Navigated Weighting to Improve Inverse Probability Weighting for Missing Data Problems and Causal Inference*

Hiroto Katsumata†

*I thank Tomoya Sasaki, Yiqing Xu, Teppei Yamamoto, and Soichiro Yamauchi for their helpful comments and suggestions. This work was supported by JSPS Grant-in-Aid for JSPS Research Fellow Grant Number JP17J03508.

†JSPS Post-doctoral Research Fellow at Gakushuin University and Massachusetts Institute of Technology, hrt.katsumata@gmail.com

Abstract

The inverse probability weighting (IPW) is broadly utilized in dealing with missing data problems including causal inference. Under the conditional ignorability assumption, it utilizes the inverse probability of non-missingness weights to eliminate confounding without relying on correct specification of outcome models. Despite its theoretical appeal, the IPW has difficulty in estimating propensity scores to construct inverse probability weights. Existing research has found that the IPW can have an excessively large variance due to extreme estimated weights and be highly vulnerable to the misspecification of the propensity score model. To solve these problems, I propose an estimation method called the NAvigated WeighTing (NAWT), which conveys efficiency and reduces biases due to propensity score model misspecification by tweaking the score function of propensity scores estimation depending on a pre-specified estimand. The NAWT includes the standard IPW and the covariate balancing propensity score as special cases and can further be made robust to the propensity score model misspecification by incorporating covariate balance conditions. I investigate large-sample theoretical properties of the NAWT and demonstrate that it improves the performance of estimation through simulation studies. It is also ready to be combined with cutting-edge machine learning techniques. An easy-to-use R package which implements the NAWT is being developed.

Keywords: Missing data problem; Causal inference; Inverse probability weighting; Propensity score

1 Introduction

Missing data problems, where outcomes of some observations are not observed, are quite common problems in social sciences and public health. A simple example is the situation where we want to know the average outcome for a certain population and have data of a sample. Missing data problems even cover the fundamental problem in causal inference: we can only observe one of the potential outcomes with or without treatment and the other ones are missing. In dealing with these missing data problems, the inverse probability weighting (IPW) method is broadly utilized, which uses the propensity score defined as the probability of missing conditional on covariates to construct weights (Rosenbaum 1987). The resulting inverse probability weights eliminate confounding under the conditional ignorability assumption without relying on correct specification of outcome models. When the propensity score is utilized in causal inference, it is defined as the probability of receiving treatment given covariates. If the treatment is randomly assigned to units in a large sample, we can assume that the propensity score is the same across units and may not need to use the IPW. On the other hand, in observational studies where the treatment assignment is not randomized, we estimate the propensity score first and then use it for the IPW. Even in experimental studies, there are many relevant cases where randomization of treatment assignment alone cannot eliminate confounding and the IPW helps to unbiasedly estimate treatment effects in such cases as the mediation analysis, panel attrition problems, and the generalization of experimental results.

Although the IPW has broad applicability and nice theoretical properties, it has difficulty in estimating propensity scores and inverse probability weights. Existing research has found that it suffers from an excessively large variance due to extreme estimated weights and is highly vulnerable to propensity score model misspecification. To solve these problems, I propose an estimation method called the NAvigated WeighTing (NAWT) method, which conveys efficiency and reduces bias due to the propensity score model misspecification by tailoring the

score function of propensity scores estimation for a pre-specified estimand. Specifically, the NAWT tweaks the propensity score estimation by weighting the score function depending on the estimand. Though this modification increases variances in estimated propensity score, but it does reduce variances in the estimation of estimand itself. This result is somewhat counter-intuitive, but this type of phenomenon is not altogether unknown one. In the IPW, it is well-known that using estimated propensity scores reduces variances in the estimation of estimand compared to the estimation with true propensity scores. In this case, using estimated propensity scores, of course, has larger estimation variances in propensity scores than using true propensity scores which has no estimation variances in propensity scores. This suggests that increasing estimation variances in propensity scores does not necessarily increase and may possibly decrease estimation variances of estimand.

Therefore, the NAWT navigates the propensity score estimation for estimating a pre-specified estimand correctly and efficiently by tweaking the score function depending on the estimand at the cost of the accuracy in the propensity score estimation. Among many types of estimands, this study focuses on weighted average treatment effects (WATE), which is one of the most important and broadly utilized estimands, including the average treatment effects (ATE), average treatment effects on the treated (ATT), average treatment effects on the controlled (ATC), and average treatment effects for overlap population (ATO), for example (Li, Morgan, and Zaslavsky 2018). This study also focuses on estimation of the average outcome when we cannot observe outcomes of all units due to missing but can observe covariates of all units and calls it the MAR estimation because it relies on the missing-at-random assumption.

On the other hand, when using the IPW, applied researchers almost always estimate propensity scores via the standard logistic, or maybe the probit, regression irrespective of their estimand. The non-parametric propensity score estimation, though it is proved to be asymptotically efficient, are rarely utilized because it is difficult to use and does not necessarily work well with a finite sample. Since the NAWT is an extension of the standard parametric

IPW, it is as simple and easy to use as the standard IPW, but also more robust and efficient than the standard IPW like the IPW with the non-parametric propensity score estimation.

To get intuition why the NAWT is efficient than the standard IPW, consider the simple MAR estimation case where the estimand is average outcome but some units' outcome are missing. As explained later, this result can be easily extended to the case where the estimand is WATE. Here, I concentrate on a brief and intuitive explanation and details are shown in Section 2. In this case, the Horvitz–Thompson estimator uses $1/(1-\pi_i)$ for non-missing units i where π_i represents propensity scores of missing, and assume that we know the correct propensity score model. If we use the standard IPW, it has smaller variances for estimand than the IPW with true propensity scores but larger variances than the IPW with the non-parametric propensity score estimation with a large sample. To improve estimation, the NAWT weights the score function of the maximum likelihood estimation for propensity scores by monotonically increasing functions of the propensity score $\omega(\pi)$. This weighting makes estimated propensity scores for units with larger propensity scores to approach to the non-parametric estimates and those for units with smaller propensity scores to depart from the non-parametric estimates. Note that the latter are still consistent and variances for estimand get closer to those with true propensity scores. Since non-parametric estimates reduces variances for estimand more for units with larger propensity scores unless the conditional mean outcome correlates with propensity scores,[1] this weighting results in variance reduction for the estimand in total by largely decreasing variances for units with larger propensity scores and slightly increasing variances for units with smaller propensity scores.

The key idea of the NAWT is to improve the estimation by weighting the score function for the propensity score estimation depending on the estimand while being totally agnostic about outcome values nor models. This is different from the methods recently proposed in the framework of empirical likelihoods which incorporate outcome models for efficiency and other

1. See Appendix A for the proof.

desirable properties (Graham, De Xavier Pinto, and Egel 2012; Tan 2010). This also differs from another robust and efficient propensity scores estimation method, the covariate balancing propensity score (CBPS) method, which estimates propensity scores so that covariates are balanced between treatment groups (Imai and Ratkovic 2014; Zhao 2019). The CBPS does not exploit outcome values in estimating propensity scores but its robustness and efficiency depend on how close the linear combination of balanced covariates approximates the true outcome model (Fan et al. 2016). In contrast, the NAWT gains efficiency and robustness without assuming any outcome models nor exploiting outcome values in the propensity score estimation, which is in line with the original spirit of propensity scores (Rubin 2007).

The NAWT has following several attractive characteristics. First, it is more robust and efficient than the standard parametric IPW as I briefly explained above. Second, it can be applied generally to the IPW estimation with various estimators for the estimand, such as the Horvitz–Thompson estimator, weighted difference-in-means estimator, doubly robust IPW estimator, and weighted least squares estimator (Horvitz and Thompson 1952; Imai and Ratkovic 2014). Third, the NAWT is still a consistent estimator for propensity scores and does not shrink the coefficients, so that they are easy to interpret. Fourth, since the NAWT is a simple and intuitive extension of the standard IPW, it is easy to understand and utilize it. Fifth, its workflow is quite simple. In the standard IPW, we first estimate propensity scores via the standard logistic regression regardless of the estimand and then estimate the estimand by using estimated inverse probability weights. THe NAWT adds only one simple step before this procedure and slightly modifies the first step, where we first define the estimand, then estimate propensity scores by the weighted score function for the specific estimand, and estimate the estimand using estimated inverse probability weights. Sixth, basically, no hyper-parameter tuning is required, though we can tune it to improve estimation as explained later in Section 3. Lastly, it can be combined with other attractive methods, such as the CBPS, kernel balancing, ridge regression, and LASSO (Zhao 2019).

The rest of the paper is organized as follows. First, I propose the navigated weighting (NAWT) and investigate its large sample theoretical properties in Section 2. In Section 3, I extend the NAWT in two directions. The first extension incorporates covariate balancing conditions and the second extension considers the hyper-parameter tuning to improve estimation further. In Section 4, I conduct simulation studies to demonstrate that the NAWT dramatically improves the standard IPW in efficiency and robustness to propensity scores model misspecification and it also outperforms the IPW with the CBPS in terms of the bias when the propensity score model is misspecified. The final section concludes and discusses future research directions.

2 Proposed methodology

Suppose we have a simple random sample of n units ($i = 1, 2, \ldots, n$) from a population. For each unit i, we observe a k-dimensional vector of pretreatment covariates $\mathbf{x_i} \in \mathbb{R}^k$. We consider following two cases.

Missing data problem The first case is the missing data problem where each unit has an outcome $y_i \in \mathbb{R}$ but we cannot observe outcomes of all the units and a missingness indicator $m_i \in \{0, 1\}$ is introduced to denote the missing units as $m_i = 1$ and non-missing units as $m_i = 0$. Note that we can observe covairates of all the units, including the missing units. The estimand here is the average outcome (MAR):

$$\mu \equiv \mathbb{E}[y_i]. \tag{1}$$

To estimate this, the inverse probability weighting (IPW) usually makes the following assumptions. The first one is the conditional ignorability of missing assumption, or the missing-at-random assumption, that the missing is ignorable conditional on the observed covariates and this implies that the missing and non-missing units have the same expected outcome

conditional on the covariates.

Assumption 1 (Conditional ignorability of missing)

$$m_i \perp\!\!\!\perp y_i \mid \mathbf{x}_i = \mathbf{x}. \tag{2}$$

Hence, covariates which are associated with both outcomes and missingness should be conditioned, and those associated only with outcomes can be conditioned. However, we do not have to, and actually should not, condition on the covariates associated only with missingness due to the instrument variable problem (Bhattacharya and Vogt 2012; Brookhart et al. 2006).

The IPW produces the pseudo-population that would have been observed if there had been no missingness by re-weighting non-missing units with the inverse probability of non-missing conditional on the covairates $w(\mathbf{x}) \equiv 1/(1-\pi(\mathbf{x}))$, where $\pi(\mathbf{x}) \equiv \Pr(m_i = 1 \mid \mathbf{x}_i = \mathbf{x})$ represents the propensity score for missing given covariates. Here, we need the positivity assumption that the probability of non-missing is bounded away from 0.

Assumption 2 (Positivity of the non-missing probability)

$$1 - \pi(\mathbf{x}) > 0. \tag{3}$$

Causal Inference The second case is the causal inference where each unit has two potential outcomes $y_i(0), y_i(1) \in \mathbb{R}$, only one of which $y_i(t_i)$ is realized and observed depending on the binary treatment $t_i \in \{0, 1\}$ the unit gets. The estimand here is the weighted average treatment effects (WATE):

$$\tau_{\text{WATE}} \equiv \frac{\int \mathbb{E}[y_i(1) - y_i(0) \mid \mathbf{x}_i = \mathbf{x}] h(\mathbf{x}) dF(\mathbf{x})}{\int h(\mathbf{x}) dF(\mathbf{x})}, \tag{4}$$

where $h(\cdot)$ is a known function of covariates. This includes the average treatment effects (ATE):

$$\tau_{\text{ATE}} \equiv \mathbb{E}[y_i(1) - y_i(0)], \tag{5}$$

the average treatment effects on the treated (ATT):

$$\tau_{\text{ATT}} \equiv \mathbb{E}[y_i(1) - y_i(0) \mid t_i = 1], \tag{6}$$

and the average treatment effects on the controlled (ATC):

$$\tau_{\text{ATC}} \equiv \mathbb{E}[y_i(1) - y_i(0) \mid t_i = 0], \tag{7}$$

as the special case depending on the choice of $h(\cdot)$ and this study focuses on the ATE and ATT.

In causal inference, the IPW needs the conditional ignorability of treatment assumption that the treatment is ignorable conditional on the observed covariates, which implies that units with and without treatment have the same expected potential outcomes both with and without treatment conditional on the covariates.

Assumption 3 (Conditional ignorability of treatment)

$$t_i \perp\!\!\!\perp \{y_i(1), y_i(0)\} \mid \mathbf{x}_i = \mathbf{x}. \tag{8}$$

Again, covariates associated only with outcomes or with both outcomes and missingness can/should be conditioned but covariates associated only with missingness should not be conditioned.

The IPW re-weights units with the inverse probability weights depending on the estimand. When the estimand is the ATE, weights are $w(1, \mathbf{x}) \equiv 1/\pi(\mathbf{x})$ for treated units and $w(0, \mathbf{x}) \equiv 1/(1 - \pi(\mathbf{x}))$ for controlled units, where $\pi(\mathbf{x}) \equiv \Pr(t_i = 1 \mid \mathbf{x}_i = \mathbf{x})$ represents the propensity score for treatment given covariates. The positivity assumption that the probability of non-missing is bounded away from 0 and 1 needs to be satisfied.

Assumption 4 (Positivity of the treatment probability)

$$0 < \pi(\mathbf{x}) < 1. \tag{9}$$

This second case (causal inference) can be seen as the extension of the first case (missing data problem) because, in causal inference, we cannot observe the potential outcomes with

treatment $y_i(1)$ for controlled units $t_i = 0$ nor those without treatment $y_i(0)$ for treated units $t_i = 1$. Since the estimand in the causal inference is a function of the estimand in the missing data problem, e.g. $\tau_{\text{ATE}} = \mu_1 - \mu_0$ where μ_1 represents average potential outcome with treatment and and μ_0 represents average potential outcome without treatment, we can estimate the estimand in causal inference by estimating both the average potential outcomes with and without treatment. Or conversely, we can think of the missing data problem as the special case of the ATT estimation where the potential outcomes with treatment are 0 for all units. For simplicity, the following explanation mainly focuses on the missing data problem case but the same logic is applicable to causal inference case.

In both cases, when the propensity score is unknown, it must be estimated, and even when it is known, using estimated propensity scores improves estimation of the estimand (Hahn 1998; Hirano, Imbens, and Ridder 2003). Though non-parametric propensity score estimation is proved to be asymptotically efficient, researchers typically utilize parametric model $\pi_\beta(\mathbf{x_i})$ for the propensity score estimation because the non-parametric estimation is difficult to use and does not necessarily work well with a finite sample,

$$\Pr(t_i = 1 \mid \mathbf{x}_i) \equiv \pi(\mathbf{x_i}) = \pi_\beta(\mathbf{x_i}), \tag{10}$$

where $\beta \in B$ is a k-dimensional vector of unknown parameters. The standard IPW utilizes the logistic model

$$\pi_\beta(\mathbf{x_i}) = \frac{\exp(\mathbf{x_i}^\mathsf{T}\beta)}{1 + \exp(\mathbf{x_i}^\mathsf{T}\beta)} \tag{11}$$

and estimates parameters β by the maximum likelihood estimation (MLE), where the log-likelihood function

$$\hat{\beta}_{\text{MLE}} = \arg\max_{\beta \in B} \sum_{i=1}^{n} m_i \log(\pi_\beta(\mathbf{x_i})) + (1 - m_i) \log(1 - \pi_\beta(\mathbf{x_i})) \tag{12}$$

is maximized. Under the assumption that $\pi_\beta(\cdot)$ is twice continuously differentiable with respect

to β, the first order condition is

$$\frac{1}{n}\sum_{i=1}^{n} s_{\mathrm{MLE}}(\beta, m_i, \mathbf{x_i}) = 0 \tag{13}$$

$$s_{\mathrm{MLE}}(\beta, m_i, \mathbf{x_i}) = \left(\frac{m_i}{\pi_\beta(\mathbf{x_i})} - \frac{1-m_i}{1-\pi_\beta(\mathbf{x_i})}\right)\pi'_\beta(\mathbf{x_i}), \tag{14}$$

where $\pi'_\beta(\mathbf{x_i}) = \partial\pi_\beta(\mathbf{x_i})/\partial\beta^{\mathsf{T}}$.

This MLE estimation is the best unbiased estimation for β when the propensity score model is correctly specified and justified even when the model is misspecified as it minimizes the Kullback-Leibler distance between the true model and misspecified model. However, it does not necessarily imply that this MLE estimation is the best parametric model for the estimation of the estimand μ or τ. As I explain later, we can actually gain efficiency and robustness by tweaking the score function for the propensity score estimation.

To see how and why this tweaking works, we first need to examine estimators for the estimand. Among many types of IPW estimators, this study focuses on the Horvitz–Thompson estimator as it is the most simple one but the same logic is applicable to other IPW estimators, such as weighted difference-in-means estimator, doubly robust IPW estimator, and weighted least squares estimator (See Subsection 2.2). These estimators utilizes inverse probability of propensity score weights depending on the estimand, which is summarized in Table 1 where $\hat{\tau}_{\mathrm{HT,ATE}}$ is the estimator for the ATE, $\hat{\tau}_{\mathrm{HT,ATT}}$ is the estimator for the ATT, $\hat{\mu}_{\mathrm{HT}}$ is the estimator for the MAR, $\hat{\tau}_{\mathrm{HT,WATE}}$ is the estimator for the WATE, and $n_1 = \sum m_i$. Table 1 also shows the relative weights which indicates the relative weights both for units $m_i = 1$ and $m_i = 0$ when the weights in the ATE estimator is 1. Interestingly, Relative weights for the MAR can be thought of as $1/\pi_\beta(\mathbf{x}_i)$, which is the same as the ones for the ATT estimator. This is because the estimation of the average outcome (MAR) can be seen as the estimation of the average outcome for $m_i = 1$ and $m_i = 0$, which can be estimated by estimating average outcome for $m_i = 1$ with non-missing units $m_i = 0$ with inverse probability weights $\pi_\beta(\mathbf{x}_i)/(1-\pi_\beta(\mathbf{x}_i))$ and estimating average outcome for $m_i = 0$ with non-missing units $m_i = 0$ with (inverse

Table 1: Inverse probability weighting estimators

Estimand	The Horvitz–Thompson Estimator	Weights $m_i = 1$	Weights $m_i = 0$	Relative weights
ATE	$\hat{\tau}_{\text{HT,ATE}} = \frac{1}{n}\sum_{i=1}^{n} \frac{m_i y_i}{\hat{\pi}_\beta(\mathbf{x_i})} - \frac{(1-m_i)y_i}{1-\hat{\pi}_\beta(\mathbf{x_i})}$	$\frac{1}{\hat{\pi}_\beta(\mathbf{x}_i)}$	$\frac{1}{1-\hat{\pi}_\beta(\mathbf{x}_i)}$	1
ATT	$\hat{\tau}_{\text{HT,ATT}} = \frac{1}{n_1}\sum_{i=1}^{n} m_i y_i - \frac{(1-m_i)\hat{\pi}_\beta(\mathbf{x_i})y_i}{1-\hat{\pi}_\beta(\mathbf{x_i})}$	1	$\frac{\hat{\pi}_\beta(\mathbf{x}_i)}{1-\hat{\pi}_\beta(\mathbf{x}_i)}$	$\hat{\pi}_\beta(\mathbf{x}_i)$
MAR	$\hat{\mu}_{\text{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-m_i)y_i}{1-\hat{\pi}_\beta(\mathbf{x_i})}$	0	$\frac{1}{1-\hat{\pi}_\beta(\mathbf{x}_i)}$	$\hat{\pi}_\beta(\mathbf{x}_i)$
WATE	$\hat{\tau}_{\text{HT,WATE}} = \frac{1}{n}\sum_{i=1}^{n} \frac{m_i h(\mathbf{x}_i) y_i}{\hat{\pi}_\beta(\mathbf{x_i})} - \frac{(1-m_i)h(\mathbf{x}_i)y_i}{1-\hat{\pi}_\beta(\mathbf{x_i})}$	$\frac{h(\mathbf{x}_i)}{\hat{\pi}_\beta(\mathbf{x}_i)}$	$\frac{h(\mathbf{x}_i)}{1-\hat{\pi}_\beta(\mathbf{x}_i)}$	$h(\mathbf{x}_i)$

probability) weights 1:

$$\hat{\mu}_{\text{HT}} = \frac{1}{n}\sum_{i=1}^{n}(1-m_i)\left(\frac{\hat{\pi}_\beta(\mathbf{x_i})}{1-\hat{\pi}_\beta(\mathbf{x_i})}+1\right)y_i. \tag{15}$$

This implies that we should use the same propensity score estimation for the ATT and MAR, but not for the ATE and MAR. This implication is quite simple but important, and sometimes ignored as we will see in Section 4.

Table 1 also demonstrates that the larger these inverse probability weights are, the more influence those units have on the estimation of estimand. Since the non-parametric propensity score estimation is asymptotically efficient, this implies that we can enjoy great efficiency gain when we estimate propensity scores which approximates the non-parametric propensity score estimation for units with large weights. To gain this efficiency, I propose the NAvigated WeighTing (NAWT), which puts more weights on units who should have large weights in the IPW estimator when we estimate propensity scores by tweaking the score function. On the other hand, this makes estimated weights for units with small weights to get closer to true propensity scores and variances increase for these units. Thus, the NAWT appropriately tweaks the score function to balance these advantage and disadvantage of putting more importance on units which should have large inverse probability weights in propensity score

estimation.

In general, the NAWT weights the score function by a function of propensity scores $\omega(\pi_\beta(\mathbf{x}_i))$. In this study, the NAWT utilizes a power function of propensity scores $\pi_\beta(\mathbf{x}_i)^\alpha$ to weight the score function for the ATT and MAR estimation, where $\alpha \geq 0$, and and the weights for the ATE estimation is discussed in Subsection 2.4. This simple specification has several attractive characteristics. First, it includes the standard IPW estimation, which uses the standard MLE for the propensity score estimation, as the special case where $\alpha = 0$. Second, it asymptotically more efficient than the standard IPW when $\alpha = 2$ irrespective of the covariate distribution, which implies we need not tune this hyper-parameter α to gain efficiency. Third, we can easily tune this hyper-parameter α to improve the NAWT for a specific data distribution, which is explained in Section 3 as an extension.

The general form of the NAWT also includes the just-identified covariate balancing propensity score (CBPS) with the logistic model as the special case where $\omega(\pi_\beta(\mathbf{x}_i)) = 1/(\pi_\beta(\mathbf{x}_i)(1 - \pi_\beta(\mathbf{x}_i)))$ for the ATE estimation and $\omega(\pi_\beta(\mathbf{x}_i)) = 1/(1 - \pi_\beta(\mathbf{x}_i))$ for the ATT estimation:

$$s_{\text{CBPS,ATE}}(\beta, m_i, \mathbf{x_i}) = \left(\frac{m_i}{\pi_\beta(\mathbf{x_i})} - \frac{1 - m_i}{1 - \pi_\beta(\mathbf{x_i})} \right) \mathbf{x_i} \tag{16}$$

$$s_{\text{CBPS,ATT}}(\beta, m_i, \mathbf{x_i}) = \left(m_i - \frac{(1 - m_i)\pi_\beta(\mathbf{x_i})}{1 - \pi_\beta(\mathbf{x_i})} \right) \mathbf{x_i}, \tag{17}$$

considering $\pi'_\beta(\mathbf{x_i}) = \pi_\beta(\mathbf{x_i})(1 - \pi_\beta(\mathbf{x_i}))\mathbf{x_i}$ for the logistic model. This demonstrates that the CBPS tweaks the score function in the same spirits as the NAWT so that it puts more importance on units which should have large estimated inverse probability weights. However, the CBPS has never been justified from this perspective nor its weights for the score function are appropriate as a kind of the NAWT, and the over-identified CBPS uses the unweighted score of the standard MLE (14) as its score condition, which is not in line with the NAWT. Although the CBPS has not been justified as the technique to tweak the score function, considering it as a special case of the NAWT help understand why the CBPS gains efficiency and robustness even when the true outcome model is not a linear combination of covariates

balanced via the CBPS.

2.1 The navigated weighting

The nawt weights the score function for the propensity score estimation by a function of propensity scores $\omega(\pi_\beta(\mathbf{x}_i))$. In this section, I focus on the particular choice of the function $\pi_\beta(\mathbf{x}_i)^\alpha$ to weight the score function of the logistic model for the MAR estimation for simplicity. The weighted score function is:

$$s_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \equiv \left(\frac{m_i}{\pi_\beta(\mathbf{x_i})} - \frac{1 - m_i}{1 - \pi_\beta(\mathbf{x_i})} \right) \pi_\beta(\mathbf{x}_i)^\alpha \pi'_\beta(\mathbf{x_i}) \tag{18}$$

$$= \left(m_i - \pi_\beta(\mathbf{x}_i)\right) \pi_\beta(\mathbf{x}_i)^\alpha \mathbf{x}_i. \tag{19}$$

I introduce the pseudo-log-likelihood which integrates the score with respect to β:

$$l_{\mathrm{MAR}}(\beta, \mathbf{m}, \mathbf{X}) \equiv \int \sum_{i=1}^{n} s_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) d\beta \tag{20}$$

$$= \sum_{i=1}^{n} \pi_\beta(\mathbf{x}_i)^\alpha \left(\frac{m_i}{\alpha} - \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i) {}_2F_1(1, 1 + \alpha, 2 + \alpha, \pi_\beta(\mathbf{x}_i))}{1 + \alpha} \right), \tag{21}$$

where ${}_2F_1(a, b, c, z)$ is a hyper-geometric function, $\mathbf{m}$ is a vector of the missing indicator m_i, and $\mathbf{X}$ is a matrix of the covariates $\mathbf{x}_i$. We can estimate β via the M-estimation by maximizing the pseudo-log-likelihood $l_{\mathrm{MAR}}(\beta, \mathbf{m}, \mathbf{X})$:

$$\hat{\beta}_{\mathrm{MAR}} = \arg\max_{\beta \in B} l_{\mathrm{MAR}}(\beta, \mathbf{m}, \mathbf{X}). \tag{22}$$

As the properties of the M-estimator, $\hat{\beta}_{\mathrm{MAR}}$ is consistently estimated.

$$\hat{\beta}_{\mathrm{MAR}} \xrightarrow{p} \beta, \tag{23}$$

which implies that the NAWT does not shrink coefficients toward 0 and thus they are easy to interpret. The asymptotic distribution of $\hat{\beta}_{\mathrm{MAR}}$ is:

$$\sqrt{n}(\hat{\beta}_{\mathrm{MAR}} - \beta) \underset{d}{\to} \mathcal{N}(0, \mathbf{H}_{\beta\beta}^{-1} \mathbf{\Sigma}_{\beta\beta} \mathbf{H}_{\beta\beta}^{-1}), \tag{24}$$

where hessian $\mathbf{H}_{\beta\beta}$ is

$$\begin{aligned}\mathbf{H}_{\beta\beta} &\equiv \mathbb{E}\left[\frac{\partial l_{\text{MAR}}(\beta, \mathbf{m}, \mathbf{X})}{\partial \beta^{\mathsf{T}}}\right] \\ &= \mathbb{E}\left[-((1+\alpha)\pi_\beta(\mathbf{x}_i) - \alpha m_i)\pi_\beta(\mathbf{x}_i)^\alpha(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right] \\ &= \mathbb{E}\left[-\pi_\beta(\mathbf{x}_i)^{1+\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]\end{aligned} \tag{25}$$

and $\mathbf{\Sigma}_{\beta\beta}$ is

$$\begin{aligned}\mathbf{\Sigma}_{\beta\beta} &\equiv \mathbb{E}\left[s_{\text{MAR}}(\beta, m_i, \mathbf{x_i}) s_{\text{MAR}}(\beta, m_i, \mathbf{x_i})^{\mathsf{T}}\right] \\ &= \mathbb{E}\left[((m_i - \pi_\beta(\mathbf{x}_i))\pi_\beta(\mathbf{x}_i)^\alpha)^2 \mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right] \\ &= \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^{1+2\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right].\end{aligned} \tag{26}$$

To convey the intuition behind the NAWT, Figure 1 presents the (pseudo-) log-likelihoods of the NAWT for the ATT and MAR estimation with $\alpha = 2$ and standard logistic regression in the left panel, and the expected (pseudo-) log-likelihoods of the NAWT with $\alpha = 2$ and standard logistic regression, respectively in the center and right panels. In the left panel, the green and purple curves represent the (pseudo-) log-likelihood for the missing units and the red and blue curves represent the log-likelihood for the non-missing units whose estimated propensity scores are shown along the x-axis from 0.01 to 0.99. The (pseudo-) log-likelihoods for missing units are

$$l_{\text{MAR},1}(\pi_\beta(\mathbf{x}_i)) = \frac{\pi_\beta(\mathbf{x}_i)^\alpha}{\alpha} \tag{27}$$

$$l_{\text{MLE},1}(\pi_\beta(\mathbf{x}_i)) = \log(\pi_\beta(\mathbf{x_i})), \tag{28}$$

the (pseudo-) log-likelihoods for non-missing units are

$$l_{\text{MAR},0}(\pi_\beta(\mathbf{x}_i)) = -\frac{\pi_\beta(\mathbf{x}_i)^{1+\alpha}{}_2F_1(1, 1+\alpha, 2+\alpha, \pi_\beta(\mathbf{x}_i))}{1+\alpha} \tag{29}$$

$$l_{\text{MLE},0}(\pi_\beta(\mathbf{x}_i)) = \log(1-\pi_\beta(\mathbf{x_i})), \tag{30}$$

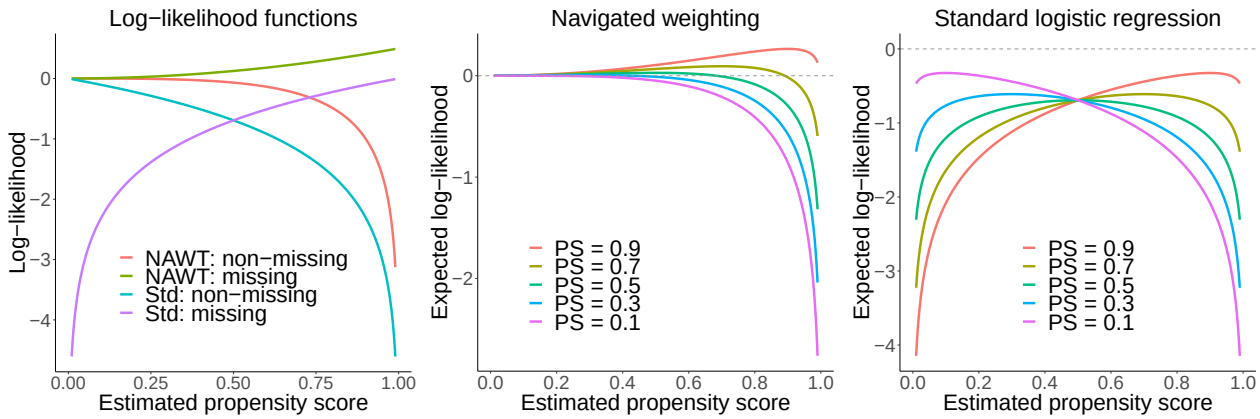

Figure 1: The left panel of the figure shows the (pseudo-) log-likelihoods of the navigated weighting (NAWT) for the ATT and MAR estimation with $\alpha = 2$ and standard logistic regression, and the center and right panels show the expected (pseudo-) log-likelihoods of the NAWT with $\alpha = 2$ and standard logistic regression, respectively. In the left panel, the green and purple curves represent the (pseudo-) log-likelihood for the missing units and the red and blue curves represent the (pseudo-) log-likelihood for the non-missing units whose estimated propensity scores are shown along the x-axis from 0.01 to 0.99. In the center and right panels, the expected (pseudo-) log-likelihood functions for the units whose true missing propensity scores are 0.9, 0.7, 0.5, 0.3, and 0.1 are represented by the red, yellow, green, blue, and purple curves, respectively, along with estimated propensity scores on the x-axis.

and the expected (pseudo-) log-likelihoods are

$$l_{\mathrm{MAR},e}(\pi_\beta(\mathbf{x}_i)) = \pi_\beta(\mathbf{x}_i)\; l_{\mathrm{MAR},1}(\pi_\beta(\mathbf{x}_i)) + (1-\pi_\beta(\mathbf{x}_i))\; l_{\mathrm{MAR},0}(\pi_\beta(\mathbf{x}_i)) \tag{31}$$

$$l_{\mathrm{MLE},e}(\pi_\beta(\mathbf{x}_i)) = \pi_\beta(\mathbf{x}_i)\; l_{\mathrm{MLE},1}(\pi_\beta(\mathbf{x}_i)) + (1-\pi_\beta(\mathbf{x}_i))\; l_{\mathrm{MLE},0}(\pi_\beta(\mathbf{x}_i)). \tag{32}$$

In the left panel, the green curve which represents the pseudo-log-likelihood of the NAWT for missing units does not increase much even when estimated propensity scores increase, which demonstrates that the NAWT for the MAR estimation is influenced by the missing units quite a little. In contrast, the pseudo-log-likelihood for non-missing units, the red curve, steeply decreases as estimated propensity scores approach 1, which indicates that the estimation of the NAWT is dominated by the non-missing units. On the other hand, the standard MLE symmetrically places importance on missing and non-missing units as shown in its symmetric curves in the left and right panels.

In the center and right panels, the expected (pseudo-) log-likelihood functions for the units whose true missing propensity scores are 0.9, 0.7, 0.5, 0.3, and 0.1 are represented by the red, yellow, green, blue, and purple curves, respectively, along with estimated propensity scores on the x-axis. the center panel shows that the NAWT places more weights on units with *large estimated propensity scores* and its estimation is also dominated by units with *small true propensity scores* as they heavily drop where estimated propensity scores are large, which may sound counter-intuitive. However, these results are reasonable because the probability of $m_i = 0$ for units with small true propensity scores are much higher than units with large true propensity scores and thus it heavily decreases the pseudo-log-likelihood if their propensity scores are estimated as large. These results imply that the estimation of the NAWT for the MAR estimation is anchored by units with small true propensity scores so that they do not have large inverse probability weights, which leads to robust and efficient estimation.

2.2 The Inverse probability weighting estimators with the navigated weighting

The NAWT can be combined with various types of estimators for the estimand, such as the Horvitz–Thompson estimator (HT), weighted difference-in-means estimator (IPW), doubly robust IPW estimator (DR), and weighted least squares estimator (WLS):

$$\hat{\mu}_{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-m_i)y_i}{1-\hat{\pi}_\beta(\mathbf{x_i})} \tag{33}$$

$$\hat{\mu}_{\mathrm{IPW}} = \sum_{i=1}^{n}\frac{(1-m_i)y_i}{1-\hat{\pi}_\beta(\mathbf{x_i})}\Bigg/\sum_{i=1}^{n}\frac{1-m_i}{1-\hat{\pi}_\beta(\mathbf{x_i})} \tag{34}$$

$$\hat{\mu}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x_i}^{\mathsf{T}}\hat{\gamma}_{\mathrm{OLS}} + \frac{(1-m_i)(y_i - \mathbf{x_i}^{\mathsf{T}}\hat{\gamma}_{\mathrm{OLS}})}{1-\hat{\pi}_\beta(\mathbf{x_i})}\right), \quad \hat{\gamma}_{\mathrm{OLS}} = \frac{\sum_{i=1}^{n}(1-m_i)\mathbf{x_i}y_i}{\sum_{i=1}^{n}(1-m_i)\mathbf{x_i}\mathbf{x_i}^{\mathsf{T}}} \tag{35}$$

$$\hat{\mu}_{\mathrm{WLS}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}^{\mathsf{T}}\hat{\gamma}_{\mathrm{WLS}}, \quad \hat{\gamma}_{\mathrm{WLS}} = \sum_{i=1}^{n}\frac{(1-m_i)\mathbf{x_i}y_i}{1-\hat{\pi}_\beta(\mathbf{x_i})}\Bigg/\sum_{i=1}^{n}\frac{(1-m_i)\mathbf{x_i}\mathbf{x_i}^{\mathsf{T}}}{1-\hat{\pi}_\beta(\mathbf{x_i})}. \tag{36}$$

Using these estimators, we can also estimate the estimand via the M-estimation. The joint

conditions for estimating the MAR are

$$\sum_{i=1}^{n} s_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) = 0 \tag{37}$$

$$\sum_{i=1}^{n} q_{\mathrm{MAR}}(\mu, \beta, m_i, \mathbf{x_i}) = 0 \tag{38}$$

and those for the ATT are

$$\sum_{i=1}^{n} s_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) = 0 \tag{39}$$

$$\sum_{i=1}^{n} q_{\mathrm{ATT}}(\tau_{\mathrm{ATT}}, \beta, m_i, \mathbf{x_i}) = 0, \tag{40}$$

where $q_{\mathrm{MAR}}(\mu, \beta, m_i, \mathbf{x_i})$ and $q_{\mathrm{ATT}}(\tau_{\mathrm{ATT}}, \beta, m_i, \mathbf{x_i})$ depend on the estimator for the estimand. Here, we consider the following two popular estimators: The Horvitz–Thompson estimator

$$q_{\mathrm{MAR}}(\mu, \beta, m_i, \mathbf{x_i}) \equiv \frac{(1 - m_i) y_i}{1 - \pi_\beta(\mathbf{x}_i)} - \mu \tag{41}$$

$$q_{\mathrm{ATT}}(\tau_{\mathrm{ATT}}, \beta, m_i, \mathbf{x_i}) \equiv \frac{n}{n_1} \left(m_i y_i - \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i) y_i}{1 - \pi_\beta(\mathbf{x}_i)} \right) - \tau_{\mathrm{ATT}} \tag{42}$$

and weighted difference-in-means estimator

$$q_{\mathrm{MAR}}(\mu, \beta, m_i, \mathbf{x_i}) \equiv \frac{(1 - m_i)(y_i - \mu)}{1 - \pi_\beta(\mathbf{x}_i)} \tag{43}$$

$$q_{\mathrm{ATT}}(\tau_{\mathrm{ATT}}, \beta, m_i, \mathbf{x_i}) \equiv \frac{n}{n_1} m_i y_i - n \left(\sum_{i=1}^{n} \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i)}{1 - \pi_\beta(\mathbf{x}_i)} \right)^{-1} \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i) y_i}{1 - \pi_\beta(\mathbf{x}_i)} - \tau_{\mathrm{ATT}} \tag{44}$$

$$= \frac{n}{n_1} m_i \left(y_i - \mu_1 \right) - n \left(\sum_{i=1}^{n} \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i)}{1 - \pi_\beta(\mathbf{x}_i)} \right)^{-1} \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i)(y_i - \mu_0)}{1 - \pi_\beta(\mathbf{x}_i)} \tag{45}$$

where

$$\mu_1 \equiv \sum_{i=1}^{n} \frac{1}{n_1} m_i y_i \tag{46}$$

$$\mu_0 \equiv \left(\sum_{i=1}^{n} \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i)}{1 - \pi_\beta(\mathbf{x}_i)} \right)^{-1} \frac{(1 - m_i)\pi_\beta(\mathbf{x}_i) y_i}{1 - \pi_\beta(\mathbf{x}_i)}. \tag{47}$$

As we can see, the weighted difference-in-means estimator is quite similar to the Horvitz–Thompson

estimator.

The estimand as well as β can be consistently estimated.

$$\hat{\beta}_{\mathrm{MAR}} \xrightarrow{p} \beta \tag{48}$$

$$\hat{\mu} \xrightarrow{p} \mu \tag{49}$$

$$\hat{\tau}_{\mathrm{ATT}} \xrightarrow{p} \tau_{\mathrm{ATT}}. \tag{50}$$

Asymptotic distribution (MAR) The asymptotic distribution of the NAWT for the MAR estimation is

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{\mathrm{MAR}} - \beta \\ \hat{\mu} - \mu \end{pmatrix} \xrightarrow[d]{} \mathcal{N}(0, \mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}), \tag{51}$$

where hessian $\mathbf{H}$ is

$$\mathbf{H} \equiv \begin{pmatrix} \mathbf{H}_{\beta\beta} & 0 \\ \mathbf{H}_{\mu\beta} & \mathbf{H}_{\mu\mu} \end{pmatrix}, \tag{52}$$

where $\mathbf{H}_{\mu\beta}$ is

$$\mathbf{H}_{\mu\beta} \equiv \mathbb{E}\left[\frac{\partial q_{\mathrm{MAR}}(\mu, \beta, \mathbf{m}, \mathbf{X})}{\partial \beta}\right], \tag{53}$$

which is for the Horvitz–Thompson estimator

$$\begin{aligned} \mathbf{H}_{\mu\beta} &= \mathbb{E}\left[\frac{(1-m_i)\pi_\beta(\mathbf{x}_i)y_i}{1-\pi_\beta(\mathbf{x}_i)}\mathbf{x}_i\right] \\ &= \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i\right], \end{aligned} \tag{54}$$

and for the weighted difference-in-means estimator

$$\begin{aligned} \mathbf{H}_{\mu\beta} &= \mathbb{E}\left[\frac{(1-m_i)\pi_\beta(\mathbf{x}_i)(y_i-\mu)}{1-\pi_\beta(\mathbf{x}_i)}\mathbf{x}_i\right] \\ &= \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)(y_i-\mu)\mathbf{x}_i\right], \end{aligned} \tag{55}$$

and $\mathbf{H}_{\mu\mu}$ is

$$\mathbf{H}_{\mu\mu} \equiv \mathbb{E}\left[\frac{\partial q_{\mathrm{MAR}}(\mu, \beta, \mathbf{m}, \mathbf{X})}{\partial \mu}\right], \tag{56}$$

which is for the Horvitz–Thompson estimator

$$\mathbf{H}_{\mu\mu} = -1, \tag{57}$$

and for the weighted difference-in-means estimator

$$\begin{aligned} \mathbf{H}_{\mu\mu} &= \mathbb{E}\left[-\frac{1-m_i}{1-\pi_\beta(\mathbf{x}_i)}\right] \\ &= -1, \end{aligned} \tag{58}$$

and $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} \equiv \begin{pmatrix} \mathbf{\Sigma}_{\beta\beta} & \mathbf{\Sigma}_{\beta\mu} \\ \mathbf{\Sigma}_{\mu\beta} & \mathbf{\Sigma}_{\mu\mu} \end{pmatrix} \tag{59}$$

$$= \mathbb{E}\left[\left(s^{\mathrm{ATT}}(\beta, m_i, \mathbf{x_i}),\ q_{\mathrm{MAR}}(\mu, \beta, m_i, \mathbf{x_i})\right)\left(s^{\mathrm{ATT}}(\beta, m_i, \mathbf{x_i}),\ q_{\mathrm{MAR}}(\mu, \beta, m_i, \mathbf{x_i})\right)^{\mathsf{T}}\right], \tag{60}$$

where $\mathbf{\Sigma}_{\beta\mu}$ and $\mathbf{\Sigma}_{\mu\beta}$ are

$$\mathbf{\Sigma}_{\beta\mu} = (\mathbf{\Sigma}_{\mu\beta})^{\mathsf{T}}. \tag{61}$$

Asymptotic distribution (ATT) The asymptotic distribution of the NAWT for the ATT estimation is

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{\mathrm{MAR}} - \beta \\ \hat{\tau}_{\mathrm{ATT}} - \tau \end{pmatrix} \underset{d}{\to} \mathcal{N}(0, \mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}), \tag{62}$$

where hessian $\mathbf{H}$ is

$$\mathbf{H} \equiv \begin{pmatrix} \mathbf{H}_{\beta\beta} & 0 \\ \mathbf{H}_{\tau\beta} & \mathbf{H}_{\tau\tau} \end{pmatrix}, \tag{63}$$

where $\mathbf{H}_{\tau\beta}$ is

$$\mathbf{H}_{\tau\beta} \equiv \mathbb{E}\left[\frac{\partial q_{\mathrm{ATT}}(\tau_{\mathrm{ATT}}, \beta, \mathbf{m}, \mathbf{X})}{\partial \beta}\right], \tag{64}$$

which is for the Horvitz–Thompson estimator

$$\begin{aligned}\mathbf{H}_{\tau\beta} &= -\frac{n}{n_1}\mathbb{E}\left[\frac{(1-m_i)\pi_\beta(\mathbf{x}_i)y_i}{1-\pi_\beta(\mathbf{x}_i)}\mathbf{x}_i\right] \\ &= -\frac{n}{n_1}\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i\right],\end{aligned} \tag{65}$$

and for the weighted difference-in-means estimator

$$\begin{aligned}\mathbf{H}_{\tau\beta} &= -n\mathbb{E}\left[\frac{1}{\sum\frac{(1-m_i)\pi_\beta(\mathbf{x}_i)}{1-\pi_\beta(\mathbf{x}_i)}}\frac{(1-m_i)\pi_\beta(\mathbf{x}_i)(y_i-\mu_0)}{1-\pi_\beta(\mathbf{x}_i)}\mathbf{x}_i\right] \\ &= -\frac{n}{n_1}\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)(y_i-\mu_0)\mathbf{x}_i\right],\end{aligned} \tag{66}$$

and $\mathbf{H}_{\tau\tau}$ is

$$\mathbf{H}_{\tau\tau} \equiv \mathbb{E}\left[\frac{\partial q_{\text{ATT}}(\tau_{\text{ATT}},\beta,\mathbf{m},\mathbf{X})}{\partial\tau}\right], \tag{67}$$

which is for the Horvitz–Thompson estimator

$$\mathbf{H}_{\tau\tau} = -1, \tag{68}$$

and for the weighted difference-in-means estimator:

$$\mathbf{H}_{\tau\tau} = -1, \tag{69}$$

and $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} \equiv \begin{pmatrix} \mathbf{\Sigma}_{\beta\beta} & \mathbf{\Sigma}_{\beta\tau} \\ \mathbf{\Sigma}_{\tau\beta} & \mathbf{\Sigma}_{\tau\tau} \end{pmatrix} \tag{70}$$

$$= \mathbb{E}\left[\left(s_{\text{ATT}}(\beta,m_i,\mathbf{x_i}),\ q_{\text{ATT}}(\tau_{\text{ATT}},\beta,m_i,\mathbf{x_i})\right)\left(s_{\text{ATT}}(\beta,m_i,\mathbf{x_i}),\ q_{\text{ATT}}(\tau_{\text{ATT}},\beta,m_i,\mathbf{x_i})\right)^{\mathsf{T}}\right], \tag{71}$$

where $\mathbf{\Sigma}_{\beta\tau}$ and $\mathbf{\Sigma}_{\tau\beta}$ are

$$\mathbf{\Sigma}_{\beta\tau} = (\mathbf{\Sigma}_{\tau\beta})^{\mathsf{T}}. \tag{72}$$

2.3 Efficiency

Since the efficiency gains from using the NAWT with the Horvitz–Thompson estimator can be easily extended to the NAWT with the weighted difference-in-means estimator and the efficiency for the MAR estimation is also easily extended to the ATT estimation, I focus on the MAR estimation with the Horvitz–Thompson estimator from now on.

First, as the property of the sequential M-estimation, the asymptotic distribution of $\hat{\mu}$ is the following.

$$\begin{aligned}\sqrt{n}(\hat{\mu}-\mu) \underset{d}{\to} \mathcal{N}(0,&\mathbf{H}_{\mu\mu}^{-1}\mathbf{\Sigma}_{\mu\mu}\mathbf{H}_{\mu\mu}^{-1}+\mathbf{H}_{\mu\mu}^{-1}\mathbf{H}_{\mu\beta}\mathbf{H}_{\beta\beta}^{-1}\mathbf{\Sigma}_{\beta\beta}\mathbf{H}_{\beta\beta}^{-1}\mathbf{H}_{\mu\beta}^{\mathrm{T}}\mathbf{H}_{\mu\mu}^{-1} \\ &-\mathbf{H}_{\mu\mu}^{-1}\mathbf{\Sigma}_{\mu\beta}\mathbf{H}_{\beta\beta}^{-1}\mathbf{H}_{\mu\beta}^{\mathrm{T}}\mathbf{H}_{\mu\mu}^{-1}-\mathbf{H}_{\mu\mu}^{-1}\mathbf{H}_{\mu\beta}\mathbf{H}_{\beta\beta}^{-1}\mathbf{\Sigma}_{\beta\mu}\mathbf{H}_{\mu\mu}^{-1}).\end{aligned} \tag{73}$$

Considering that the estimation of β does not depend on the estimation of μ,

$$
\begin{aligned}
n\mathbb{V}\left[\hat{\mu}\right] &= n\mathbb{E}\left[\mathbb{V}\left[\hat{\mu} \mid \mathbf{x}\right]\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right] \\
&= \mathbb{E}\left[\mathbf{H}_{\mu\mu}^{-1}\boldsymbol{\Sigma}_{\mu\mu}\mathbf{H}_{\mu\mu}^{-1} + \mathbf{H}_{\mu\mu}^{-1}(\mathbf{H}_{\mu\beta} \mid \mathbf{x})\mathbf{H}_{\beta\beta}^{-1}\boldsymbol{\Sigma}_{\beta\beta}\mathbf{H}_{\beta\beta}^{-1}(\mathbf{H}_{\mu\beta}^{\mathsf{T}} \mid \mathbf{x})\mathbf{H}_{\mu\mu}^{-1}\right. \\
&\qquad \left. - \mathbf{H}_{\mu\mu}^{-1}(\boldsymbol{\Sigma}_{\mu\beta} \mid \mathbf{x})\mathbf{H}_{\beta\beta}^{-1}(\mathbf{H}_{\mu\beta}^{\mathsf{T}} \mid \mathbf{x})\mathbf{H}_{\mu\mu}^{-1} - \mathbf{H}_{\mu\mu}^{-1}(\mathbf{H}_{\mu\beta} \mid \mathbf{x})\mathbf{H}_{\beta\beta}^{-1}(\boldsymbol{\Sigma}_{\beta\mu} \mid \mathbf{x})\mathbf{H}_{\mu\mu}^{-1}\right] \\
&\quad + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right] \\
&= \mathbb{E}\left[(\mathbf{H}_{\mu\beta} \mid \mathbf{x})\mathbf{H}_{\beta\beta}^{-1}\boldsymbol{\Sigma}_{\beta\beta}\mathbf{H}_{\beta\beta}^{-1}(\mathbf{H}_{\mu\beta}^{\mathsf{T}} \mid \mathbf{x})\right] - 2\mathbb{E}\left[(\boldsymbol{\Sigma}_{\mu\beta} \mid \mathbf{x})\mathbf{H}_{\beta\beta}^{-1}(\mathbf{H}_{\mu\beta}^{\mathsf{T}} \mid \mathbf{x})\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\Sigma}_{\mu\mu} \mid \mathbf{x}\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i \mid \mathbf{x}\right]\mathbb{E}\left[-\pi_\beta(\mathbf{x}_i)^{1+\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]^{-1}\right. \\
&\qquad \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^{1+2\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right] \\
&\qquad\quad \left.\mathbb{E}\left[-\pi_\beta(\mathbf{x}_i)^{1+\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]^{-1}\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i \mid \mathbf{x}\right]^{\mathsf{T}}\right] \\
&\quad - 2\mathbb{E}\left[\mathbb{E}\left[-\pi_\beta(\mathbf{x}_i)^{1+\alpha}y_i\mathbf{x}_i \mid \mathbf{x}\right]\mathbb{E}\left[-\pi_\beta(\mathbf{x}_i)^{1+\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]^{-1}\right. \\
&\qquad\quad \left.\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i \mid \mathbf{x}\right]^{\mathsf{T}}\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\Sigma}_{\mu\mu} \mid \mathbf{x}\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right] \\
&= \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i\ \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^\alpha\mathbf{D}\right]^{-1}\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^{2\alpha}\mathbf{D}\right]\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^\alpha\mathbf{D}\right]^{-1}\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i^{\mathsf{T}}\right] \\
&\quad - 2\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^\alpha\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i\ \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^\alpha\mathbf{D}\right]^{-1}\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i^{\mathsf{T}}\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\Sigma}_{\mu\mu} \mid \mathbf{x}\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right],
\end{aligned}
\tag{74}
$$

where $\mathbf{D} = \pi_\beta(\mathbf{x}_i)(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}$.

When we use the standard IPW where $\alpha = 0$, the variance becomes

$$
\begin{aligned}
n\mathbb{V}\left[\hat{\mu}_{\alpha=0}\right] &= -\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i\ \mathbb{E}\left[\mathbf{D}\right]^{-1}\pi_\beta(\mathbf{x}_i)y_i\mathbf{x}_i^{\mathsf{T}}\right] + \mathbb{E}\left[\boldsymbol{\Sigma}_{\mu\mu} \mid \mathbf{x}\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right] \\
&\leq \mathbb{E}\left[\boldsymbol{\Sigma}_{\mu\mu} \mid \mathbf{x}\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right],
\end{aligned}
\tag{75}
$$

which is equal to or smaller than the variance using the true propensity scores, where the equality is satisfied when the true propensity scores are constant, or equivalently, $\pi_\beta(\mathbf{x}_i) = \pi_\beta(\mathbf{x}_{i'})$ for any i and i'.

When we utilize the NAWT with $\alpha = 2$, the variance becomes

$$
\begin{aligned}
n\mathbb{V}\left[\hat{\mu}_{\alpha=2}\right] &= \mathbb{E}\left[\pi_\beta(\mathbf{x}_i) y_i \mathbf{x}_i \, \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^2 \mathbf{D}\right]^{-1} \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^4 \mathbf{D}\right] \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^2 \mathbf{D}\right]^{-1} \pi_\beta(\mathbf{x}_i) y_i \mathbf{x}_i^\mathsf{T}\right] \\
&\quad - 2\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^2 \pi_\beta(\mathbf{x}_i) y_i \mathbf{x}_i \, \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^2 \mathbf{D}\right]^{-1} \pi_\beta(\mathbf{x}_i) y_i \mathbf{x}_i^\mathsf{T}\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\Sigma}_{\mu\mu} \mid \mathbf{x}\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right] \\
&\leq -\mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^2 \pi_\beta(\mathbf{x}_i) y_i \mathbf{x}_i \, \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^2 \mathbf{D}\right]^{-1} \pi_\beta(\mathbf{x}_i) y_i \mathbf{x}_i^\mathsf{T}\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\Sigma}_{\mu\mu} \mid \mathbf{x}\right] + n\mathbb{V}\left[\mathbb{E}\left[\hat{\mu} \mid \mathbf{x}\right]\right] \\
&\leq n\mathbb{V}\left[\hat{\mu}_{\alpha=0}\right],
\end{aligned}
\tag{76}
$$

implying that it is equal to or smaller than the variance using the standard IPW and the one with true propensity scores, where the equality is satisfied when the true propensity scores are constant, or equivalently, $\pi_\beta(\mathbf{x}_i) = \pi_\beta(\mathbf{x}_{i'})$ for any i and i'. These results imply that we do not need to tune the hyper-parameter α to gain efficiency because the NAWT with $\alpha = 2$ is always more efficient than the standard IPW and the IPW with the true propensity scores.

On the other hand, the NAWT with $\alpha = 2$ may not be the most efficient among the NAWT with $\omega(\pi_\beta(\mathbf{x}_i)) = \pi_\beta(\mathbf{x}_i)^\alpha$, and the most efficient NAWT is the one with α being slightly larger than 2. Since the exact value of α for the most efficient estimation depends on the data, we cannot specify it without looking at data. Hyper-parameter tuning to improve estimation is considered later as an extension of the NAWT in Section 3.

2.4 Estimation of the ATE

The estimation of the ATE is not as simple as the estimation of the MAR or the ATT because the ATE estimation includes two different missing data problems. One of them is the average potential outcomes of the treated without treatment $\mathbb{E}[y_i(0) \mid t_i = 1]$ and the other is the average potential outcomes of the controlled with treatment $\mathbb{E}[y_i(1) \mid t_i = 0]$, neither of which can be observed. This naturally leads to the separate propensity score estimation for the potential outcomes with and without treatment. For estimating propensity scores

to estimate the average potential outcomes without treatment by (inversely) weighting the controlled units, the NAWT utilizes $\omega(\pi_\beta(\mathbf{x}_i)) = \pi_\beta(\mathbf{x}_i)^\alpha$, whereas it utilizes $\omega(\pi_\beta(\mathbf{x}_i)) = (1 - \pi_\beta(\mathbf{x}_i))^\alpha$ for estimating propensity scores to estimate the average potential outcomes with treatment by (inversely) weighting the treated units. This separate estimation produces two estimated propensity scores for each combination of covariates $\mathbf{x}$, one of which is for estimating average potential outcomes with treatment $\hat{\pi}^1_\beta(\mathbf{x}_i)$ and the other is for estimating that without treatment $\hat{\pi}^0_\beta(\mathbf{x}_i)$. In general, these two estimated propensity scores are not equal $\hat{\pi}^1_\beta(\mathbf{x}_i) \neq \hat{\pi}^0_\beta(\mathbf{x}_i)$ except for the standard IPW case where $\alpha = 0$. Although this requires a little caution to interpret estimated propensity scores and coefficients for them, the NAWT with the separate propensity score estimation for the ATE estimation is efficient as it is shown in the previous subsection for the MAR estimation.

Alternatively, for the ease of interpretation, we can combine the two score functions and estimate one propensity score for each combination of covariates $\mathbf{x}$, using $\omega(\pi_\beta(\mathbf{x}_i)) = \pi_\beta(\mathbf{x}_i)^\alpha + (1 - \pi_\beta(\mathbf{x}_i))^\alpha$. This combined estimation has an advantage in interpretation of estimated propensity scores, but it is not efficient as shown in Appendix C. This combined propensity score estimation is utilized in the CBPS for the ATE estimation $1/(\pi_\beta(\mathbf{x}_i)(1 - \pi_\beta(\mathbf{x}_i))) = 1/\pi_\beta(\mathbf{x}_i) + 1/(1 - \pi_\beta(\mathbf{x}_i))$, which implies that it balances covariates between the treated and controlled but not between the treated and combined nor between the controlled and combined (Chan, Yam, and Zhang 2016).

3 Extensions

3.1 The NAWT with covariate balance conditions

The NAWT method can be made robust to propensity score model misspecification by incorporating covariate balance conditions like the CBPS. In addition to the (tweaked) score

condition, we can utilize the following covariate balance conditions.

$$\sum c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) = 0 \tag{77}$$

$$c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \equiv \left(m_i - \frac{(1-m_i)\pi_\beta(\mathbf{x}_i)}{1-\pi_\beta(\mathbf{x}_i)} \right) \tilde{\mathbf{x}}_i \tag{78}$$

$$= \left(1 - \frac{(1-m_i)}{1-\pi_\beta(\mathbf{x}_i)} \right) \tilde{\mathbf{x}}_i \tag{79}$$

$$\sum c_{\mathrm{ATE}}(\beta, m_i, \mathbf{x_i}) = 0 \tag{80}$$

$$c_{\mathrm{ATE}}(\beta, m_i, \mathbf{x_i}) \equiv \left(\frac{m_i}{\pi_\beta(\mathbf{x}_i)} - \frac{1-m_i}{1-\pi_\beta(\mathbf{x}_i)} \right) \tilde{\mathbf{x}}_i, \tag{81}$$

where $\tilde{\mathbf{x}}_i$ are some functions of covariates to balance, which are typically $\mathbf{x}_i$ themselves. For simplicity, I focus on the estimation of the MAR with the Horvitz–Thompson estimator in the following. Using the score and covariate balance conditions results in more conditions than the parameters to be estimated and they can be estimated via the over-identified GMM estimation:

$$\hat{\beta} \equiv \arg\min_{\beta \in B} \bar{g}(\beta, \mathbf{m}, \mathbf{X})^{\mathsf{T}} \mathbf{A} \bar{g}(\beta, \mathbf{m}, \mathbf{X}), \tag{82}$$

where

$$\bar{g}(\beta, \mathbf{m}, \mathbf{X}) \equiv \frac{1}{n} \sum g(\beta, m_i, \mathbf{x}_i) \tag{83}$$

$$g(\beta, m_i, \mathbf{x}_i) = \begin{pmatrix} s_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \\ c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \end{pmatrix}, \tag{84}$$

and $\mathbf{A}$ is some positive definite symmetric weighting matrix. Note that when we utilize only the covariate balance conditions, it becomes the same method as the just-identified CBPS, where $g(\beta, m_i, \mathbf{x}_i) = c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i})$ and $\mathbf{A} = \mathbf{I}$ whose gradient is

$$\frac{\partial}{\partial \beta} \left(\frac{1}{n} \sum c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \right)^{\mathsf{T}} \frac{1}{n} \sum c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \tag{85}$$

$$= \frac{2}{n^2} \sum c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \frac{\partial}{\partial \beta} \sum c_{\mathrm{MAR}}(\beta, m_i, \mathbf{x_i}) \tag{86}$$

$$= \frac{2}{n^2} \sum \left(1 - \frac{(1-m_i)}{1-\pi_\beta(\mathbf{x}_i)} \right) \tilde{\mathbf{x}}_i \sum \frac{(m_i - \pi_\beta(\mathbf{x}_i))^2 \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}}{\pi_\beta(\mathbf{x}_i)(1-\pi_\beta(\mathbf{x}_i))}. \tag{87}$$

For efficiency, we can use the inverse of estimated covariance as the weighting matrix $\mathbf{A} = \hat{\mathbf{\Sigma}}(\beta, \mathbf{m}, \mathbf{X})^{-1}$. The continuously updating inverse covariance weight is

$$\hat{\mathbf{\Sigma}}(\beta, \mathbf{m}, \mathbf{X}) = \frac{1}{n} \sum \mathbb{E}[g(\beta, m_i, \mathbf{x}_i) g(\beta, m_i, \mathbf{x}_i)^{\mathsf{T}} \mid \mathbf{x}_i] \tag{88}$$

$$= \frac{1}{n} \sum \begin{pmatrix} \pi_\beta(\mathbf{x}_i)^{1+2\alpha}(1 - \pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}} & \pi_\beta(\mathbf{x}_i)^{1+\alpha}\mathbf{x}_i\tilde{\mathbf{x}}_i^{\mathsf{T}} \\ \pi_\beta(\mathbf{x}_i)^{1+\alpha}\tilde{\mathbf{x}}_i\mathbf{x}_i^{\mathsf{T}} & \pi_\beta(\mathbf{x}_i)/(1 - \pi_\beta(\mathbf{x}_i))\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^{\mathsf{T}} \end{pmatrix} \tag{89}$$

where m_i is integrated out conditional on the covariates $\mathbf{x}_i$, or we can utilize the two-step GMM (Imai and Ratkovic 2014).

As the properties of the GMM estimation, $\hat{\beta}_{\text{MAR}}$ is consistently estimated.

$$\hat{\beta}_{\text{MAR}} \xrightarrow{p} \beta, \tag{90}$$

which implies that the NAWT with the covariate balance conditions also does not shrink coefficients toward 0 and thus they are easy to interpret. The asymptotic distribution of $\hat{\beta}_{\text{MAR}}$ is:

$$\sqrt{n}(\hat{\beta}_{\text{MAR}} - \beta) \underset{d}{\to} \mathcal{N}(0, (\mathbf{G}^{\mathsf{T}}\mathbf{A}\mathbf{G})^{-1}\mathbf{G}^{\mathsf{T}}\mathbf{A}\mathbf{\Sigma}\mathbf{A}\mathbf{G}(\mathbf{G}^{\mathsf{T}}\mathbf{A}\mathbf{G})^{-1}) \tag{91}$$

where

$$\mathbf{G} \equiv \mathbb{E}\left[\frac{\partial}{\partial\beta}\bar{g}(\beta, m_i, \mathbf{x}_i)\right] \tag{92}$$

$$= \begin{pmatrix} \mathbb{E}\left[-((1+\alpha)\pi_\beta(\mathbf{x}_i) - \alpha m_i)\pi_\beta(\mathbf{x}_i)^\alpha(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right] \\ \mathbb{E}\left[(1-m_i)\pi_\beta(\mathbf{x}_i)(1-\pi_\beta(\mathbf{x}_i))^{-1}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\mathsf{T}\right] \end{pmatrix} \tag{93}$$

$$= \begin{pmatrix} \mathbb{E}\left[-\pi_\beta(\mathbf{x}_i)^{1+\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right] \\ \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\mathsf{T}\right] \end{pmatrix} \tag{94}$$

$$\mathbf{\Sigma} \equiv \mathbb{E}[g(\beta, m_i, \mathbf{x}_i)g(\beta, m_i, \mathbf{x}_i)^\mathsf{T}] \tag{95}$$

$$= \begin{pmatrix} \mathbb{E}\left[\left((m_i - \pi_\beta(\mathbf{x}_i))\pi_\beta(\mathbf{x}_i)^\alpha\right)^2\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right] \\ \mathbb{E}\left[\left(1-(1-m_i)(1-\pi_\beta(\mathbf{x}_i))^{-1}\right)^2\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\mathsf{T}\right] \end{pmatrix} \tag{96}$$

$$= \begin{pmatrix} \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)^{1+2\alpha}(1-\pi_\beta(\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right] \\ \mathbb{E}\left[\pi_\beta(\mathbf{x}_i)(1-\pi_\beta(\mathbf{x}_i))^{-1}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\mathsf{T}\right]. \end{pmatrix} \tag{97}$$

After estimating propensity scores, we can estimate the estimand by using the IPW estimators, such as the Horvitz–Thompson estimator, weighted difference-in-means estimator, doubly robust IPW estimator, and weighted least squares estimator, and its variance can be estimated by bootstrapping.

The advantage of using both the (tweaked) score conditions and covariate balance conditions is that it may perform better when the propensity score model is misspecified thanks to its over-identified GMM property, but the disadvantage is that it may not perform well with a small sample. Note that the NAWT with the covariate balance conditions for the MAR estimation utilizes the covariate balance conditions for the MAR, or the ATT, estimation, which is different from the original CBPS method which utilizes those for the ATE estimation when estimating the MAR.

3.2 Hyper-parameter tuning

Even though it is not necessary to tune the hyper-parameter α when using the NAWT, but we may gain additional efficiency by tuning it. The simplest way is to select α which minimizes the estimated variance of the estimand. Since the NAWT consistently estimates propensity scores, it also minimizes the root mean squared errors (RMSEs) as long as the propensity score model is correctly specified. Note that when the propensity score model is misspecified, this procedure does not necessarily choose α which has the smallest RMSE but only chooses one with the smallest variance and the bias may be larger than the NAWT with another value of α.

Some may be concerned about over-fitting in hyper-parameter tuning, but over-fitting is not a problem because we tune hyper-parameter α for estimating propensity scores. Over-fitting in estimating propensity scores may actually improve estimation of the estimand, as we know that the estimation with estimated propensity scores performs better than that with true propensity scores if the propensity score model is correct.

Again, the NAWT does not require the hyper-parameter tuning and we can always gain efficiency by using the NAWT with $\alpha = 2$. Hyper-parameter tuning by minimizing the estimated variance of the estimand may improve estimation further and over-fitting is not problematic in this situation, but it may not be in line with the original spirit of the propensity score and may not minimize the RMSE when propensity score model is misspecified.

4 Simulation studies

In this section, I apply the proposed method, the NAWT, to simulation data, to demonstrate how much it improves the performance of the standard IPW estimation and to compare with the performance of the IPW using the CBPS to estimate propensity scores. The IPW method has large variances and it is also vulnerable to propensity score model misspecification, and

the CBPS is proposed to solve these problems. Therefore, I conduct a simulation in which: (a) correct propensity score model, and two types of propensity score model misspecification (b) and (c) to test the validity of the NAWT for each of the MAR, ATT, and ATE estimation.

Specifically, I use the following data-generating process. There are $1,000$ units and each unit i has four covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$, each of which is independently and identically distributed according to the standard normal distribution. In the MAR estimation, some units are missing $m_i = 1$ and we cannot observe their outcomes though we can still observe their covariates. In the ATT and ATE estimation, some units are assigned treatment $m_i = 1$ and the others are not $m_i = 0$, and we can observe outcomes of all of them. The true outcome model is $y_i \sim \mathcal{N}(\mu, 1)$ and $\mu = 210 + \tau m_i + 27.4x_{i1} + 13.7x_{i2} + 13.7x_{i3} + 13.7x_{i4}$, where $\tau = 0$ for the MAR estimation and $\tau = 10$ for the ATT and ATE estimation. The estimand in the MAR estimation is μ and that in the ATT and ATE estimation is τ.

The true missing model in the MAR estimation and true treatment assignment model in the ATT and ATE estimation are $\Pr(m_i = 1) \sim \text{Bernoulli}(\pi_i)$ and $\pi_i = \text{logit}^{-1}(x_{i1} - 0.5x_{i2} + 0.25x_{i3} + 0.1x_{i4}))$ for scenarios (a) and (b) and $\pi_i = \text{logit}^{-1}(-x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4})$ for scenario (c), where $\text{logit}^{-1}(\cdot) = 1/(1 + \exp(-(\cdot)))$, resulting $\Pr(m_i = 1) = 0.5$. Finally, in scenarios (b) and (c), only the non-linear transforms of the covariates $\mathbf{x}_i^\star = (x_{i1}^\star, x_{i2}^\star, x_{i3}^\star, x_{i4}^\star) = (\exp(x_{i1}/2), x_{i2}/(1+\exp(x_{i1}))+10, (x_{i1}x_{i3}/25+0.6)^3, (x_{i1}+x_{i4}+20)^2)$ can be observed, which results in exactly the same misspecification in scenario (b) as that used in the existing studies (Imai and Ratkovic 2014; Kang and Schafer 2007). In scenarios (b) and (c), the propensity score model is misspecified because the true propensity score model is not a logistic function with $\mathbf{x}_i^*$ but one with $\mathbf{x}_i$ as linear predictors. Hence, the estimates are expected to be biased, but the NAWT and IPW with the CBPS are expected to mitigate this bias. To replicate the simulation study of Imai and Ratkovic (2014), I use the CBPS with the ATE covariate balance condition for the MAR estimation, but this covariate balance condition is not the best one but the covariate balance condition for the ATT is the best as I menthoned in Section 2. For the

ATE estimation, the NAWT utilizes the separate propensity score estimation for the potential outcomes with and without treatment, which performs better than the combined estimation (See Section 2 for details and Appendix C for the results of the combined estimation).

I use the weighted difference-in-means estimator for the estimand and conduct $2,000$ Monte Carlo simulations and calculate the bias and root-mean-squared error (RMSE) for each propensity score estimator (the NAWT with $\alpha = 2$, standard IPW, IPW with the just-identified CBPS) in each scenario ((a), (b), and (c)) for each estimand (the MAR, ATT, and ATE).

The summary of the results is shown in Figures 2–4, and Figures 5–7 in the Appendix B show the distribution of the estimates. The results demonstrate that the NAWT has negligible biases when the propensity score model is correctly specified and dramatically improves the estimates compared with the standard IPW. The NAWT has smaller RMSEs than the standard IPW when (a) the propensity score model is correctly specified and smaller biases than the standard IPW when (b)(c) the propensity score model is misspecified. Since the covariate balance condition of the CBPS for the MAR estimation is not the best one, the IPW with the CBPS performs poorly for the MAR estimation especially for scenario (c) though it performs well for scenario (b) by chance. In terms of the RMSE, it depends on the situation whether the NAWT works better than the IPW with the CBPS or not because the performance of the CBPS depends on how close the linear combination of the balanced covariates can approximate the true outcome model. Since the true outcome model is linear combination of the covariates in scenario (a), the IPW with the CBPS works better than the NAWT, but when the propensity score model is misspecified and the outcome model is not the linear combination of the covariates (scenarios (b) and (c)), the NAWT works better than the IPW with the CBPS in scenario (b) but not in scenario (c).

Surprisingly, when the propensity score model is misspecified in scenarios (b) and (c), the NAWT overwhelms the IPW with the CBPS in terms of the bias. Considering that the CBPS is proposed and used to mitigate the bias due to propensity score model misspecification,

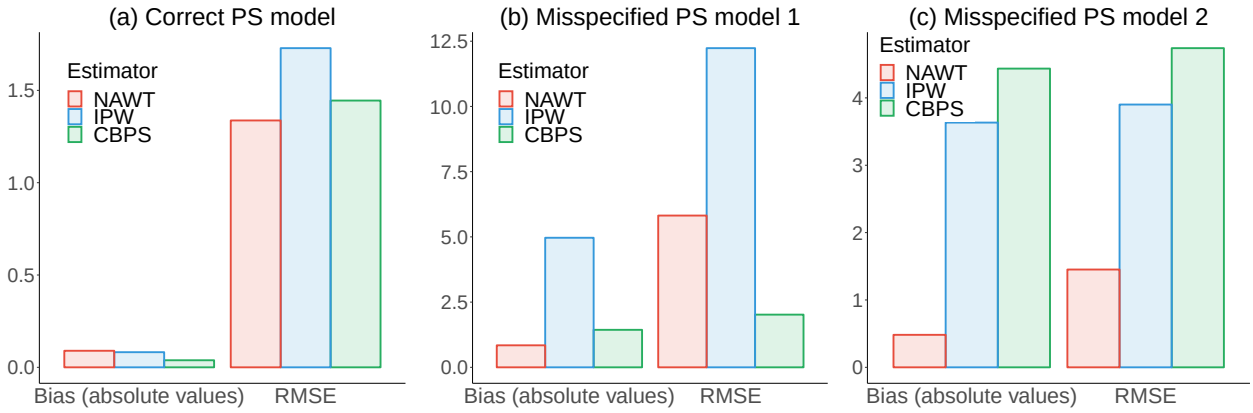

Figure 2: The bias in absolute values and the RMSE for the MAR estimation using the NAWT, standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c). The NAWT outperforms the standard IPW in terms of the RMSE in all the scenarios, and it depends on the situation whether the NAWT works better than the IPW with CBPS in terms of both the bias and RMSE. Note that the covariate balance condition of the CBPS for the MAR is not the best one.

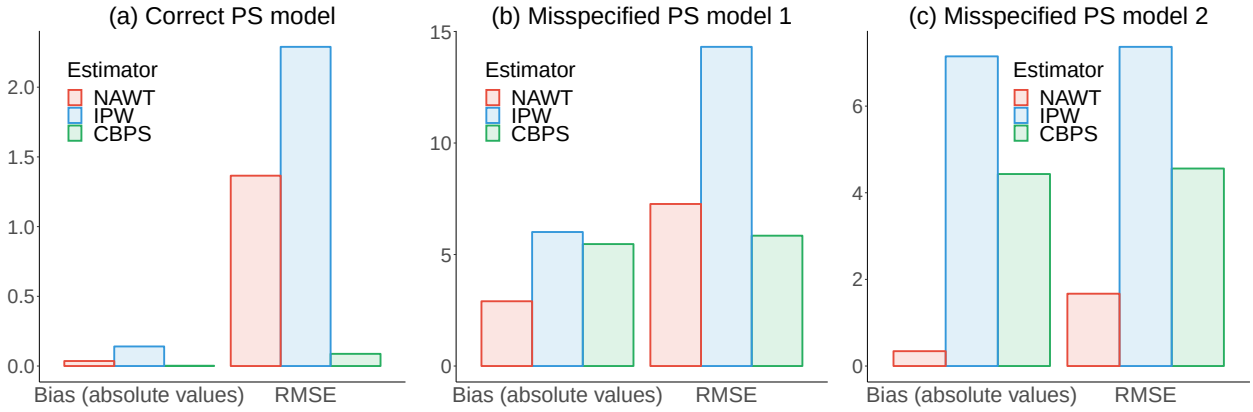

Figure 3: The bias in absolute values and the RMSE for the ATT estimation using the NAWT, standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c). The NAWT outperforms the standard IPW in terms of the RMSE in all the scenarios, and it depends on the situation whether the NAWT works better than the IPW with CBPS in terms of both the bias and RMSE.

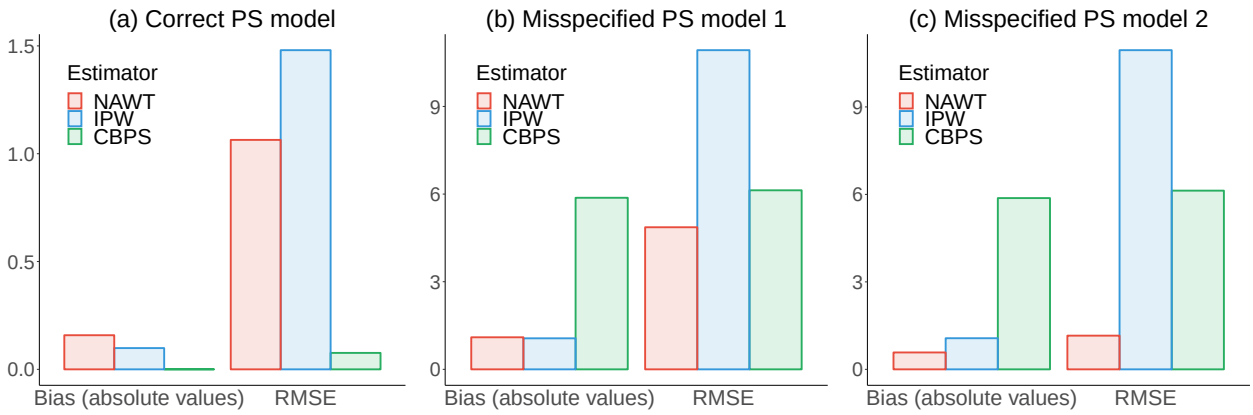

Figure 4: The bias in absolute values and the RMSE for the ATE estimation using the NAWT (the separate estimation), standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c). The NAWT outperforms the standard IPW in terms of the RMSE in all the scenarios, and it depends on the situation whether the NAWT works better than the IPW with CBPS in terms of both the bias and RMSE.

these results impressively confirm the robustness of the NAWT. The reason why the NAWT can reduce the bias due to propensity score model misspecification is similar to the reason of its efficiency. First, the bias can be decomposed into the bias due to the violation of the conditional ignorability (the omitted variable bias) and the bias due to parametric propensity score model misspecification. Since the non-parametric propensity score estimation asymptotically eliminates the latter bias and the NAWT approximates this estimation for units who have large estimated inverse probability weights and thus have large impact on the estimation of the estimand, it reduces the latter bias greatly for these units and, therefore, in total.

For the ATE estimation, the separate propensity score estimation for the potential outcomes with and without treatment slightly outperforms the combined estimation in terms of the RMSE when the propensity score model is correctly specified. Moreover, the separate estimation has much smaller RMSEs than the combined estimation when the propensity score model is misspecified, demonstrating that the separate estimation is preferable to the combined estimation.

In summary, the NAWT has negligible biases when the propensity score model is correctly

specified and is much robust and efficient than the standard IPW. In addition, it is also much more robust than the IPW with the CBPS which is proposed to mitigate the bias due to the propensity score model misspecification.

5 Conclusions

The IPW is broadly utilized in dealing with missing data problems including causal inference because it can eliminate confounding without relying on correct specification of outcome models under the conditional ignorability assumption. However, existing research has pointed out that the IPW may have an excessively large variance due to extreme estimated weights and be highly vulnerable to the misspecification of the propensity score model. To solve these problems, this study proposed the NAvigated WeighTing (NAWT), which conveys efficiency and robustness by tweaking the score function of propensity scores estimation depending on a pre-specified estimand. The NAWT includes the standard IPW and the covariate balancing propensity score (CBPS) as special cases and can further be made robust to propensity score model misspecification by incorporating covariate balance conditions. Its impressive improvement of the performance of estimation compared with the standard IPW and IPW with the CBPS was demonstrated through simulation studies.

The key idea of the NAWT is that it improves estimation of the estimand at the cost of accuracy in propensity score estimation. It put more importance on units who should have large inverse probability weights when estimating propensity scores so that it approximates the non-parametric propensity score estimation for units who have large impact on the estimation of the estimand. Since the IPW with the non-parametric propensity score estimation is proved to be asymptotically efficient and has no bias due to propensity score model misspecification, this approximation makes the NAWT to be robust and efficient. At the same time, since the NAWT is a simple extension of the standard IPW, it is efficient with a finite sample, no tuning

is required, and it is easy to use and interpret. These characteristics enable the NAWT to enjoy the best of both worlds. Note that it still needs the conditional ignorability assumption and careful specification of the propensity score model even though it is much robust and efficient than the standard IPW.

Finally, I show some future directions for the improvement and the application of the NAWT. As I demonstrated that the NAWT can incorporate covariate balance conditions, it is natural to extend it to incorporate kernel balance conditions, which makes the NAWT more flexible (Hazlett 2016; Wong and Chan 2017; Zhao 2019). Since the NAWT is an extension of the standard IPW, it may be combined with attractive methods for the standard IPW, such as estimating propensity scores with regularization via the ridge or LASSO. Although this study focuses on the IPW, the spirits of the NAWT may be applicable to the propensity score matching and stratification, where we can gain efficiency and robustness by tweaking the propensity score estimation. This study also concentrate itself on the standard logistic regression for the propensity score estimation, the idea of the NAWT may be applied to recently proposed machine learning techniques for the propensity score estimation, such as the decision tree, random forest, and generalized random forest, where we can tweak the algorithm via weighting a purity measure, such as the Gini impurity, by a function of propensity scores when splitting units. I am currently exploring these potential directions and other application of the NAWT and developing an easy-to-use R package which implements the NAWT.

References

Bhattacharya, Jay, and William B Vogt. 2012. "Do Instrumental Variables Belong in Propensity Scores?" *International Journal of Statistics & Economics* 9 (A12): 107–27.

Brookhart, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology* 163 (12): 1149–56.

Chan, Kwun Chuen Gary, Sheung Chi Phillip Yam, and Zheng Zhang. 2016. "Globally Efficient Non-parametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78 (3): 673–700.

Fan, Jianqing, Kosuke Imai, Han Liu, Yang Ning, and Xiaolin Yang. 2016. "Improving Covariate Balancing Propensity Score: A Doubly Robust and Efficient Approach."

Graham, Bryan S., Cristine Campos De Xavier Pinto, and Daniel Egel. 2012. "Inverse Probability Tilting for Moment Condition Models with Missing Data." *Review of Economic Studies* 79 (3): 1053–79.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66 (2): 315–32.

Hazlett, Chad. 2016. "Kernel Balancing: A Flexible Non-parametric Weighting Procedure for Estimating Causal Effects." *arXiv preprint arXiv:1605.00155.*

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–89.

Horvitz, D. G., and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47 (260): 663–85.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 243–63.

Kang, Joseph D. Y., and Joseph L. Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22 (4): 523–39.

Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400.

Rosenbaum, Paul R. 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82 (398): 387–94.

Rubin, Donald B. 2007. "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36.

Tan, Zhiqiang. 2010. "Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting." *Biometrika* 97 (3): 661–82.

Wong, Raymond K. W., and Kwun Chuen Gary Chan. 2017. "Kernel-based Covariate Functional Balancing for Observational Studies." *Biometrika* 105 (1): 199–213.

Zhao, Qingyuan. 2019. "Covariate Balancing Propensity Score by Tailored Loss Functions." *The Annals of Statistics* 47 (2): 965–93.

Appendix

A Proof

The non-parametric propensity score estimation reduces variances for estimand by the following value given covariates $\mathbf{x}$:

$$\mathbb{E}\left[\left(\frac{\mu_0(\mathbf{x})}{1-\pi(\mathbf{x})}(\mathbf{m}-\pi(\mathbf{x}))\right)^2 \mid \mathbf{x}\right] = \mathbb{E}\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\mu_0(\mathbf{x})^2 \mid \mathbf{x}\right], \tag{98}$$

where $\mu_0(\mathbf{x})$ represents the mean outcome given covariates $\mathbf{x}$, $\pi(\mathbf{x})$ represents propensity scores given covariates $\mathbf{x}$, and $\mathbf{m}$ is a missingness indicator which takes 1 when outcome are missing and 0 otherwise. This value is larger for units with larger propensity scores unless the conditional mean outcome correlates with propensity scores.

B Distribution of estimates in the simulation studies

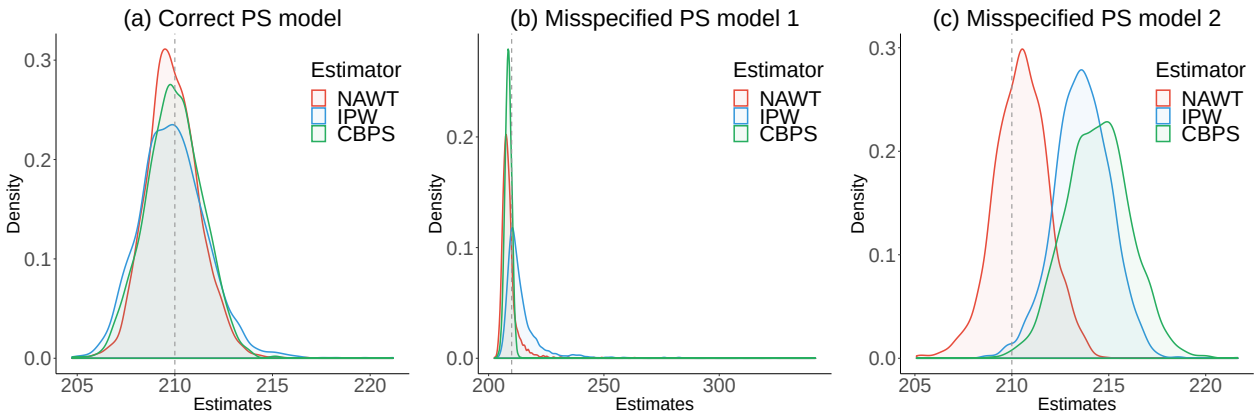

Figure 5: The distribution of estimates for the MAR estimation using the NAWT, standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c).

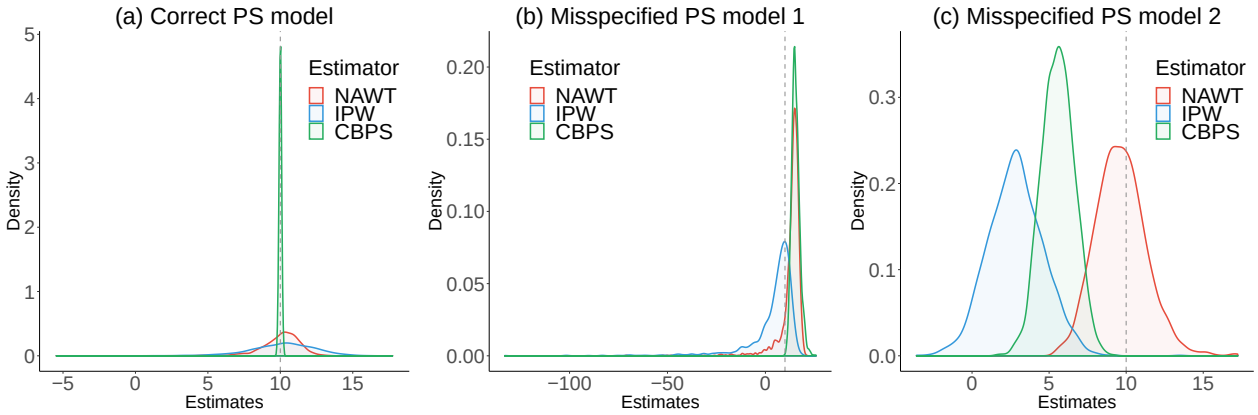

Figure 6: The distribution of estimates for the ATT estimation using the NAWT, standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c).

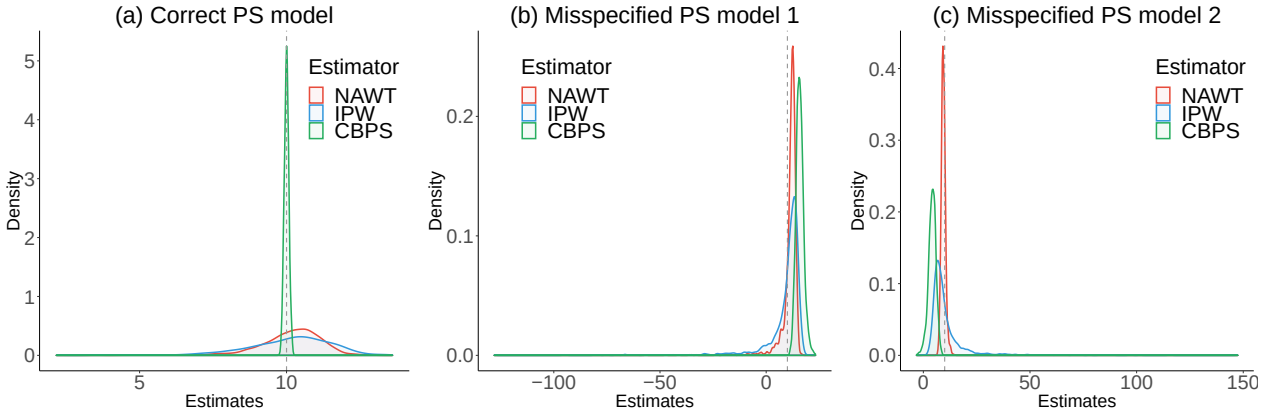

Figure 7: The distribution of estimates for the ATE estimation using the NAWT (the separate estimation), standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c).

C The results of the NAWT with the combined propensity score estimation for the ATE estimation

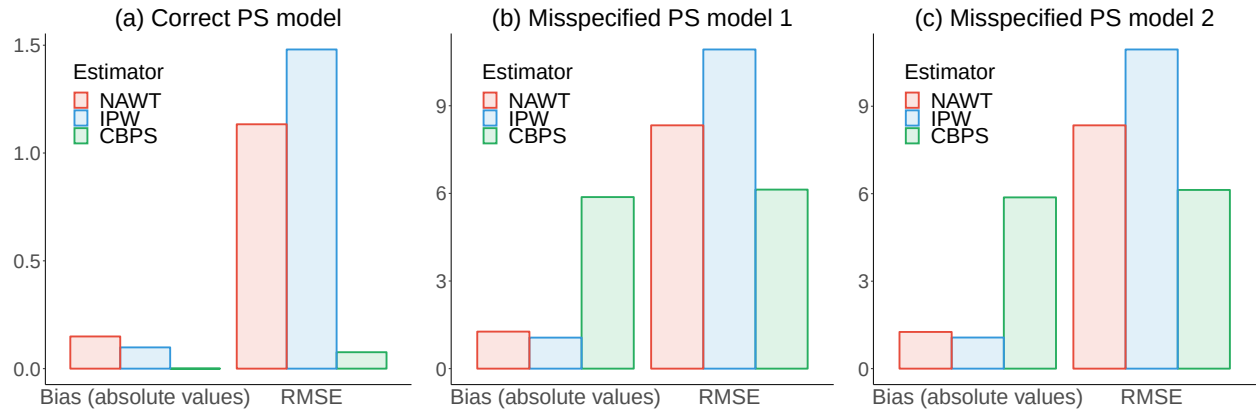

Figure 8: The bias in absolute values and the RMSE for the ATE estimation using the NAWT (the combined estimation), standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c). The NAWT outperforms the standard IPW in terms of the RMSE in all the scenarios, and it depends on the situation whether the NAWT works better than the IPW with CBPS in terms of both the bias and RMSE.

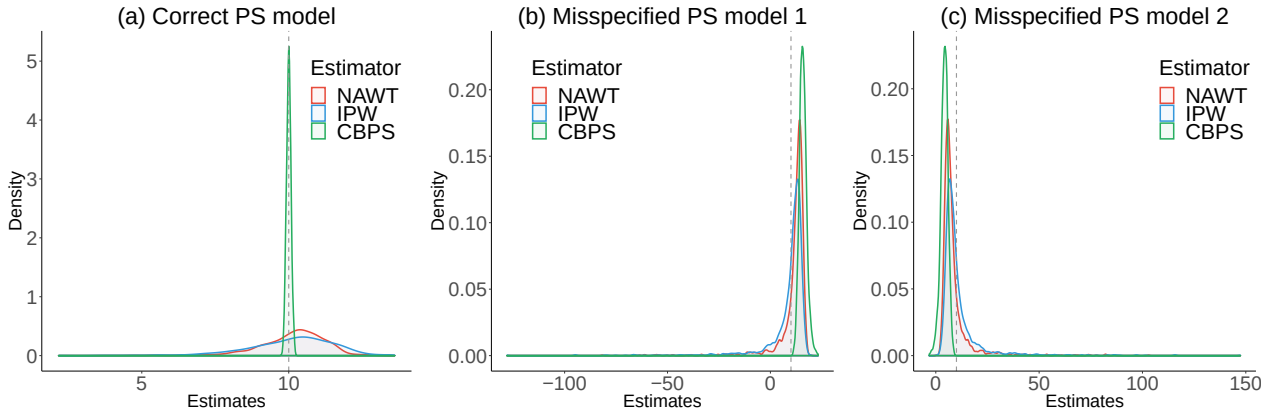

Figure 9: The distribution of estimates for the ATE estimation using the NAWT (the combined estimation), standard IPW, and IPW with the CBPS under the following scenarios: (a) correct propensity score model and two types of propensity score model misspecification (b) and (c).