0} \times \left(\alpha_+(tx)\mathbb{1}_{x\ge 0} + \alpha_-(tx)\mathbb{1}_{x<0}\right)\right].$$

(ii.2) Heavy trimming: if $\gamma_0 < 2$ and $b_n a_n \to \infty$, (1) holds with $a_{n,b_n} = \sqrt{n\mathbb{V}[DY/e(X)\mathbb{1}_{e(X)\ge b_n}]}$, and the asymptotic distribution is standard Gaussian.

To provide some insight, we first consider the untrimmed IPW estimator ($b_n = 0$), whose large-sample properties are summarized in part (i) and (ii.1). Theorem 1 demonstrates how a non-Gaussian asymptotic distribution can emerge when the untrimmed IPW estimator does not have a finite variance ($\gamma_0 < 2$). The asymptotic distribution in this case is also known as the Lévy stable distribution, which has the following equivalent representation,

$$\psi(\zeta) = -|\zeta|^{\gamma_0}\frac{\Gamma(3-\gamma_0)}{\gamma_0(\gamma_0-1)}\left[-\cos\left(\frac{\gamma_0\pi}{2}\right) + i\frac{\alpha_+(0)-\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\mathrm{sgn}(\zeta)\sin\left(\frac{\gamma_0\pi}{2}\right)\right],$$

where $\Gamma(\cdot)$ is the gamma function and $\mathrm{sgn}(\cdot)$ is the sign function. From this alternative form, we deduce several properties of the asymptotic Lévy stable distribution. First, this distribution is not symmetric unless $\alpha_+(0) = \alpha_-(0)$. Second, the characteristic function has a sub-exponential tail, meaning that the limiting Lévy stable distribution has a smooth density function (although in general it does not have a closed-form expression). Finally, the above characteristic function is continuous in γ_0, in the sense that as $\gamma_0 \uparrow 2$, and it reduces to the standard Gaussian characteristic function.

Theorem 1 also shows how the convergence rate of the untrimmed IPW estimator depends on the tail index γ_0. For $\gamma_0 > 2$, the IPW estimator converges at the usual parametric rate $n/a_{n,b_n} \asymp \sqrt{n}$. This extends to the $\gamma_0 = 2$ case, except that an additional slowly varying factor is present in the convergence rate. For $\gamma_0 < 2$, the convergence rate is only implicitly defined from a truncated second moment, and generally does not have an explicit formula. One can consider the special case that the probability weights have an approximately polynomial tail: $\mathbb{P}[e(X) \le x] \asymp x^{\gamma_0 - 1}$, for which a_{n,b_n} can be set to n^{1/γ_0}. As a result, the untrimmed IPW estimator will have a slower convergence rate if the probability weights have a heavier tail at zero (i.e., smaller γ_0). Fortunately, the (unknown) convergence rate is captured by self-normalization (Studentization), which we employ in our robust inference procedure.

Now we discuss the impact of trimming, a strategy commonly employed in practice in response to small probability weights entering the IPW estimator. We distinguish among three cases: light trimming ($b_n a_n \to 0$), moderate trimming ($b_n a_n \to t \in (0,\infty)$), and heavy trimming ($b_n a_n \to \infty$). For light trimming, b_n shrinks to zero fast enough so that asymptotically trimming becomes negligible, and the asymptotic distribution is Lévy stable as if there were no trimming. For heavy trimming, the trimming threshold shrinks to zero slowly, hence most of the small probability weights are excluded. This leads to a Gaussian asymptotic distribution. The moderate trimming scenario lies between the two extremes. On the one hand, a nontrivial number of small probability weights are discarded, making the limit no longer the Lévy stable distribution. On the other hand, the trimming is not heavy enough to restore asymptotic Gaussianity. The asymptotic distribution in this case is quite complicated, and depends on two (infinitely dimensional) nuisance parameters, $\alpha_+(\cdot)$ and $\alpha_-(\cdot)$. For this reason, inference is quite challenging.

Despite the asymptotic distribution taking on a complicated form, moderate trimming as in Theorem 1(ii.1) is highly relevant. In Section 3.1, we discuss how the trimming threshold can be chosen to balance the bias and variance (i.e., to minimize the MSE), which corresponds to this moderate trimming scenario. In addition, unless one employs a very large trimming threshold, it is unclear how well the Gaussian approximation performs in samples of moderate size.

2.3. Estimated Probability Weights

The probability weights are usually unknown and are estimated in a first step, which are then plugged into the IPW estimator, making it a two-step estimation problem. Estimating the

probability weights in a first step can affect large-sample properties of the IPW estimator through two channels: the estimated weights enter the final estimator both through inverse weighting and through the trimming function. In this subsection, we first impose high-level assumptions and discuss the impact of employing estimated probability weights. Then we verify those high-level assumptions for Logit and Probit models, which are widely used in applied work.

Assumption 3 (First-step estimation). The probability weights are parametrized as $e(X,\pi)$ with $\pi \in \Pi$, and $e(\cdot)$ is continuously differentiable with respect to π. Let $e(X) = e(X,\pi_0)$ and $\hat{e}(X) = e(X,\hat{\pi}_n)$. Further, (i) $\sqrt{n}(\hat{\pi}_n - \pi_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h(D_i,X_i) + o_{\mathrm{p}}(1)$, where $h(D_i,X_i)$ is mean zero and has a finite variance; and (ii) for some $\varepsilon > 0$, $\mathbb{E}\left[\sup_{\pi:|\pi-\pi_0|\le\varepsilon}\left|\frac{e(X_i)}{e(X_i,\pi)^2}\frac{\partial e(X_i,\pi)}{\partial\pi}\right|\right] < \infty$.

Assumption 4 (Trimming threshold). The trimming threshold satisfies $c_n\sqrt{b_n\mathbb{P}[e(X_i)\le b_n]} \to 0$, where c_n is a positive sequence such that, for any $\varepsilon > 0$,

$$c_n^{-1}\max_{1\le i\le n}\sup_{|\pi-\pi_0|\le\varepsilon/\sqrt{n}}\left|\frac{1}{e(X_i)}\frac{\partial e(X_i,\pi)}{\partial\pi}\right| = o_{\mathrm{p}}(1).$$

Now we state the analogue of Theorem 1 but with the probability weights estimated in a first step.

Proposition 1 (Asymptotic distribution with estimated probability weights). Assume Assumptions 1–4 hold with $\alpha_+(0)+\alpha_-(0) > 0$. Let a_n be defined such that

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{DY}{e(X)} - \theta_0 - A_0 h(D,X)\right|^2 \mathbb{1}_{|DY/e(X)-A_0h(D,X)|\le a_n}\right] \to 1,$$

where $A_0 = \mathbb{E}\left[\frac{\mu_1(e(X))}{e(X)}\frac{\partial e(X,\pi)}{\partial\pi}\Big|_{\pi=\pi_0}\right]$. Then the IPW estimator has the following linear representation:

$$\frac{n}{a_{n,b_n}}\left(\hat{\theta}_{n,b_n} - \theta_0 - \mathbf{B}_{n,b_n}\right) = \frac{1}{a_{n,b_n}}\sum_{i=1}^n\left(\frac{D_iY_i}{e(X_i)}\mathbb{1}_{e(X_i)\ge b_n} - \theta_0 - \mathbf{B}_{n,b_n} - A_0h(D_i,X_i)\right) + o_{\mathrm{p}}(1),$$

and the conclusions of Theorem 1 hold with estimated probability weights.

To understand Proposition 1, we consider two cases. In the first case, $\mathbb{V}[DY/e(X)] < \infty$, and estimating the probability weights in a first step will contribute to the asymptotic variance. The second case corresponds to $\mathbb{V}[DY/e(X)] = \infty$, implying that the final estimator, $\hat{\theta}_{n,b_n}$, has a slower convergence rate compared to the first-step estimated probability weights. As a result, the two definitions of the scaling factor a_n (in Theorem 1 and in Proposition 1) are asymptotically equivalent, and the asymptotic distribution will be the same regardless of whether the probability weights are known or estimated. In addition, Proposition 1 shows that, despite the estimated probability weights entering both the denominator and the trimming function, the second channel is asymptotically negligible under an additional assumption, which turns out to be very mild in applications.

Assumption 3(i) is standard. In the following remark we provide primitive conditions to justify Assumption 3(ii) and Assumption 4 in Logit and Probit models.

Remark 2 (Logit and Probit models). Assuming a Logit model for the probability weights, we show in the supplementary materials that a sufficient condition for Assumption 3(ii) is the covariates having a sub-exponential tail: $\mathbb{E}[e^{\varepsilon|X|}] < \infty$ for some (small) $\varepsilon > 0$. This condition is fully compatible with Assumption 1, as we show in Remark 1 that for Assumption 1 to hold in a Logit model, the index $X^{\mathrm{T}}\pi_0$ needs to have a sub-exponential left tail. As for the Probit model, Assumption 3(ii) is implied by a sub-Gaussian tail of the covariates: $\mathbb{E}[e^{\varepsilon|X|^2}] < \infty$ for some (small) $\varepsilon > 0$. Again, it is possible to show that Assumption 1 implies a sub-Gaussian left tail for the index $X^{\mathrm{T}}\pi_0$.

To verify Assumption 4, it suffices to set $c_n = \log^2(n)$ for Logit and Probit models. Therefore, we only require the trimming threshold shrinking to zero faster than a logarithmic rate. See the supplementary materials for details.

3. Robust Inference

In the previous section, we show that non-Gaussian asymptotic distributions may arise as a result of small probability weights entering the IPW estimator and trimming. In this section, we first study the trimming bias, and show that this bias is usually nonnegligible for inference purpose. Together, these findings explain why the point estimate is sensitive to the choice of the trimming threshold, and more importantly, why conventional inference procedures based on the standard Gaussian approximation perform poorly.

As a remedy, we first introduce a method to choose the trimming threshold by minimizing an empirical MSE, and discuss how our trimming threshold selector can be modified in a disciplined way if the researcher prefers to discard more observations. Then we propose to combine subsampling with a novel bias correction technique, where the latter employs local polynomial regression to approximate the trimming bias.

3.1. Bias, Variance, and Trimming Threshold Selection

If the sole purpose of trimming is to stabilize the IPW estimator, one can argue that only a fixed trimming rule, $b_n = b \in (0,1)$, should be used. Such practice, however, completely ignores the bias introduced by trimming, and forces the researcher to change the target estimand and reinterpret the estimation/inference result. Practically, the trimming threshold can be chosen by minimizing the asymptotic MSE. We first characterize the bias and variance of the (trimmed) IPW estimator.

Lemma 2 (Bias and variance of $\hat{\theta}_{n,b_n}$). Assume Assumptions 1 and 2 hold with $\gamma_0 < 2$. Further, assume that $\mu_1(\cdot)$ and $\mu_2(\cdot)$ do not vanish near 0. Then the bias and variance of $\hat{\theta}_{n,b_n}$ are

$$\mathbf{B}_{n,b_n} = -\mu_1(0)\mathbb{P}\left[e(X)\le b_n\right](1+o(1)),$$
$$\mathbf{V}_{n,b_n} = \mu_2(0)\frac{1}{n}\mathbb{E}\left[e(X)^{-1}\mathbb{1}_{e(X)\ge b_n}\right](1+o(1)).$$

In addition, $\mathbf{B}_{n,b_n}^2/\mathbf{V}_{n,b_n} \asymp nb_n\mathbb{P}[e(X)\le b_n]$.

To balance the bias and variance, minimizing the leading MSE with respect to the trimming threshold leads to

$$b_n^\dagger \cdot \mathbb{P}[e(X) \le b_n^\dagger] = \frac{1}{2n}\frac{\mu_2(0)}{\mu_1(0)^2}.$$

The MSE-optimal trimming $b_n^\dagger$ helps understand the three scenarios in Theorem 1: light, moderate, and heavy trimming. More importantly, it helps clarify whether (and when) the trimming bias features in the asymptotic distribution. (The trimming bias $\mathbf{B}_{n,b_n}$ vanishes as long as $b_n \to 0$. What matters for inference, however, is the asymptotic bias, which is defined as the trimming bias scaled by the convergence rate: $\frac{n}{a_{n,b_n}}\mathbf{B}_{n,b_n}$. This asymptotic bias may not be negligible even in large samples.) $b_n^\dagger$ corresponds to the moderate trimming scenario, and since it balances the leading bias and variance, the asymptotic distribution of the trimmed IPW estimator is not centered at the target estimand (i.e., it is asymptotically biased). A trimming threshold that shrinks more slowly than the optimal one corresponds to the heavy trimming scenario, where the bias dominates in the asymptotic distribution. The only scenario in which one can ignore the trimming bias for inference purposes is when light trimming is used. That is, the trimming threshold shrinks faster than $b_n^\dagger$.

The following theorem shows that, under very mild regularity conditions, the MSE-optimal trimming threshold can be implemented in practice by solving a sample analogue. It also provides a disciplined method for choosing the trimming threshold if the researcher prefers to employ a heavy trimming.

Theorem 2 (Trimming threshold selection). Assume Assumption 1 holds, and $0 < \mu_2(0)/\mu_1(0)^2 < \infty$. For any $s > 0$, define b_n and $\hat{b}_n$ as

$$b_n^s \mathbb{P}[e(X) \le b_n] = \frac{1}{2n}\frac{\mu_2(0)}{\mu_1(0)^2},$$
$$\hat{b}_n^s \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{e(X)\le \hat{b}_n}\right) = \frac{1}{2n}\frac{\hat{\mu}_2(0)}{\hat{\mu}_1(0)^2},$$

where $\hat{\mu}_1(0)$ and $\hat{\mu}_2(0)$ are some consistent estimates of $\mu_1(0)$ and $\mu_2(0)$, respectively. Then $\hat{b}_n$ is consistent for b_n: $\hat{b}_n/b_n \xrightarrow{\mathrm{p}} 1$. Therefore, for $0 < s < 1$, $s = 1$ and $s > 1$, $\hat{b}_n/b_n^\dagger$ converges in probability to 0, 1 and ∞, respectively.

In addition to Assumption 3, if we have for any $\varepsilon > 0$,

$$\max_{1\le i\le n} \sup_{|\pi-\pi_0|\le \varepsilon/\sqrt{n}} \left|\frac{1}{e(X_i)}\frac{\partial e(X_i,\pi)}{\partial \pi}\right| = o_{\mathrm{p}}\left(\sqrt{\frac{n}{\log(n)}}\right),$$

then $\hat{b}_n$ can be constructed with estimated probability weights.

This theorem states that, as long as one can construct a consistent estimator for the ratio $\mu_2(0)/\mu_1(0)^2$, the optimal trimming threshold can be implemented in practice with the unknown distribution $\mathbb{P}[e(X) \le x]$ replaced by the standard empirical distribution function. In addition, Theorem 2 allows the use of estimated probability weights to construct $\hat{b}_n$. The extra condition turns out to be quite weak, and is easily satisfied if the probability weights are estimated in a Logit or Probit model. (See Remark 2 and the supplementary materials for further discussion.)

In the following, we show that distributional convergence of the IPW estimator is unaffected by the use of data-driven trimming threshold. Establishing distributional convergence results with data-driven tuning parameters tends to be quite difficult in general. In our setting, however, incorporating the estimated trimming threshold does not require additional (strong) assumptions, as it is possible to exploit the specific structure that the trimming threshold enters only through an indicator function.

Proposition 2 (Asymptotic distribution with data-driven trimming threshold). Assume the assumptions of Theorem 2 hold. Then Theorem 1 and Proposition 1 hold with the data-driven trimming threshold $b_n = \hat{b}_n$.

3.2. Bias Correction

To motivate our bias correction technique, recall that the bias is $\mathbf{B}_{n,b_n} = -\mathbb{E}[\mu_1(e(X))\mathbb{1}_{e(X)\le b_n}]$, where $\mu_1(\cdot)$ is the expectation of the outcome Y conditional on the probability weight and $D = 1$. Next, we replace the expectation by a sample average, and the unknown conditional expectation by a pth order polynomial expansion, which is then estimated by local polynomial regressions (Fan and Gijbels 1996). To be precise, one first implements a pth order local polynomial regression of the outcome variable on the probability weight using the $D = 1$ subsample in a region $[0, h_n]$, where $(h_n)_{n\ge1}$ is a bandwidth sequence. The estimated bias is then constructed by replacing the unknown conditional expectation function and its derivatives by the first-step estimates. Following is the detailed algorithm, which is also illustrated in Figure 1.

Algorithm 1 (Bias estimation).

Step 1. With the $D = 1$ subsample, regress the outcome variable Y_i on the (estimated) probability weight in a region $[0, h_n]$:

$$\left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right]' = \underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^n D_i\left[Y_i - \sum_{j=0}^p \beta_j \hat{e}(X_i)^j\right]^2 \mathbb{1}_{\hat{e}(X_i)\le h_n}.$$

Step 2. Construct the bias correction term as

$$\hat{\mathbf{B}}_{n,b_n} = -\frac{1}{n}\sum_{i=1}^n \left(\sum_{j=0}^p \hat{\beta}_j \hat{e}(X_i)^j\right)\mathbb{1}_{\hat{e}(X_i)\le b_n},$$

so that the bias-corrected estimator is $\hat{\theta}^{\mathrm{bc}}_{n,b_n} = \hat{\theta}_{n,b_n} - \hat{\mathbf{B}}_{n,b_n}$.

By inspecting the bias-corrected estimator, our procedure can be understood as a "local regression adjustment," since we replace the trimmed observations by their conditional expectation, which is further approximated by a local polynomial. In the local polynomial regression step, it is possible to incorporate other kernel functions: we use the uniform kernel to avoid introducing additional notation, but all the main conclusions continue to hold with other commonly employed kernel functions. As for the order of local polynomial regression, common choices are $p = 1$ and 2, which reduce the bias to a satisfactory level without introducing too much additional variation.

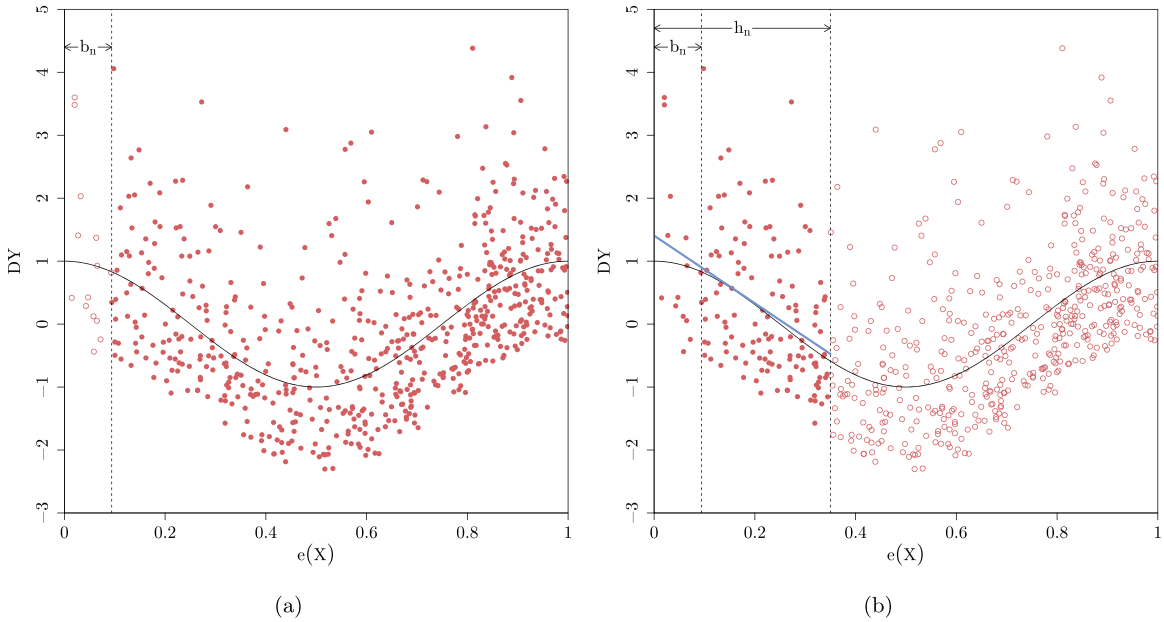

Figure 1. Trimming and local polynomial bias estimation. (a) Illustration of trimming. Circles: trimmed observations. Solid dots: observations included in the estimator. Solid curve: conditional expectation function $\mathbb{E}[Y|e(X), D = 1]$. (b) Illustration of the local polynomial regression. Solid dots: observations used in the local polynomial regression. Solid straight line: local linear regression function.

Standard results from the local polynomial regression literature require the density of the design variable to be bounded away from zero, which is not satisfied in our context. In the $D = 1$ subsample which we use for the local polynomial regression, the distribution of the probability weights quickly vanishes near the origin. (More precisely, $\mathbb{P}[e(X) \leq x|D = 1] \prec x$ as $x \downarrow 0$, meaning that the conditional density of the probability weights (if it exists) tends to zero: $f_{e(X)|D=1}(0) = 0$.) As a result, nonstandard scaling is needed to derive large-sample properties of $\hat{\mu}_1^{(j)}(0)$. See the supplementary materials for a precise statement.

Theorem 3 (Large-sample properties of the estimated bias). Assume Assumptions 1 and 2 (and in addition Assumptions 3 and 4 with estimated probability weights) hold. Further, assume (i) $\mu_1(\cdot)$ is $p + 1$ times continuously differentiable; (ii) $\mu_2(0) - \mu_1(0)^2 > 0$; (iii) the bandwidth sequence satisfies $nh_n^{2p+3}\mathbb{P}[e(X) \leq h_n] \asymp 1$; (iv) $nb_n^{2p+3}\mathbb{P}[e(X) \leq b_n] \to 0$. Then the bias correction is valid, and does not affect the asymptotic distribution: $\hat{\theta}^{\mathbf{bc}}_{n,b_n} - \theta_0 = (\hat{\theta}_{n,b_n} - \mathbf{B}_{n,b_n} - \theta_0)(1 + o_{\mathrm{P}}(1))$.

Theorem 3 has several important implications. First, our bias correction is valid for a wide range of trimming threshold choices, as long as the trimming threshold does not shrink to zero too slowly: $nb_n^{2p+3}\mathbb{P}[e(X) \leq b_n] \to 0$. However, fixed trimming $b_n = b \in (0, 1)$ is ruled out. This is not surprising, since with fixed trimming the correct scaling is $\sqrt{n}$, and generally the bias cannot be estimated at this rate without additional parametric assumptions.

Second, it gives a guidance on how the bandwidth for the local polynomial regression can be chosen. In practice, this is done by solving $n\hat{h}_n^{2p+3}\hat{\mathbb{P}}[e(X) \leq \hat{h}_n] = c$ for some $c > 0$, so that the resulting bandwidth makes the (squared) bias and variance of the local polynomial regression the same order. A simple strategy is to set $c = 1$. It is also possible to construct a bandwidth that minimizes the leading MSE of the local polynomial regression, for which c has to be estimated in a pilot step (see the supplementary materials for a characterization of the leading bias and variance).

Third, it shows how trimming and bias correction together can help improve the convergence rate of the (untrimmed) IPW estimator. From Theorem 1(ii), we have $|\hat{\theta}_{n,b_n} - \theta_0 - \mathbf{B}_{n,b_n}| = O_{\mathrm{P}}((n/a_{n,b_n})^{-1})$, where the convergence rate $n/a_{n,b_n}$ is typically faster when a heavier trimming is employed. This, however, should not be interpreted as a real improvement, as the trimming bias can be quite large. With bias correction, it is possible to achieve a faster rate of convergence for the target estimand, since under the assumptions of Theorem 3, one has $|\hat{\theta}^{\mathbf{bc}}_{n,b_n} - \theta_0| = O_{\mathrm{P}}((n/a_{n,b_n})^{-1})$, which is valid for a wide range of trimming threshold choices.

Finally, we note that when Assumption 1 is violated, bias correction may not be feasible. For example, if in some region of the covariate distribution there are lots of observations from one group but not the other, there will be a spike very close to zero (or at zero) in the probability weights distribution. As a result, bias correction in either case requires extrapolating a local polynomial regression, which can be unreliable.

3.3. Robust Inference

The asymptotic distribution of the IPW estimator can be quite complicated and depend on multiple nuisance parameters which are usually difficult to estimate. We propose the use of subsampling, which is a powerful data-driven method for distributional approximation. It draws samples of size $m \ll n$ and recomputes the statistic with each subsample. Together with our bias correction technique, subsampling can be employed to conduct statistical inference and to construct confidence intervals that are valid for the target estimand. Although Theorem 3 states that estimating the bias does not have a first order

contribution to the asymptotic distribution, it may still introduce additional variability in finite samples (Calonico, Cattaneo, and Farrell 2018). Therefore, we recommend subsampling the bias-corrected statistic.

Algorithm 2 (Robust inference). Let $\hat{\theta}^{\mathbf{bc}}_{n,b_n}$ be defined as in Algorithm 1, and

$$T_{n,b_n} = \frac{\hat{\theta}^{\mathsf{bc}}_{n,b_n} - \theta_0}{S_{n,b_n}/\sqrt{n}},$$
$$S_{n,b_n} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\hat{e}(X_i)}\mathbb{1}_{\hat{e}(X_i)\geq b_n} - \hat{\theta}_{n,b_n}\right)^2}.$$

Step 1. Draw $m \ll n$ observations from the original data without replacement, denoted by $(Y_i^\star, D_i^\star, X_i^\star)$, $i = 1, 2, \ldots, m$.
Step 2. Construct the trimmed IPW estimator and the bias correction term from the new subsample, and write the bias-corrected and self-normalized statistic as

$$T^\star_{m,b_m} = \frac{\hat{\theta}^{\star\mathbf{bc}}_{m,b_m} - \hat{\theta}^{\mathbf{bc}}_{n,b_n}}{S^\star_{m,b_m}/\sqrt{m}},$$
$$S^\star_{m,b_m} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{D_i^\star Y_i^\star}{\hat{e}^\star(X_i^\star)}\mathbb{1}_{\hat{e}^\star(X_i^\star)\geq b_m} - \hat{\theta}^\star_{m,b_m}\right)^2}.$$

Step 3. Repeat Steps 1 and 2, and a $(1-\alpha)$%-confidence interval can be constructed as

$$\left[\hat{\theta}^{\mathbf{bc}}_{n,b_n} - q_{1-\frac{\alpha}{2}}(T^\star_{m,b_m})\frac{S_{n,b_n}}{\sqrt{n}},\quad \hat{\theta}^{\mathbf{bc}}_{n,b_n} - q_{\frac{\alpha}{2}}(T^\star_{m,b_m})\frac{S_{n,b_n}}{\sqrt{n}}\right],$$

where $q_{(\cdot)}(T^\star_{m,b_m})$ denotes the quantile of the statistic $T^\star_{m,b_m}$.

Subsampling validity typically relies on the existence of an asymptotic distribution (Politis and Romano 1994; Romano and Wolf 1999). We follow this approach and justify our robust inference procedure by showing that the self-normalized statistic T_{n,b_n} converges in distribution. Under $\gamma_0 > 2$, S_n converges in probability and T_{n,b_n} converges to a Gaussian distribution. Asymptotic Gaussianity of T_{n,b_n} continues to hold for $\gamma_0 = 2$. Under $\gamma_0 < 2$, T_{n,b_n} still converges in distribution, although the limit will depend on the trimming threshold. Establishing the asymptotic distribution in the heavy trimming case is relatively easy (Lindeberg–Feller central limit theorem). With light or moderate trimming, however, the asymptotic distribution of T_{n,b_n} is quite complicated. This technical by-product generalizes Logan et al. (1973). (To be precise, with light trimming, we obtain the same distribution as in Logan et al. (1973), while the asymptotic distribution of T_{n,b_n} with moderate trimming is new.) We leave the details to the supplementary materials.

Theorem 4 (Validity of robust inference). Assume the assumptions of Theorem 1 (or Proposition 1 with estimated probability weights) and Theorem 3 hold, $m \to \infty$, and $m/n \to 0$. Then $\sup_{t\in\mathbb{R}} |\mathbb{P}[T_{n,b_n} \leq t] - \mathbb{P}^\star[T^\star_{m,b_m} \leq t]| \xrightarrow{\mathrm{p}} 0$.

Before closing this section, we address two practical issues when applying the robust inference procedure. First, it is desirable to have an automatic and adaptive procedure to capture the possibly unknown convergence rate $n/a_{n,b_n}$, as the convergence rate depends on the tail index γ_0. In the subsampling algorithm, this is achieved by self-normalization (Studentization). Second, one has to choose the subsample size m. Some suggestions have been made in the literature: Arcones and Giné (1991) suggested to use $m = \lfloor n/\log\log(n)^{1+\varepsilon}\rfloor$ for some $\varepsilon > 0$, although they consider the m-out-of-n bootstrap. Romano and Wolf (1999) proposed a calibration technique. We use $m = \lfloor n/\log(n)\rfloor$ in our simulation study in the supplementary materials, which performs reasonably well.

4. Empirical Illustration

In this section, we revisit a dataset from the National Supported Work (NSW) program. Our aim is not to discuss to what extent experimental estimates can be recovered by non-experimental methods. Rather, we use this dataset to showcase our robust inference procedure. The NSW program is a labor training program implemented in 1970s by providing work experience to selected individuals. It has been analyzed in multiple studies since LaLonde (1986). We use the same dataset employed in Dehejia and Wahba (1999). Our sample consists of the treated individuals in the NSW experimental group ($D = 1$, sample size $n_1 = 185$), and a nonexperimental comparison group from the Panel Study of Income Dynamics (PSID, $D = 0$, sample size $n_0 = 1157$). The outcome variable Y is the post-intervention earning measured in 1978. The covariates X include information on age, education, marital status, ethnicity, and earnings in 1974 and 1975. We refer interested readers to Dehejia and Wahba (1999, 2002) and Smith and Todd (2005) for more details on variable definition and sample inclusion. We follow the literature and focus on the treatment effect on the treated (ATT),

$$\hat{\tau}^{\mathrm{ATT}}_{n,b_n} = \frac{1}{n_1}\sum_{i=1}^{n}\left[D_i Y_i - \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(1-D_i)Y_i\mathbb{1}_{1-\hat{e}(X_i)\geq b_n}\right],$$

which requires weighting observations from the comparison group by $\hat{e}(X)/(1-\hat{e}(X))$. As a result, probability weights that are close to 1 can pose a challenge to both estimation and inference. (We discuss in the supplementary materials how our IPW framework can be generalized to provide robust inference for treatment effect estimands.)

The probability weight is estimated in a Logit model with `age`, `education`, `earn1974`, `earn1975`, `age`^{2}, `education`^{2}, `earn1974`^{2}, `earn1975`^{2}, three indicators for `married`, `black` and `hispanic`, and an interaction term `black` × `u74`, where `u74` is the unemployment status in 1974. Figure 2 plots the distribution of the estimated probability weights, which clearly exhibits a heavy tail near 1. Since $\gamma_0 = 2$ roughly corresponds to uniformly distributed probability weights, the tail index in this dataset should be well below 2, suggesting that standard inference procedures based on the Gaussian approximation may not perform well.

In Figure 3, we plot the bias-corrected ATT estimates (solid triangles) and the robust 95% confidence intervals (solid vertical lines) with different trimming thresholds. For comparison, we also show conventional point estimates and confidence intervals (solid dots and dashed vertical lines, based on the Gaussian approximation) using the same trimming thresholds. Without trimming, the point estimate is $1451 with a confidence interval

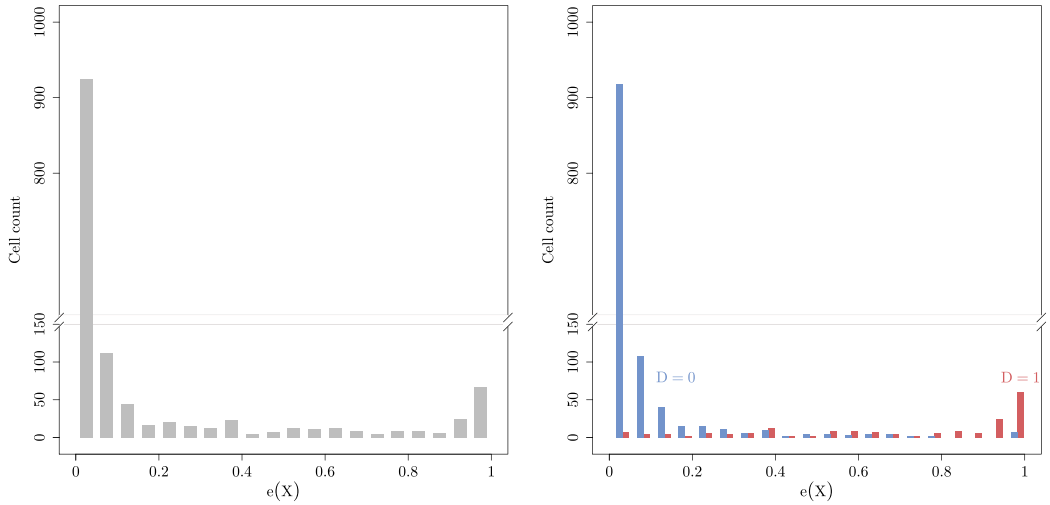

Figure 2. Histogram of the estimated probability weights.

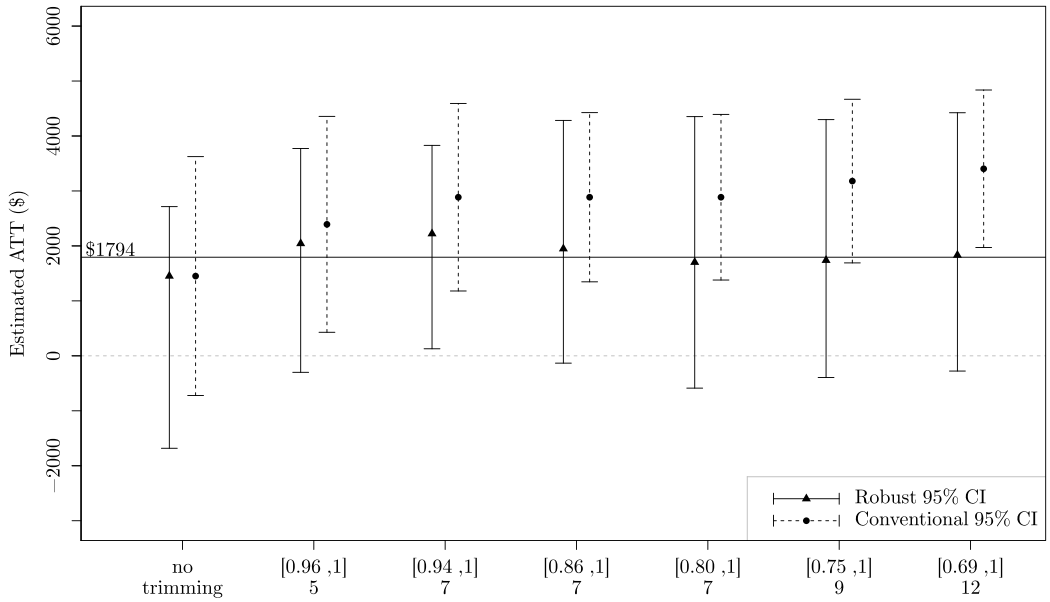

Figure 3. Treatment effect on the treated estimation and inference. Numbers below the horizontal axis show the trimming threshold/region and the effective number of observations trimmed from the comparison group. The experimental benchmark ($1794) is indicated by the solid horizontal line.

[−1763, 2739]. The robust confidence interval is asymmetric around the point estimate, which is a feature also predicted by our theory. For the trimmed IPW estimator, the trimming thresholds are chosen following Theorem 2, and the region used for local polynomial bias estimation is [0.71, 1], corresponding to a bandwidth $h_n = 0.29$. Under the MSE optimal trimming, units in the comparison group with probability weights above 0.96 (five observations) are discarded. Compared to the untrimmed case, the robust confidence interval becomes more symmetric.

In this empirical example, a noteworthy feature of our method is that both the bias-corrected point estimates and the robust confidence intervals remain quite stable for a range of trimming threshold choices, and the point estimates are very close to the experimental benchmark ($1794). This is in stark contrast to conventional confidence intervals that rely on Gaussian approximation. First, conventional confidence intervals fail to adapt to the non-Gaussian asymptotic distributions we documented in Theorem 1, and are overly optimistic/narrow. Second, by ignoring the trimming bias, they are only valid for a pseudo-true parameter implicitly defined by the trimming threshold. As a result, the researcher changes the target estimand each time a different trimming threshold is used, making conventional confidence intervals very sensitive to b_n.

5. Conclusion

We study the large-sample properties of the IPW estimator. We show that, in the presence of small probability weights, this estimator may have a slower-than-$\sqrt{n}$ convergence rate and a non-Gaussian asymptotic distribution. We also study the effect of discarding observations with small probability weights, and show that such trimming not only complicates the asymptotic distribution, but also causes a nonnegligible bias. We propose an inference procedure that is robust not only to small probability weights entering the IPW estimator but also to a range of trimming threshold choices. The "two-way robustness" is achieved by combining resampling with a novel local polynomial-based

bias-correction technique. We also propose a method to choose the trimming threshold, and show that our inference procedure remains valid with the use of a data-driven trimming threshold.

Supplementary Materials

The supplementary materials contain auxiliary lemmas, additional results, simulation evidence, and all proofs. It also discusses how the IPW framework in the main article can be generalized to provide robust inference for treatment effect estimands and parameters defined by general (nonlinear) estimating equations.

Acknowledgments

The authors are deeply grateful to Matias Cattaneo for the comments and suggestions that significantly improved the manuscript. The authors also thank Sebastian Calonico, Max Farrell, Yingjie Feng, Andreas Hagemann, Xuming He, Michael Jansson, Lutz Kilian, Jose Luis Montiel Olea, Kenichi Nagasawa, Rocío Titiunik, Gonzalo Vazquez-Bare, the editor, an associate editor, and two referees for their valuable feedback and thoughtful discussions.

ORCID

Xinwei Ma https://orcid.org/0000-0001-8827-9146
Jingshen Wang https://orcid.org/0000-0001-9498-1185

References

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [1852]

——— (2005), "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72, 1–19. [1852]

Abadie, A., and Cattaneo, M. D. (2018), "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503. [1852]

Arcones, M. A., and Giné, E. (1991), "Additions and Correction to 'The Bootstrap of the Mean With Arbitrary Bootstrap Sample Size'," *Annals of the Institute Henri Poincaré*, 27, 583–595. [1858]

Athey, S., Imbens, G. W., and Wager, S. (2018), "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions," *Journal of the Royal Statistical Society*, Series B, 80, 597–623. [1852]

Athreya, K. B. (1987), "Bootstrap of the Mean in the Infinite Variance Case," *Annals of Statistics*, 15, 724–731. [1852]

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018), "High-Dimensional Econometrics and Regularized GMM," arXiv no. 1806.01888. [1852]

Busso, M., DiNardo, J., and McCrary, J. (2014), "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *Review of Economics and Statistics*, 96, 885–897. [1851]

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018), "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113, 767–779. [1858]

Cattaneo, M. D. (2010), "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability," *Journal of Econometrics*, 155, 138–154. [1852]

Cattaneo, M. D., and Jansson, M. (2018), "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," *Econometrica*, 86, 955–995. [1852]

Cattaneo, M. D., Jansson, M., and Ma, X. (2019), "Two-Step Estimation and Inference With Possibly Many Included Covariates," *Review of Economic Studies*, 86, 1095–1122. [1852]

Chaudhuri, S., and Hill, J. B. (2016), "Heavy Tail Robust Estimation and Inference for Average Treatment Effects," Working Paper. [1852]

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2018), "Locally Robust Semiparametric Estimation," arXiv no. 1608.00033. [1852]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing With Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [1852]

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluations of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [1858]

——— (2002), "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics*, 84, 151–161. [1858]

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996), "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044. [1852]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, New York: Chapman and Hall. [1852,1856]

Farrell, M. H. (2015), "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," *Journal of Econometrics*, 189, 1–23. [1852]

Farrell, M. H., Liang, T., and Misra, S. (2018), "Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands," arXiv no. 1809.09953. [1852]

Feller, W. (1991), *An Introduction to Probability Theory and Its Applications* (Vol. II, 2nd ed.), New York: Wiley. [1853]

Hernán, M. A., and Robins, J. M. (2019), *Causal Inference*, Boca Raton, FL: Chapman and Hall. [1852]

Hill, J. B., and Renault, E. (2012), "Variance Targeting for Heavy Tailed Time Series," Working Paper. [1852]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [1852]

Hong, H., Leung, M., and Li, J. (2018), "Inference on Finite Population Treatment Effects Under Limited Overlap," ssrn no. 3128546. [1852]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, New York: Cambridge University Press. [1852]

Khan, S., and Nekipelov, D. (2013), "On Uniform Inference in Nonlinear Models With Endogeneity," ssrn no. 2331552. [1852]

Khan, S., and Tamer, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042. [1852]

Knight, K. (1989), "On the Bootstrap of the Sample Mean in the Infinite Variance Case," *Annals of Statistics*, 17, 1168–1175. [1852]

LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, 76, 604–620. [1858]

Logan, B. F., Mallows, C. L., Rice, S. O., and Shepp, L. A. (1973), "Limit Distributions of Self-Normalized Sums," *Annals of Probability*, 1, 788–809. [1858]

Politis, D. N., and Romano, J. P. (1994), "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions," *Annals of Statistics*, 22, 2031–2050. [1852,1858]

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. [1852]

Romano, J. P., and Wolf, M. (1999), "Subsampling Inference for the Mean in the Heavy-Tailed Case," *Metrika*, 50, 55–69. [1858]

Sasaki, Y., and Ura, T. (2018), "Estimation and Inference for Moments of Ratios With Robustness Against Large Trimming Bias," arXiv no. 1709.00981. [1852]

Smith, J. A., and Todd, P. E. (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125, 305–353. [1858]

Vaynman, I., and Beare, B. K. (2014), "Stable Limit Theory for the Variance Targeting Estimator," in *Advances in Econometrics: Essays in Honor of Peter CB Phillips* (Vol. 33), eds. Y. Chang, T. B. Fomby, and J. Y. Park, Bingley: Emerald Group Publishing Limited, pp. 639–672. [1852]

Wooldridge, J. M. (2007), "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141, 1281–1301. [1852]