Statistica Sinica Preprint No: SS-2018-0416	
Title	Sufficient Dimension Reduction for Feasible and Robust Estimation of Average Causal Effect
Manuscript ID	SS-2018-0416
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0416
Complete List of Authors	Trinetri Ghosh Yanyuan Ma and Xavier de Luna
Corresponding Author	Trinetri Ghosh
E-mail	tbg5133@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

Sufficient Dimension Reduction for Feasible and Robust Estimation of Average Causal Effect

Trinetri Ghosh, Yanyuan Ma and Xavier de Luna

Pennsylvania State University, Pennsylvania State University and Umeå University

Abstract:

To estimate the treatment effect in an observational study, we use a semiparametric locally efficient dimension-reduction approach to assess the treatment assignment mechanisms and average responses in both the treated and the non-treated groups. We then integrate our results using imputation, inverse probability weighting, and doubly robust augmentation estimators. Doubly robust estimators are locally efficient, and imputation estimators are super-efficient when the response models are correct. To take advantage of both procedures, we introduce a shrinkage estimator that combines the two. The proposed estimators retains the double robustness property, while improving on the variance when the response model is correct. We demonstrate the performance of these estimators using simulated experiments and a real data set on the effect of maternal smoking on baby birth weight.

Key words and phrases: Average Treatment Effect, Double Robust Estimator, Efficiency, Inverse Probability Weighting, Shrinkage Estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

1. Introduction

Dimension reduction is a major methodological issue in observational studies that estimate the causal effect of a non-randomized treatment. This is largely because of the increased availability of health and administrative registers, giving access to high-dimensional pre-treatment information sets that can help identifying causal effects of interest. To better estimate the average causal effect of a treatment under possibly high-dimensional covariates, while maintaining flexibility in terms of the model assumptions, we propose and study new estimators. These estimators are based on semiparametric sufficient dimension-reduction methods, together with various well-known missing-data approaches, including imputation, inverse probability weighting (IPW) and doubly robust augmentation estimators. To take advantage of the various estimators' properties, we propose a new shrinkage-based procedure to estimate the average causal effect. The resulting estimator is consistent in estimating the causal effect, even when the treatment assignment model or one of the outcome models in the treated and untreated groups is misspecified. Furthermore, its asymptotic variance is no larger than that of any single approach.

Dimension reduction for feasible nonparametric and semiparametric causal inference has recently been formalized, with most contributions fo-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

cusing on covariate selection, that is, methods that determine which covariates are confounders that need to be controlled for; see, for example, Gruber & van der Laan (2010), de Luna et al. (2011), Farrell (2015), and Shortreed & Ertefaie (2017). Dimension reduction must consider nuisance conditional models, that is, the probability of treatment given the covariates (propensity score), and models for the two potential responses (i.e., responses under two possible levels of a binary treatment) given the covariates (de Luna et al. 2011). Sufficient dimension reduction (Li 1991, Li & Duan 1991, Cook 1998, Xia et al. 2002, Xia 2007, Ma & Zhu 2012) constitutes an alternative to covariate selection, and has the advantage that, in addition to considering covariates in isolation as confounders, it can accommodate linear combinations of the whole covariate set. Such methods have recently attracted attention in semiparametric causal inference. For example, Liu et al. (2018) considered sufficient dimension reduction when estimating the propensity score alone, and Luo et al. (2017) considered that when estimating the response models alone. In contrast, Ma et al. (2018) considered classical sufficient dimension in all nuisance models.

In this study, we take a general approach to estimating the average causal effect. We first use efficient semiparametric sufficient dimension-reduction methods (Ma & Zhu 2013, 2014) in all nuisance models to explain

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

the potential responses and the treatment assignment. Then, we combine these into classical imputation (IMP) and IPW estimators. Although our semiparametric sufficient dimension-reduction model is very flexible, nuisance models may still be misspecified. Thus, a doubly robust estimator (augmented inverse probability weighting (AIPW) estimator) is also considered, which allows for the misspecification of one of the nuisance models. The AIPW estimator is locally efficient, in the sense that it reaches efficiency at the true nuisance models. The imputation estimator is super-efficient, in the sense that if the true response model is known, then this knowledge yields a lower asymptotic efficiency bound than that which the AIPW estimator may reach (Tan 2007). We therefore propose a novel estimator that shrinks the imputation and AIPW estimators toward each other. The shrinkage estimator is also doubly robust. Furthermore, it is asymptotically equivalent to the AIPW estimator if the response model is misspecified; if all nuisance models are correctly specified, it shrinks toward the imputation estimator, which is more efficient than the AIPW in this case. In general, the variability of the estimator is no larger than that of the AIPW or IMP.

The remainder of the paper is organized as follows. Section 2 introduces the semiparametric sufficient dimension-reduction structures and their estimations for the nuisance models. Section 3 proposes estimators of average

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

causal effect using the models and estimations pressented in Section 2. This section also provides the asymptotic properties of the imputation, IPW, AIPW, and shrinkage estimators. Section 4 examines the finite-sample performance of the estimators for different designs, including well-specified and misspecified situations. A real data example on the effect of smoking on birth weight illustrates the use of the methods proposed in Section 5. Section 6 concludes the paper.

2. Model and Dimension Reduction

Let Y_T be the treatment response under treatment T, where $T = 1$ if the treatment of interest is applied, and $T = 0$ if some alternative treatment (e.g., a placebo or no treatment) is applied. Let $\mathbf{X} \in \mathcal{R}^p$ be the set of pre-treatment covariates. We observe a random sample $\{\mathbf{X}_i, T_i, Y_{1i}T_i + Y_{0i}(1 - T_i)\}$, for $i = 1, \ldots, n$. In particular, Y_{ti} is observed only for unit i, such that $T_i = t$, and is therefore called a potential response. Our goal is to estimate the average causal effect of the treatment, here $D = E(Y_1 - Y_0)$. We assume $0 < \text{pr}(T = 1 \mid Y_0, Y_1, \mathbf{X}) = \text{pr}(T = 1 \mid \mathbf{X}) < 1$ throughout. This assumption is often called strong ignorability of the treatment assignment, and yields the parameter D under the above sampling scheme (e.g., Rosenbaum & Rubin 1983).

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

We now describe the flexible dimension-reduction structures that we combine into different semiparametric estimators for D. First, the treatment assignment probability, also called the propensity score in the literature, can be modeled as

$$\mathrm{pr}(T = 1 \mid \mathbf{X} = \mathbf{x}) = e^{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})}/\{1 + e^{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})}\}, \tag{2.1}$$

where $\eta(\cdot)$ is an unknown function that is smooth and bounded from both above and below to guarantee that the propensity is strictly in $(0, 1)$, and $\boldsymbol{\alpha}$ is an unknown index vector or matrix with dimension $p \times d_\alpha$, for $p > d_\alpha$.

Further, we model Y_1 given $\mathbf{X} = \mathbf{x}$ using the flexible dimension-reduction model

$$Y_1 = m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + \epsilon_1, \tag{2.2}$$

where $E(\epsilon_1 \mid \mathbf{x}) = 0$. Similarly, we model Y_0 given $\mathbf{X} = \mathbf{x}$ as

$$Y_0 = m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) + \epsilon_0, \tag{2.3}$$

where $E(\epsilon_0 \mid \mathbf{x}) = 0$. Here, $m_1(\cdot)$ and $m_0(\cdot)$ are unknown functions, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ are unknown index vectors or matrices with dimensions $p \times d_1$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

$p \times d_0$, respectively, for $p > d_1$ and $p > d_0$, respectively.

The models (2.1), (2.2), and (2.3) separately describe the probability of receiving treatment and the mean potential responses, respectively, without imposing a relation between these models. Indeed, unless prior knowledge suggests otherwise, the three processes are irrelevant to each other and, hence, should be modeled separately. Conceptually, when the structural dimension (d_α, d_1 or d_0) is p, dimension-reduction modeling includes nonparametric modeling; hence, using the dimension-reduction models in (2.1), (2.2), and (2.3) provides large flexibility in practice. Using each of the three models, we can estimate the corresponding unknown parameters and unknown functions separately using a random sample. We can then combine these estimators in various ways to estimate the treatment effect $D = E(Y_1 - Y_0)$.

2.1 Estimation of Response Models

We first consider (2.2). Because of the ignorability of the treatment assignment assumption, the treated subsample forms a random sample from which to fit model (2.2). Thus, we can directly implement the semiparametric method of Ma & Zhu (2014) for the estimations of $\boldsymbol{\beta}_1$ and $m_1(\cdot)$ based on the subset of the data with $T_i = 1$. For identifiability purposes,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

we adopt the parameterization of Ma & Zhu (2014), fix the upper $d_1 \times d_1$ submatrix of $\boldsymbol{\beta}_1$ as the identity matrix, and leave the lower $(p-d_1)\times d_1$ submatrix arbitrary. Thus, the locally efficient estimator of $\boldsymbol{\beta}_1$ is obtained by solving

$$\sum_{i=1}^{n} t_i\{y_{1i} - \widehat{m}_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\beta}_1)\}\widehat{\mathbf{m}}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\beta}_1) \otimes \{\mathbf{x}_{Li} - \widehat{E}(\mathbf{X}_{Li} \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}, \quad (2.4)$$

where the Nadaraya–Watson kernel estimator is used to obtain $\widehat{E}(\mathbf{X}_L \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$, and the local linear estimator is used to obtain $\widehat{m}_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}_1)$ and $\widehat{\mathbf{m}}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}_1)$, where $\mathbf{X}_L$ represents the subvector of $\mathbf{X}$ formed by the lower $p-d_1$ components. Specifically, in (2.4), $\widehat{E}(\mathbf{X}_L \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) = \sum_{i=1}^{n} \mathbf{x}_{Li} K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) / \sum_{i=1}^{n} K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$ and $\widehat{m}_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}_1) = c_0, \widehat{\mathbf{m}}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}_1) = \mathbf{c}_1$ are the solution to

$$\min_{c_0, \mathbf{c}_1} \sum_{i=1}^{n} t_i\{y_{1i} - c_0 - \mathbf{c}_1^{\mathrm{T}}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})\}^2 K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}). \quad (2.5)$$

Many kernel functions can be used, for example, the Epanechnikov kernel $(1-u^2)3/4I(|u| \leqslant 1)$, the quartic kernel $(1-u^2)^2 15/16 I(|u| \leqslant 1)$, and so

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

on. It is easy to verify that the minimizer of (2.5) has the explicit form

$$\begin{aligned}\widehat{m}_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x},\boldsymbol{\beta}_1) &= A_{11}-\mathbf{A}_{13}^{\mathrm{T}}(\mathbf{A}_{14}-\mathbf{A}_{13}\mathbf{A}_{13}^{\mathrm{T}})^{-1}(\mathbf{A}_{12}-\mathbf{A}_{13}A_{11}), \quad (2.6)\\ \widehat{\mathbf{m}}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x},\boldsymbol{\beta}_1) &= (\mathbf{A}_{14}-\mathbf{A}_{13}\mathbf{A}_{13}^{\mathrm{T}})^{-1}(\mathbf{A}_{12}-\mathbf{A}_{13}A_{11}),\end{aligned}$$

where $A_{11} = \sum_{i=1}^n t_i y_{1i} K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})/\sum_{i=1}^n t_i K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$, $\mathbf{A}_{12} = \sum_{i=1}^n t_i y_{1i}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})/\sum_{i=1}^n t_i K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$, $\mathbf{A}_{13} = \sum_{i=1}^n t_i(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})/\sum_{i=1}^n t_i K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$, $\mathbf{A}_{14} = \sum_{i=1}^n t_i(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})^{\otimes 2}K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})/\sum_{i=1}^n t_i K_h(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ throughout the text. Note that the above description is a typical profiling estimation procedure for $\boldsymbol{\beta}_1$. Once we obtain $\widehat{\boldsymbol{\beta}}_1$, we then estimate m_1 using $\widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x},\widehat{\boldsymbol{\beta}}_1)$ given in (2.6). Note that the incorporation of the kernel-based nonparametric estimation enables us to perform the dimension reduction without assuming the frequently adopted linearity or constant variance conditions.

Theorem 1 of Ma & Zhu (2014) established the property of the above estimator. Specifically, the estimator $\widehat{\boldsymbol{\beta}}_1$ satisfies

$$\begin{aligned}\sqrt{n_1}\mathrm{vecl}(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1) &= -\mathbf{B}_1 n_1^{-1/2}\sum_{i=1}^n t_i\{y_{1i}-m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}\mathrm{vec}[\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\\ &\quad\otimes\{\mathbf{x}_{Li}-E(\mathbf{X}_{Li}\mid\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}]+o_p(1), \quad (2.7)\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

where $n_1 = \sum_{i=1}^{n} T_i$, $\mathrm{vecl}(\boldsymbol{\beta}_1)$ is the vector formed by the lower $(p-d_1)\times d_1$ submatrix of $\boldsymbol{\beta}_1$, and

$$\begin{aligned} &\mathbf{B}_1 \\ \equiv\ &\left\{E\left(\frac{\partial \mathrm{vec}[T_i\{Y_{1i}-m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)\}\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)\otimes\{\mathbf{X}_{Li}-E(\mathbf{X}_{Li}\mid\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)\}]}{\partial \mathrm{vecl}(\boldsymbol{\beta}_1)^{\mathrm{T}}}\right)\right\}^{-1}. \end{aligned} \tag{2.8}$$

We can estimate $\boldsymbol{\beta}_0$ and m_0 in a similar manner using the subset of the data set corresponding to $T_i = 0$. Then, implementing Theorem 1 from Ma & Zhu (2014), the asymptotic behavior of the efficient estimator $\widehat{\boldsymbol{\beta}}_0$ is given by

$$\begin{aligned} \sqrt{n_0}\mathrm{vecl}(\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0) &= -\mathbf{B}_0 n_0^{-1/2}\sum_{i=1}^{n}(1-t_i)\{y_{0i}-m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}\mathrm{vec}[\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) \\ &\quad\otimes\{\mathbf{x}_{Li}-E(\mathbf{X}_{Li}\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}]+o_p(1), \end{aligned} \tag{2.9}$$

where $n_0 = n - n_1$, and

$$\begin{aligned} &\mathbf{B}_0 \\ \equiv\ &\left\{E\left(\frac{\partial \mathrm{vec}[(1-T_i)\{Y_{0i}-m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\}\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\otimes\{\mathbf{X}_{Li}-E(\mathbf{X}_{Li}\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\}]}{\partial \mathrm{vecl}(\boldsymbol{\beta}_0)^{\mathrm{T}}}\right)\right\}^{-1}. \end{aligned} \tag{2.10}$$

When the mean function models are correct, the meanings of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_0$, m_1 and m_0 are easy to understand. When the models are incorrect, as we allow

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

in the following, we can understand $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_0$, m_1, and m_0 as quantities that satisfy

$$E[T\{Y_1 - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_1)\}\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_1) \otimes \{\mathbf{X}_L - E(\mathbf{X}_L \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})\}] = \mathbf{0},$$

$$E[(1-T)\{Y_0 - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_0)\}\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_0) \otimes \{\mathbf{X}_L - E(\mathbf{X}_L \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})\}] = \mathbf{0},$$

where $m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) = E(Y_1 \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) \neq E(Y_1 \mid \mathbf{x})$, and $m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = E(Y_0 \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) \neq E(Y_0 \mid \mathbf{x})$.

2.2 Estimation of Propensity Score Model

The estimation of $\boldsymbol{\alpha}, \eta$ has been studied previously (Liu et al. 2018, Ma & Zhu 2013). Hence, we provide the five-step algorithm here, for completeness and clarity.

Step 1. Form the Nadaraya–Watson estimator of $E(\mathbf{X}_i \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)$ to obtain $\widehat{E}(\mathbf{X}_i \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)$.

Step 2. Solve $\sum_{i=1}^n \mathrm{vecl}(\{\mathbf{x}_i - \widehat{E}(\mathbf{X}_i \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}[t_i - 1 + 1/\{1+\exp(\mathbf{1}_d^{\mathrm{T}}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}]\mathbf{1}_d^{\mathrm{T}}) = \mathbf{0}$ to obtain a consistent initial estimator $\widetilde{\boldsymbol{\alpha}}$.

Step 3. Obtain the local linear estimators of $\eta(\mathbf{z}, \boldsymbol{\alpha})$ and its first derivative

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

$\boldsymbol{\eta}'(\mathbf{z}, \boldsymbol{\alpha})$ by solving

$$\sum_{i=1}^{n} \left[t_i - \frac{\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mathbf{z})\}}{1 + \exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mathbf{z})\}} \right] K_h(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mathbf{z}) = 0 \quad (2.11)$$

$$\sum_{i=1}^{n} \left[t_i - \frac{\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mathbf{z})\}}{1 + \exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mathbf{z})\}} \right] (\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mathbf{z}) K_h(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mathbf{z}) \quad = \mathbf{0},$$

for $b_0, \mathbf{b}_1$ at $\mathbf{z} = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_1, \ldots, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_n$. Write the resulting estimator as $\widehat{\eta}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\alpha})$ and $\widehat{\boldsymbol{\eta}}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\alpha})$.

Step 4. Insert $\widehat{\eta}(\cdot, \boldsymbol{\alpha})$, $\widehat{\boldsymbol{\eta}}'(\cdot, \boldsymbol{\alpha})$ and $\widehat{E}(\cdot)$ into the estimating equation

$$\sum_{i=1}^{n} \{\mathbf{x}_{Li} - \widehat{E}(\mathbf{X}_{Li} \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\} \left[t_i - \frac{\exp\{\widehat{\eta}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}}{1 + \exp\{\widehat{\eta}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \widehat{\boldsymbol{\eta}}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} = \mathbf{0},$$

and solve to obtain the efficient estimator $\widehat{\boldsymbol{\alpha}}$, using the starting value $\widetilde{\boldsymbol{\alpha}}$.

Step 5. Repeat Step 3 at $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}$ to obtain the final estimator of $\eta(\cdot)$.

We then have $\widehat{\mathrm{pr}}(T = 1 \mid \mathbf{X} = \mathbf{x}) = \exp\{\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x})\}/[1 + \exp\{\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x})\}]$, which we use in the final calculation of the average causal effect. Let us write

$$p_i = \frac{\exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}}{1 + \exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}}, P_i = \frac{\exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\}}{1 + \exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\}}, \widehat{p}_i = \frac{\exp\{\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\}}{[1 + \exp\{\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\}]},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

and define

$$\mathbf{B} \equiv \left\{ E\left(\frac{\partial \mathrm{vec}\left[\{\mathbf{X}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\}(T_i - P_i)\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\right]}{\partial \mathrm{vecl}(\boldsymbol{\alpha})^{\mathrm{T}}} \right) \right\}^{-1} \tag{2.12}$$

Then, using Lemma 2 from Liu et al. (2018), we have

$$\begin{aligned} & \sqrt{n}\mathrm{vecl}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \qquad (2.13) \\ = \; & -\mathbf{B}n^{-1/2}\sum_{i=1}^{n}(t_i - p_i)\mathrm{vec}[\{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}] + o_p(1). \end{aligned}$$

When the propensity score model is correct, the meaning of $\boldsymbol{\alpha}$ and η is clear. When the model is incorrect, as we shall allow in the following, $\boldsymbol{\alpha}$ and η are quantities that satisfy

$$E[\{\mathbf{X}_L - E(\mathbf{X}_L \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X})\}\left[T - \frac{\exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X})\}}{1 + \exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X})\}}\right]\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}] = \mathbf{0},$$

where $[1 + \exp\{\eta(-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})\}]^{-1} = E(T \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) \neq E(T \mid \mathbf{x})$.

3. Average Causal Effect: Estimators and Properties

We are now ready to propose several estimators for estimating the average treatment effect, based on the semiparametric modeling and estimators described in Section 2. These propositions all take advantage of existing

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

methods in missing-at-random problems, including imputation and weighting; hence, they inherit the properties expected. We also introduce a novel shrinkage estimator that combines imputation and weighting, and has an optimal property. Let $y_i = t_i y_{1i} + (1-t_i)y_{0i}$ be the observed response value.

3.1 Imputation Estimators

First, we estimate the average causal effect using an imputation approach, as proposed in the context of missing data (Rubin 1978b). The imputation approach is semiparametric in spirit, similar to the nonparametric imputation (Wang et al. 2012). Specifically, we construct $\widehat{E}(Y_1) = n^{-1}\sum_{i=1}^n \{t_i y_i + (1-t_i)\widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)\}$, $\widehat{E}(Y_0) = n^{-1}\sum_{i=1}^n \{(1-t_i)y_i + \ t_i\widehat{m}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}_i)\}$, and then form the imputation estimator IMP as $\widehat{D}_{\mathrm{IMP}} = \widehat{E}(Y_1) - \widehat{E}(Y_0)$.

We further consider an alternative imputation estimator that uses the model-predicted values, while ignoring the observed responses, even when they are available. Specifically, we still form $\widehat{D}_{\mathrm{IMP2}} \equiv \widehat{E}(Y_1) - \widehat{E}(Y_0)$ for the treatment effect, using $\widehat{E}(Y_1) = n^{-1}\sum_{i=1}^n \widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)$ and $\widehat{E}(Y_0) = n^{-1}\sum_{i=1}^n \widehat{m}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}_i)$ to obtain the imputation estimator IMP2. The latter is sometimes called the outcome regression estimator; see, for example, Tan (2007).

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

3.2 (Augmented) IPW Estimators

Robins et al. (1994) proposed a class of semiparametric estimators based on IPW estimating equations, borrowing from Horvitz & Thompson (1952) in the survey sampling literature. Later, Liu et al. (2018) implemented an IPW estimator that uses semiparametric modeling to assess the propensity score function. Following this procedure, the IPW estimator first constructs $\widehat{E}(Y_1) = n^{-1}\sum_{i=1}^{n} t_i y_i/\widehat{p}_i$ and $\widehat{E}(Y_0) = n^{-1}\sum_{i=1}^{n}(1-t_i)y_i/(1-\widehat{p}_i)$, and then estimates the average causal effect $\widehat{D}_{\mathrm{IPW}} \equiv \widehat{E}(Y_1) - \widehat{E}(Y_0)$.

If at least one of the mean function models, $m_1(\cdot)$ and $m_0(\cdot)$, is incorrectly specified, the IMP and IMP2 estimators will be inconsistent. Similarly, if $\eta(\cdot)$ is incorrectly specified, the IPW will not be consistent. As a result, we use the more flexible semiparametric dimension-reduction models instead of fully parametric models. This reduces, but does not completely eliminate, the chance of model misspecification. Thus, we still need protection against either misspecification using the doubly robust estimator (Robins et al. 1994). This leads to the AIPW estimator, which is consistent when either the mean models are correctly specified or the propensity score model is correctly specified. The estimate of the average causal effect is still $\widehat{D}_{\mathrm{AIPW}} \equiv \widehat{E}(Y_1) - \widehat{E}(Y_0)$, where now $\widehat{E}(Y_1) = n^{-1}\sum_{i=1}^{n}\left\{t_i y_i/\widehat{p}_i + (1-t_i/\widehat{p}_i)\,\widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)\right\}$ and $\widehat{E}(Y_0) = n^{-1}\sum_{i=1}^{n}\left[(1-t_i)y_i/(1-\widehat{p}_i) + \{1-(1-t_i)/\right.$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

$(1-\widehat{p}_i)\}\widehat{m}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}_i)\big]$. An improved version of the AIPW estimator was proposed in Robins et al. (1995), providing extra protection against deteriorated estimation variability. Based on this idea, Tan (2006) later developed a nonparametric likelihood estimator. Adopting this idea in the treatment effect estimation framework, we construct the estimator $\widehat{E}(Y_1) = n^{-1}\sum_{i=1}^n \big\{t_i y_i/\widehat{p}_i + \widehat{\gamma}_1\,(1-t_i/\widehat{p}_i)\,\widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)\big\}$, $\widehat{E}(Y_0) = n^{-1}\sum_{i=1}^n \big[(1-t_i)y_i/(1-\widehat{p}_i) + \widehat{\gamma}_0\{1-(1-t_i)/(1-\widehat{p}_i)\}\widehat{m}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}_i)\big]$, and estimate the average causal effect by $\widehat{D}_{\mathrm{IAIPW}} \equiv \widehat{E}(Y_1) - \widehat{E}(Y_0)$. Here, $\widehat{\gamma}_1 = \mathrm{cov}\big\{\widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)t_i/\widehat{p}_i, (1-t_i/\widehat{p}_i)\,\widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)\big\}^{-1}\mathrm{cov}\big\{t_i y_i/\widehat{p}_i, (1-t_i/\widehat{p}_i)\,\widehat{m}_1(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)\big\}$ and $\widehat{\gamma}_0 = \mathrm{cov}\big[(1-t_i)/(1-\widehat{p}_i)\widehat{m}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}_i), \{1-(1-t_i)/(1-\widehat{p}_i)\}\,\widehat{m}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}_i)\big]^{-1}\mathrm{cov}\big[(1-t_i)y_i/(1-\widehat{p}_i), \{1-(1-t_i)/(1-\widehat{p}_i)\}\,\widehat{m}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}_i)\big]$.

3.3 The Shrinkage Estimator

The ideas of imputation and weighting are quite different, and each has its own advantages and drawbacks. For example, when the treatment mean models $m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})$ and $m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ are correct, regardless of whether or not the propensity score model is correct, the IMP and AIPW are both consistent; however, it is unclear which estimator is more efficient. When the treatment mean models $m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})$ and $m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ are not both correct, the AIPW is still consistent as long as the propensity score model is correct, but the IMP

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

methods will be inconsistent. Of course, if both the mean models and the propensity models are incorrect, then neither method will provide a consistent estimation. In practice, we typically do not know which scenario we are in, making it difficult to determie which method to employ. Therefore, in order to take advantage of both methods, we use the idea of a shrinkage estimator (Mukherjee & Chatterjee 2008) to construct a weighted average between the IMP and the AIPW.

The general observation is that if the IMP is consistent, then the AIPW will be consistent as well, but not vice versa. However, it is not generally clear which estimator is more efficient. We construct the following shrinkage estimator: Let $\sqrt{n}(\widehat{D}_{\mathrm{AIPW}} - D_{\mathrm{AIPW}}) \to N(0, v_{\mathrm{AIPW}})$ in distribution and $\sqrt{n}(\widehat{D}_{\mathrm{IMP}} - D_{\mathrm{IMP}}) \to N(0, v_{\mathrm{IMP}})$ in distribution, and let $\mathrm{cov}\{\sqrt{n}(\widehat{D}_{\mathrm{AIPW}} - D_{\mathrm{AIPW}}), \sqrt{n}(\widehat{D}_{\mathrm{IMP}} - D_{\mathrm{IMP}})\} \to v_{\mathrm{AI}}$. We form $w = \{(\widehat{D}_{\mathrm{AIPW}} - \widehat{D}_{\mathrm{IMP}})^2 + (v_{\mathrm{IMP}} - v_{\mathrm{AI}})/\sqrt{n}\}/\{(\widehat{D}_{\mathrm{AIPW}} - \widehat{D}_{\mathrm{IMP}})^2 + (v_{\mathrm{IMP}} + v_{\mathrm{AIPW}} - 2v_{\mathrm{AI}})/\sqrt{n}\}$, and form the shrinkage estimator $\widehat{D} = w\widehat{D}_{\mathrm{AIPW}} + (1-w)\widehat{D}_{\mathrm{IMP}}$, where we replace $v_{\mathrm{AIPW}}, v_{\mathrm{IMP}}$, and v_{AI} with their estimated versions. This construction has the property that when the IMP is inconsistent and the AIPW is consistent, $w \to 1$, and we essentially obtain the AIPW; that is, the shrinkage estimator is doubly robust. On the other hand, when both estimators are consistent, $w \to w_0$, where $w_0 \equiv (v_{\mathrm{IMP}} - v_{\mathrm{AI}})/(v_{\mathrm{IMP}} + v_{\mathrm{AIPW}} - 2v_{\mathrm{AI}})$ in probability, which

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

yields the optimal combination of the two estimators in terms of the final estimation variability. Of course, when both estimators are inconsistent, the weighted average is still inconsistent.

To construct the shrinkage estimator described above, we derive the asymptotic variances and covariances of the estimators in Section 3.4.

Note that one may also choose to shrink the IMP2 and AIPW, or either of the two versions of the imputation estimator and the improved AIPW, in a similar fashion.

3.4 Asymptotic Properties of the Treatment Effect Estimators

In this section, we discuss the asymptotic properties of the proposed average treatment effect estimators. These properties are developed under the following conditions:

C1 The univariate mth-order kernel function $K(\cdot)$ is symmetric and Lipschitz continuous on its support $[-1, 1]$, which satisfies $\int K(u)du = 1$, $\int u^i K(u)du = 0$, for $1 \leqslant i \leqslant m-1$, $0 \neq \int u^m K(u)du < \infty$.

C2 The bandwidths satisfy $nh^{2m} \to 0$ and $nh^{2d} \to \infty$.

C3 The probability density functions of $\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}$, $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$, denoted by $f\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)$, $f\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\right)$, and $f\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\right)$, respectively, with an abuse of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

notation, are bounded away from zero and ∞.

Let the true average causal effect be $D = E(Y_1 - Y_0)$. Then, we have the following results.

Theorem 3.1. *Under the regularity conditions C1–C3, when $n \to \infty$, the IMP estimator $\widehat{D}_{\mathrm{IMP}}$ satisfies $\sqrt{n}(\widehat{D}_{\mathrm{IMP}} - D) \xrightarrow{d} \mathrm{N}(0, v_{\mathrm{IMP}})$, where, using the results for $\widehat{E}(Y_1)$ and $\widehat{E}(Y_0)$ in the Supplementary Material S3,*

$$\begin{aligned} v_{\mathrm{IMP}} \;=\; & E\Big(\big\{ m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) - E(Y_1) + E(Y_0) \big\} \\ & + E[1 + \exp\{-\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\} \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i] t_i \{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\} \\ & - E[1 + \exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i](1 - t_i)\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\} \\ & - E[(1 - P_i)\mathrm{vec}\{\mathbf{X}_{Li}\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}]^{\mathrm{T}}\mathbf{B}_1 t_i \{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\} \\ & \times \mathrm{vec}[\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) \otimes \{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}] \\ & + E[P_i \mathrm{vec}\{\mathbf{X}_{Li}\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}]^{\mathrm{T}}\mathbf{B}_0(1 - t_i)\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\} \\ & \times \mathrm{vec}[\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) \otimes \{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}]\Big)^2, \end{aligned} \tag{3.1}$$

where $\mathbf{B}_1$ and $\mathbf{B}_0$ are defined in (2.8) and (2.10), respectively.

In the variance expression v_{IMP}, the first term captures the treatment effect estimation variability due to the different covariates. The second term is related to the variability of the outcome, given the covariates in

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

the treated group, weighted by the treatment probability. The third term resembles the second term, but applies to the non-treated group. The fourth term compensates for the second term to fully capture the variability due to the imputation and dimension reduction in the treated group. Similarly, the fifth term compensates for the third term in the non-treated group.

Theorem 3.2. *Under the regularity conditions C1–C3, when $n \to \infty$, the IMP2 estimator $\widehat{D}_{\mathrm{IMP2}}$ satisfies $\sqrt{n}(\widehat{D}_{\mathrm{IMP2}} - D) \xrightarrow{d} \mathrm{N}(0, v_{\mathrm{IMP2}})$, where using the results for $\widehat{E}(Y_1)$ and $\widehat{E}(Y_0)$ from the Supplementary Material S4,* $v_{\mathrm{IMP2}} = E\big(\{m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) - E(Y_1) + E(Y_0)\} + E(P_i^{-1} \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)t_i\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\} - E\{(1-P_i)^{-1} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i\}(1-t_i)\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}$ $- E[\mathrm{vec}\{\mathbf{X}_{Li}\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}]^{\mathrm{T}}\mathbf{B}_1 t_i\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\} \times \mathrm{vec}[\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) \otimes \{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}] + E[\mathrm{vec}\{\mathbf{X}_{Li}\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}]^{\mathrm{T}}\mathbf{B}_0(1-t_i)\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\} \times \mathrm{vec}[\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) \otimes \{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}]\big)^2$, *where $\mathbf{B}_1$ and $\mathbf{B}_0$ are defined in (2.8) and (2.10), respectively.*

Note that the first three terms in v_{IMP2} are identical to those in v_{IMP}. The only difference between v_{IMP2} and v_{IMP} is in the P_i component in the last two terms, reflecting the difference due to the imputation method.

Theorem 3.3. *Under the regularity conditions C1–C3, when $n \to \infty$, the IPW estimator $\widehat{D}_{\mathrm{IPW}}$ satisfies $\sqrt{n}(\widehat{D}_{\mathrm{IPW}} - D) \xrightarrow{d} \mathrm{N}(0, v_{\mathrm{IPW}})$, where using the results for $\widehat{E}(Y_1)$ and $\widehat{E}(Y_0)$ in the Supplementary Material S1, $v_{\mathrm{IPW}} =$*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

$$E\Bigg(\{t_i y_{1i}/p_i - E(Y_1) - (1-t_i)y_{0i}/(1-p_i) + E(Y_0)\}$$
$$+ (1 - t_i/p_i)\, E\left\{m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i\right\} - (t_i - p_i)/(1-p_i) E\left\{m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i\right\}$$
$$+ \left(E\left[\{m_{1i}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)(1-P_i) + m_{0i}(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)P_i\}\mathrm{vec}\{\mathbf{X}_{Li}\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}\right]\right)^{\mathrm{T}} \mathbf{B} \times (t_i -$$
$$p_i)\mathrm{vec}[\{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}]\Bigg)^2,$$ *where* $\mathbf{B}$ *is defined in (2.12).*

The variance v_{IPW} has a very different form to those from the imputation methods, partially reflecting the difference in how the methods handle the missing outcomes. The first three terms of v_{IPW} can be rewritten as $E\left\{m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i\right\} - E\left\{m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i\right\} - E(Y_1) + E(Y_0)$, $t_i p_i^{-1}\big[y_{1i} - E\left\{m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i\right\}\big]$, and $-(1-t_i)(1-p_i)^{-1}\big[y_{0i} - E\{m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i\}\big]$. We can view the first term as the variability in the treatment effect due to the covariates, and the second term as the variability in the inversely weighted individual treatment effect in the treatment group. The third term is similar to the second term, but applies to the non-treated group. The last term compensates for the combined variability due to the way in whch the IPW handles the missing outcomes.

Theorem 3.4. *Under the regularity conditions C1–C3, when* $n \to \infty$*, the AIPW estimator* $\widehat{D}_{\mathrm{AIPW}}$ *satisfies* $\sqrt{n}(\widehat{D}_{\mathrm{AIPW}} - D) \xrightarrow{d} \mathrm{N}(0, v_{\mathrm{AIPW}})$*, where*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

v_{AIPW}, *derived in the Supplementary Material S2, is*

$$\begin{aligned} v_{\mathrm{AIPW}} \;=\; & E\left(\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}t_i[1+\exp\{-\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}] + \{m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) - E(Y_1)\}\right. \\ & -\mathbf{C}_1\mathbf{B}_1 t_i\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}\mathrm{vec}[\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\otimes\{\mathbf{x}_{Li} - E(\mathbf{X}_{Li}\mid\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}] \\ & +\mathbf{D}_1\mathbf{B}(t_i-p_i)\mathrm{vec}[\{\mathbf{x}_{Li} - E(\mathbf{X}_{Li}\mid\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}] \\ & -\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}(1-t_i)[1+\exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}] \\ & -\{m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) - E(Y_0)\} + \mathbf{C}_0\mathbf{B}_0(1-t_i) \\ & \times\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}\mathrm{vec}[\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\otimes\{\mathbf{x}_{Li} - E(\mathbf{X}_{Li}\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}] \\ & \left.+\mathbf{D}_0\mathbf{B}(t_i-p_i)\mathrm{vec}[\{\mathbf{x}_{Li} - E(\mathbf{X}_{Li}\mid\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}]\right)^2, \end{aligned} \tag{3.2}$$

where $\mathbf{C}_1 \equiv E\left[\{\partial m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)/\partial vecl(\boldsymbol{\beta}_1)^{\mathrm{T}}\}(1-T_i/P_i)\right]$, $\mathbf{D}_1 \equiv E\big[\{Y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)\}$ $T_i\exp\{-\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\}\mathrm{vec}\{\mathbf{X}_{Li}\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}\big]$, $\mathbf{C}_0 \equiv E\big[\{\partial m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)/\partial vecl(\boldsymbol{\beta}_0)^{\mathrm{T}}\}$ $\{1-(1-T_i)/(1-P_i)\}\big]$, *and* $\mathbf{D}_0 \equiv E\big[\{Y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\}(1-T_i)\exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\}$ $\mathrm{vec}\{\mathbf{X}_{Li}\boldsymbol{\eta}'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}\big]$. *Note that* $\mathbf{C}_1$, $\mathbf{C}_0$, $\mathbf{D}_1$, *and* $\mathbf{D}_0$ *will degenerate to zero if the relevant model is correct. Then,*

$$\begin{aligned} v_{\mathrm{AIPW}} \;=\; & E\Big[\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}t_i/p_i + m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) - E(Y_1) \\ & - \{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}(1-t_i)/(1-p_i) - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) + E(Y_0)\Big]^2. \end{aligned} \tag{3.3}$$

The expression for v_{AIPW} is closely realated to that for v_{IMP2}. In fact,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

the third and seventh terms in v_{AIPW} are refinements of the fourth and fifth terms, respectively, in v_{IMP2}. Furthermore, we have an additional fourth and eighth term in v_{AIPW} to provide extra protection against a treatment assignment model misspecification. When the outcome models and assignment models are correct, as seen from (3.3), the variability contains only two parts, that due to the covariate variability, and that due to the incomplete outcomes and random errors.

Noting that $(1 - t_i/p_i)\, m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)$ and $\{1 - (1 - t_i)/(1 - p_i)\}m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)$ have mean zero, it is straightforward to show that the improved AIPW estimator has the same asymptotic expansion as the AIPW estimator when all three models are correct. Thus, despite their different finite-sample performance, the expansion in (3.3) also applies to the improved AIPW estimator. Therefore, the following result holds.

Theorem 3.5. *Under the regularity conditions C1–C3, and assuming all models are correct, then when $n \to \infty$, the improved AIPW estimator $\widehat{D}_{\mathrm{IAIPW}}$ satisfies $\sqrt{n}(\widehat{D}_{\mathrm{IAIPW}} - D) \xrightarrow{d} \mathrm{N}(0, v_{\mathrm{AIPW}})$, where v_{AIPW} is given by (3.3).*

Finally, when both estimators $\hat{D}_{IMP}$ and $\hat{D}_{AIPW}$ are consistent, we have $\sqrt{n}(\widehat{D} - D) = \sqrt{n}w_0(\widehat{D}_{AIPW} - D) + \sqrt{n}(1 - w_0)(\widehat{D}_{IMP} - D) + o_p(1)$, as noted above.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

Theorem 3.6. *Under the regularity conditions C1–C3, when* $\widehat{D}_{\mathrm{AIPW}}$ *and* $\widehat{D}_{\mathrm{IMP}}$ *are consistent and* $n \to \infty$*, the shrinkage estimator* $\widehat{D}$ *satisfies* $\sqrt{n}(\widehat{D} - D) \xrightarrow{d} \mathrm{N}(0, v_{\mathrm{shrinkage}})$*, where* $v_{\mathrm{shrinkage}} = w_0^2 v_{\mathrm{AIPW}} + (1-w_0)^2 v_{\mathrm{IMP}} + 2w_0(1-w_0)v_{\mathrm{AI}}$*, with* $v_{\mathrm{AI}} = E\Big\{\big(\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}t_i/p_i + m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) - E(Y_1) - \{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}(1-t_i)/(1-p_i) - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) + E(Y_0)\big) \times \Big(t_i y_{1i} - (1-t_i)y_{0i} + (1-t_i)m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) - t_i m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) - E(Y_1) + E(Y_0) + E[\exp\{-\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\} \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i]t_i\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\} - E[\exp\{\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_i)\} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i](1-t_i)\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\} - E[(1-P_i)\mathrm{vec}\{\mathbf{X}_{Li}\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}]^{\mathrm{T}}\mathbf{B}_1 t_i\{y_{1i} - m_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\} \times \mathrm{vec}[\mathbf{m}_1'(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i) \otimes \{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i)\}] + E[P_i\mathrm{vec}\{\mathbf{X}_{Li}\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\}]^{\mathrm{T}}\mathbf{B}_0(1-t_i)\{y_{0i} - m_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\} \times \mathrm{vec}[\mathbf{m}_0'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) \otimes \{\mathbf{x}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\}]\Big)\Big\}$.

The v_{AI} term is a simple result of the correlation between the AIPW estimator and IMP estimator. When $\widehat{D}_{IMP}$ is not consistent owing to a misspecification of at least one of the treatment mean models $m_1(\cdot)$ and $m_0(\cdot)$, $w \to 1$; thus, $\sqrt{n}(\widehat{D} - D) \xrightarrow{d} \sqrt{n}(\widehat{D}_{\mathrm{AIPW}} - D)$.

4. Simulation Study

We conducted a simulation study to compare the performance of the estimators discussed in Section 3. We used a sample size $n = 1000$ and covariate dimension $p = 6$ with 1000 replicates.

Specifically, the covariate vector $\mathbf{X} = (X_1, \ldots, X_6)^{\mathrm{T}}$ is generated as

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

follows. First, X_1 and X_2 are generated independently from $N(1,1)$ and $N(0,1)$, respectively. We let $X_4 = 0.015X_1 + u_1$, where u_1 is uniformly distributed in $(-0.5, 0.5)$. Then, X_3 and X_5 are generated independently from the Bernoulli distributions with success probabilities $0.5 + 0.05X_2$ and $0.4 + 0.2X_4$, respectively. We let $X_6 = 0.04X_2 + 0.15X_3 + 0.05X_4 + u_2$, where $u_2 \sim N(0,1)$. We set $\boldsymbol{\beta}_1 = (1, -1, 1, -2, -1.5, 0.5)^{\mathrm{T}}$, $\boldsymbol{\beta}_0 = (1, 1, 0, 0, 0, 0)^{\mathrm{T}}$, and $\boldsymbol{\alpha} = (-0.27, 0.2, -0.15, 0.05, 0.15, -0.1)^{\mathrm{T}}$.

4.1 Study 1

Our first study examines the estimators when the response and the propensity score models are correctly specified. We generated the response variables based on $Y_1 = 0.7(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})^2 + \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + \epsilon_1$ and $Y_0 = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x} + \epsilon_0$. Here, ϵ_1 and ϵ_0 are normally distributed with mean zero and variances 0.5 and 0.2, respectively. We let $\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$. Thus, the treatment indicator T is generated from the logistic model $\mathrm{pr}(T = 1|\mathbf{X}) = \exp(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})/\{1 + \exp(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})\}$.

We implemented the six estimators described in Section 3. In the nonparametric estimations of $\eta(\cdot)$ and the mean functions $m_1(\cdot)$ and $m_0(\cdot)$, we used a local linear regression with an Epanechnikov kernel and a bandwidth chosen as $c\sigma n^{-1/3}$, where σ^2 is the estimated variance of the corresponding index, and c is a constant ranging from 0.1 to 3.5. As is frequently observed

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

in semiparametric estimations, the final estimator is relatively insensitive to the bandwidth used for the nuisance estimation, because this bandwidth has no first-order effect as long as it satisfies Condition C2. When needed, we extrapolated the local linear fit at the boundary of the support. For comparison, we also computed $\sum_{i=1}^n T_i Y_{1i}/(\sum_{i=1}^n T_i) - \sum_{i=1}^n (1-T_i) Y_{0i}/(n - \sum_{i=1}^n T_i)$ as a naive sample average estimator.

From the results summarized in Figure 1 and Table 1, we can see that the naive estimator is obviously severely biased. As expected, all six methods yield a small bias, and the IMP2 and IPW provide the smallest and largest, respectively, variability and mean squared error (MSE). The estimator that shrinls the IMP and the AIPW improves slightly on the latter with respect to the variability and MSE. The estimated standard deviation (based on the asymptotics) matches fairly well with the empirical variability of the estimators.

4.2 Study 2

The second study compares the performance of the estimators when the mean functions $m_1(\cdot)$ and $m_0(\cdot)$ are misspecified. We kept the data-generation procedure identical to that of Study 1, except that we generated the response variables based on the models $Y_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})^2 + \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + (\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{x})^2 + \epsilon_1$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

and $Y_0 = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x} + \sin(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{x}) + \epsilon_0$, where $\boldsymbol{\gamma}_1 = (0, 1, 1, 0, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{\gamma}_0 = (0, 1, -0.75, 0, -1, 0)^{\mathrm{T}}$. Here, ϵ_1 and ϵ_0 are normally distributed with mean zero and variance 0.5 and 0.2, respectively. Note that the mean functions no longer have the single index forms.

When we implemented the six estimators described in Section 3, we still treated $m_1(\cdot)$ and $m_0(\cdot)$ as functions of $\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$, respectively; hence, the mean function models we used are misspecified. The same nonparametric estimation procedures as in Study 1 were used to estimate $\eta(\cdot)$, $m_1(\cdot)$, and $m_0(\cdot)$.

From the results in Figure 2 and Table 2, we can see that the IMP and IMP2 estimators are biased, along with the severely biased naive estimator, whereas the IPW, AIPW, IAIPW and shrinkage methods yield a small bias, even when $m_1(\cdot)$ and $m_0(\cdot)$ are misspecified, as expected. Although the IMP is biased, it provides the smallest variability, whereas the IPW yields the largest variability. Here, the shrinkage estimator that combines the IMP and AIPW is able to down-weight the IMP and inherit the lower bias and variability from the AIPW. Again estimated standard deviations match the empirical variability of the estimators.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

4.3 Study 3

In the third simulation study, we compare the performance of the estimators when the model of the propensity score function is misspecified. We followed the same data-generation procedure as that in Section 4.1, but the true function inside the logistic link here is $\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) = (\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + 0.45/\{(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^2 + 0.5\}$, where $\boldsymbol{\gamma} = (1, 0.5, -1, 0.5, -1, -3)^{\mathrm{T}}$. Thus, $\eta(\cdot)$ is no longer a function of a single index. The treatment indicator T is generated from

$$\mathrm{pr}(T = 1|\mathbf{X}) = \frac{\exp[(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + 0.45/\{(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^2 + 0.5\}]}{1 + \exp[(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + 0.45/\{(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^2 + 0.5\}]}.$$

In implementing the six estimators described in Section 3, we considered $\eta(\cdot)$ as a function of $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$ only; thus, the propensity score used to estimate the average causal effect is misspecified. Furthermore, we used the same nonparametric approach as in Studies 1 and 2 to estimate $m_1(\cdot)$, $m_0(\cdot)$, and $\eta(\cdot)$.

The results in Figure 3 and Table 3 show that, except for the naive estimator, which is significantly biased, all six estimators yield small biases. Whereas the small biases of IMP, IMP2, AIPW, IAIPW, and the shrinkage estimator are within our expectation, the IPW performs better than anticipated by the theory. Here, the IMP2 has the smallest variabil-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

ity and MSE, whereas the IPW performs worst. As in Study 1, the IMP and AIPW are both consistent in this design, and the shrinkage estimator is again as good as the AIPW. By construction, we expect the shrinkage estimator to have a lower variability in this situation. However, thi sis not evident, probably qwing to the difficulty in obtaining precise estimates of the asymptotic variances used to compute the shrinkage weight. On the other hand, the variance estimates are sufficiently good to yield satisfactory empirical coverages for the confidence intervals constructed.

4.4 Study 4

In this last study, we consider a scenario in which all models, $m_1(\cdot)$, $m_0(\cdot)$, and $\eta(\cdot)$ are misspecified. Here, the covariate $\mathbf{X}$ is generated as in the previous studies, the response variables Y_1 and Y_0 are generated as in Section 4.2, and the treatment assignment is as described in Section 4.3. While implementing the estimators described in Section 3, we still treat $m_1(\cdot)$, $m_0(\cdot)$, and $\eta(\cdot)$ as functions of $\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}$, $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$, and $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$, respectively, and use the same nonparametric estimation procedure as in earlier sections.

From Figure 4 and Table 4, we can see that the misspecification of the mean function models means the IMP and IMP2 estimators are biased, as is the naive estimator. As in Study 3, although $\eta(\cdot)$ is misspecified, the IPW

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

estimator yields quite a small bias. Consequently, the AIPW, IAIPW, and shrinkage estimators are not affected significantly by the misspecifications of the various models. The IMP2 and IMP have the lowest variability, followed by the IAIPW and AIPW, and IPW has the largest variance, as in the earlier cases. Because the IMP has a much larger bias than that of the AIPW, the shrinkage estimator mimics the AIPW, as the theory predicts.

Following the request of a referee, we also conducted the simulation study using sample sizes of $n = 100, 200$, and 500. The results are provided in the Supplementary Material S5 to Section S7. The results show that as the sample size increases, the bias and variance (and thus the MSEs) decrease for all of the estimators.

5. Data Analysis

We now apply the proposed methods to estimate the average causal effect of maternal smoking during pregnancy on birth weight. The data consist of the birth weights (in grams) of 4642 singleton births in Pennsylvania, USA (Almond et al. 2005), for which several covariates are observed: mother's age, mother's marital status, an indicator variable for alcohol consumption during pregnancy, an indicator variable for a previous birth in which the infant died, mother's education, father's education, number of prenatal care

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

visits, months since last birth, mother's race, and an indicator variable for the first-born child. The data set also contains the maternal smoking habit during pregnancy, which we view as our treatment, T_i (1 = Smoking, 0 = Non-Smoking). This data set was first used by Almond et al. (2005) to study the economic cost of low birth weights on society, and was further analyzed in Cattaneo (2010) and Liu et al. (2018). The data set is available at `http://www.stata-press.com/data/r13/cattaneo2.dta`.

To determine the structural dimension of the two response models and the propensity score function model, we use the validated information criterion (VIC) (Ma & Zhang 2015), where the true reduced space dimension corresponds to the smallest VIC value. We conducted the VIC calculation separately for all three models to determine their suitable dimensions. When we consider the mean response model for the non-treated group, the VIC value at $d = 1$ is 84.43, and is 201.86 at $d = 2$, after which it continues to increase with d. Hence, we select $d = 1$ for this model, and fit a single index structure. Similarly, when we conducted the VIC method on the mean response model for the treated group, the smallest VIC value was also obtained at $d = 1$. Finally, the same is true for the propensity score model, where the VIC value at the single index case is the smallest. Thus, we apply the single index structure in all three dimension-reduction

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

models. Of the 4642 observations, 864 of the mothers smoked ($T = 1$) and 3778 non-smoking ($T = 0$). The naive estimator (without the covariate adjustment) yields an effect of -275 g. We used a local linear regression with an Epanechnikov kernel in the nonparametric estimations of the propensity score function, $\eta(\cdot)$, and the mean functions, $m_1(\cdot)$ and $m_0(\cdot)$, where the bandwidth was selected as $c\sigma n^{-1/3}$, with σ^2 the estimated variance of the corresponding index and c a constant. In our analysis, we find that the results are not sensitive to the value of c; for example, when we vary c from 0.01 to 5, the results barely change. Applying the six estimators studied in Section 3 yields estimated effects of smoking of between -259 and -296 g. These are displayed in Table 5, together with the estimated standard deviations and the 95% confidence intervals. The IPW stands out, with an estimated effect larger than the naive value. This is because some observations have propensity scores close to zero, leading to very large weights, and thus much larger standard error. Overall, there is evidence that smoking results in lower birth weight, given the assumption that we have observed all confounders.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

6. Conclusion

We have introduced feasible and robust estimators for the average causal effect of a nonrandomized treatment. Nuisance models are fitted using semiparametric sufficient dimension-reduction methods. The parameter estimation in these nuisance models is locally efficient, which is important when combining the IPW and IMP estimators. The AIPW estimators are efficient and their asymptotic distributions do not depend on the fit of the nuisance parameters, as long as the nuisance models are well specified and the estimations are consistent (e.g., Farrell 2015, Belloni et al. 2014). The proposed shrinkage estimator combines the AIPW and IMP, thus improving the efficiency when the nuisance model for the response is correctly specified. When the latter model is misspecified, the shrinkage estimator is asymptotically equivalent to the AIPW, and nothing is lost. Numerical experiments show that the shrinkage estimator performs at least as well as the AIPW, although no improvement could be observed over the AIPW for well-specified response models, possibly because the weights estimates are insufficient for the sample size considered. As is the case for the IMP, the shrinkage estimator is super-efficient and its asymptotic inference is not expected to be uniform.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

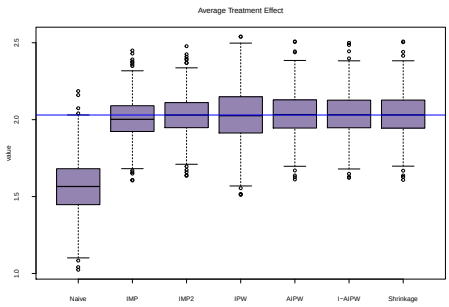

Figure 1: Box plot of Naive, IMP, IMP2, IPW, AIPW, IAIPW, and Shrinkage estimators for Study 1. The blue horizontal line is the true average causal effect (ACE), here 2.030.

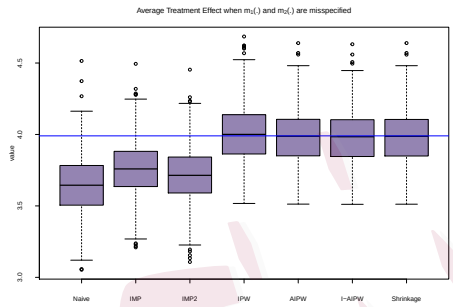

Figure 2: Box plot of Naive, IMP, IMP2, IPW, AIPW, IAIPW, and Shrinkage estimators for Study 2, where $m_1(\cdot)$ and $m_0(\cdot)$ are misspecified. The blue horizontal line is the true ACE, here 3.990.

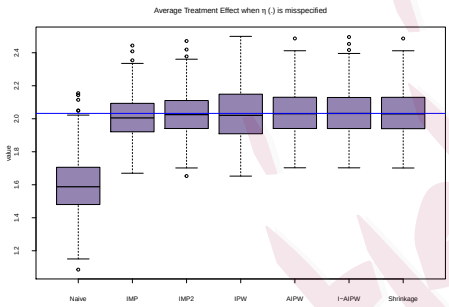

Figure 3: Box plot of Naive, IMP, IMP2, IPW, AIPW, IAIPW, and Shrinkage estimators for Study 3, where $\eta(\cdot)$ is misspecified. The blue horizontal line is the true ACE, here 2.033.

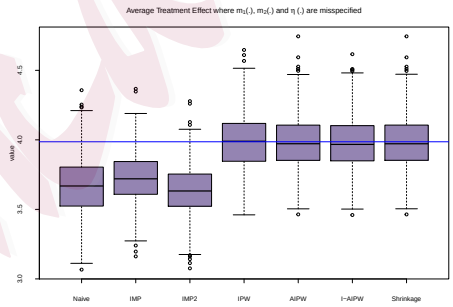

Figure 4: Box plot of Naive, IMP, IMP2, IPW, AIPW, IAIPW, and Shrinkage estimators for Study 4, where $m_1(\cdot)$, $m_0(\cdot)$ and $\eta(\cdot)$ are misspecified. The blue horizontal line is the true ACE, here 3.986.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

Table 1: Results for Study 1 based on 1000 replicates; Full gives the average causal effect and corresponding standard deviation (sd) based on all potential responses, including the counterfactual ones not observable in practice; Naive provides the same statistics, but based only on the observed potential responses. For the different estimators, we also compute the mean of the estimated sd (based on asymptotics, row $\widehat{\text{sd}}$), empirical coverage obtained with confidence intervals based on these estimated sd (95% cvg), and mean squared error (mse).

Estimators	Full	Naive	IMP	IMP2	IPW	AIPW	IAIPW	Shrinkage
mean	2.030	1.569	2.007	2.032	2.029	2.037	2.036	2.036
sd	0.118	0.172	0.123	0.122	0.168	0.131	0.130	0.131
$\widehat{\text{sd}}$	-	-	0.134	0.130	0.176	0.146	0.146	0.138
95% cvg	-	-	96.1%	96%	96.5%	97.8%	98%	97.5%
mse	-	-	0.016	0.015	0.028	0.017	0.017	0.017

Table 2: Results for Study 2, where $m_1(\cdot)$ and $m_0(\cdot)$ are misspecified; see also the caption of Table 1.

Estimators	Full	Naive	IMP	IMP2	IPW	AIPW	IAIPW	Shrinkage
mean	3.990	3.647	3.761	3.716	4.005	3.984	3.979	3.983
sd	0.137	0.202	0.187	0.189	0.207	0.188	0.189	0.188
$\widehat{\text{sd}}$	-	-	0.188	0.193	0.211	0.195	0.195	0.194
95% cvg	-	-	79%	74.7%	95.8%	94.9%	94.9%	94.9%
mse	-	-	0.087	0.111	0.043	0.035	0.036	0.035

Table 3: Results for Study 3, where $\eta(\cdot)$ is misspecified; see also the caption of Table 1.

Estimators	Full	Naive	IMP	IMP2	IPW	AIPW	IAIPW	Shrinkage
mean	2.033	1.596	2.009	2.029	2.030	2.037	2.037	2.036
sd	0.122	0.165	0.123	0.122	0.169	0.135	0.134	0.135
$\widehat{\text{sd}}$	-	-	0.140	0.140	0.160	0.143	0.143	0.142
95% cvg	-	-	96.8%	97.6%	94.5%	96%	96.3%	95.8%
mse	-	-	0.016	0.015	0.029	0.018	0.018	0.018

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

Table 4: Results for Study 4, where $m_1(\cdot)$, $m_0(\cdot)$, and $\eta(\cdot)$ are misspecified; see also the caption of Table 1.

Estimators	Full	Naive	IMP	IMP2	IPW	AIPW	IAIPW	Shrinkage
mean	3.986	3.665	3.727	3.637	3.987	3.980	3.977	3.980
sd	0.135	0.198	0.175	0.173	0.202	0.186	0.184	0.186
$\widehat{\text{sd}}$	-	-	0.194	0.205	0.207	0.191	0.191	0.191
95% cvg	-	-	78.5%	66.7%	95.4%	95.5%	96.1%	95.5%
mse	-	-	0.098	0.152	0.041	0.035	0.034	0.035

Table 5: Estimated average causal effect of maternal smoking on birth weight, including standard error and confidence interval, for the estimators introduced.

Estimator	Estimate	se	95% CI
naive	-275.3	-	-
IMP	-259.8	22.2	(-303.3,-216.3)
IMP2	-262.6	23.1	(-307.8,-217.4)
IPW	-296.5	85.5	(-464.2,-128.9)
AIPW	-264.6	22.2	(-308.1,-221.1)
IAIPW	-264.7	22.2	(-308.3,-221.2)
Shrinkage	-264.6	22.2	(-308.1,-221.1)

Supplementary Material

The online Supplementary Material contains proofs for Theorems 3.1–3.4.

Acknowledgments

This research was supported by the National Science Foundation, the National Institute of Health, and the Marianne and Marcus Wallenberg Foundation.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

References

Almond, D., Chay, K. Y. & Lee, D. S. (2005), 'The costs of low birth weight', *The Quarterly Journal of Economics* **120**, 1031–1083.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls†', *The Review of Economic Studies* **81**, 608–650.

Cattaneo, M. D. (2010), 'Efficient semiparametric estimation of multi-valued treatment effects under ignorability', *Journal of Econometrics* **155**, 138 – 154.

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, Wiley, New York.

de Luna, X., Waernbaum, I. & Richardson, T. S. (2011), 'Covariate selection for the nonparametric estimation of an average treatment effect', *Biometrika* **98**, 861–875.

Farrell, M. (2015), 'Robust inference on average treatment effects with possibly more covariates than observations.', *Journal of Econometrics* **189**, 1–23.

Gruber, S. & van der Laan, M. J. (2010), 'An application of collaborative targeted maximum likelihood estimation in causal inference and genomics', *The International Journal of Biostatistics* **6**.

Horvitz, D. G. & Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**, 663–685.

Li, K.-C. (1991), 'Sliced inverse regression for dimension reduction', *Journal of the American*

Statistical Association **86**, 316–327.

Li, K. C. & Duan, N. (1991), 'Regression analysis under link violation', *Annals of Statistics* **17**, 1009–1052.

Liu, J., Ma, Y. & Wang, L. (2018), 'An alternative robust estimator of average treatment effect in causal inference', *Biometrics* **74**, doi.org/10.1111/biom.12859.

Luo, W., Zhu, Y. & Ghosh, D. (2017), 'On estimating regression-based causal effects using sufficient dimension reduction', *Biometrika* **104**, 51–65.

Ma, S., Zhu, L., Zhang, Z., Tsai, C. & Carroll, R. (2018), 'A robust and efficient approach to causal inference based on sparse sufficient dimension reduction', *Annals of Statistics* (Published on-line ahead of print).

Ma, Y. & Zhang, X. (2015), 'A validated information criterion to determine the structural dimension in dimension reduction models', *Biometrika* **102**(2), 409–420.

URL: *https://doi.org/10.1093/biomet/asv004*

Ma, Y. & Zhu, L. (2012), 'A semiparametric approach to dimension reduction', *Journal of the American Statistical Association* **107**, 168–179.

Ma, Y. & Zhu, L. (2013), 'Efficient estimation in sufficient dimension reduction', *The Annals of Statistics* **41**, 250–268.

Ma, Y. & Zhu, L. (2014), 'On estimation efficiency of the central mean subspace', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 885–901.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

Mukherjee, B. & Chatterjee, N. (2008), 'Exploiting gene-environment independence for analysis of case-control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency', *Biometrics* **64**, 685–694.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association* **89**, 846–866.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association* **90**, 106–121.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**, 41–55.

Rubin, D. B. (1978b), 'Multiple imputations in sample surveys: A phenomenological bayesian approach to nonresponse (with discussion)', *American Statistical Association Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA* pp. 20–34.

Shortreed, S. & Ertefaie, A. (2017), 'Outcome-adaptive lasso: Variable selection for causal inference', *Biometrics* **73**, 1111–1122.

Tan, Z. (2006), 'A distributional approach for causal inference using propensity scores', *Journal of the American Statistical Association* **101**, 1619–1637.

Tan, Z. (2007), 'Comment: Understanding or, ps and dr', *Statistical Science* **22**, 560–568.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0416

Wang, Y., Garcia, T. P. & Ma, Y. (2012), 'Nonparametric estimation for censored mixture data with application to the cooperative huntington's observational research trial', *Journal of the American Statistical Association* **107**, 1324–1338.

Xia, Y. C. (2007), 'A constructive approach to the estimation of dimension reduction directions', *Annals of Statistics* **35**, 2654–2690.

Xia, Y., Tong, H., Li, W. K. & Zhu, L. X. (2002), 'An adaptive estimation of dimension reduction space (with discussion)', *Journal of the royal statistical society, series B* **64**, 363–410.

Pennsylvania State University

E-mail: tbg5133@psu.edu

Pennsylvania State University

E-mail: yzm63@psu.edu

Umeå University

E-mail: xavier.de.luna@umu.se