Biometrika (2018), **105**, 2, *pp.* 487–493
Printed in Great Britain
doi: 10.1093/biomet/asy008
Advance Access publication 12 March 2018

Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores

By S. YANG

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, North Carolina 27695, U.S.A.
syang24@ncsu.edu

and P. DING

Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, U.S.A.
pengdingpku@berkeley.edu

Summary

Causal inference with observational studies often relies on the assumptions of unconfoundedness and overlap of covariate distributions in different treatment groups. The overlap assumption is violated when some units have propensity scores close to 0 or 1, so both practical and theoretical researchers suggest dropping units with extreme estimated propensity scores. However, existing trimming methods often do not incorporate the uncertainty in this design stage and restrict inference to only the trimmed sample, due to the nonsmoothness of the trimming. We propose a smooth weighting, which approximates sample trimming and has better asymptotic properties. An advantage of our estimator is its asymptotic linearity, which ensures that the bootstrap can be used to make inference for the target population, incorporating uncertainty arising from both design and analysis stages. We extend the theory to the average treatment effect on the treated, suggesting trimming samples with estimated propensity scores close to 1.

Some key words: Bootstrap; Limited overlap; Nonsmooth estimator; Potential outcome; Unconfoundedness.

1. Introduction

In the potential outcomes framework, there is an extensive literature on estimating causal effects based on the assumptions of unconfoundedness and overlap of the covariate distributions (Rosenbaum & Rubin, 1983; Angrist & Pischke, 2008; Imbens & Rubin, 2015). Unfortunately, it is common to have limited overlap in covariates between the treatment and control groups, which affects the credibility of all methods attempting to estimate causal effects for the population (King & Zeng, 2005; Imbens, 2015). Consequently, extreme estimated propensity scores induce large weights, which can result in a large variance and poor finite-sample properties (Kang & Schafer, 2007; Khan & Tamer, 2010). Therefore, it may seem desirable to modify the estimand to averaging only over that part of the covariate space with treatment probabilities bounded away from 0 and 1. For example, in a medical study of a particular chemotherapy for breast cancer, because patients with stage I breast cancer have never been treated with chemotherapy, clinicians then redefine the study population to be patients with stage II to stage IV breast cancer, omitting patients with stage I breast cancer for whom the propensity scores are zero. This effectively alters the estimand by changing the reference population to a different target population. Petersen et al. (2012) used a projection function to define the target parameter within a marginal structural working model. Li et al. (2018) proposed a general representation for the target population.

Trimming observational studies based on estimated propensity scores was first used in medical applications (e.g., Vincent et al., 2002; Grzybowski et al., 2003; Kurth et al., 2005) and then formalized by Crump et al. (2009), who suggested dropping units from the analysis which have estimated propensity scores outside an interval $[\alpha_1, \alpha_2]$, so that the average treatment effect for the target population can be estimated with the smallest asymptotic variance. Other methods, e.g., those of Traskin & Small (2011) and Fogarty et al. (2016), construct the study population based on covariates themselves. But with moderate- or high-dimensional covariates, these rules for discarding units become complicated. In these cases, dimension reduction, for example seeking a scalar summary of the covariates, seems important. This was the original motivation of the propensity score (Rosenbaum & Rubin, 1983), which is arguably the most interpretable scalar function of the covariates.

Existing methods rarely incorporate the uncertainty in this design stage and restrict inference to the trimmed sample. We incorporate uncertainty in both the design and the analysis stages. The nonsmooth nature of trimming renders the target causal estimand not root-n estimable (Crump et al., 2009), so, instead of making a binary decision to include or exclude units from analysis, we propose to use a smooth weight function to approximate the existing sample trimming. This allows us to derive the asymptotic properties of the corresponding causal effect estimators using conventional linearization methods for two-step statistics. We show that the new weighting estimators are asymptotically linear, so the bootstrap can be used to construct confidence intervals.

2. POTENTIAL OUTCOMES, CAUSAL EFFECTS AND ASSUMPTIONS

For each unit i, the treatment is $A_i \in \{0, 1\}$, where 0 and 1 are labels for control and treatment. There are two potential outcomes, one for treatment and the other for control, denoted by $Y_i(1)$ and $Y_i(0)$, respectively. The observed outcome is $Y_i = Y_i(A_i)$. Let X_i be the observed pre-treatment covariates. We assume that $\{A_i, X_i, Y_i(1), Y_i(0)\}_{i=1}^N$ are independent draws from the distribution of $\{A, X, Y(1), Y(0)\}$. Given the observed covariates, the conditional average causal effect is $\tau(X) = E\{Y(1) - Y(0) \mid X\}$. The average treatment effect is $\tau = E\{Y(1) - Y(0)\} = E\{\tau(X)\}$. The common assumptions to identify τ are as follows (Rosenbaum & Rubin, 1983).

Assumption 1 (Unconfoundedness). For $a = 0, 1$, $Y(a)$ is independent of $A \mid X$.

Assumption 2 (Overlap). There exist constants c_1 and c_2 such that with probability 1, $0 < c_1 \leqslant e(X) \leqslant c_2 < 1$, where $e(X) = \text{pr}(A = 1 \mid X)$ is the propensity score.

In observational studies, the propensity score is not known and therefore must be estimated from data. Following Rosenbaum & Rubin (1983) and most of the empirical literature, we assume that the propensity score is correctly specified by a generalized linear model $e(X) = e(X^{\mathrm{T}}\theta^*)$. We focus on $\hat{\theta}$, the maximum likelihood estimator of the true parameter θ^*, although our method is also applicable to other asymptotically linear estimators of θ^*. Then, a simple weighting estimator of τ is $N^{-1}\sum_{i=1}^N \hat{\tau}(X_i)$, where

$$\hat{\tau}(X_i) = \frac{A_i Y_i}{e(X_i^{\mathrm{T}}\hat{\theta})} - \frac{(1 - A_i)Y_i}{1 - e(X_i^{\mathrm{T}}\hat{\theta})}.$$

If we further estimate $\mu(a, X) = E(Y \mid A = a, X)$ by $\hat{\mu}(a, X)$ and obtain the residual $\hat{R}_i = Y_i - \hat{\mu}(A_i, X_i)$, then the augmented weighting estimator is $N^{-1}\sum_{i=1}^N \hat{\tau}^{\text{aug}}(X_i)$ (Lunceford & Davidian, 2004; Bang & Robins, 2005), where

$$\hat{\tau}^{\text{aug}}(X_i) = \left\{\frac{A_i\hat{R}_i}{e(X_i^{\mathrm{T}}\hat{\theta})} + \hat{\mu}(1, X_i)\right\} - \left\{\frac{(1 - A_i)\hat{R}_i}{1 - e(X_i^{\mathrm{T}}\hat{\theta})} + \hat{\mu}(0, X_i)\right\}.$$

The augmented weighting estimator features a double robustness property in the sense that under Assumptions 1 and 2, it is consistent for τ if either $e(X)$ or $\mu(a, X)$ is correctly specified.

The weighting estimators may be variable when Assumption 2 is violated or nearly violated. When there is limited overlap, define the set with adequate overlap to be $\mathcal{O} = \{X : \alpha_1 \leqslant e(X) \leqslant \alpha_2\}$, where α_1 and α_2 are fixed cut-off values, e.g., $\alpha_1 = 0{\cdot}1$ and $\alpha_2 = 0{\cdot}9$ (Crump et al., 2009). The target population is then represented by $\mathcal{O}$, and the estimand of interest becomes $\tau(\mathcal{O}) = E\{\tau(X) \mid X \in \mathcal{O}\}$. The trimmed sample based on the estimated propensity score is $\hat{\mathcal{O}} = \{X : \alpha_1 \leqslant e(X^{\mathrm{T}}\hat{\theta}) \leqslant \alpha_2\}$. Correspondingly, the inclusion weight is

$$\omega(X_i^{\mathrm{T}}\hat{\theta}) = 1\{\alpha_1 \leqslant e(X_i^{\mathrm{T}}\hat{\theta}) \leqslant \alpha_2\}, \tag{1}$$

where $1(\cdot)$ is the indicator function, and the weighting estimators of $\tau(\mathcal{O})$ become

$$\hat{\tau} = \hat{\tau}(\hat{\theta}) = \left\{\sum_{i=1}^{N} \omega(X_i^{\mathrm{T}}\hat{\theta})\right\}^{-1} \sum_{i=1}^{N} \omega(X_i^{\mathrm{T}}\hat{\theta})\hat{\tau}(X_i), \tag{2}$$

$$\hat{\tau}^{\mathrm{aug}} = \hat{\tau}^{\mathrm{aug}}(\hat{\theta}) = \left\{\sum_{i=1}^{N} \omega(X_i^{\mathrm{T}}\hat{\theta})\right\}^{-1} \sum_{i=1}^{N} \omega(X_i^{\mathrm{T}}\hat{\theta})\hat{\tau}^{\mathrm{aug}}(X_i). \tag{3}$$

The main question we address is how the estimated support affects the inference. To make inference for $\tau(\mathcal{O})$, we need to take into account the sampling variability in $\hat{\theta}$, which induces variability of the estimated set $\hat{\mathcal{O}}$, and the sampling variability in $\hat{\tau}$ and $\hat{\tau}^{\mathrm{aug}}$. We cannot directly apply conventional asymptotic linearization methods because the weight function (1) is nonsmooth, so we consider a smooth weight function

$$\omega_\epsilon(X_i^{\mathrm{T}}\hat{\theta}) = \Phi_\epsilon\left\{e(X_i^{\mathrm{T}}\hat{\theta}) - \alpha_1\right\}\Phi_\epsilon\left\{\alpha_2 - e(X_i^{\mathrm{T}}\hat{\theta})\right\}, \tag{4}$$

where $\Phi_\epsilon(z)$ is the normal cumulative distribution with mean zero and variance ϵ^2. The normal distribution can be changed to any differentiable distribution whose variance increases with ϵ. As $\epsilon \to 0$, (4) converges to the indicator weight function (1). Both functions include units with nonextreme propensity scores with probability 1. In contrast, another smooth weight function, the overlap weight function $\omega\{e(X)\} = e(X)\{1 - e(X)\}$ recently proposed by Li et al. (2018), overweighs units with propensity scores close to 0·5 and thus does not target $\tau(\mathcal{O})$.

3. Main results for the average causal effect

We derive the asymptotic results for the smooth weighting estimators. Based on data $\{(A_i, X_i)\}_{i=1}^N$, let the score function and the Fisher information matrix of θ be

$$S(\theta) = \frac{1}{N}\sum_{i=1}^{N} X_i \frac{A_i - e(X_i^{\mathrm{T}}\theta)}{e(X_i^{\mathrm{T}}\theta)\{1 - e(X_i^{\mathrm{T}}\theta)\}} f(X_i^{\mathrm{T}}\theta), \quad \mathcal{I}(\theta) = E\left[\frac{f(X^{\mathrm{T}}\theta)^2}{e(X^{\mathrm{T}}\theta)\{1 - e(X^{\mathrm{T}}\theta)\}} XX^{\mathrm{T}}\right],$$

where $f(t) = \mathrm{d}e(t)/\mathrm{d}t$. Let $\sigma^2(a, X) = \mathrm{var}(Y \mid A = a, X)$ for $a = 0, 1$. Let $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\mathrm{aug}}$ denote the weighting estimators (2) and (3) with the smooth weight function (4), respectively. Let $\tau_\epsilon = E\{\omega_\epsilon(X^{\mathrm{T}}\theta^*)\tau(X)\}$ and $\omega_\epsilon(\theta) = E\{\omega_\epsilon(X^{\mathrm{T}}\theta)\}$. We show that $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\mathrm{aug}}$ are consistent for τ_ϵ. Moreover, the discrepancy between τ_ϵ and the target estimand $\tau(\mathcal{O})$ can be made arbitrarily small by choosing a small ϵ.

Theorem 1. *Under Assumption 1, $\hat{\tau}_\epsilon$ is asymptotically linear. Moreover,*

$$N^{1/2}(\hat{\tau}_\epsilon - \tau_\epsilon) \to \mathcal{N}\left\{0, \sigma_\epsilon^2 + b_{1,\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}b_{1,\epsilon} - b_{2,\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon}\right\}$$

in distribution as $N \to \infty$, *where*

$$b_{1,\epsilon} = E\left[\frac{\partial}{\partial\theta}\left\{\omega_\epsilon(\theta^*)^{-1}\omega_\epsilon(X^{\mathrm{T}}\theta^*)\right\}\tau(X)\right],$$

$$b_{2,\epsilon} = \omega_\epsilon(\theta^*)^{-1}E\left\{\omega_\epsilon(X^{\mathrm{T}}\theta^*)f(X^{\mathrm{T}}\theta^*)\left[\frac{E\{X\mu(1,X)\mid e(X)\}}{e(X)} + \frac{E\{X\mu(0,X)\mid e(X)\}}{1-e(X)}\right]\right\},$$

$$\begin{aligned}\sigma_\epsilon^2 &= \omega_\epsilon(\theta^*)^{-2}E[\omega_\epsilon(X^{\mathrm{T}}\theta^*)^2\mathrm{var}\{\tau(X)\}]\\ &\quad + \omega_\epsilon(\theta^*)^{-2}E\left\{\omega_\epsilon(X^{\mathrm{T}}\theta^*)^2\left[\left\{\frac{1-e(X)}{e(X)}\right\}^{1/2}\mu(1,X) + \left\{\frac{e(X)}{1-e(X)}\right\}^{1/2}\mu(0,X)\right]^2\right\}\\ &\quad + \omega_\epsilon(\theta^*)^{-2}E\left[\omega_\epsilon(X^{\mathrm{T}}\theta^*)^2\left\{\frac{\sigma^2(1,X)}{e(X)} + \frac{\sigma^2(0,X)}{1-e(X)}\right\}\right].\end{aligned}$$

Remark 1. We show in the Supplementary Material that $b_{1,\epsilon} \to 0$ as $\epsilon \to 0$. Therefore, the increased variability due to estimating the support, $b_{1,\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}b_{1,\epsilon}$, is close to 0 with a small ϵ.

Remark 2. The term $-b_{2,\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon}$ implies that the estimated propensity score increases the precision of the simple weighting estimator of τ based on the true propensity score, a phenomenon that has previously appeared in the causal inference literature (e.g., Rubin & Thomas, 1992; Hahn, 1998; Abadie & Imbens, 2016).

THEOREM 2. *Under Assumption* 1, $\hat{\tau}_\epsilon^{\mathrm{aug}}$ *is asymptotically linear. Moreover,*

$$N^{1/2}(\hat{\tau}_\epsilon^{\mathrm{aug}} - \tau_\epsilon) \to \mathcal{N}\left\{0, \tilde{\sigma}_\epsilon^2 + b_{1\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}b_{1\epsilon} + (C_0+C_1)^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}(C_0+C_1) + \tilde{B}^{\mathrm{T}}(C_0-C_1)\right\}$$

in distribution as $N \to \infty$, *where* $b_{1,\epsilon}$ *is defined in Theorem* 1,

$$\tilde{\sigma}_\epsilon^2 = \omega_\epsilon(\theta^*)^{-2}E[\omega_\epsilon(X^{\mathrm{T}}\theta^*)^2\mathrm{var}\{\tau(X)\}] + \omega_\epsilon(\theta^*)^{-2}E\left[\omega_\epsilon(X^{\mathrm{T}}\theta^*)^2\left\{\frac{\sigma^2(1,X)}{e(X)} + \frac{\sigma^2(0,X)}{1-e(X)}\right\}\right],$$

$$C_a = E\left\{X\omega_\epsilon(X^{\mathrm{T}}\theta^*)f(X^{\mathrm{T}}\theta^*)\frac{\tilde{\mu}(a,X)-\mu(a,X)}{\mathrm{pr}(A=a\mid X)}\right\} \quad (a=0,1),$$

with $\hat{\mu}(a,X) \to \tilde{\mu}(a,X)$ *in probability for* $a = 0, 1$ *and* $\tilde{B} = b_{1,\epsilon} - C_0 - C_1$.

Remark 3. If the outcome model is correctly specified, then $\tilde{\mu}(a,X) = \mu(a,X)$ and thus $C_0 = C_1 = 0$. Consequently, the asymptotic variance of $\hat{\tau}_\epsilon^{\mathrm{aug}}$ reduces to $\tilde{\sigma}_\epsilon^2 + b_{1\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}b_{1\epsilon}$, which is smaller than the asymptotic variance of $\hat{\tau}_\epsilon$. Intuitively, by regressing Y on X and A, we use the residual as the new outcome, which in general has a smaller variance than Y.

Remark 4. Because $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\mathrm{aug}}$ are asymptotically linear, the bootstrap can be used to estimate the variances of $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\mathrm{aug}}$ (Shao & Tu, 2012). We evaluate the finite-sample properties of the bootstrap variance estimator by simulation in the Supplementary Material. Let $\mathcal{S} = \{X : e(X^{\mathrm{T}}\theta^*) = \alpha_1 \text{ or } \alpha_2\}$. We also show that if $\mathrm{pr}(X \in \mathcal{S}) = 0$, the bootstrap works for the weighting estimator with the indicator function, which is confirmed by simulation.

Remark 5. Although some robust nonparametric methods (Hirano et al., 2003; Lee et al., 2010, 2011) can be used for propensity score estimation, the majority of the literature uses parametric generalized linear models. When the propensity score model is misspecified, the weighting estimators are not consistent for the causal effect defined on the target population $\mathcal{O} = \{X : \alpha_1 \leqslant e(X) \leqslant \alpha_2\}$. However, our estimators can still be helpful to inform treatment effects for the population defined as $\mathcal{O}^* = \{X : \alpha_1 \leqslant e(X^{\mathrm{T}}\theta^*) \leqslant \alpha_2\}$,

Table 1. *Estimate, standard error based on* 100 *bootstrap replicates, and* 95% *confidence interval*

	ϵ	Estimate	s.e.	95% c.i.		Estimate	s.e.	95% c.i.
$\hat{\tau}(\hat{\theta})$	–	0·646	0·135	(0·376, 0·916)	$\hat{\tau}^{\text{aug}}(\hat{\theta})$	0·765	0·107	(0·552, 0·978)
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-4}	0·661	0·124	(0·412, 0·909)	$\hat{\tau}_\epsilon^{\text{aug}}(\hat{\theta})$	0·763	0·105	(0·554, 0·973)
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-5}	0·632	0·133	(0·366, 0·899)	$\hat{\tau}_\epsilon^{\text{aug}}(\hat{\theta})$	0·754	0·105	(0·543, 0·964)

s.e., standard error; c.i., confidence interval.

where $e(X^{\mathrm{T}}\theta^*)$ is the propensity score projected to the generalized linear model family. This new study population is defined as being between two hyperplanes of the covariate space, which is slightly more complicated than the study population defined by the trees in Traskin & Small (2011) or by the intervals of covariates in Fogarty et al. (2016). Moreover, the smooth weighting estimators are still asymptotically linear, and again the bootstrap can be used for constructing confidence intervals. See the Supplementary Material for more details.

Remark 6. An important issue regarding the smooth weight function is the choice of ϵ, which involves a bias-variance trade-off. On the one hand, the discrepancy between τ_ϵ and the target parameter $\tau(\mathcal{O})$ is $E([\omega_\epsilon(X^{\mathrm{T}}\theta^*) - 1\{\alpha_1 \leqslant e(X^{\mathrm{T}}\theta^*) \leqslant \alpha_2\}]\tau(X))$. Assuming that $\tau(X)$ is integrable, by the dominated convergence theorem, τ_ϵ converges to $\tau(\mathcal{O})$ as $\epsilon \to 0$. This implies that based on $\hat{\tau}_\epsilon$ or $\hat{\tau}_\epsilon^{\text{aug}}$, we can draw inference for $\tau(\mathcal{O})$ by choosing a small ϵ. On the other hand, as $\epsilon \to 0$, the smooth weight function (4) becomes closer to the indicator weight function (1), which increases the variance of the weighting estimators. In practice, we recommend a sensitivity analysis varying ϵ over a grid, for example, $10^{-4}, 10^{-5}, \ldots$, as illustrated in the Supplementary Material and the application in the next section.

4. National Health and Nutrition Examination Survey data

We examine a dataset from the 2007–2008 U.S. National Health and Nutrition Examination Survey to estimate the causal effect of smoking on blood lead levels (Hsu & Small, 2013). The dataset includes 3340 subjects consisting of 679 smokers, denoted by $A = 1$, and 2661 nonsmokers, denoted by $A = 0$. The outcome variable Y is the measured level of lead in the subject's blood, with the observed range being from 0·18 μg/dl to 33·10 μg/dl. The covariates are age, income-to-poverty level, gender, education and race.

The propensity score is estimated by a logistic regression model with linear predictors including all covariates. To help address the lack of overlap, for the average smoking effect, because there is little overlap for the propensity score less than 0·05 or greater than 0·6, we restrict our estimand to the target population $\mathcal{O} = \{X : 0{\cdot}05 \leqslant e(X) \leqslant 0{\cdot}6\}$. The truncation of the propensity score at 0·6 is because there are few subjects with propensity score above 0·6. This removes 794 subjects, including 111 smokers and 683 non-smokers. Thus, the final analysis sample includes 2546 subjects, with 568 smokers and 1978 non-smokers. In the Supplementary Material, we display the summary statistics of the covariates and give a more detailed interpretation of the target population.

We consider the weighting estimators using both the indicator and the smooth weight functions with $\epsilon = 10^{-4}$ and $\epsilon = 10^{-5}$. For the augmented weighting estimator, we use a linear outcome model adjusting for all covariates, separately for $A = 0, 1$. Table 1 shows the results. The weighting estimators with the smooth weight function are close to their counterparts with the indicator weight function, but have slightly smaller estimated standard errors. The smooth weighting estimators are insensitive to the choice of ϵ. From the results, on average, smoking increases the lead level in blood by at least 0·65 μg/dl over the target population with $0{\cdot}05 \leqslant e(X) \leqslant 0{\cdot}6$.

5. Extension to the average treatment effect on the treated

Another estimand of interest is the average treatment effect for the treated, $\tau_{\text{ATT}} = E\{Y(1) - Y(0) \mid A = 1\} = E\{\tau(X) \mid A = 1\}$. Similar to Crump et al. (2009), if $\sigma^2(1, X) = \sigma^2(0, X)$, we can show that

the optimal overlap for estimating τ_{ATT} is of the form $\mathcal{O} = \{X : 1 - e(X) \geqslant \alpha\}$ for some α, for which the estimators have the smallest asymptotic variance. Intuitively, for the treated units with $e(X)$ close to 1, there are few similar units in the control group that can provide information to infer their $Y(0)$ values. Therefore, it is reasonable to drop these units with $e(X)$ close to 1 when inferring τ_{ATT}. We give a formal discussion in the Supplementary Material.

By restricting to the subpopulation $\mathcal{O} = \{X : 1 - e(X) \geqslant \alpha\}$, the estimand of interest becomes $\tau_{\text{ATT}}(\mathcal{O}) = E\{\tau(X) \mid A = 1, X \in \mathcal{O}\}$. We propose two estimators with smooth inclusion weights $\omega_{\text{ATT},\epsilon}(X^{\mathrm{T}}\hat{\theta}) = \Phi_\epsilon\{1 - \alpha - e(X_i^{\mathrm{T}}\hat{\theta})\}e(X_i^{\mathrm{T}}\hat{\theta})$:

$$\hat{\tau}_{\text{ATT},\epsilon} = \frac{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X^{\mathrm{T}}\hat{\theta})\hat{\tau}(X_i)}{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X^{\mathrm{T}}\hat{\theta})}, \quad \hat{\tau}^{\text{aug}}_{\text{ATT},\epsilon} = \frac{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X^{\mathrm{T}}\hat{\theta})\hat{\tau}^{\text{aug}}(X_i)}{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X^{\mathrm{T}}\hat{\theta})},$$

which are (2) and (3) with $\omega_\epsilon(X^{\mathrm{T}}\hat{\theta})$ replaced by $\omega_{\text{ATT},\epsilon}(X^{\mathrm{T}}\hat{\theta})$. Even without sample trimming, the augmented weighting estimator is different from the existing estimators in the literature (e.g., Mercatanti & Li, 2014; Shinozaki & Matsuyama, 2015; Zhao & Percival, 2017). We provide the motivation in the Supplementary Material. The asymptotic properties of $\hat{\tau}_{\text{ATT},\epsilon}$ and $\hat{\tau}^{\text{aug}}_{\text{ATT},\epsilon}$ can be derived similarly to the results in Theorems 1 and 2. In particular, the asymptotic linearity of these two estimators enables use of the bootstrap for inference.

Define $\tilde{b}_{1,\epsilon}$ and $\tilde{b}_{2,\epsilon}$ as the analogues of $b_{1,\epsilon}$ and $b_{2,\epsilon}$ with weights $\omega_{\text{ATT},\epsilon}(X^{\mathrm{T}}\hat{\theta})$. In contrast to Remark 1, for τ_{ATT}, the term $\tilde{b}_{1,\epsilon}$ does not converge to 0 as $\epsilon \to 0$. The correction term in the asymptotic variance formula due to the estimated propensity score instead of the true propensity score, $\tilde{b}_{1,\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}\tilde{b}_{1,\epsilon} - \tilde{b}_{2,\epsilon}^{\mathrm{T}}\mathcal{I}(\theta^*)^{-1}\tilde{b}_{2,\epsilon}$, can be negative, zero, or positive. Ignoring the uncertainty in the estimated propensity score, the inference can be either conservative or anticonservative for τ_{ATT}, which differs from the inference for τ. This fundamental difference also appeared for matching estimators (Abadie & Imbens, 2016), which highlights the importance of incorporating the uncertainty in the design stage especially for τ_{ATT}.

ACKNOWLEDGEMENT

We benefited from the insightful comments from the associate editor and two reviewers. Peng Ding was partially supported by the U.S. Institute of Education Sciences and National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs, a simulation study, an extension, and more details on the application.

REFERENCES

ABADIE, A. & IMBENS, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84**, 781–807.

ANGRIST, J. D. & PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

BANG, H. & ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–73.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. & MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–99.

FOGARTY, C. B., MIKKELSEN, M. E., GAIESKI, D. F. & SMALL, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Am. Statist. Assoc.* **111**, 447–58.

GRZYBOWSKI, M., CLEMENTS, E. A., PARSONS, L., WELCH, R., TINTINALLI, A. T., ROSS, M. A. & ZALENSKI, R. J. (2003). Mortality benefit of immediate revascularization of acute ST-segment elevation myocardial infarction in patients with contraindications to thrombolytic therapy: A propensity analysis. *J. Am. Med. Assoc.* **290**, 1891–8.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–31.
Hirano, K., Imbens, G. W. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–89.
Hsu, J. Y. & Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69**, 803–11.
Imbens, G. W. (2015). Matching methods in practice: Three examples. *J. Hum. Resour.* **50**, 373–419.
Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.
Kang, J. D. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523–39.
Khan, S. & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica* **78**, 2021–42.
King, G. & Zeng, L. (2005). The dangers of extreme counterfactuals. *Polit. Anal.* **14**, 131–59.
Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K. & Robins, J. M. (2005). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am. J. Epidemiol.* **163**, 262–70.
Lee, B. K., Lessler, J. & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statist. Med.* **29**, 337–46.
Lee, B. K., Lessler, J. & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS One*. **6**, e18174.
Li, F., Morgan, K. L. & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *J. Am. Statist. Assoc.*, doi: 10.1080/01621459.2016.1260466.
Lunceford, J. K. & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statist. Med.* **23**, 2937–60.
Mercatanti, A. & Li, F. (2014). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Ann. Appl. Statist.* **8**, 2485–508.
Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. & Van Der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. **21**, 31–54.
Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
Rubin, D. B. & Thomas, N. (1992). Affinely invariant matching methods with ellipsoidal distributions. *Ann. Statist.* **20**, 1079–93.
Shao, J. & Tu, D. (2012). *The Jackknife and Bootstrap*. New York: Springer.
Shinozaki, T. & Matsuyama, Y. (2015). Doubly robust estimation of standardized risk difference and ratio in the exposed population. *Epidemiology* **26**, 873–77.
Traskin, M. & Small, D. S. (2011). Defining the study population for an observational study to ensure sufficient overlap: A tree approach. *Statist. Biosci.* **3**, 94–118.
Vincent, J. L., Baron, J.-F., Reinhart, K., Gattinoni, L., Thijs, L., Webb, A., Meier-hellmann, A., Nollet, G. & Peres-bota, D. (2002). Anemia and blood transfusion in critically ill patients. *J. Am. Med. Assoc.* **288**, 1499–507.
Zhao, Q. & Percival, D. (2017). Entropy balancing is doubly robust. *J. Causal Infer.* **5**, doi: 10.1515/jci–2016–0010.

[*Received on* 3 *April* 2017. *Editorial decision on* 7 *January* 2018]