Simple Robust Estimating Method for Propensity Score Models and its Application to Some Causal Estimands

Shunichiro Orihara

Graduate School of Data Science, Yokohama City University, Kanagawa, Japan

Abstract

To estimate causal effects accurately, adjusting covariates is one of the important steps in observational study. When all covariates are observed, the covariates can be adjusted, and an unbiased estimator for causal effects can be obtained. In this situation, the propensity score has the central role to estimate the causal effects. Recently, some causal estimands, the target population of causal effects estimation, are considered which depend on the “true” propensity score. A point to note that an interested estimands might have some bias if a propensity score model was misspecified. In this paper, we consider a multiply robust estimator for the propensity score. In brief, we prepare some candidate models, and construct an estimating equation including the candidate models at once. Some theoretical properties are proved, and we consider application examples for propensity score estimations when the average treatment effects for the overlap population is central interest.

Keywords: Causal inference, Overlap weight, Propensity score, Robust estimation
1 Introduction

To estimate causal effects accurately, adjusting covariates (or confounders; hereafter, we call “covariates”) is one of the important steps in observational study. When all covariates are observed, the covariates can be adjusted, and an unbiased estimator for causal effects can be obtained; the situation of “no unmeasured confounding” (c.f. Hernán and Robins, 2020). In this situation, the propensity score has the central role to estimate the causal effects (Rosenbaum and Rubin, 1983). However, the propensity score is usually unknown, and needs to be estimated by appropriate procedures. A point to note that an interested estimator might have some bias if a propensity score model was misspecified (Kang and Schafer, 2007); valid model construction is important.

Recently, some causal estimands, the target population of causal effects estimation, are considered. One of the well-known estimands is the average treatment effects (ATE; c.f. Imbens and Rubin, 2015). The ATE assumes the causal effects that the overall population is treated “counterfactual” study treatment and control treatment. In short, the ATE is interested in the overall causal effects. A different estimand called the average treatment effects for the overlap population (ATO) is proposed by Li et al., 2018. The ATO has several attractive features compared with the ATE (c.f. Li et al., 2018); for instance, the ATO may become more reasonable causal effects in the sense that subjects who could change their actual treatment to the “counterfactual” treatment have larger weights than subjects who could not change, and is less sensitive to extreme propensity scores (i.e. low propensity scores in the treatment group, or high propensity scores in the control group). Another estimand is defined by the incremental propensity score (Kennedy, 2019). The estimands are defined how change in the intervention probability (i.e., the propensity score) affects the causal effects. There are some attractive estimands, however, there is a crucial problem; some estimands depend on the “true” propensity score. For instance, as mentioned Mao et al., 2019, the ATO has some bias when the propensity score model is misspecified. Therefore, the similar problem mentioned in Kang and Schafer, 2007 may be occurred.
There is a critical problem when the propensity score model is misspecified. To overcome the problem, a variety of interesting methods have been proposed. Covariate balancing and calibration methods (Han, 2014, Imai and Ratkovic, 2014, Zubizarreta, 2015, Fong et al., 2018, and Orihara and Hamada, 2019) are one of the solutions. Propensity score subclassification (e.g., Wang et al., 2016 and Orihara and Hamada, 2021) is another solution since the propensity scores in each subclass are regarded as “smoothed” (c.f. Imbens and Rubin, 2015). In other words, the impact of misspecification is also smoothed. However, to the best of my knowledge, the properties of the methods are derived for the causal effects estimator, not for the estimator of a propensity score itself. Machine learning procedures such as classification trees and random forest are important procedures since the propensity score model needs not to be assumed. The (slight) trouble is that these methods require careful pre-preparation (i.e. hyper parameter chooning; see Cannas and Arpino, 2019).

In this paper, we consider another procedure: a multiply robust estimator for the propensity score (note that “multiply robust” is slightly different from Han, 2014). In brief, we prepare some candidate models, and construct an estimating equation including the candidate models at once (do not have to select just one). If the correct model was included in the candidate models, the parameter estimator would have the consistency. Also, we consider the situation where the true model is not included in the candidate models. Under this situation, we confirm the condition where the estimator becomes a “valid” in the sense of the true Kullback-Leibler (KL) divergence. By using the procedure, we will consider a robust estimating method to estimate a “valid” causal effects without any model/variable selection methods. Therefore, our proposed method may overcome the problem of a variable selection of a propensity score model (Brookhart et al., 2006 and Austin et al., 2007). For instance, a researcher claims that some covariates have to be included in a propensity score model, whereas, another researcher claims that the covariates do not have to be included. Commonly, in this situation, one of the researchers has to compromise, or the researchers use some model/variable selection methods. Meanwhile, our proposed method provides a
simple solution: the both models can be included in a one integrated model.

The remainder of the paper proceeds as follows. In section 2, we propose a new robust generalized linear model (GLM) estimator and confirm their properties with and without the condition where the true model is included in the candidate models. Since the propensity score model is assumed as not only a logistic regression model but also a regression model with normal error under general treatment regimes (Hirano and Imbens, 2004 and Imai and van Dyk, 2004), the general situation is considered. In section 3, we consider application examples to some statistical methods, and show simulation data examples when the ATO is central interest. Some materials are found in the appendix.

2 Constructing a robust estimating method

Let $n$ be the sample size, and assume that $i = 1, 2, \ldots, n$ are i.i.d. samples. $t \in T \subset \mathbb{R}$ and $X \in \mathbb{R}^p$ denote an outcome and a vector of covariates, respectively. We assume that an outcome has a distribution in an exponential family (McCullagh and Nelder, 2019):

$$f(t; \theta, \phi) = \exp \left\{ \frac{t \theta - b(\theta)}{a(\phi)} + c(t, \phi) \right\}, \quad (2.1)$$

where $(\theta, \phi)$ are parameters; in particular, we are interested in the parameter $\theta$. Note that a normal distribution and a Bernoulli distribution are included in an exponential family. For a normal distribution (i.e., for the generalized propensity score), each function and parameter in (2.1) become

$$\theta = \mu \ (mean), \ a(\phi) = \sigma^2 \ (variance), \ b(\theta) = \frac{\theta^2}{2}, \ c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2).$$
For a Bernoulli distribution (i.e., for the propensity score), each function and parameter in (2.1) become

$$\theta = \log \left( \frac{p}{1 - p} \right) \quad (p \text{ is a binomial probability}), \quad a(\phi) = 1, \quad b(\theta) = \log(1 + e^\theta), \quad c(y, \phi) = 0.$$ 

To construct a GLM, we consider a relationship between an expectation of an outcome and its model structure $\varphi(x; \beta)$. Specifically, $\varphi(x; \beta) = g(p)$, where a monotonic function $g$ is called “link function”. For a Bernoulli distribution, when we select a link function as $g = \log (\text{logit link})$ and assume a linear relationship between $x$ and $\beta$ ($\varphi(x; \beta) = x^\top \beta$),

$$p = \frac{\exp \left\{ x^\top \beta \right\}}{1 + \exp \left\{ x^\top \beta \right\}} \quad (\text{logistic regression model}).$$

Whereas, when we select a link function as $g = \Phi^{-1}$ (a normal distribution function; probit link) and assume a linear relationship between $x$ and $\beta$,

$$p = \Phi (x^\top \beta) \quad (\text{probit model}).$$

Note that $\theta$ becomes a function of $x$ and $\beta$ clearly from the above example: $\theta = \theta(x; \beta)$. Hereinafter, the model (2.1) will be considered for a while.

From here, we introduce a proposed procedure that many working models are combined into a one integrated model. At first, we consider working models; $k (k = 1, 2, \ldots, K)$ denotes each model:

$$f(t|x; \beta_k) = \exp \left\{ \frac{t \theta_k(x; \beta_k) - b(\theta_k(x; \beta_k))}{a(\phi)} + c(t, \phi) \right\}.$$

Note that we assume that $\phi$ is known and needs not to be estimated, and the functions $a$, $b$, and $c$ are common to each model. For a normal distribution, these assumptions imply that variance parameters are known and common to each model. To estimate parameter $\beta_k$, we
consider a KL divergence $D_k(\beta_k)$:

$$D_k(\beta_k) = E[\log g(T|X)] - E[\log f(T|X; \beta_k)], \quad (2.2)$$

where $g(t|x)$ is the unknown true probability distribution. Whereas we would like to estimate the parameters so that $D_k(\beta_k)$ is small, we cannot handle (2.2) directly. Therefore, we consider an estimator of the second term of (2.2) to estimate parameter $\beta_k$:

$$\ell(\beta_k) = \frac{1}{n} \sum_{i=1}^{n} \log f(t_i|x_i; \beta_k) \propto \frac{1}{n} \sum_{i=1}^{n} [t_i \theta_k(x_i; \beta_k) - b(\theta_k(x_i; \beta_k))],$$

where $\ell$ is a log-likelihood function. Therefore, a score function $S(\beta_k)$ becomes

$$S(\beta_k) = \frac{\partial}{\partial \beta_k} \ell(\beta_k) = \frac{1}{n} \sum_{i=1}^{n} \left[ t_i - \hat{b}(\theta_k(x_i; \beta_k)) \right] \frac{\partial}{\partial \beta_k} \theta_k(x_i; \beta_k),$$

where $\hat{b}$ is the first differentiable of $b$. The solution of $S(\beta_k) = 0$ becomes a maximum likelihood estimator (MLE) $\hat{\beta}_k$. Next, by using the MLEs $\hat{\beta}_k$, we consider an integrated model:

$$f(t|x; \gamma, \hat{\beta}) = \exp \left\{ \frac{t \sum_{k=1}^{K} \gamma_k \theta_k(x; \hat{\beta}_k) - b \left( \sum_{k=1}^{K} \gamma_k \theta_k(x; \hat{\beta}_k) \right)}{a(\phi)} + c(t, \phi) \right\},$$

where $\sum_{k=1}^{K} \gamma_k = 1$ and $\gamma_k \geq 0$. The same discussion as above, a KL divergence, a log-likelihood function, and a score function become

$$D(\gamma, \beta) = E[\log g(T|X)] - E[\log f(T|X; \gamma, \beta)], \quad (2.3)$$

$$\ell(\gamma, \hat{\beta}) \propto \frac{1}{n} \sum_{i=1}^{n} \left[ t_i \sum_{k=1}^{K} \gamma_k \theta_k(x_i; \hat{\beta}_k) - b \left( \sum_{k=1}^{K} \gamma_k \theta_k(x_i; \hat{\beta}_k) \right) \right],$$
The solution of \( S(\gamma, \hat{\beta}) = 0 \) becomes a MLE \( \hat{\gamma} \), and the integrated model \( f(t|x; \hat{\gamma}, \hat{\beta}) \) is used for a subsequent inference.

In the next step, we confirm properties of our proposed estimating procedure. At first, we consider the situation where the true model is included in the candidate models \( (k = 1, 2, \ldots, K) \). Concretely, we assume that \( k = 1 \) is the true model. Under this setting, the following theorem is proved:

**Theorem 1.**

*Under the regularity conditions from C.1 to C.3,*

\[
\hat{\gamma}_1 \xrightarrow{P} 1 \quad \text{and} \quad \hat{\gamma}_k \xrightarrow{P} 0, \ k = 2, \ldots, K. \tag{2.5}
\]

*Therefore, from the continuous mapping theorem,*

\[
f(t|x; \hat{\gamma}, \hat{\beta}) = \exp \left\{ \frac{t \sum_{k=1}^{K} \hat{\gamma}_k \theta_k(x; \hat{\beta}_k) - b \left( \sum_{k=1}^{K} \hat{\gamma}_k \theta_k(x; \hat{\beta}_k) \right)}{a(\phi)} + c(t, \phi) \right\} \xrightarrow{P} \exp \left\{ \frac{t \theta_1(x; \beta_1^0) - b \left( \beta_1^0 \right)}{a(\phi)} + c(t, \phi) \right\} = f(t|x; \beta_1^0),
\]

*where the superscript “0” of parameters means the true value of parameters.*

Proof is in the appendix A. From the Theorem 1, the integrated model is consistent with the true model when the true model is included in the candidate models, and \( n \to \infty \). Also, assuming additional regularity condition, the \( \sqrt{n} \)-consistency is also proved (see Andrews, 1999).
Theorem 2.

Under the regularity conditions from C.1 to C.4,

$$\sqrt{n}(\hat{\gamma} - \gamma^0) = O_p(1),$$

where $\gamma^0 = (1, 0, \ldots, 0)^\top$.

Proof is straightforward from Theorem 1 and the result of Andrews, 1999.

Next, we consider the situation where the true model is not included in the candidate models; the assumption of Theorem 1 and 2 is not hold. Even in this situation, the following theorem is hold:

Theorem 3.

Under the regularity condition C.2, the following inequality is hold for $\forall \gamma$:

$$D(\gamma, \beta^*) \leq \sum_{k=1}^{K} \gamma_k D_k(\beta^*_k) \quad (2.6)$$

Proof is in the appendix A. The Theorem 3 means that the integrated model becomes better than each candidate model in the sense of the true KL divergence even if the true model is not included in the candidate models. In this sense, the integrated model is better option than any model selection properties when we are not interested in the selection of the “valid” model. Note that it is necessary to choose no hyper parameters in our proposed method; this is a different point from any machine learning procedures.

The objective of the proposed estimating procedure is similar as model averaging procedures (MA; c.f. Hoeting et al., 1999, Hjort and Claeskens, 2003, and Xie et al., 2019), however, the constructions are different. Concretely, the former is to construct a weighting average of linear predictors; the latter is to construct a weighting average of interested parameter estimators. Also, properties of these methods are different. Regarding MA, predictors of model averaging become the best in the sense that the mean squared error becomes minimum
when the prior distribution is detected correctly, as discussed in Raftrey and Zheng, 2003. Whereas, our proposed method has another properties as discussed above. In the following simulations, we will confirm performance of our proposed method and a MA method.

3 Application of proposed estimator

3.1 Application to the ATO with simulation/real data analysis

Let $n$ be the sample size. $T_i \in \{0, 1\}$, $X_i \in \mathbb{R}^p$ and $(Y_{1i}, Y_{0i}) \in \mathcal{Y} \subset \mathbb{R}^2$ denote the treatment, a vector of covariates measured prior to treatment, and potential outcomes, respectively. $Y_i := T_i Y_{1i} + (1 - T_i) Y_{0i}$ denotes an observed outcome and we assume that $i = 1, 2, \ldots, n$ are i.i.d. samples and the stable unit treatment value assumption (c.f. Rosenbaum and Rubin, 1983) holds. Next, we introduce a general class of estimands called “weighted average treatment effect” (WATE) in Hirano et al., 2003 and Li et al., 2018. The ATE conditional on $x$ is defined as $\tau(x) := \mathbb{E}[Y_1 - Y_0|x]$, and the WATE is defined as

$$\tau_h := \frac{\mathbb{E}[\tau(X)h(X)]}{\mathbb{E}[h(X)]},$$

where $F(\cdot)$ and $h(\cdot)$ are the cumulative distribution function and a weight function for $x$, respectively. When $h(x) \equiv 1$, the WATE becomes the ATE: $\tau_{ATE} := \mathbb{E}[Y_1 - Y_0]$ obviously. Under the strongly ignorable treatment assignment $T \perp \perp (Y_1, Y_0)|X$, we can estimate the ATE using the propensity score $e(X_i) := \mathbb{P}(T_i = 1|X_i)$.

From here, we focus on the ATO. When $h(x) = e(x)(1 - e(x))$, the WATE becomes the ATO:

$$\tau_{ATO} := \frac{\mathbb{E}[\tau(X)e(X)(1 - e(X))]}{\mathbb{E}[e(X)(1 - e(X))]}.$$

Since the function $e(1 - e)$, $e \in (0, 1)$ is the convex and symmetric function at 1/2, subjects who could change their actual treatment to the “counterfactual” treatment (i.e. $e \approx 1/2$) have larger weights than subjects who could not change (i.e. $e \approx 0$ or 1). Also, as obviously,
the ATO depends on the true propensity score; if the estimated propensity score model is misspecified, the ATO has some bias (see also Mao et al., 2019). To estimate the ATO, we consider three estimators: the IPW-type, the AIPW-type, and BR-type estimator.

- **IPW-type estimator** (Li et al., 2018)

\[
\hat{\tau}_{\text{ATO}}^{\text{ipw}} := \frac{\sum_{i=1}^{n} T_i (1 - \hat{e}_i) Y_i}{\sum_{i=1}^{n} T_i (1 - \hat{e}_i)} - \frac{\sum_{i=1}^{n} (1 - T_i) \hat{e}_i Y_i}{\sum_{i=1}^{n} (1 - T_i) \hat{e}_i}.
\]

- **AIPW-type estimator** (Mao et al., 2019)

\[
\hat{\tau}_{\text{ATO}}^{\text{aug}} := \frac{\sum_{i=1}^{n} e_i (1 - \hat{e}_i) (\hat{m}(1, X_i; \hat{\beta}) - \hat{m}(0, X_i; \hat{\beta}))}{\sum_{i=1}^{n} e_i (1 - \hat{e}_i)} + \frac{\sum_{i=1}^{n} T_i (1 - \hat{e}_i) (Y_i - \hat{m}(1, X_i; \hat{\beta}))}{\sum_{i=1}^{n} T_i (1 - \hat{e}_i)} - \frac{\sum_{i=1}^{n} (1 - T_i) \hat{e}_i (Y_i - \hat{m}(0, X_i; \hat{\beta}))}{\sum_{i=1}^{n} (1 - T_i) \hat{e}_i},
\]

where \(\hat{m}(T_i, X_i; \beta)\) is a outcome model.

- **BR-type estimator** (see appendix C)

\[
\hat{\tau}_{\text{ATO}}^{\text{br}} := \frac{1}{n} \sum_{i=1}^{n} \tilde{m}(1, X_i; \hat{\beta}, \hat{\phi}) - \frac{1}{n} \sum_{i=1}^{n} \tilde{m}(0, X_i; \hat{\beta}, \hat{\phi})
\]

where

\[
\tilde{m}(T_i, X_i; \beta, \phi) = \hat{m}(T_i, X_i; \beta) + \frac{T_i (1 - \hat{e}_i)}{\sum_{i=1}^{n} T_i (1 - \hat{e}_i)} \phi_1 + \frac{(1 - T_i) \hat{e}_i}{\sum_{i=1}^{n} (1 - T_i) \hat{e}_i} \phi_0.
\]

Note that \(\hat{e}_i\) is any propensity score estimate. In the following analyses, we confirm performance of these estimators with multiply robust propensity score estimates.

### 3.1.1 Simulation data analysis

A part of simulation settings refer to Mao et al., 2019. The number of iterations for all simulations are 1,000. The simulation data is derived from the following distributions:
Covariates

\[
\begin{pmatrix}
X_{i1} \\
X_{i2}
\end{pmatrix} \sim N \left( \begin{pmatrix}
-X_{i3} + X_{i4} + 0.5X_{i3}X_{i4} \\
X_{i3} - X_{i4} + 0.5X_{i3}X_{i4}
\end{pmatrix}, \begin{pmatrix}
1 & 0.5 \\
1 & 1
\end{pmatrix} + (1 - X_{i3}) \begin{pmatrix}
2 \\
0.25
\end{pmatrix} \right),
\]

\[X_{i3} \sim Ber(0.6X_{i4} + 0.4(1 - X_{i4})), \quad X_{i4} \overset{i.i.d.}{\sim} Ber(0.5)\]

Treatment variable

\[T_i \sim Ber(e_i), \quad e_i = \expit \{(1, X_{i1}, X_{i2}, X_{i3}, X_{i4})\beta\},\]

where \(\beta = (-1.5, 0.5, -0.75, 2, -0.5)^\top\).

- This is “good overlap” situation in Mao et al., 2019

Outcome

\[Y_i = (1, T_i, T_i \times X_{i1}, T_i \times X_{i2}, X_i^\top) \alpha + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} N(0, 1),\]

where \(\beta = (0.5, 1, 1, -1, 1, 0.6, 2.2, -1.2)^\top\).

- This is similar as “heterogeneous treatment effect” situation in Mao et al., 2019

Under the data generating process, we consider two situations: an outcome model is mis-specified, and

1) a propensity score model is misspecified, and the true model is included in the candidate when our proposed method or a MA are applied

2) a propensity score model is misspecified, and the true model is not included in the candidates when our proposed method or a MA are applied

Note that we do not consider the situation where both models are correct or incorrect since the results are obvious. In the simulations, the following model is used for the outcome
model estimation:

\[ E[Y|x_i] = (1, T_i, T_i \times X_{i1}, T_i \times X_{i2}, X_{i3}) \alpha'. \]

In the situation 1), the following model is used for the propensity score model estimation except for our proposed method or a MA:

\[ e_i(\beta') = \expit \{(1, X_{i3})\beta'\} . \] (3.1)

In addition to (3.1), the true model is included in the candidate for our proposed method or a MA. Whereas, the following model is included in the candidate in the situation 2):

\[ e_i(\beta'') = \expit \{(1, X_{i4})\beta''\} . \]

Obviously, the propensity score model is misspecified in the situation 2).

The results are summarized in the Table 1 and from Figure 1 to Figure 4.
Table 1: Summary of estimates for the ATO

| Situation | Estimator | Propensity score estimation | Small sample \( n = 250 \) | | Large sample \( n = 500 \) |
|-----------|-----------|-----------------------------|-----------------------------|-----------------------------|
|           |           | Mean(SD) | Median(Range) | RMSE | Mean(SD) | Median(Range) | RMSE |
| 1)        | IPW-type  | Naive     | 1.553 (0.326) | 1.559 (0.55, 2.52) | 0.534 | 1.535 (0.235) | 1.534 (0.81, 2.33) | 0.468 |
|           |           | Proposed  | 1.137 (0.207) | 1.133 (0.52, 1.71) | 0.207 | 1.120 (0.149) | 1.119 (0.65, 1.66) | 0.149 |
|           |           | Model ave. | 1.137 (0.207) | 1.133 (0.52, 1.71) | 0.207 | 1.120 (0.149) | 1.119 (0.65, 1.66) | 0.149 |
|           | AIPW-type | Naive | 0.728 (0.225) | 0.733 (0.06, 1.45) | 0.461 | 0.715 (0.154) | 0.714 (0.20, 1.16) | 0.443 |
|           |           | Proposed | 1.136 (0.204) | 1.128 (0.47, 1.73) | 0.204 | 1.122 (0.146) | 1.120 (0.65, 1.58) | 0.146 |
|           |           | Model ave. | 1.136 (0.204) | 1.128 (0.47, 1.73) | 0.204 | 1.122 (0.146) | 1.120 (0.65, 1.58) | 0.146 |
|           | BR-type  | Naive     | 0.823 (0.233) | 0.824 (0.13, 1.49) | 0.386 | 0.810 (0.161) | 0.809 (0.15, 1.29) | 0.358 |
|           |           | Proposed  | 1.136 (0.205) | 1.128 (0.47, 1.74) | 0.205 | 1.122 (0.146) | 1.120 (0.65, 1.58) | 0.146 |
|           |           | Model ave. | 1.136 (0.205) | 1.128 (0.47, 1.74) | 0.205 | 1.122 (0.146) | 1.120 (0.65, 1.58) | 0.146 |
| 2)        | IPW-type  | Proposed | 1.559 (0.320) | 1.561 (0.57, 2.49) | 0.536 | 1.552 (0.233) | 1.552 (0.81, 2.37) | 0.482 |
|           |           | Model ave. | 1.566 (0.322) | 1.572 (0.58, 2.50) | 0.542 | 1.555 (0.236) | 1.552 (0.83, 2.37) | 0.486 |
|           | AIPW-type | Proposed | 0.849 (0.222) | 0.860 (0.16, 1.55) | 0.358 | 0.830 (0.161) | 0.836 (0.31, 1.34) | 0.340 |
|           |           | Model ave. | 0.849 (0.225) | 0.860 (0.16, 1.52) | 0.360 | 0.835 (0.171) | 0.843 (0.20, 1.34) | 0.342 |
|           | BR-type  | Proposed | 0.856 (0.204) | 0.858 (0.23, 1.49) | 0.341 | 0.843 (0.146) | 0.844 (0.15, 1.28) | 0.322 |
|           |           | Model ave. | 0.853 (0.206) | 0.856 (0.22, 1.44) | 0.346 | 0.837 (0.148) | 0.839 (0.33, 1.27) | 0.328 |

Note: The results of “Naive” in situation 2) are the same as situation 1).
3.2 Another application examples

As described in Introduction, our proposed method can be applied usual propensity score based procedures. For instance, IPW estimator, matching estimator, and subclassification estimator for the ATE. However, one of the important notations is that the weight estimators for each candidates models \( \hat{\gamma} \) only have the \( \sqrt{n} \)-consistency; not the asymptotic normality. Theoretically, this difference is important since the propensity score based procedures have different property compared with using only one propensity score estimator. Applicationally, constructing confidence intervals by bootstrap method is necessary.

Since our proposed method is based on GLM, the application for the prognostic score (Hansen, 2008) is also considered. As described in Antonelli et al., 2018, the doubly robust matching estimator, in the sense that the estimator becomes the consistency either the propensity score model or the prognostic score model is correctly specified, is proposed. By applying our proposed method, the multiply robust matching estimator can be constructed.

4 Conclusion

In this paper, we propose the novel estimator for the generalized linear model. Construction of the estimator is very simple: some candidate model is summarized to one integrated model, and we can use the integrated model for the following analysis or inference. The estimator has the multiple robustness in the sense that the true model can be detected accurately when the true model is included in the candidate models without model selection. Also, even if the true model is not included, the estimator is better than each candidate in the sense that the Kullback-Leibler divergence is smaller. Application for the propensity score is considered, and confirm its performance through simulation datasets.

As mentioned section 3, there are many applications to apply the proposed method. Especially, comparing with some machine learning procedures, easy to use is an advantage of our proposed method. As mentioned Imbens and Rubin (2015), the prior analysis phase
such as estimating the propensity score is called “design phase”. In design phase, we think the machine learning procedures may be hard to implement since we need to take care not only properties as the propensity score but also the tuning parameters. Whereas, regarding our proposed method, we need to take care only the former point (since parametric model is assumed implicitly). Even if there are some candidate models, we need not to discuss; the models are summarized to the integrated model. Therefore, our proposed method can contribute “simple” causal inference.
References

[1] Andrews, D. W. (1999). Estimation when a parameter is on a boundary. *Econometrica, 67*(6), 1341-1383.

[2] Antonelli, J., Cefalu, M., Palmer, N., and Agniel, D. (2018). Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics, 74*(4), 1171-1179.

[3] Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine, 26*(4), 734-753.

[4] Bang, H., and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*(4), 962-973.

[5] Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology, 163*(12), 1149-1156.

[6] Cunhas, M., and Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal, 61*(4), 1049-1072.

[7] Chu, A. T., Holt, S. K., Wright, J. L., Ramos, J. D., Grivas, P., Yu, E. Y., and Gore, J. L. (2019). Delays in radical cystectomy for muscle-invasive bladder cancer. *Cancer, 125*(12), 2011-2017.

[8] Fong, C., Hazlett, C., and Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics, 12*(1), 156-177.

[9] Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association, 109*(507), 1159-1173.
[10] Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2), 481-488.

[11] Hernán, M. A., and Robins, J. M. (2020). *Causal inference: what if*. Chapman & Hill/CRC.

[12] Hirano, K., and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164, 73-84.

[13] Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161-1189.

[14] Hjort, N. L., and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464), 879-899.

[15] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors). *Statistical science*, 14(4), 382-417.

[16] Imai, K., and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243-263.

[17] Imai, K., and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467), 854-866.

[18] Imbens, G. W., and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[19] Kang, J. D., and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 523-539.
[20] Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association, 114*(526), 645-656.

[21] Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association, 113*(521), 390-400.

[22] Mao, H., Li, L., and Greene, T. (2019). Propensity score weighting analysis and treatment effect discovery. *Statistical methods in medical research, 28*(8), 2439-2454.

[23] McCullagh, P., and Nelder, J. A. (2019). *Generalized linear models*. Routledge.

[24] Orihara, S., and Hamada, E. (2019). Double robust estimator in general treatment regimes based on Covariate-balancing. *Communications in Statistics-Theory and Methods, 48*(3), 462-478.

[25] Orihara, S., and Hamada, E. (2021). Determination of the optimal number of strata for propensity score subclassification. *Statistics & Probability Letters, 168*, 108951.

[26] Patel, V. G., Oh, W. K., and Galsky, M. D. (2020). Treatment of muscle-invasive and advanced bladder cancer in 2020. *CA: a cancer journal for clinicians, 70*(5), 404-423.

[27] Raftery, A. E., and Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association, 98*(464), 931-938.

[28] Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

[29] Shimizu, F., Muto, S., ..., and Horie, S. (2017). Effectiveness of platinum-based adjuvant chemotherapy for muscle-invasive bladder cancer: A weighted propensity score analysis. *International Journal of Urology, 24*(5), 367-372.

[30] Wang, L., Zhang, Y., Richardson, T. S., and Zhou, X. H. (2016). Robust estimation of propensity score weights via subclassification. *arXiv preprint arXiv:1602.06366*.
[31] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, 1-25.

[32] Xie, Y., Zhu, Y., Cotton, C. A., and Wu, P. (2019). A model averaging approach for estimating propensity scores by optimizing balance. *Statistical methods in medical research*, 28(1), 84-101.

[33] Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511), 910-922.
A Regularity Conditions

C.1 For all $k \in \{1, \ldots, K\}$, $\hat{\beta} \overset{P}{\rightarrow} \beta^*$, where $\beta^*$ is the “true” parameter value (see White, 1982).

C.2 For all $\theta \in \mathbb{R}$, $\ddot{b}(\theta) > 0$, where $\ddot{b}$ is the second differentiales of $b$.

C.3 $\exists x \in \mathbb{R}^p$ s.t. $\theta_k(x; \beta^*_k) \neq \theta_{k'}(x; \beta^*_{k'})$ for all combinations of $k \neq k'$.

C.4 $\forall \delta_n \rightarrow 0,$

$$\sup_{\gamma : \|\gamma - \gamma^0\| < \delta_n} \frac{|R(\gamma)|}{(1 + \|\sqrt{n}(\gamma - \gamma^0)\|)^2} = o_p(1),$$

where $\gamma^0 = (1, 0, \ldots, 0)^T$, and $R(\gamma)$ is 3rd remainder term of the taylor expansion of the likelihood function of the integrated model around $\gamma^0$.

B Proofs

B.1 Proof of Theorem 1

From C.1 and the continuous mapping theorem, (2.4) becomes

$$S(\gamma, \hat{\beta}) = S(\gamma, \beta^*) + o_p(1).$$

Also, from (2.4),

$$\frac{\partial^2}{\partial \gamma \partial \beta^*} \ell(\gamma, \beta^*) = \frac{\partial}{\partial \gamma} S^T(\gamma, \beta^*) = -\frac{1}{n} \sum_{i=1}^n \ddot{b} \left( \sum_{k'=1}^K \gamma_{k'} \theta_{k'}(x_i; \beta^*_{k'}) \right) \left( \begin{array}{c} \theta_1(x_i; \beta^*_1) \\ \theta_2(x_i; \beta^*_2) \\ \vdots \\ \theta_K(x_i; \beta^*_K) \end{array} \right).$$

(B.1)
Therefore, from C.2, (B.1) becomes the negative definite:

$$\frac{\partial^2}{\partial \gamma^2} \ell(\gamma, \beta^*) < 0.$$  \hspace{1cm} (B.2)

From (B.2), (2.4) has the unique solution \( \hat{\gamma} \xrightarrow{P} \gamma^0 \) satisfying

$$E \left[ \begin{bmatrix} T - \hat{b} \left( \sum_{k' = 1}^{K} \gamma_{k'}^0 \theta_{k'}(X; \beta_{k'}^*) \right) \end{bmatrix} \begin{pmatrix} \theta_1(X; \beta_1^0) \\ \theta_2(X; \beta_2^*) \\ \vdots \\ \theta_K(X; \beta_K^*) \end{pmatrix} \right] = 0$$ \hspace{1cm} (B.3)

Next, the concrete values of \( \gamma^0 \) is confirmed. The expectation regarding \( T|\mathbf{x} \) of (B.3) is

$$E \left[ T - \hat{b} \left( \sum_{k' = 1}^{K} \gamma_{k'}^0 \theta_{k'}(\mathbf{x}; \beta_{k'}^*) \right) | \mathbf{x} \right]$$ \hspace{1cm} (B.4)

From the property of GLM (see McCullagh and Nelder, 2019),

$$E[ T | \mathbf{x} ] = \hat{b}(\theta_1(\mathbf{x}; \beta_1^0)).$$

Therefore, (B.4) becomes

$$\hat{b}(\theta_1(\mathbf{x}; \beta_1^0)) - \hat{b} \left( \sum_{k' = 1}^{K} \gamma_{k'}^0 \theta_{k'}(\mathbf{x}; \beta_{k'}^*) \right).$$

From the above, (B.3) becomes

$$E \left[ \begin{bmatrix} \hat{b}(\theta_1(X; \beta_1^0)) - \hat{b} \left( \sum_{k' = 1}^{K} \gamma_{k'}^0 \theta_{k'}(X; \beta_{k'}^*) \right) \end{bmatrix} \begin{pmatrix} \theta_1(X; \beta_1^0) \\ \theta_2(X; \beta_2^*) \\ \vdots \\ \theta_K(X; \beta_K^*) \end{pmatrix} \right] = 0.$$ \hspace{1cm} (B.5)
From C.3, there exists the unique value of $\gamma$ such that

$$
\gamma_1^0 = 1 \quad \text{and} \quad \gamma_k^0 = 0, \quad k = 2, \ldots, K.
$$

Therefore, (2.5) is obtained.

### B.2 Proof of Theorem 3

It is sufficient to show that

$$
- \mathbb{E} \left[ \log f(T|X; \gamma, \beta^*) \right] + \sum_{k=1}^{K} \gamma_k \mathbb{E} \left[ \log f(T|X; \beta_k^*) \right] \leq 0,
$$

i.e.,

$$
- \int \left( t \sum_{k=1}^{K} \gamma_k \theta_k(x; \beta_k^*) - b \left( \sum_{k=1}^{K} \gamma_k \theta_k(x; \beta_k^*) \right) \right) f(y, x) dy dx \\
+ \sum_{k=1}^{K} \gamma_k \left[ \int (t \theta_k(x; \beta_k^*) - b(\theta_k(x; \beta_k^*))) f(t, x) dy dx \right]
= \int \left[ b \left( \sum_{k=1}^{K} \gamma_k \theta_k(x; \beta_k^*) \right) \right] f(x) dx - \sum_{k=1}^{K} \gamma_k \int (t \theta_k(x; \beta_k^*) - b(\theta_k(x; \beta_k^*))) f(x) dx
= \int \left[ b \left( \sum_{k=1}^{K} \gamma_k \theta_k(x; \beta_k^*) \right) \right] f(x) dx \leq 0. \tag{B.6}
$$

From C.2, by using the property of convex functions,

$$
b \left( \sum_{k=1}^{K} \gamma_k \theta_k \right) \leq \sum_{k=1}^{K} \gamma_k b(\theta_k)
$$

for all $\theta_k$ and $\gamma$. Therefore, (B.6) is hold; (2.6) is obtained.
C Construction of BR-type estimator

To estimate the ATO, we introduce the previous two estimators: the IPW-type and the AIPW-type estimators.

- IPW-type estimator (Li et al., 2018)

\[
\hat{\tau}_{\text{IPW}}^{\text{ATO}} := \frac{\sum_{i=1}^{n} T_i (1 - e_i) Y_i}{\sum_{i=1}^{n} T_i (1 - e_i)} - \frac{\sum_{i=1}^{n} (1 - T_i) e_i Y_i}{\sum_{i=1}^{n} (1 - T_i) e_i}.
\]

- AIPW-type estimator (Mao et al., 2019)

\[
\hat{\tau}_{\text{AIPW}}^{\text{ATO}} := \frac{\sum_{i=1}^{n} e_i (1 - e_i) (m_{1i} - m_{0i})}{\sum_{i=1}^{n} e_i (1 - e_i)} + \frac{\sum_{i=1}^{n} T_i (1 - e_i) (Y_i - m_{1i})}{\sum_{i=1}^{n} T_i (1 - e_i)} - \frac{\sum_{i=1}^{n} (1 - T_i) e_i (Y_i - m_{0i})}{\sum_{i=1}^{n} (1 - T_i) e_i}
\]

(C.1)

where \( m_{ti} = E[Y_t|X = x_i] \), \( t \in \{0, 1\} \) is the true outcome model.

Focusing on the first term of (C.1), the denominator becomes

\[
\frac{1}{n} \sum_{i=1}^{n} e_i (1 - e_i) (m_{1i} - m_{0i}) \xrightarrow{P} E[e(X)(1 - e(X))(m_1(X) - m_0(X))]
\]

\[
= E[e(X)(1 - e(X))E[Y_1 - Y_0|X]].
\]

Therefore,

\[
\frac{\sum_{i=1}^{n} e_i (1 - e_i) (m_{1i} - m_{0i})}{\sum_{i=1}^{n} e_i (1 - e_i)} \xrightarrow{P} \tau_{\text{ATO}}.
\]

From the above, AIPW-type estimator (C.1) can be considered as

\[
\hat{\tau}_{\text{AIPW}}^{\text{ATO}} = \text{“Main term” + “Augmentation term”}.
\]

The augmentation term is involved in the efficiency improvement of IPW-type estimator.

From here, we consider the boundedness; when considering the property, we assume
$Y = \{0, 1\}^\otimes 2$. Actually, the AIPW-estimator (C.1) does not have the property. Obviously, the first term falls within $(-1, 1)$. Whereas, the augmentation terms cause the break-down of the boundedness. To confirm it, we consider an extreme example: almost all subjects are $T_i = 1$, $e_i \approx 0$, $Y_i = 0$, and $m_{1i} \approx 1$. Then, the second and third terms become

$$\sum_{i=1}^{n} T_i (1 - e_i) (Y_i - m_{1i}) \approx -1, \quad \sum_{i=1}^{n} (1 - T_i) e_i (Y_i - m_{0i}) \approx 0.$$  

This is very rare or unusual situation, however, the AIPW-type estimator may fail the boundedness. To overcome the problem, the novel BR-type estimator is derived. The main flow is the same as the Bang and Robins’ estimator for the ATE; more precisely, we consider the outcome regression based doubly robust estimator. To derive the estimator, we consider the following estimating equation:

$$\sum_{i=1}^{n} \left( \frac{\partial}{\partial \beta} \tilde{m}(T_i, X_i; \beta) \right) \left( Y_i - \tilde{m}(T_i, X_i; \beta, \phi) \right) = 0. \quad (C.2)$$

where

$$\tilde{m}(T_i, X_i; \beta) = \hat{m}(T_i, X_i; \beta) + \frac{T_i (1 - e(X_i; \hat{\alpha}))}{\sum_{i=1}^{n} T_i (1 - e(X_i; \hat{\alpha}))} \phi_1 + \frac{(1 - T_i) e(X_i; \hat{\alpha})}{\sum_{i=1}^{n} (1 - T_i) e(X_i; \hat{\alpha})} \phi_0,$$

$\hat{m}(T_i, X_i; \beta)$ is an outcome model, and $e(X_i; \hat{\alpha})$ are estimated propensity scores. In other words, estimators $\hat{\beta}$ and $\hat{\phi}$ are the solutions of (C.2). To estimate the ATO, constructing the following estimator:

- BR-type estimator (proposed estimator)

$$\tau_{ATO}^{br} := \frac{1}{n} \sum_{i=1}^{n} \tilde{m}(1, X_i; \hat{\beta}, \hat{\phi}) - \frac{1}{n} \sum_{i=1}^{n} \tilde{m}(0, X_i; \hat{\beta}, \hat{\phi}) \quad (C.3)$$

The proposed BR-type estimator is known as a “plug-in” type estimator since treatment
values $t = 0$ and $1$ are plugged into the estimated outcome models $\tilde{m}(t, X_i; \hat{\beta}, \hat{\phi})$. This type estimator is easy to calculate, and will be used not only in academia but also industry (see FDA guidance, 2019). Note that $\hat{\tau}_{ATO}^{br}$ has the boundedness obviously. From here, we consider the properties of the proposed BR-type estimator.

**Theorem 4.**

If propensity score models $e(X_i; \alpha)$ are correctly specified, then $\hat{\tau}_{ATO}^{br} \xrightarrow{P} \tau_{ATO}$. Additionally, outcome models $\tilde{m}(T_i, X_i; \beta)$ are correctly specified, then $\hat{\tau}_{ATO}^{br}$ has the semiparametric efficiency.

The statement of Theorem 4 is different from the ordinary double robustness: the propensity score models need to be specified. Since the proof is the same flow as p.963-964 of Bang and Robins (2005), we do not provide it in this paper.

D Simulation results
Figure 1: Simulation results (1/4)
Figure 2: Simulation results (2/4)
Figure 3: Simulation results (3/4)

Situation: 2) & N = 250
Figure 4: Simulation results (4/4)