Submitted to Statistical Science

Comment: Stabilizing the doubly-robust estimators of the average treatment effect under positivity violations

Fan Li

Department of Biostatistics, Yale School of Public Health

Abstract. Doubly-robust estimators within the one-step and TMLE frameworks could exhibit finite-sample bias and excess variability under positivity violations. We comment on how the application of a stabilization factor may improve the efficiency property of one-step estimator and TMLE, and the comparisons with their collaborative counterparts using the adaptive propensity scores.

Key words and phrases: Efficient influence function, one-step estimation, stabilized weight, targeted maximum likelihood estimation.

1. INTRODUCTION

We congratulate Benkeser, Cai and van der Laan on their valuable and timely article which contributes an improved approach to estimate the average treatment effect (ATE) in observational studies. While the traditional semiparametric doubly-robust methods—the one-step and the targeted minimum loss estimation (TMLE)—are locally efficient when the relevant nuisance parameters are consistently estimated, they could exhibit finite-sample bias and excess variability in challenging scenarios when the propensity score distribution has a long tail. Using the adaptive propensity score, defined as the conditional probability of assignment given the conditional mean outcome, in the construction of the semiparametric estimators, the authors demonstrated super-efficiency over their locally efficient counterparts. As the adaptive propensity score may be inconsistent to the true propensity score, the proposed collaborative estimators trade off the double robustness property for greater stability and efficiency.

In this discussion, we would like to contribute our thoughts on alternative methods that may improve the finite-sample behaviour of the semiparametric estimators, in challenging scenarios when there is lack of overlap (or positivity violations) and the causal estimand is only weakly identifiable. For inverse probability weighting (IPW), a common remedy for lack of overlap is to trim observations with extreme propensity scores and restrict the analysis to the subpopulation where the average causal effect is better identified. To avoid arbitrary selection

Fan Li is Assistant Professor, Department of Biostatistics, Yale School of Public Health, 135 College Street, Suite 200, New Haven, Connecticut, 06510 (e-mail: fan.f.li@yale.edu).

of the trimming threshold, Crump et al. (2009) developed the optimal symmetric trimming rule where the asymptotic variance of the weighted estimator is minimized among all symmetric trimming rules. By continuously down-weighting influential observations at the propensity tails, Li, Morgan and Zaslavsky (2018) introduced the overlap weights which target the average causal effect among the subpopulation with substantial equipoise. The overlap weights are also efficient in a sense that they minimize the asymptotic variance of the weighted estimator among the larger family of balancing weights. While these alternative weighting schemes address limitations of IPW, it is less clear whether they could improve the performance of the doubly-robust estimators in challenging scenarios. In what follows, we introduce these two weighting schemes to the one-step and TMLE framework for stabilization purposes, and evaluate their operating characteristics under lack of overlap.

The remainder of this discussion is organized as follows. Section 2 and 3 introduce notations and the alternative weighting schemes. In Section 4, we compare these alternative methods with the standard one-step estimator and TMLE, as well as their collaborative counterparts based on the adaptive propensity score. Concluding remarks are summarized in Section 5.

2. NOTATIONS

Throughout we will be using the notations of Benkeser et al. to ensure a consistent presentation. Consider a sample of n independent units, each receiving one of the two treatments. Let $A = 1$ if the unit receives treatment and $A = 0$ otherwise. Without loss of generality, we denote the observed outcome for each unit by $Y \in [0, 1]$, and the set of pre-treatment covariates by $W \in \mathcal{W}$. Under the Stable Unit Treatment Value Assumption, each unit has two potential outcomes, $Y(1)$ and $Y(0)$, mapped to each level of treatment. However, only one of them is observed corresponding to the actual assignment. The target estimand, ATE, is defined as $E_{P_0^1}[Y(1)] - E_{P_0^0}[Y(0)]$, where P_0^a $(a = 0, 1)$ is the true probability distribution of the potential outcome $Y(a)$. The observed data triplet for each unit is $O = (W, A, Y)$, and we define P_0 as the probability distribution of O. We assume the standard *unconfoundedness*: $\{Y(1), Y(0)\} \perp A|W$, and *overlap*: $\text{pr}_{P_0}\{0 < \text{pr}_{P_0}(A = 1|W) < 1)\} = 1$ in order to identify the ATE. To simplify the presentation, we focus on the estimation of the treatment-specific average $\psi_0^1 = E_{P_0^1}[Y(1)] = E_{P_0}[E_{P_0}(Y|A = 1, W)]$, since the development for $\psi_0^0 = E_{P_0^0}[Y(0)]$ is completely symmetric.

Write the true outcome regression function as $\bar{Q}_0^1(w) = E_{P_0}[Y|A = 1, W = w]$, and its estimate $\bar{Q}_n^1(w)$ for each $w \in \mathcal{W}$. The distribution function of pre-treatment covariates is defined as $Q_{0,W}(w)$, and the corresponding empirical distribution function is $Q_{n,W} = n^{-1}\sum_{i=1}^n \mathbb{1}(W_i \leq w)$. Define the true propensity score for each $w \in \mathcal{W}$ by $\bar{G}_0(w) = \text{pr}_{P_0}(A = 1|W = w)$. Based on these notations, the efficient influence function that leads to the construction of the one-step estimator and TMLE has the form

$$D^1(o|\bar{Q}_0^1, Q_{0,W}, \bar{G}_0) = \frac{a}{\bar{G}_0(w)}[y - \bar{Q}_0^1(w)] + \bar{Q}_0^1(w) - \psi_0^1, \tag{2.1}$$

for a typical observation o. In particular, the leading term can be interpreted as a mean-zero residual bias-correction to the outcome regression function $\bar{Q}_0^1(w)$ for

the estimation of ψ_0^1. As the bias-correction term concerns the inverse probability weights, $a/\bar{G}_0(w)$, the overlap assumption is necessary to ensure the existence of variance of the influence function, or equivalently the semiparametric variance lower bound.

3. ALTERNATIVE WEIGHTING SCHEMES

Even though the overlap assumption should hold in theory, practical lack of overlap can arise in finite samples due to a number of reasons (Petersen et al., 2012), leading to poor identification of the treatment-specific averages. In this case, the estimated propensity score $\bar{G}_n(w)$ may be close to zero for some design point $w \in \mathcal{W}$, and an estimate for the residual bias-correction term would involve an exploding weight, resulting in bias and excess variability (Li, Thomas and Li, 2019). As a potential remedy, we consider the following influence function

$$D_h^1(o|\bar{Q}_0^1, Q_{0,W}, \bar{G}_0) = \frac{ah(w)}{\bar{G}_0(w)}[y - \bar{Q}_0^1(w)] + \bar{Q}_0^1(w) - \psi_0^1, \tag{3.1}$$

where $h(w)$ is a stabilization factor that depends only on w and satisfies

$$\int h(u) dQ_{0,W}(u) = 1.$$

Influence function (3.1) is similar to (2.1) except that the bias-correction term includes a stabilized weight, $ah(w)/\bar{G}_0(w)$. In fact, this bias-correction term in (3.1) is identical to that of the efficient influence function for the estimation of the *weighted average treatment effect* (WATE, Hirano, Imbens and Ridder, 2003). In the latter context, $h(w)$ is a tilting function that redefines the target population and causal estimand with potentially improved identifiability (Li and Li, 2019). In the current setting, however, we still focus on the ATE and only consider $h(w)$ to avoid unnecessary bias-correction based on extremely small propensity score values. The influence functions, $D_h^1(\cdot|\bar{Q}_0^1, Q_{0,W}, \bar{G}_0)$ and $D^1(\cdot|\bar{Q}_0^1, Q_{0,W}, \bar{G}_0)$, are equivalent when $h(w) = 1$ for all $w \in \mathcal{W}$, but not necessarily so when $h(w)$ deviates from unity. Similar to the collaborative estimators based on the adaptive propensity score, the specification of a stabilization factor trades off robustness for efficiency in estimating the ATE. Under the assumption that $\bar{Q}_n^1$ is consistent to $\bar{Q}_0^1$ (also assumed for the collaborative estimators), both the stabilized one-step estimator and TMLE based on influence function (3.1) converge to ψ_0^1, and therefore different choices of $h(w)$ only affect efficiency. With an inconsistent initial outcome model estimate, however, the stabilized estimators could be biased to the true ATE even when the propensity score is consistently estimated.

In what follows, we will assume that the outcome model can be consistently estimated (possibly by using machine learning methods) and explore the efficiency property under two different specifications of $h(w)$. The specification associated with the symmetric trimming due to Crump et al. (2009) corresponds to

$$h(w) = \frac{\mathbb{1}(\delta \leq \bar{G}_0(w) \leq 1 - \delta)}{\int \mathbb{1}(\delta \leq \bar{G}_0(u) \leq 1 - \delta) dQ_{0,W}(u)}, \tag{3.2}$$

for some threshold $\delta \in [0, 0.5]$. For units with propensity scores bounded between δ and $1 - \delta$, $h(w) \propto$ constant, while for units with propensities outside of that

range, $h(w) = 0$ and thus effectively removes the influential observations at the tails during the residual bias-correction step. To avoid arbitrary trimming decisions, we also explore the optimal trimming strategy suggested in Crump et al. (2009). The optimal trimming threshold δ_n (if exists) is defined as the solution of δ to the sample analogue of

$$\frac{1}{\delta(1-\delta)} = 2E_{P_0}\left[\frac{1}{\bar{G}_0(W)(1-\bar{G}_0(W))}\middle|\frac{1}{\bar{G}_0(W)(1-\bar{G}_0(W))} \leq \frac{1}{\delta(1-\delta)}\right]. \tag{3.3}$$

Focusing on the optimizable part of the asymptotic variance, the optimal trimming threshold maximizes the efficiency of IPW among all trimming rules under homoscedasticity. Here, we are interested in whether such implications for IPW translate into optimal efficiency gain for the one-step estimator and TMLE.

While the trimming specification corresponds to setting $h(w)$ as a step function, the overlap specification smoothly down-weights the influential observations at the tails. This is achieved by defining

$$h(w) = \frac{\bar{G}_0(w)(1-\bar{G}_0(w))}{\int \bar{G}_0(u)(1-\bar{G}_0(u))dQ_{0,W}(u)}. \tag{3.4}$$

Clearly, when the propensity score $\bar{G}_0(w)$ is close to 0.5 (point of equipoise), the stabilization factor reaches its maximum. On the contrary, when the propensity becomes extreme, $h(w)$ gradually reduces to zero, and therefore mimics a smooth trimming operator without arbitrary decisions on the trimming threshold. For traditional propensity score weighted estimators, the overlap weights approximately minimize the asymptotic variance among all possible choices of $h(w)$ (Li, Morgan and Zaslavsky, 2018), and often leads to a more efficient weighted estimator than trimming (Li, Thomas and Li, 2019). Finally, both specifications of $h(w)$ are somewhat adaptive to the distributions of propensity scores. For example, under adequate overlap such that the majority of $\bar{G}_0(w)$ is close to 0.5, $h(w) \approx 1$ and the influence function (3.1) approximates (2.1). Under lack of overlap, $h(w)$ removes extreme observations at the tails and stabilizes the residual bias-correction step. Weight truncation is another approach that could improve efficiency under lack of overlap, and has been well-studied in Bembom and van der Laan (2008); Petersen et al. (2012) and Ju, Schwab and van der Laan (2019) for one-step estimation, TMLE and collaborative TMLE. We do not further explore truncation in this discussion as truncation implies a stabilization factor that depends on both a and w, corresponding to an influence function outside of the class given by (3.1).

Denote $h_n(w)$ as an estimate of the stabilization factor, where $\bar{G}_0(w)$ and $Q_{0,W}(w)$ are replaced by their estimates, $\bar{G}_n(w)$ and $Q_{n,W}(w)$. Based on (3.1), the stabilized one-step estimator for ψ_0^1 is given by

$$\psi_{n,+}^1 = \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n^1(W_i) + \frac{1}{n}\sum_{i=1}^{n}D_{h_n}^1(O_i|\bar{Q}_n^1, Q_{n,W}, \bar{G}_n). \tag{3.5}$$

The stabilized TMLE algorithm based on (3.1) can proceed as follows.

1. *estimate outcome model*: regress Y on W among units receiving treatment $A = 1$ to obtain the outcome model estimate $\bar{Q}_n^1$;

TABLE 1

Comparisons of one-step (OS) estimators for the ATE. Bias: absolute bias (×100); RE: relative efficiency defined as the ratio between the empirical variance of the standard one-step and that of the estimator of interest; CP: empirical coverage probability (%).

		$\gamma = 0$			$\gamma = 3$			$\gamma = 6$		
	Estimator	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP
$n = 100$	OS	0.4	1.00	92.7	0.3	1.00	90.2	2.4	1.00	70.1
	COS	0.1	1.19	88.7	1.0	2.04	82.2	1.7	3.04	62.0
	OS_{δ_n}	0.2	1.09	93.7	0.7	1.66	93.2	3.0	1.93	85.7
	$\text{OS}_{\text{overlap}}$	0.2	1.21	93.6	0.8	1.98	93.7	2.1	2.71	88.5
	g-comp	0.2	1.21	–	0.8	1.98	–	2.1	2.71	–
$n = 500$	OS	0.3	1.00	94.4	0.6	1.00	93.7	1.0	1.00	91.6
	COS	0.2	1.07	93.1	0.0	1.56	87.5	0.1	2.65	64.7
	OS_{δ_n}	0.2	1.02	94.1	0.2	1.28	94.2	0.3	1.86	92.8
	$\text{OS}_{\text{overlap}}$	0.2	1.10	95.1	0.0	1.55	94.4	0.0	2.60	93.2
	g-comp	0.2	1.10	–	0.0	1.55	–	0.0	2.60	–
$n = 1000$	OS	0.3	1.00	94.3	0.3	1.00	94.2	0.5	1.00	92.1
	COS	0.3	1.02	93.5	0.2	1.59	88.1	0.5	2.39	65.4
	OS_{δ_n}	0.2	1.01	94.3	0.0	1.29	94.3	0.3	1.64	92.8
	$\text{OS}_{\text{overlap}}$	0.3	1.08	95.0	0.2	1.59	94.8	0.5	2.34	93.1
	g-comp	0.3	1.08	–	0.2	1.59	–	0.5	2.34	–

2. *predict potential outcomes*: use the estimated outcome model to predict $\bar{Q}_n^1(W_i)$ for each unit, $i = 1, \ldots, n$;
3. *estimate propensity score model*: regress A on W to obtain the propensity score estimate $\bar{G}_n$;
4. *predict propensity scores*: predict the propensity score, $\bar{G}_n(W_i)$, for each unit, $i = 1, \ldots, n$;
5. *fit fluctuation working model*: fit logistic regression of outcome Y on the stabilized clever covariate, $H_n(A, W) = Ah_n(W)/\bar{G}_n(W)$, with offset $\text{logit}[\bar{Q}_n^1(W)]$; denote by ϵ_n^1 the estimated coefficient;
6. *target outcome estimates*: use the outcome working model to obtain a prediction $\bar{Q}_{n,*}^1 = \text{expit}\{\text{logit}[\bar{Q}_n^1(W_i)] + \epsilon_n^1 H_n(1, W_i)\}$ for each unit, $i = 1, \ldots, n$;
7. *compute final estimate*: the stabilized TMLE is $\psi_{n,*}^1 = n^{-1} \sum_{i=1}^n \bar{Q}_{n,*}^1(W_i)$.

Similar to the standard TMLE, it can be shown that the score of the stabilized fluctuation model at zero fluctuation ($\epsilon = 0$) spans the influence curve (3.1) at the initial estimator, and also that the final TMLE estimate $\psi_{n,*}^1$ solves the desired influence curve estimating equations.

4. COMPARISONS VIA SIMULATIONS

We replicate the simulation experiments carried out in Benkeser et al. to explore whether including the stabilization factor $h(w)$ could improve the finite-sample behaviour of the one-step estimator and TMLE, and if so, how the improvement compares with that due to the collaborative estimators. In the first simulation, seven confounders $(W_1, \ldots, W_7)$ are independently generated from Uniform$[-1.5, 1.5]$, and an instrumental variable is generated from Bernoulli(0.5). The true propensity score model is $\text{logit}[\bar{G}_0(w)] = \gamma/2 - \gamma W_8 + \sum_{j=2}^7 2^{1-j} W_j$ and the true potential outcome model is $Y(a) = a - \sum_{j=2}^7 2^{1-j} W_j + e$, with $e \sim N(0, 1)$; the true ATE is unity. The coefficient of the instrumental variable in the propen-

TABLE 2

Comparisons of TMLE estimators for the ATE. Bias: absolute bias (×100); RE: relative efficiency defined as the ratio between the empirical variance of the standard TMLE and that of the estimator of interest; CP: empirical coverage probability (%).

		$\gamma = 0$			$\gamma = 3$			$\gamma = 6$		
	Estimator	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP
$n = 100$	TMLE	0.2	1.00	91.2	6.2	1.00	75.8	27.1	1.00	42.3
	CTMLE	0.0	1.13	88.5	1.0	2.45	82.3	1.6	6.48	62.0
	TMLE_{δ_n}	0.1	1.09	93.2	0.8	2.29	92.1	1.1	5.92	82.7
	$\text{TMLE}_{\text{overlap}}$	0.1	1.17	93.7	0.7	2.38	92.7	1.5	5.78	85.2
	g-comp	0.2	1.18	–	1.1	2.50	–	1.7	6.47	–
$n = 500$	TMLE	0.2	1.00	94.5	0.7	1.00	91.8	4.3	1.00	71.7
	CTMLE	0.2	1.05	93.0	0.0	1.49	87.5	0.1	3.58	64.7
	TMLE_{δ_n}	0.2	1.02	93.9	0.2	1.33	94.8	0.1	3.20	93.1
	$\text{TMLE}_{\text{overlap}}$	0.2	1.09	95.1	0.0	1.48	94.4	0.3	3.40	92.6
	g-comp	0.2	1.09	–	0.0	1.49	–	0.1	3.59	–
$n = 1000$	TMLE	0.3	1.00	94.4	0.2	1.00	92.9	0.7	1.00	83.4
	CTMLE	0.3	1.01	93.5	0.2	1.55	88.0	0.5	2.70	65.4
	TMLE_{δ_n}	0.2	1.02	94.2	0.1	1.38	95.5	0.3	2.34	94.2
	$\text{TMLE}_{\text{overlap}}$	0.3	1.08	94.9	0.2	1.55	94.8	0.4	2.64	92.9
	g-comp	0.2	1.08	–	0.2	1.58	–	0.5	2.70	–

sity model is varied $\gamma \in \{0, 3, 6\}$ to induce increasing lack of overlap. Three sample sizes $n \in \{100, 500, 1000\}$ are considered. In addition to the standard and collaborative one-step estimators and TMLE (COS and CTMLE), we include the stabilized estimators associated with the optimal trimming threshold (OS_{δ_n} and TMLE_{δ_n}) and overlap weighting ($\text{OS}_{\text{overlap}}$ and $\text{TMLE}_{\text{overlap}}$). Estimation of all nuisance parameters are based on correctly-specified parametric models except for the adaptive propensity score, which is estimated by the highly adaptive loss minimum loss estimator (HAL-MLE, van der Laan, 2017; Benkeser and van der Laan, 2016). As the outcome model is correctly specified, we also include the outcome regression estimator without the targeting step (g-comp) as a benchmark for efficiency. We report the absolute bias, relative efficiency and empirical coverage across 1000 simulations. The relative efficiency is defined as the ratio between the empirical variance of the standard estimator (one-step or TMLE) and that of the estimator of interest; values larger than 1 indicate higher efficiency compared to the standard implementation. For each simulation, the 95% confidence intervals are constructed based on the non-cross-validated sample variance of the corresponding influence function. We do not report the coverage of the outcome regression estimator as we mainly focus on its efficiency.

Table 1 summarizes the results for the one-step estimators. Including the stabilization factor improves the efficiency of the standard one-step estimator, across all degrees of overlap and sample sizes considered. The relative efficiency gain is greater with increasing lack of overlap. Particularly, the overlap stabilization factor leads to the same efficiency as the outcome regression estimator based on the g-computation formula, and is more efficient than using the optimal trimming stabilization factor. While the overlap stabilization offers a slight advantage over the collaborative estimator under adequate overlap ($\gamma = 0$), the collaborative estimator remains slightly more efficient with increasing lack of overlap ($\gamma = 3$ and 6). The interval estimates from the stabilized estimators demonstrate the best

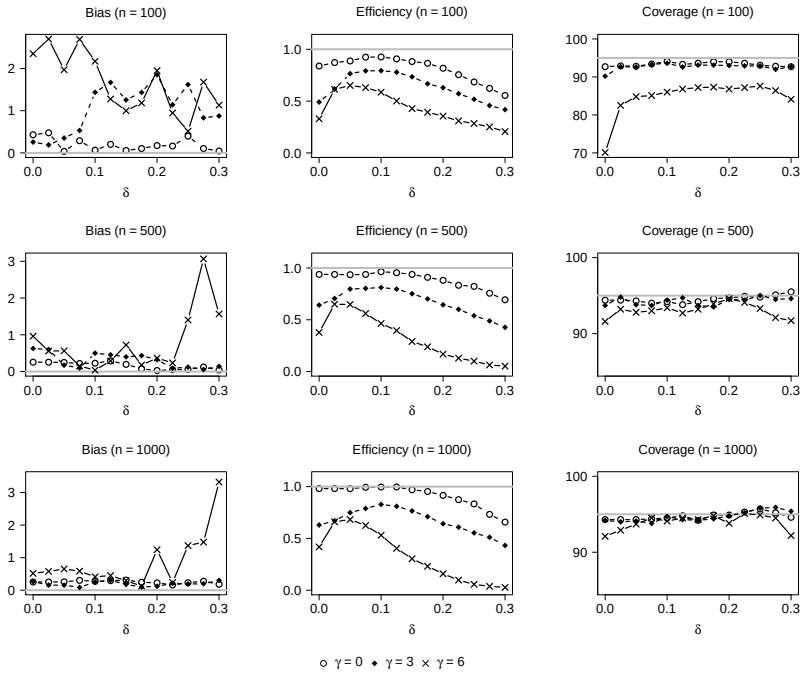

FIG 1. *Performance of the trimmed one-step estimators as a function of threshold* $\delta \in [0, 0.3]$. *Bias: absolute bias* ($\times 100$)*; Efficiency: defined as the ratio between the empirical variance of the collaborative one-step estimator and that of the trimmed one-step estimator; CP: empirical coverage probability (%).*

empirical coverage, which is fairly close to 95% except when the sample size is small and the lack of overlap is substantial. In contrast, the interval estimates for the collaborative estimator present notable under-coverage. The findings for the stabilized TMLE are similar (Table 2). It is interesting to observe that, with increasing lack of overlap, the relative efficiency gain over the standard TMLE due to the adaptive propensity score or stabilization factor appears more substantial than that in the one-step framework.

We realize that the optimal trimming threshold δ_n is derived for the IPW estimator, and thus may not be truly optimal within the one-step or TMLE framework. To study whether other deterministic trimming rules could potentially provide better efficiency, we study the estimator's performance as a function of the trimming threshold $\delta \in [0, 0.3]$ defined in (3.2). We present the absolute bias, efficiency and coverage of the stabilized one-step estimator and TMLE in Figure 1 and 2. Here, we redefine the efficiency as the ratio between the empirical variance of the collaborative estimator and that of the stabilized estimator. An interesting dichotomy between the one-step and TMLE frameworks emerges as to how the efficiency varies over δ. For example, as the threshold moves away from 0, the efficiency of the one-step estimator tends to first increase but then decrease, while the TMLE estimator monotonically increases efficiency, eventually approximating the super-efficiency of the CTMLE. This suggests that the stabilized TMLE with trimming has potential to achieve a competitive level of efficiency advantage just as the CTMLE, and the truly optimal threshold is further away from zero than δ_n.

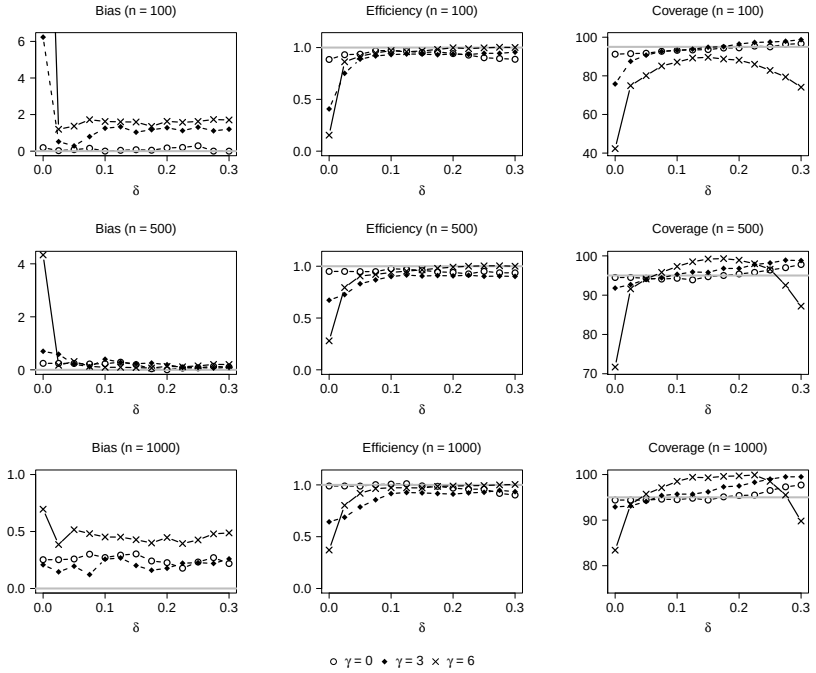

FIG 2. *Performance of the trimmed TMLE estimators as a function of threshold* $\delta \in [0, 0.3]$. *Bias: absolute bias* ($\times 100$)*; Efficiency: defined as the ratio between the empirical variance of collaborative TMLE and that of the trimmed TMLE; CP: empirical coverage probability (%).*

TABLE 3

Comparisons of estimators for estimating ATE in the Kang and Schafer (2007) simulation design. Bias: absolute bias ($\times 100$)*; RE: relative efficiency defined as the ratio between the empirical variance of the standard estimator (OS or TMLE) and that of the estimator of interest; CP: empirical coverage probability (%).*

	Estimator	Bias	RE	CP	Estimator	Bias	RE	CP
	OS	0.89	1.00	16.8	TMLE	0.97	1.00	14.6
	COS	0.89	1.00	17.9	CTMLE	0.89	1.13	17.9
	$\text{OS}_{[\delta=0.05]}$	0.89	1.00	17.3	$\text{TMLE}_{[\delta=0.05]}$	0.93	1.10	15.4
$n = 100$	$\text{OS}_{[\delta=0.15]}$	0.89	1.00	18.1	$\text{TMLE}_{[\delta=0.15]}$	0.90	1.12	17.9
	$\text{OS}_{[\delta=0.25]}$	0.89	0.99	19.0	$\text{TMLE}_{[\delta=0.25]}$	0.88	1.13	19.6
	$\text{OS}_{\text{overlap}}$	0.89	1.00	17.7	$\text{TMLE}_{\text{overlap}}$	0.91	1.11	16.2
	g-comp	0.86	1.01	–	g-comp	0.86	1.13	–
	OS	0.21	1.00	16.1	TMLE	0.21	1.00	16.1
	COS	0.21	1.00	16.2	CTMLE	0.21	1.00	16.4
	$\text{OS}_{[\delta=0.05]}$	0.21	1.00	16.2	$\text{TMLE}_{[\delta=0.05]}$	0.21	1.00	16.3
$n = 500$	$\text{OS}_{[\delta=0.15]}$	0.21	1.00	16.4	$\text{TMLE}_{[\delta=0.15]}$	0.21	1.00	16.5
	$\text{OS}_{[\delta=0.25]}$	0.21	1.00	17.2	$\text{TMLE}_{[\delta=0.25]}$	0.21	1.00	16.5
	$\text{OS}_{\text{overlap}}$	0.21	1.00	16.2	$\text{TMLE}_{\text{overlap}}$	0.21	1.00	16.4
	g-comp	0.21	1.00	–	g-comp	0.21	1.00	–
	OS	0.11	1.00	19.9	TMLE	0.11	1.00	19.3
	COS	0.11	1.00	20.6	CTMLE	0.11	1.01	20.3
	$\text{OS}_{[\delta=0.05]}$	0.11	1.00	20.5	$\text{TMLE}_{[\delta=0.05]}$	0.11	1.01	19.9
$n = 1000$	$\text{OS}_{[\delta=0.15]}$	0.11	1.00	22.0	$\text{TMLE}_{[\delta=0.15]}$	0.11	1.01	22.2
	$\text{OS}_{[\delta=0.25]}$	0.11	1.00	23.9	$\text{TMLE}_{[\delta=0.25]}$	0.11	1.01	23.6
	$\text{OS}_{\text{overlap}}$	0.11	1.00	21.0	$\text{TMLE}_{\text{overlap}}$	0.11	1.01	20.8
	g-comp	0.11	1.01	–	g-comp	0.11	1.01	–

We also replicate the bias and efficiency results based on the Kang and Schafer (2007) simulation design. Details of the data generating process are described in Benkeser et al. Similar to the first simulation design, we summarize in Table 3 the absolute bias, relative efficiency and empirical coverage corresponding to various one-step estimators and TMLE for estimating the ATE; the true ATE is zero. In both the one-step and TMLE frameworks, we consider the trimming stabilization with $\delta \in \{0.05, 0.15, 0.25\}$ and the overlap stabilization, beyond the standard and collaborative estimators. Given the true propensity score and outcome model are highly nonlinear functions of the observed covariates, these models are estimated by the flexible regression tool, HAL-MLE. Neither the collaborative one-step nor the stabilized one-step estimators improves the efficiency of the standard one-step estimator, regardless of sample sizes considered. However, both the collaborative and the stabilized TMLE provide modest efficiency benefits at the smallest sample size $n = 100$. In this scenario, the trimming stabilization with $\delta = 0.25$ reaches the same efficiency as the CTMLE and the non-targeted outcome regression estimates. Due to the non-negligible bias and the underestimation of true variance (via the non-cross-validated sample variance of the influence function), the coverage of all interval estimates are substantially lower than nominal.

5. CONCLUDING REMARKS

Benkeser et al. have made an important contribution by providing a super-efficient estimator of ATE with the adaptive propensity score. Under the assumption that the initial outcome model is consistently estimated, the proposed estimator leads to greater efficiency than its standard locally efficient counterpart, especially under lack of overlap. In our discussion, we have explored alternative methods to improve efficiency of the standard implementation through a stabilization factor $h(w)$. While the adaptive propensity score could eliminate the deleterious effect of instrumental variables or nonconfounders in estimating the assignment mechanism (avoiding the Z-bias), the stabilization factor directly removes or down-weights the influential observations with extreme propensity scores under lack of overlap. In both simulation designs, we observe an efficiency advantage over the standard one-step estimator and TMLE by including the stabilization factor. In particular, with an appropriately selected threshold δ, the trimming stabilization has potential to achieve the same super-efficiency as the collaborative estimator, but only within the TMLE framework. It would be desirable to automate the selection of δ for TMLE in a data-adaptive fashion along the lines of Bembom and van der Laan (2008) and Ju, Schwab and van der Laan (2019). Finally, we conjecture that both the collaborative and stabilized estimators would also have an efficiency advantage for estimating the pairwise ATEs with multiple treatments, via the formulation of an adaptive generalized propensity score (Imbens, 2000). The lack of overlap is a frequent challenge for estimating average causal effects with multiple treatments, and it would be worthwhile to explore the super-efficiency property of the collaborative estimators and stabilized estimators corresponding to multinomial trimming and generalized overlap specifications of $h(w)$ (Li and Li, 2019).

ACKNOWLEDGEMENTS

Li's work is partially supported by the National Institutes of Health (NIH) Common Fund through cooperative agreement U24AT009676 from the Office of Strategic Coordination within the Office of the NIH Director and cooperative agreement UH3DA047003 from the National Institute on Drug Abuse. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

REFERENCES

Bembom, O. and van der Laan, M. (2008). Data-adaptive selection of the truncation level for Inverse-Probability-of-Treatment-Weighted estimators. *Technical Report 230, Division of Biostatstics, University of California: Berkeley*.

Benkeser, D. and van der Laan, M. (2016). The highly adaptive lasso estimator. *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016* 689–696.

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.

Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710.

Ju, C., Schwab, J. and van der Laan, M. J. (2019). On adaptive propensity score truncation in causal inference. *Statistical Methods in Medical Research* **28** 1741–1760.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22** 523–539.

Li, F. and Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics* **4** 2389–2415.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113** 390–400.

Li, F., Thomas, L. E. and Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology* **1** 250–257.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. and van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* **21** 31–54.

van der Laan, M. (2017). A Generally Efficient Targeted Minimum Loss Based Estimator based on the Highly Adaptive Lasso. *International Journal of Biostatistics* **13** 1–35.